Genetica
DOI 10.1007/s10709-011-9606-9

Comparison of SSRs and SNPs in assessment of genetic relatedness in maize

Xiaohong Yang · Yunbi Xu · Trushar Shah · Huihui Li · Zhenhai Han · Jiansheng Li · Jianbing Yan

Received: 7 December 2009 / Accepted: 25 August 2011

Abstract Advances in high-throughput SNP genotyping and genome sequencing technologies have enabled genome-wide association mapping in dissecting the genetic basis of complex quantitative traits. In this study, 82 SSRs and 884 SNPs with minor allele frequencies (MAF) over 0.20 were used to compare their ability to assess population structure, principal component analysis (PCA) and relative kinship in a maize association panel consisting of 154 inbred lines. Compared to SNPs, SSRs provided more information on genetic diversity. The expected heterozygosity (He) of SSRs and SNPs averaged 0.65 and 0.44, and the polymorphic information content of these two markers was 0.61 and 0.34 in this panel, respectively. Additionally, SSRs performed better at clustering all lines into groups using STRUCTURE and PCA approaches, and estimating relative kinship. For both marker systems, the same clusters were observed based on PCA and the first two eigenvectors accounted for similar percentage of genetic variations in this panel. The correlation coefficients of each eigenvector from SSRs and SNPs decreased sharply when the eigenvector varied from 1 to 3, but kept around 0 when the eigenvector were over 3. The kinship estimates based on SSRs and SNPs were moderately correlated ($r^2 = 0.69$). All these results suggest that SSR markers with moderate density are more informative than SNPs for assessing genetic relatedness in maize association mapping panels.

Keywords SSR · SNP · Population structure · Kinship · PCA

X. Yang · J. Li · J. Yan (✉)
National Maize Improvement Center of China, Beijing Key Laboratory of Crop Genetic Improvement, China Agricultural University, No 2 Yuanmingyuan West Road, Haidian, Beijing 100193, China
e-mail: yjianbing@gmail.com

Y. Xu
International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico

T. Shah
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Hyderabad, Andhra Pradesh 502324, India

H. Li
Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing 100081, China

Z. Han
College of Agriculture and Biotechnology, China Agricultural University, No 2 Yuanmingyuan West Road, Beijing 100193, China

J. Yan
National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, 430070 Wuhan, China

Introduction

Genome-wide association studies (GWAS) have been widely applied in identifying the causal variants affecting complex disease in humans (Altshuler et al. 2008). The availability of complete genome sequences in some model species, the advances in rapid and cost-effective genotyping technologies and the development of statistical methods have allowed association mapping (AM) to be a powerful tool to dissect the genetic basis of quantitative traits in plants (Yu and Buckler 2006; Buckler and Gore 2007). Two types of strategies are used in association studies: candidate-gene association (CGA) and GWA (Yang et al. 2007; Zhu et al. 2008). CGA is a hypothesis-driven approach that surveys the polymorphisms in

Published online: 09 September 2011

selected candidate genes associated with phenotypic variation (Mackay 2001). Candidate genes are selected based on the knowledge of metabolic pathways, linkage analysis, expression profile and biochemistry. This approach has recently been widely applied to identify the functional variations in plants (Zhu et al. 2008). Due to the lack of genomic data and the high cost of genotyping, only a few GWAS were conducted in plants regardless of its wide application in humans. Recently, the GWAS have been performed using medium level-density markers in Arabidopsis (Aranzana et al. 2005; Zhao et al. 2007) and maize (Beló et al. 2008). However, the rapid development of genomic technologies made the large-scale GWAS available in marker-trait associations. Up to now, more than 30 commercial single nucleotide polymorphisms (SNPs) detection platforms were developed (Gupta et al. 2008). In addition, several high-density platforms are now available that can simultaneously genotype up to 384 DNA samples across 96–1 M SNPs (Gupta et al. 2008). One million SNPs were developed in Arabidopsis by resequencing 20 wide diverse accessions, among which 250,000 nonredundant SNPs were used to genotype 1,000 accessions for GWAS (Clark et al. 2007), and more recently, a real GWAS was conducted in Arabidopsis (Chan et al. 2009). In maize, 27 foundation lines of nested AM were genotyped using next-generation sequencing technology and over one million SNPs were developed (Gore et al. 2009).

