Theor Appl Genet (2013) 126:133–141
DOI 10.1007/s00122-012-1968-6

ORIGINAL PAPER

Genetic diversity analysis of elite European maize (*Zea mays* L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs

Elisabetta Frascaroli · Tobias A. Schrag · Albrecht E. Melchinger

Received: 24 May 2012 / Accepted: 14 August 2012 / Published online: 4 September 2012

Abstract Recent advances in high-throughput sequencing technologies have triggered a shift toward single-nucleotide polymorphism (SNP) markers. A systematic bias can be introduced if SNPs are ascertained in a small panel of genotypes and then used for characterizing a larger population (ascertainment bias). With the objective of evaluating a potential ascertainment bias of the Illumina MaizeSNP50 array with respect to elite European maize dent and flint inbred lines, we compared the genetic diversity among these materials based on 731 amplified fragment length polymorphisms (AFLPs), 186 simple sequence repeats (SSRs), 41,434 SNPs of the MaizeSNP50 array (SNP-A), and two subsets of it, i.e., 30,068 Panzea (SNP-P) and 11,366 Syngenta markers (SNP-S). We evaluated the bias effects on major allele frequency, allele number, gene diversity, modified Roger's distance (MRD), and on molecular variance (AMOVA). We revealed ascertainment bias in SNP-A, compared to AFLPs and SSRs. It affected especially European flint lines analyzed with markers (SNP-S) specifically developed to maximize differences among North American dent germplasm. The bias affected all genetic parameters, but did not substantially alter the relative distances between inbred lines within groups. For these reasons, we conclude that the SNP markers of the MaizeSNP50 array can be employed for breeding purposes in the investigated material. However, attention should be paid in case of comparisons between genotypes belonging to different heterotic groups. In this case, it is advisable to prefer a marker subset with potentially low ascertainment bias, like in our case the SNP-P marker set.

Elisabetta Frascaroli and Tobias A. Schrag contributed equally to this work.

Communicated by M. Xu.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-012-1968-6) contains supplementary material, which is available to authorized users.

E. Frascaroli
Department of Agroenvironmental Sciences and Technologies, University of Bologna, Viale Fanin 44, 40127 Bologna, Italy

T. A. Schrag · A. E. Melchinger (✉)
Institute of Plant Breeding, Seed Science, Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany
e-mail: melchinger@uni-hohenheim.de

T. A. Schrag
e-mail: schrag@uni-hohenheim.de

Introduction

Genotyping with molecular markers has become crucial for understanding the genetic variation in many organisms, including plant species. Reliable and cost-effective marker technologies nowadays allow a better characterization of genetic resources and breeding materials. This will help to maintain genetic diversity and sustain long-term selection gains.

Different marker systems can be utilized for assessing crop genomic diversity and performing marker-assisted selection. Recent advances in sequencing technologies have triggered a shift toward single-nucleotide polymorphism (SNP) markers in many species, particularly for model organisms with substantial genomic resources. SNP markers are biallelic and have lower information content in comparison with the multiallelic simple sequence repeats (SSRs), but occur at much higher density in the genome and are amenable to high-throughput methods such as genotyping arrays (Rafalski 2002). SNPs have many

advantages over other marker systems, including the availability of high numbers of annotated markers, improved results for poor quality samples, a simple mutation model, and the ability to examine both neutral variation and regions under selection. For these reasons, they offer unparalleled potential for extended screening of genomes and large sample sizes also from natural populations (Seeb et al. 2011). Very large numbers of SNP markers are now available for detailed analysis of genome structure, genome-wide association studies, and precision breeding, especially for those animals and plants for which high-density genotyping arrays are commercially produced (e.g., Ramos et al. 2009; Ganal et al. 2011).

Despite the numerous advantages of SNPs, their use in genetic diversity studies has been criticized due to the fact that SNPs are first discovered in a small panel of sequenced individuals and subsequently used in arrays for genotyping of much larger panels (Ramìrez-Soriano and Nielsen 2009). Although this procedure provides a fast and cheap way of generating data, it may also lead to an ascertainment bias. Such a systematic bias may result from the criteria used to select the individuals in which genetic variation is assayed during SNP discovery (Clark et al. 2005). According to Schlötterer (2004), all SNP isolation strategies result in a notable bias for various parameters such as Wright's among-population fixation index F_{ST}, the allele frequency distribution, and linkage disequilibrium (Nielsen and Signorovitch 2003). As a consequence, ascertainment bias has been widely studied in different animal (e.g., Albrechtsen et al. 2010; Seeb et al. 2011) and plant species (Hamblin et al. 2007; Moragues et al. 2010).

In maize (*Zea mays* L.), development of SNP-based markers brought a new level of resolution to the analysis of genetic diversity and superseded other genetic marker categories for most applications. Nevertheless, in maize too, ascertainment bias is a concern in population-genetic analyses and in population-based genetic association studies. As an example, Rafalski (2011) reported that, with SNPs developed in elite maize inbreds for investigating non-adapted germplasm, genetic distances between genotypes determined in the ascertainment population were always larger in comparison with distances between genotypes in the non-ascertained population. On the contrary, ascertainment bias does not have an impact when SNPs are used for genotyping biparental segregating populations, because the only markers employed in this case are those polymorphic between the two parents and their allelic frequencies in the original population are known.