During the past two decades, linkage analysis has been well-developed and a number of friendly softwares were available, such as WinQTLCart (Wang et al. 2005). However, the methods for AM are still at exploration stages due to its recent application in plants. The presence of genetic relatedness in an association panel, often generating spurious associations (Yu and Buckler 2006; Yu et al. 2006; Yang et al. 2007), is one of the key factors affecting the application of this statistical analysis. Currently, most statistical methods focus on how to exclude the effects of genetic relatedness for AM, and the first generation of methods are genome control (GA) and structure association (SA). GA method uses random markers to estimate and adjust the effects of population structure, assuming that such a population structure has similar effects at all loci (Zheng et al. 2005; Devlin and Roeder 1999). SA analysis uses random markers to estimate population structure (Q matrix) by the program STRUCTURE (Pritchard et al. 2000; Falush et al. 2003) and then incorporates it into further statistical analysis. However, STRUCTURE assumes the individuals in a population are unrelated and all loci within a population are at Hardy–Weinberg equilibrium. In real situations, few data agree well with this assumption. Furthermore, estimating population structure using STRUCTURE is computationally intense. Recently, principal component analysis (PCA) has been suggested to infer population structure as it is fast, makes no assumptions of populations and loci (Price et al. 2006; Zhu and Yu 2009). The PCA method infers the observed variations across all markers into a few variables, which were used to analyze the relationships among individuals in association panels. However, both STRUCTURE and PCA approaches may not well capture the differences between individuals as most individuals have complex relatedness that cannot be described by a few axes of differentiation (Myles et al. 2009). An alternative is to use random molecular makers to estimate pairwise relatedness between all individuals (K matrix) in a population that can be incorporated into the mixed-liner model (MLM) to correct for relatedness in AM (Yu et al. 2006). The MLM method including Q and K or PCA and K was successfully applied in AM in plants, such as maize (Yu et al. 2006; Harjes et al. 2008), wheat (Breseghello and Sorrells 2006), sorghum (Murray et al. 2009), Arabidopsis (Zhao et al. 2007) and potato (Malosetti et al. 2007).

For all statistical methods mentioned above, random markers, typically including SSRs and SNPs, were used to assess genetic relatedness. Because SSR markers are reproducible, PCR-based and informative (Smith et al. 1997), they play a predominant role in evaluating genetic diversity and relatedness in plants (Liu et al. 2003; Reif et al. 2006; Thomson et al. 2007). However, the detection of SSR genotypes is often conducted using agarose gel or polyacrylamide gel or sequencers, which is time consuming or costly. Furthermore, SNPs have a lower error rate compared with SSRs (Jones et al. 2007). SNPs have now become an ideal marker system that can be used in the same manner as other genetic markers for a variety of functions in crop improvement, including linkage map construction, genetic diversity analysis, marker-trait association and marker-assisted selection. Yan et al. (2010) compared the two marker systems in constructing linkage maps and found that an array-based SNP detection method was 100 of times faster than gel-based SSR detection method and cost was 4–5 times lower. Hamblin et al. (2007) compared the ability of SSRs and SNPs in assessment of population structure using 89 SSRs and 847 SNPs, and found that SSRs performed better at clustering individuals into populations than SNPs, but that the population structure assessed by both marker systems was consistent.

Recently, we have developed a maize association panel consisting of 155 inbred lines, which was genotyped using 82 random SSRs and 1,536 SNPs throughout the genome (Yang et al. 2010). In this study, the ability of SSRs and SNPs in assessment of population structure (Q), relative kinship (K) and PCA was compared to provide information for choosing the marker systems in evaluating genetic relatedness to correct spurious associations.

Materials and methods

Maize association panel

A set of 155 diverse maize inbred lines was used in this study: 35 high-oil lines mainly selected from American and Chinese high-oil populations (Song and Chen 2004), 91 inbred lines from the parents of commercial hybrids used widely in China in the past two decades (Teng et al. 2004), 25 inbred lines developed from landraces in China and four high provitamin A lines introduced from Illinois University in the United States. Only 154 lines were used for subsequent analysis as the SNP data of one line was missing. The detailed list and pedigree information can be found in previous studies (Yan et al. 2010; Yang et al. 2010).

SSR genotyping

A set of 82 SSRs evenly distributed throughout the maize genome was used to genotype all 155 lines. The details of SSR list and genotyping were described by Yang et al. (2010). Most of these markers were in previous studies of genetic diversity and population structure in maize (Liu et al. 2003). Among 82 SSRs, 43.9% makers were di-nucleotide repeats, 22.0% tri-nucleotide repeats and the remaining over tetra-nucleotide repeats.