For these reasons, an important question is whether maize SNPs in commercial arrays (e.g., Illumina MaizeSNP50; Ganal et al. 2011) are associated with ascertainment bias leading to an underestimation of polymorphism within elite European inbred lines, which were not included in the original process of SNP identification. To reveal ascertainment bias in an SNP marker set, results obtained for the same genotypes with other marker systems can be taken as reference. Such references could be amplified fragment length polymorphism (AFLP) markers (Vos et al. 1995) that are mostly random and thus do not present the same bias, and simple sequence repeat (SSR) markers.

The present study has the goal to evaluate a potential ascertainment bias of the MaizeSNP50 array in genotyping dent and flint inbred lines from elite European maize. In particular, our objectives were to (1) compare various parameters describing the genetic variation among these materials determined with AFLPs, SSRs, SNPs, and subsets of the SNP array, and (2) discuss the implications for the use of the MaizeSNP50 array for diversity studies and breeding applications with European elite maize germplasm.

Materials and methods

Plant material

In total, 77 elite maize inbred lines described in a previous treatise (Schrag et al. 2010) were analyzed in this study. The inbreds were developed by the breeding program of the University of Hohenheim and comprised 46 dent lines, with Iodent or Iowa Stiff Stalk Synthetic background, and 31 flint lines, with European flint or flint/Lancaster background.

Molecular markers

Each of the 77 inbred lines was genotyped with AFLP, SSR, and SNP markers. The AFLP analyses were carried out with 20 primer–enzyme combinations (Vos et al. 1995), as described in detail by Schrag et al. (2006). These analyses resulted in 910 mapped AFLP markers. The SSR analyses were carried out for 270 publicly available SSR markers, uniformly distributed across the genome according to the "IBM2 2004 neighbors" map (http://www.maizegdb.org), as described in detail by Schrag et al. (2010). For SNP genotyping, the Illumina MaizeSNP50 array was used and provided 49,585 SNPs. For the subsequent statistical analyses, markers were used only if they were polymorphic among the whole set of 77 inbred lines and if they showed less than 20 % of missing observations (which in the case of SNPs corresponds to a call rate higher than 80 %). Accordingly, the final number of markers was 731 AFLPs, 186 SSRs, and 41,434 SNPs.

The Illumina MaizeSNP50 array is composed of SNPs from several sources, mainly contributed by the Panzea project and Syngenta. The SNPs from the Panzea project (Zhao et al. 2006; http://www.panzea.org/) were developed on the basis of a larger population derived from a diverse

set of 14 maize and 16 teosinte inbreds (Wright et al. 2005). In contrast, the Syngenta SNP markers were developed in order to maximize differences among elite inbred lines, particularly those belonging to the North American dent pool (Ganal et al. 2011). For the purpose of this study, the complete set of 41,434 selected SNP markers of the whole array (hereafter referred to as SNP-A) was split into two subsets, i.e., 30,068 SNP markers contributed by Panzea and other sources (hereafter referred to as SNP-P) and 11,366 SNP markers developed by Syngenta (hereafter referred to as SNP-S).

Statistical analysis

For each of the five marker sets (SSR, AFLP, SNP-A, SNP-P, SNP-S) and two groups of lines (dent, flint), we determined the following parameters for each locus: major allele frequency, number of alleles, and gene diversity as described by Weir (1996). Subsequently, we calculated for each group of lines and each parameter the means $\bar{X}_D$ and $\bar{X}_F$ across each set of marker loci and determined the corresponding standard deviation (SD) by means of bootstrapping 1,000 rounds across the respective set of loci. Descriptive statistics were calculated with PowerMarker (Liu and Muse 2005).

In addition, we calculated for each marker set the modified Roger's distance (MRD) (Wright 1978) for all pairs of lines as well as the mean and the SD of MRD within the two groups of lines. For testing the hypotheses that genome-wide averages $\bar{X}_D$ and $\bar{X}_F$ of major allele frequency, number of alleles, gene diversity, and MRD differed significantly between the dent and flint lines ($\bar{X}_D \neq \bar{X}_F$), we obtained empirical distributions of their difference between the two groups of lines by bootstrapping 10,000 rounds based on resampling genotypes within each group (Efron 1993, p. 214).

An analysis of molecular variance (AMOVA, Weir and Cockerham 1984; Excoffier et al. 1992; Weir 1996) on the basis of the MRD values was performed to compare the relative importance of the molecular variation within each group of lines for each marker set. AMOVA-based variance components between groups (s_B^2) and within the group of dent (s_D^2) and flint (s_F^2) lines were calculated. The total molecular variance was defined as $s_T^2 = s_B^2 + s_D^2 + s_F^2$ and each variance component was reported as the proportion of the total molecular variance. The hypothesis that s_D^2 and s_F^2 differed significantly ($s_D^2 \neq s_F^2$) and that proportions of molecular variance differed between two specific marker sets were tested on the empirical distributions of the differences obtained by bootstrapping 10,000 rounds based on resampling genotypes within each group (Efron 1993, p. 214). The AMOVA-related calculations were performed using the R environment (R Development Core Team 2011) with package *pegas* (Paradis 2010).