SNP genotyping

The details of SNP genotyping were described in previous study (Yan et al. 2010). Briefly, GoldenGate assay (Illumina, San Diego, CA) containing 1,536 SNPs (http://www.panzea.org) was applied to genotype 154 lines. The SNP genotyping was performed on Illumina BeadStation 500G (Illumina, San Diego, CA) at Cornell University Life Sciences Core Laboratories Center with the protocol supported by Illumina company (Fan et al. 2006). Eight hundred and eighty-four SNPs with MAF over 0.20 and of good quality were used for further analysis.

Summary statistics analysis

PowerMarker Version 3.25 (Liu and Muse 2005) was used to calculate allele number, expected He, observed He, PIC, genetic distance based on allele sharing using different marker classes: 82 SSRs; 884 SNPs; 82 SSRs + 884 SNPs. The same marker classes were used for subsequent analysis.

Population structure analysis

A model-based program STRUCTURE 2.2 (Pritchard et al. 2000; Falush et al. 2003) was used to infer genetic relationship among individual genotypes from 154 lines. This model assumed that the number of populations was k, and the loci were independent and at Hardy–Weinberg equilibrium. Three independent runs were done by setting the number of populations (k) from 1 to 10, burn in time and Markov chain Monte Carlo (MCMC) replication number both to 500,000, and a model for admixture and correlated allele frequencies. Both LnP(D) in STRUCTURE output and its derived Δk (Evanno et al. 2005) were used to determine the k value. Lines with membership probabilities $\geq$0.75 were assigned to given clusters; lines with membership probabilities $<$0.75 were assigned to a mixed group.

To further investigate the appropriate number of SSR and SNP markers for estimating population structure, a random re-sampling approach was used to generate marker sets with ten repetitions. The number of SSR markers randomly re-sampled was 10, 20, 30, 40, 50, 60, 70 and 80, while that of SNP markers was 100, 200, 300, 400, 500, 600, 700 and 800. Three independent runs were performed for each marker sets by using STRUCTURE software. The PROC CORR in SAS Version 8.02 was used to calculate the correlation coefficients of membership probabilities for all marker sets.

Principal component analysis

PowerMarker Version 3.25 (Liu and Muse 2005) was used to create Nei's genetic matrices (Nei 1972). Distance matrices were double-centered, and used to obtain eigenvectors by the modules DCENTER and EIGEN implemented in NTSYSpc Version 2.1 (Rohlf 2000). Combining the population structure from STRUCTURE using SSR markers (Yang et al. 2010), the 2-D plots were obtained using the first two eigenvectors. To compare the ability of three marker types in performance of PCA, correlation coefficients were calculated for marker type pair at each eigenvector using PROC CORR in SAS Version 8.02.

Kinship analysis

The relative kinship was calculated by SPAGeDi software (Hardy and Vekemans 2002) with the option (Loiselle et al. 1995). All negative values between individuals were set to 0, which indicated that they were little related to each other; and the kinship matrix was multiplied by two to be integrated into the mixed model for AM (Yu et al. 2006). PROC CORR in SAS Version 8.02 was performed to

calculate the correlation coefficients of relative kinship for marker type pairs.

All the population structure (Q), principal components (PC) and kinship (K) were analyzed by SSRs and SNPs alone and combining both marker types.

Results

Statistics of SSRs and SNPs

The set of 154 maize inbred lines was genotyped using 82 SSRs and 1,536 SNPs. For SNPs, the minor allele frequencies (MAF) of 1,394 polymorphic SNPs with good quality averaged 0.26 with a range from 0.01 to 0.50. The expected He of these SNPs varied from 0.01 to 0.67 with an average of 0.36, and the PIC ranged from 0.01 to 0.59 with an average of 0.29. All 82 SSRs and the 884 SNPs with MAF over 0.20 were used to compare the performance of different marker systems on estimating genetic diversity and relatedness in the maize panel. The summary statistics of SSRs and SNPs in 154 maize inbred lines were illustrated in Table 1. Among all lines, a total number of 675 and 1,768 alleles were detected with 82 SSRs and 884 SNPs, respectively. Compared with diallelic SNPs, an average of 8.2 alleles/locus for SSRs was observed with a range from 2 to 26. The expected He of SSRs averaged 0.65 and varied from 0.27 to 0.91, and SNPs averaged 0.44 and varied from 0.32 to 0.50. The PIC of SSRs ranged from 0.25 to 0.91 with an average of 0.61, and SNPs ranged from 0.27 to 0.38 with an average of 0.34. Between any two lines, the polymorphic ratio of SSRs averaged 0.66 with a range from 0.06 to 0.86, and SNPs averaged 0.46 with a range from 0.01 to 0.68. The allele frequencies of SSRs in all lines ranged from 0.01 to 0.85, and 64.9% of SSR alleles were rare with allele frequencies lower than 0.1 (Fig. 1a). For the SNPs, all allele frequencies were lower than 0.50 and distributed evenly (Fig. 1b).

Estimation of Q matrix

The population structure of this association panel was assessed using 82 SSRs, 884 SNPs and 82 SSRs + 884 SNPs. For all marker sets, the log-likelihood value [LnP(D)] for each given k kept on increasing with the increase of k and the most significant change was observed when k was increased from one to two (Fig. 2). However, the increase rate of LnP(D) from SNP + SSR data set was the greatest, followed by SNP data set and SSR data set. (Fig. 2). For the Δk values, there was a sharp peak at $k = 2$ for both SSR and SNP data sets; but the sharp peak occurred at $k = 3$ for SSR + SNP data set (Fig. 2). According to LnP(D) and Δk values, all 154 lines were best inferred into two population groups ($k = 2$) although the Δk value at $k = 3$ was the highest. For SSR markers, the maize panel was clearly classified into non-flint (P1, 78 lines) and flint (P2, 44 lines) grain texture (Yang et al. 2010). However, only 14 lines clustered into non-flint population separated from the whole panel for SNP markers and SSR + SNP marker sets, The correlation coefficients of membership probabilities between SSRs and SNPs or SSR + SNP were mediate while the correlation

Table 1 Summary statistics of SSRs and SNPs

Markers	Loci	Alleles	Alleles/locus	He	PIC	Polymorphisms between any two lines
SSR	82	675	8.2 (2–26)	0.65 (0.27–0.91)	0.61 (0.25–0.91)	0.66 (0.06–0.85)
SNP	884	1,768	2.0 (2)	0.44 (0.32–0.50)	0.34 (0.27–0.38)	0.46 (0.01–0.68)

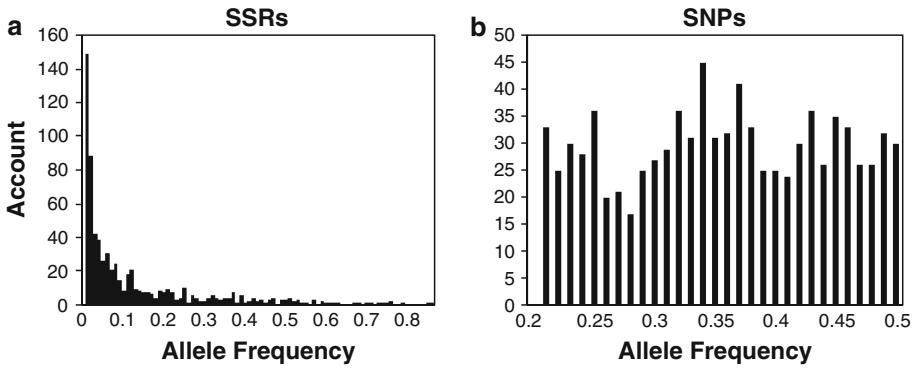

Fig. 1 Allele frequency spectra for 82 SSRs (**a**) and 884 SNPs (**b**) in 154 maize *inbred lines*