To test the hypothesis that the ratios of the values of the flint and dent lines for a given parameter (major allele frequency, number of alleles, gene diversity, MRD, AMOVA-based proportions of variance components) differed significantly between two specific marker sets, we determined the empirical distributions of the differences by bootstrapping (10,000 rounds) based on resampling of genotypes in each group of lines.

Correlations of MRD between different marker sets were determined separately for pairwise combinations among dent lines (DD, r_{DD}), among flint lines (FF, r_{FF}), and between dent and flint lines (DF, r_{DF}). The significance of each correlation was assessed by a Mantel test (Mantel 1967). A bootstrapping (10,000 rounds) based on resampling of genotypes was carried out to test the hypotheses that these correlations differed between DD and FF lines combinations ($r_{DD} \neq r_{FF}$), and between intragroup (i.e., DD, FF) and intergroup (i.e., DF) line combinations ($(r_{DD} + r_{FF})/2 \neq r_{DF}$).

For each marker set, associations among genotypes were displayed with principal coordinate analysis (PCoA, Gower 1966) based on MRD estimates. These association plots were then compared between marker sets by means of a Procrustes analysis (Jackson 1995) and significance levels were obtained based on 10,000 permutations, using *ade4* package in R (Dray and Dufour 2007).

Results

The multiallelic SSR showed a lower major allele frequency, and a higher allele number, gene diversity, and MRD than the biallelic AFLP and SNP markers (Table 1). The number of alleles averaged slightly below two for AFLPs and SNPs, and was more than three with SSRs. The frequency of observed heterozygous loci was below 2 % for all marker systems (data not shown).

No significant differences were found between dent and flint lines for allele number, gene diversity, and for MRD when analyzed with AFLP and SSR markers. In contrast, the dent and flint lines differed significantly in all parameters for SNP-A and SNP-S marker sets, and only in major allele frequency and gene diversity for SNP-P. Higher values were found in the flint lines for major allele frequency, and lower values for the other parameters.

The ratio $R = 100 \times \bar{X}_F/\bar{X}_D$ was generally below 100 % for all marker data sets and parameters, except major allele frequency (Table 1). For all parameters, the ratios did not differ between AFLPs and SSRs, but differed significantly between AFLPs and each of the three SNP marker sets. Ratios also differed significantly for gene diversity and MRD between SSRs and each of the three SNP marker sets. The ratios differed for all parameters among the three

Table 1 Means ($\bar{X}$) and standard deviations (SD) for major allele frequency, number of alleles, gene diversity and modified Roger's distance (MDR) for dent and flint lines and ratio of the means for five different sets of markers

Measure	Marker set[a]	Dent $\bar{X}_D \pm$ SD	Flint $\bar{X}_F \pm$ SD	Hypothesis Dent $\neq$ flint	Ratio (%) Flint/dent
Major allele frequency	AFLP	0.792 ± 0.004	0.812 ± 0.004	*	*102.5 a*
	SSR	0.644 ± 0.012	0.683 ± 0.011	*	*106.9 abcd*
	SNP-A	0.791 ± 0.001	0.826 ± 0.001	**	*104.4 b*
	SNP-P	0.795 ± 0.001	0.827 ± 0.003	**	*104.0 c*
	SNP-S	0.781 ± 0.001	0.823 ± 0.001	**	*105.5 d*
Allele number	AFLP	1.856 ± 0.009	1.867 ± 0.009	ns	*100.6 a*
	SSR	3.296 ± 0.082	3.177 ± 0.075	ns	*96.4 abcd*
	SNP-A	1.877 ± 0.002	1.831 ± 0.002	*	*97.6 b*
	SNP-P	1.870 ± 0.002	1.832 ± 0.002	ns	*98.0 c*
	SNP-S	1.895 ± 0.003	1.828 ± 0.004	**	*96.4 d*
Gene diversity	AFLP	0.275 ± 0.004	0.261 ± 0.004	ns	*94.7 a*
	SSR	0.454 ± 0.013	0.428 ± 0.012	ns	*94.4 a*
	SNP-A	0.276 ± 0.001	0.243 ± 0.001	**	*88.1 b*
	SNP-P	0.272 ± 0.001	0.243 ± 0.001	*	*89.3 c*
	SNP-S	0.288 ± 0.002	0.245 ± 0.002	**	*85.1 d*
MRD	AFLP	0.526 ± 0.005	0.514 ± 0.009	ns	*97.8 a*
	SSR	0.667 ± 0.008	0.651 ± 0.014	ns	*97.6 a*
	SNP-A	0.525 ± 0.007	0.492 ± 0.013	*	*93.9 b*
	SNP-P	0.520 ± 0.007	0.492 ± 0.014	ns	*94.5 c*
	SNP-S	0.536 ± 0.007	0.495 ± 0.013	**	*92.2 d*

Values followed by the same letter within a measure do not differ significantly at $P \leq 0.05$, based on bootstrapping

ns, *, **: non-significant, significant at $P \leq 0.05$ and $P \leq 0.01$, respectively, based on bootstrapping

[a] SNP-A refers to all 41,434 MaizeSNP50 markers, SNP-P to the subset of 30,068 markers from Panzea and SNP-S to the subset of 11,366 markers from Syngenta

SNP marker sets with the highest values for the SNP-P set and lowest values for the SNP-S set, except for major allele frequency, where the reverse held true.

The trends reported for all parameters mentioned above were confirmed by AMOVA (Table 2). The proportion of variance component within dent s_D^2 was greater than within flint s_F^2 with all SNP sets, as revealed by the significance of the contrast dent versus flint. The proportion of s_D^2 was higher with SNP-S markers (44.0 %) than with the other marker sets, and s_F^2 was lower with all the SNP marker sets than with AFLPs and SSRs. The ratio of the within-group variances s_F^2/s_D^2 was highest for AFLPs and SSRs (95.1 and 95.0 %) and lowest for the SNP-S set (85.0 %). The proportion of s_B^2 was lower with AFLP (19.3 %) and SNP-S (18.6 %) than with the other marker sets.

Correlations of MRD values based on AFLPs versus those based on SNP-A and SNP-S were significantly higher for the DD pairs of lines than for the FF pairs of lines (Table 3). For the remaining combinations of marker sets (i.e., AFLP vs. SSR, AFLP vs. SNP-P, and SSR vs. all SNP sets), however, no significant differences were found between dent and flint. Correlations of MRD calculated with different marker sets for the DF, i.e., pairs of lines from different groups, were always significantly lower than those for pairs from the same groups, i.e., DD and FF, with the lowest values for AFLP versus SSR (24.1 %) and AFLP versus SNP-S (24.3 %). Differences in the degree of association between MRD for FF, DD, and DF line combinations were also illustrated in the graphs of MRD for AFLP versus SNP-P, AFLP versus SNP-S, and SNP-P versus SNP-S (Fig. 1). The distributions of the DD and FF distances were more similar for AFLP versus SNP-P (Fig. 1a) than for AFLP versus SNP-S (Fig. 1b), as evident from a higher degree of overlapping between red and blue points. Figure 1b shows the comparably low correlation (24.3 %) of AFLP versus SNP-S for MRD involving one dent and one flint line (i.e., DF). The high correlations of SNP-P versus SNP-S for MRD, 98.3 % for DD, 98.1 % for FF, and 69.1 % for DF, are illustrated by Fig. 1c. The PCoA using MRD for each marker set resulted in similar

Table 2 Analyses of molecular variance (AMOVA) obtained for the different marker sets

Marker set[a]	Proportion of			Hypothesis	Ratio (%)
	s_B^2 (%)	s_D^2 (%)	s_F^2 (%)	Dent ≠ flint	Flint/dent
AFLP	19.3 *a*	41.4 *abc*	39.3 *a*	ns	95.1 *a*
SSR	21.1 *bc*	40.5 *a*	38.5 *a*	ns	95.0 *a*
SNP-A	20.5 *b*	42.2 *b*	37.2 *b*	*	88.2 *b*
SNP-P	21.2 *c*	41.6 *c*	37.2 *b*	*	89.4 *c*
SNP-S	18.6 *a*	44.0 *d*	37.4 *b*	**	85.0 *d*

Variance components corresponding to the genetic variation between (s_B^2) and within the groups of dent (s_D^2) and flint (s_F^2) lines are shown as proportion of the total variance

Values followed by the same letter within a column do not significantly differ at $P \leq 0.05$, based on bootstrapping

ns, *, **: non-significant, significant at $P \leq 0.05$ and $P \leq 0.01$, respectively, based on bootstrapping

[a] SNP-A refers to all 41,434 MaizeSNP50 markers, SNP-P to the subset of 30,068 markers from Panzea and SNP-S to the subset of 11,366 markers from Syngenta

Table 3 Correlation coefficient (*r*) between modified Roger's distances (MRD) for pairs of dent lines (DD), flint lines (FF) and dent by flint line combinations (DF) calculated with different marker sets

Marker sets[a]		DD r_{DD}	FF r_{FF}	DD versus FF	DF r_{DF}	(DD, FF) versus DF
AFLP	SSR	90.5**	86.1**	ns	24.1*	**
AFLP	SNP-A	95.7**	91.8**	*	37.2**	**
AFLP	SNP-P	95.4**	91.6**	ns	39.2**	**
AFLP	SNP-S	95.3**	91.2**	*	24.3*	**
SSR	SNP-A	93.5**	91.7**	ns	43.7**	**
SSR	SNP-P	92.6**	90.4**	ns	44.5**	**
SSR	SNP-S	94.5**	93.8**	ns	32.5**	**

ns, *, **: non-significant, significant at $P \leq 0.05$ and $P \leq 0.01$ respectively, based on bootstrapping

[a] SNP-A refers to all 41,434 MaizeNP50 markers, SNP-P to the subset of 30,068 markers from Panzea and SNP-S to the subset of 11,366 markers from Syngenta

grouping patterns of the inbred lines (Fig. 2, supplemental Fig. S1). The first principal coordinate explained from 20.0 to 22.8 % of the molecular variation, and the second principal coordinate explained from 10.2 to 11.2 %. For all marker sets, the first coordinate clearly separated the group of flint lines from the dent lines. Comparison of the PCoA patterns by means of Procrustes analysis resulted in correlations from 0.96 to 0.98 between the patterns (data not shown).

Discussion

Well-founded knowledge about the genetic diversity in a breeding program is crucial for (1) early diagnosis of genetic narrowing of heterotic pools and (2) design of efficient strategies for broadening them. It has been acknowledged that maize germplasm needs to be broadened to assure future gains in yield performance (Mikel and Dudley 2006). For this purpose, reliable measures of the genetic diversity in elite germplasm might help in identifying genomic regions in exotic germplasm or landraces that may support resolving genetic bottlenecks. These regions may contribute to increased yield performance and stability of elite germplasm. Moreover, assessment of the genetic variation and structure of diversity panels of lines represent important information for genetic analyses and identification of quantitative trait loci by means of association mapping (Mezmouk et al. 2011). In addition, reliable estimations of genetic relationships among genotypes can assist breeders in the selection of diverse parental lines in recycling breeding programs (Mikel and Dudley 2006; Lu et al. 2009). While the assignment of lines to heterotic groups requires only a rough scale of genotypic classification (Lu et al. 2009), reliable estimates of genetic distances among lines are crucial for the implementation of genomic selection, which is expected to take advantage of high-density SNP arrays (Jannink et al. 2010). The ability of a set of markers to reveal an exact picture of the genetic diversity can be impaired by the ascertainment bias (Clark et al. 2005). This bias can affect any set of markers, when they are screened and optimized in a certain panel of germplasm and then used for characterizing materials with different allele frequencies.

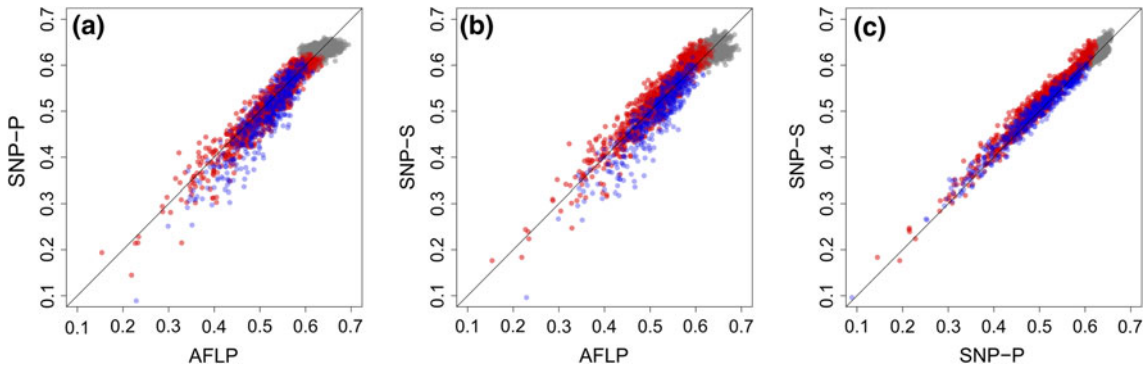

Fig. 1 Correlations between genetic distances based on different marker sets. Each point represents the modified Roger's distance (MRD) between pairs of inbred lines, for **a** AFLP versus SNP-P, **b** AFLP versus SNP-S, and **c** SNP-P versus SNP-S markers. AFLP refers to 731 mapped markers, SNP-P to the subset of MaizeSNP50 with 30,068 markers from Panzea, and SNP-S to the subset with 11,366 markers from Syngenta. Distances between pairs of *dent lines* are highlighted in *red*, between pairs of *flint lines* in *blue*, and between pairs of *dent and flint lines* in *gray*

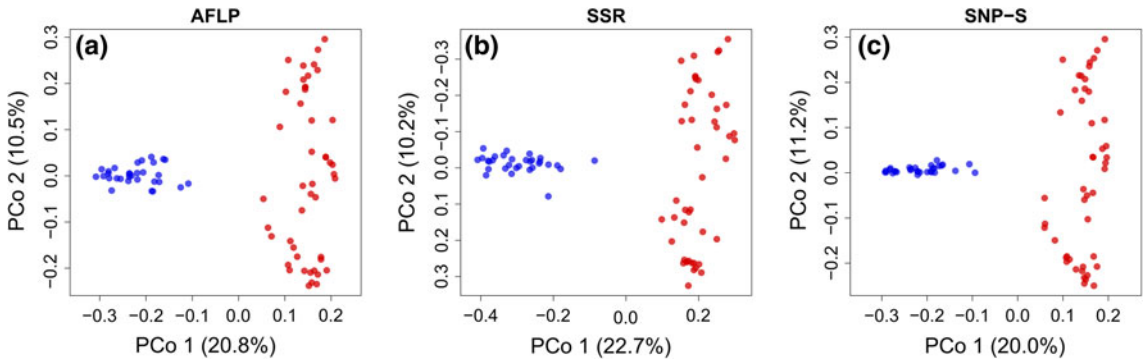

Fig. 2 Principal coordinate analysis of 77 maize inbred lines based on modified Roger's distance calculated from **a** 731 mapped AFLP, **b** 186 SSR, and **c** 11,366 SNP-S markers. Genotypes were assigned to subgroups according to their breeding group: *dent inbred lines* are represented in *red* and *flint inbred lines* in *blue*. PCo 1 and PCo 2 are the first and second principal coordinates, respectively, and *numbers* in *parentheses* refer to the proportion of variance explained by these principal coordinates

AFLP markers identify DNA polymorphisms by amplification of random genomic DNA fragments and simultaneously screen different DNA regions distributed randomly throughout the whole genome (Vos et al. 1995). Since AFLPs are mainly optimized with regard to employed restriction enzymes and adaptors, but not with regard to a specific set of lines, they are expected to yield an unbiased picture of the true genetic diversity among genotypes and can serve as a reference to test ascertainment bias in other kinds of markers such as SNPs. We are aware that AFLPs may underestimate genetic distances due to homoplasy, when identical bands in pairs of profiles correspond to different fragments, and due to collision, when different fragments of equal length appear as a single band within a profile (Gort et al. 2009). However, homoplasy should be minimized for AFLPs that had been mapped (Vekemans et al. 2002), and particularly if the maps integrated information from numerous crosses like in the case of AFLP markers chosen for the present study (Peleman et al. 2000).

Commonly, SSRs are identified through the screening of genomic libraries or by searching DNA sequence databases. SSRs can suffer from homoplasy, as already described for AFLPs, and from heteroplasy, when the same sequence is associated with different SSR lengths (Rafalski and Tingey 2008). As pointed out by these authors, homoplasy is especially important when analyzing genetically more distant germplasm sets, whereas heteroplasy may extend apparent genetic distances in more closely related individuals. The maize SSRs employed in this study were identified mainly by using the intermated B73 × Mo17 (IBM) mapping population (Sharopova et al. 2002) as a reference and for this reason ascertainment bias cannot be completely ruled out for these markers when they are used to characterize very distant germplasm (Chen et al. 2002). However, at least this latter limitation should be mitigated in SSR markers because of their multiallelic nature (Hamblin et al. 2007).

The Illumina MaizeSNP50 array employed in this study was established combining a large amount of DNA sequence information from different sources. Among the SNPs on this array, about 70 % were discovered within the Panzea project (Zhao et al. 2006; http://www.panzea.org),

about 25 % were provided by Syngenta, and the remaining 5 % were developed using sequence information from lines of diverse germplasm pools (Ganal et al. 2011). The Syngenta SNP markers were originally developed to maximize the difference among elite inbred lines, belonging to the North American dent pool, with special emphasis on B73 and Mo17 (Ganal et al. 2011). The remaining markers were developed within other projects, and those derived from the Panzea project were assembled on the basis of a larger population comprising diverse sets of maize and teosinte inbreds (Wright et al. 2005). This particular characteristic of the MaizeSNP50 array qualifies it for diversity assays in a broad range of maize genotypes.

Nevertheless, before embarking on large-scale screening of germplasm with the Illumina MaizeSNP50 array, it is prudent to test this marker set for a potential ascertainment bias. One way to empirically assess the degree and relevance of bias is by comparing the genetic diversity measures obtained with different marker sets on germplasms with differing allele frequencies. In our study, we used AFLP and SSR markers as reference for this purpose. Similar approaches were used in maize (Hamblin et al. 2007, Lu et al. 2009) for comparison of different sets of SNP markers, indicating that some bias cannot be ruled out. We provide an estimate of the relative bias of different marker subsets included in the MaizeSNP50 array, by analyzing two well-established heterotic groups of European maize and comparing the result with those obtained with AFLP and SSR markers.

For the European inbred lines considered in the present study, neither AFLPs nor SSRs revealed differences between the flint and dent lines for allele number and gene diversity. In contrast, with the SNP sets, the two groups often differed from each other. In particular, the frequency of major alleles for the SNP sets was generally higher for the flint lines than for the dent lines, while allele number and gene diversity were higher for the dent lines. These findings indicate that SNP markers were less discriminative and polymorphic in comparison with AFLPs and SSRs for the flint germplasm. Under the hypothesis of a bias, rare alleles are missed in the non-ascertained population (e.g., the flint lines) and, thus, markers which are polymorphic show a high frequency of major alleles, many markers become monomorphic and the genetic diversity is underestimated (Clark et al. 2005). Hence, our observations are consistent with the hypothesis of ascertainment bias in the SNPs considered, when used for genotyping European flint inbred lines.

Our results from AMOVA supported this conclusion, in that the ratio s_F^2/s_D^2 was lower for all SNP marker sets than for AFLPs and SSRs, whereas the corresponding proportion of the genetic variance between groups was higher (with the exception of SNP-S). This is in agreement with the findings of Albrechtsen et al. (2010) and Clark et al. (2005), who pointed out that in the presence of ascertainment bias, the proportion of variation between groups tends to increase as a consequence of an underestimation of the genetic variance within groups. Interestingly, this was not observed for the SNP-S, due to an overestimation of the proportion of s_D^2. Information about the breeding history of the germplasm provides an explanation for this result. Our dent lines trace back to numerous sources from North America used at the beginning of hybrid breeding in Europe in the 1950s (Fischer et al. 2008). Even during the last decades, new dent inbreds were often extracted by selfing late-maturing dent hybrids cultivated in Southern Europe (Technow et al. 2012). Since the SNP-S markers were developed in material with the same or a similar background (Ganal et al. 2011), they were optimized toward revealing the genetic variation in this germplasm, resulting in an overestimation of genetic diversity when used for the dent population considered. The ascertainment bias was almost negligible for the SNP-P marker set, likely because they were developed on the basis of a very diverse set of maize and teosinte inbred lines (Wright et al. 2005), thus resulting in lower bias with respect to a wider maize germplasm, including the European populations investigated.

Notwithstanding the observed bias, we generally found a very good agreement between the MRD obtained with different marker systems for line combinations within heterotic pools (DD, FF). Correlations between marker sets observed for MRD of DF line combinations were low (Table 3; Fig. 1). This was in agreement with the results of other studies (Hamblin et al. 2007; Jones et al. 2007) reporting that measures of distance based on different marker sets were well correlated only for closely related individuals (Jones et al. 2007), because in this case common marker alleles are mostly identical by descent. Thus, for choice of parent lines in recycling breeding programs, the ascertainment bias would practically have no impact. On the other hand, genetic distances between unrelated genotypes have not proven promising for choice of parents for superior hybrid combinations (Melchinger 1999) caused by only very weak relationships between genetic distances and heterosis due to differences in the linkage disequilibrium between markers and quantitative trait loci in different heterotic pools (Charcosset and Essioux 1994).

Despite some ascertainment bias introduced primarily by SNP-S, the PCoA and Procrustes analyses for the different marker systems resulted in similar graphical representations of the structure in the dent and flint populations. A similar conclusion was reached for barley (Moragues et al. 2010; Hübner et al. 2012). Accordingly, development of SNPs from a selected set of genotypes most likely introduces a bias against low-frequency SNPs in unrelated

materials. This applies especially when comparing materials that have evolved separately over a long period of time or originated even from different taxa (Rafalski 2011; Hübner et al. 2012).

Determination of the extent to which rare variants contribute to genetic distances and to trait variation is presently a major goal of genome-resequencing projects. Advances in next-generation technologies have reduced the costs of DNA sequencing and genotyping by sequencing (GBS) is affordable nowadays even for species with large genomes (Elshire et al. 2011). The consequent adoption of GBS could allow bypassing ascertainment bias. However, although GBS is fairly straightforward for small genomes, an enrichment of target regions or reduction of genome complexity (Gore et al. 2009) is necessary to ensure sufficient overlap in sequence coverage for species with large genomes. In addition, cleaning of sequence data in GBS and the use of imputation, i.e., a statistical technique to predict unobserved genotypes, still present problems, so that rare allele identification may be still a challenge.

In conclusion, our results revealed a mild effect of ascertainment bias in Illumina MaizeSNP50 array, when used to characterize European flint and dent elite inbred lines. The bias was appreciable in flint lines and for the SNP-S markers, which were specifically developed to maximize differences among US germplasm. While the bias affected the population-genetic parameters of the two germplasm pools, it did not substantially alter the information concerning distances between inbred lines within each heterotic group. For these reasons, we conclude that the MaizeSNP50 array considered here can be employed for breeding purposes in the investigated material. However, attention should be paid in comparisons between genotypes belonging to different heterotic groups. In this case, we recommend restricting the population-genetic analyses to marker subsets with no or only low ascertainment bias, like the SNP-P marker set in our study.

Acknowledgments We thank H.-P. Piepho and K. J. Schmid for their useful comments and discussion and the two anonymous reviewers for their valuable suggestions. This research was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr Synbreed—Synergistic plant and animal breeding (FKZ: 0315528d). E. Frascaroli was supported by the Deutscher Akademischer Austauschdienst (DAAD) grant A/11/04103.

References

Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evo 27:2534–2547

Charcosset A, Essioux L (1994) The effect of population-structure on the relationship between heterosis and heterozygosity at marker loci. Theor Appl Genet 89:336–343

Chen X, Cho YG, McCouch SR (2002) Sequence divergence of rice microsatellites in Oryza and other plant species. Mol Genet Genomics 268:331–343

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15:1496–1502

Dray S, Dufour AB (2007) The ade4 package: implementing the duality diagram for ecologists. J Stat Softw 22:1–20

Efron, B, Tibshirani RJ (1993). An introduction to the bootstrap. Chapman and Hall, London, p 214

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6(5):e19379. doi:10.1371/journal.pone.0019379

Excoffier L, Smouse P, Quattro J (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491

Fischer S, Möhring J, Schön CC, Piepho HP, Klein D, Schipprack W, Utz HF, Melchinger AE, Reif JC (2008) Trends in genetic variance components during 30 years of hybrid maize breeding at the University of Hohenheim. Plant Breed 127:446–451

Ganal MW, Durstewitz G, Polley A, Berard A, Buckler ES, Charcosset A, Clarke JD, Graner EM, Hansen M, Joets J, Le Paslier MC, McMullen MD, Montalent P, Rose M, Schön CC, Sun Q, Walter H, Martin OC, Falque M (2011) A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PLoS ONE 6:e28,334

Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES et al (2009) A first generation haplotype map of maize. Science 326:1115–1117

Gort G, van Hintum T, van Eeuwijk F (2009) Homoplasy corrected estimation of genetic similarity from AFLP bands, and the effect of the number of bands on the precision of estimation. Theor Appl Genet 119:397–416

Gower JC (1966) Some distance properties of latent root and vector methods in multivariate analysis. Biometrika 53:325–338

Hamblin MT, Warburton ML, Buckler ES (2007) Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. PLoS ONE 2:e1367

Hübner S, Günter T, Flavell A, Fridman E, Graner A, Korol A, Schmid KJ (2012) Islands and streams: clusters and gene flow in wild barley populations from the Levant. Mol Ecol. doi:10.1111/j.1365-294X.2011.05434.x

Jackson D (1995) PROTEST: a PROcustean randomization TEST of community environment concordance. Ecosciences 2:297–303

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9:166–177

Jones E, Sullivan H, Bhattramakki D, Smith J (2007) A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize *Zea mays* L. Theor Appl Genet 115:361–371

Liu K, Muse S (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129

Lu Y, Yan J, Guimarães CT, Taba S, Hao Z, Gao S, Chen S, Li J, Zhang S, Vivek BS, Magorokosho C, Mugo S, Makumbi D, Parentoni SN, Shah T, Rong T, Crouch JH, Xu Y (2009) Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. Theor Appl Genet 120:93–115

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220

Melchinger AE (1999) Genetic diversity and heterosis. In: Coors JG, Pandey S (eds) The genetics and exploitation of heterosis in crops. ASA-CSSA, Madison, WI, pp 99–118

Mezmouk S, Dubreuil P, Bosio M, Décousset L, Charcosset A, Praud S, Mangin B (2011) Effect of population structure corrections on the results of association mapping tests in complex maize diversity panels. Theor Appl Genet 122:1149–1160

Mikel MA, Dudley JW (2006) Evolution of North American dent corn from public to proprietary germplasm. Crop Sci 46: 1193–1205

Moragues M, Comadran J, Waugh R, Milne I, Flavell AJ, Russell JR (2010) Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. Theor Appl Genet 120:1525–1534

Nielsen R, Signorovitch J (2003) Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. Theor Popul Biol 63:245–255

Paradis E (2010) pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 26:419–420

Peleman J, van Wijk R, van Oeveren J, van Schaik R (2000) Linkage map integration: an integrated genetic map of *Zea mays* L. Poster P472. Plant and animal genome conference VIII, San Diego

R Development Core Team (2011) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria, http://www.R-project.org/, ISBN 3-900051-07-0

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100

Rafalski JA (2011) Genomic tools for the analysis of genetic diversity. Plant Genet Res Charact Util 9:159–162

Rafalski A, Tingey S (2008) SNPs and their use in maize. In: Henry RJ (ed) Plant genotyping II- SNP technology. CABI, Wallingford, Oxfordshire, UK; Cambridge, MA, pp 30–43

Ramìrez-Soriano A, Nielsen R (2009) Correcting estimators of and Tajima's D for ascertainment biases caused by the single-nucleotide polymorphism discovery process. Genetics 181:701–710

Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL et al (2009) Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS ONE 4:e6524

Schlötterer C (2004) The evolution of molecular markers—just a matter of fashion? Nat Rev Genet 5:63–69

Schrag TA, Melchinger AE, Sørensen AP, Frisch M (2006) Prediction of single-cross hybrid performance for grain yield and grain dry matter content in maize using AFLP markers associated with QTL. Theor Appl Genet 113:1037–1047

Schrag TA, Möhring J, Melchinger AE, Kusterer B, Dhillon BS, Piepho H-P, Frisch M (2010) Prediction of hybrid performance in maize using molecular markers and joint analyses of hybrids and parental inbreds. Theor Appl Genet 120:451–461

Seeb JE, Carvalho G, Hauser L, Naish K, Roberts S, Seeb LW (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in non-model organisms. Mol Ecol Res 11(Suppl 1):1–8

Sharopova N, McMullen MD, Schultz L, Schroeder S, Sanchez-Villeda H, Gardiner J, Bergstrom D, Houchins K, Melia-Hancock S, Musket T, Duru N, Polacco M, Edwards K, Ruff T, Register JC, Brouwer C, Thompson R, Velasco R, Chin E, Lee M, Woodman-Clikeman W, Long MJ, Liscum E, Cone K, Davis G, Coe EH Jr (2002) Development and mapping of SSR markers for maize. Plant Mol Biol 48:463–481

Technow F, Riedelsheimer C, Schrag TA, Melchinger AE (2012) Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet. doi:10.1007/s00122-012-1905-8

Vekemans X, Beauwens T, Lemaire M, Roldan-Ruiz I (2002) Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. Mol Ecol 11:139–151

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

Weir BS (1996) Genetic data analysis II. Sinauer Associates, Inc., Sunderland, MA

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370

Wright S (1978) Evolution and the genetics of populations. Variability within and among natural populations, vol 4. University of Chicago Press, Chicago, IL

Wright SI, Vroh Bi I, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS (2005) The effects of artificial selection on the maize genome. Science 308:1310–1314

Zhao W, Canaran P, Jurkuta R, Fulton T, Glaubitz J, Buckler E, Doebley J, Gaut B, Goodman M, Holland J, Kresovich S, McMullen M, Stein L, Ware D (2006) Panzea: a database and resource for molecular and functional diversity in the maize genome. Nucleic Acids Res 34:D752–D757