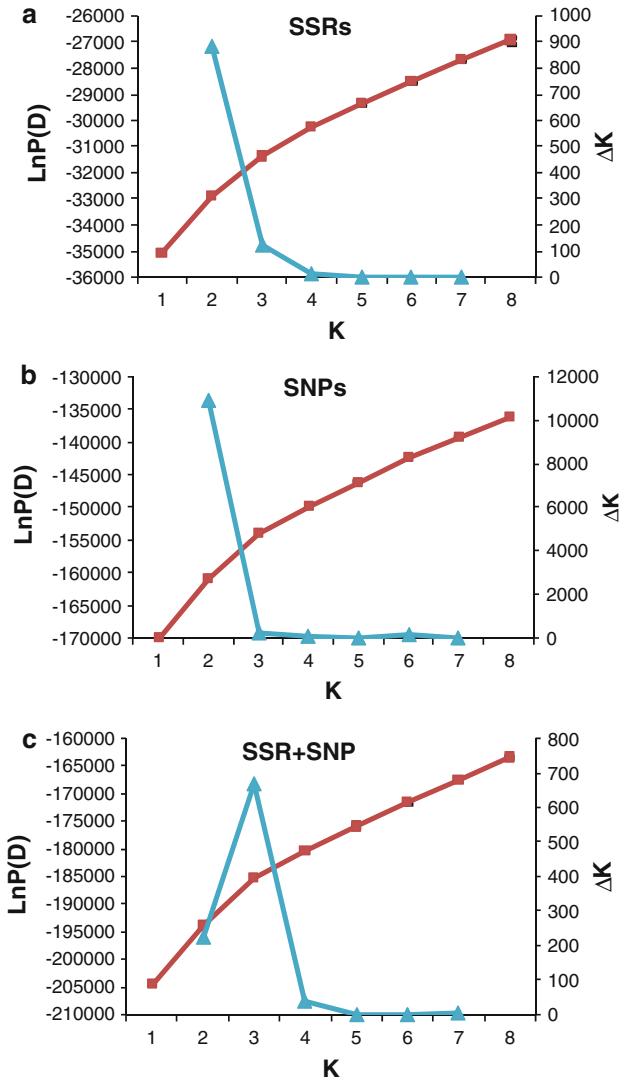

Fig. 2 Estimated LnP(D) and its derived statistics Δk for k from 1 to 8. Values of LnP(D) are from STRUCTURE run 3 times at each value of k using 82 SSRs (**a**), 884 SNPs (**b**) and 82 SSRs + 884 SNPs (**c**). The values of Δk were calculated by the equation $\Delta k = M[|L(k+1) - 2L(k) + L(k-1)|]/S[L(k)]$, where $L(k)$ represents the kth LnP(D), M is the mean of three runs, and S is the standard deviation of $L(k)$. The *diamonds* are LnP(D) and the *triangles* is Δk

coefficient is extremely high between SNPs and SNP + SSR (Table 3). Contrary to LnP(D), the assignment percentage with membership probabilities ≥0.75 for all marker sets declined from $k = 2$ to 4 and were similar from $k = 4$ to 8 (Table 2). At $k = 3$–8, the percentage of individuals assigned to populations for SSR sets was lower than that for the other two marker sets while the percentage for the SSR sets was higher than that for SNPs but slightly lower than that for SSR + SNP (Table 2).

To address the performance of various marker numbers on estimating population structure, the marker sets with eight classes were used for both SSR and SNP markers. For each pair of marker sets with various marker number of the same marker system, the correlation coefficients of membership probabilities at $k = 2$ were high while relatively lower between SSR and SNP markers (Table 3). The correlation coefficients averaged 0.91 ranging from 0.63 to 1.00 for SSR pairs with various numbers, 0.99 ranging from 0.94 to 1.00 for SNP pairs, and 0.56 ranging from 0.36 to 0.74 between SSR and SNP. For SSR markers, the percentage of individuals assigned to populations increased when the marker number increased from 10 to 70 and was similar when the marker number was 70, 80 and 82; Most of the lines assigned to given population were identity with those inferred using 82 SSRs (Table 4). For SNP markers, the percentage of individuals assigned to populations increased when the marker number increased from 100 to 300 and was similar when the marker number ranged from 300 to 884; All of the lines assigned to P1 population were identity with those inferred using 82 SSRs but over half of lines assigned to P2 population were the lines clustered in P1 population inferred using 82 SSRs when the marker number was over 300 (Table 4).

Estimation of principal components

To further investigate the population differentiation, PCA was performed using these three data sets. The first 10 eigenvectors from SSR data set accounted for 15.6, 10.6, 8.9, 8.0, 7.7, 7.0, 6.0, 5.6, 5.1 and 4.8% of the genetic variations, totaling 79.2%. For SNP or SSR + SNP data sets, the cumulative genetic variations were more than 60% when the eigenvectors reached 10, and they were similar to that for SSR when the eigenvector was 1 and 2 (Fig. 3). For all eigenvectors ≥3, the genetic variation explained by each eigenvector was the greatest for SSRs, followed by SNPs and SSR + SNP (Fig. 3).

Table 2 Percent population assignment (membership probabilities ≥0.75) based on different marker sets

Markers	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
SSRs	79.4	60.2	58.9	59.6	57.8	58.7	60.9
SNPs	73.6	43.5	34.8	27.3	26.6	34.6	31.8
SSR + SNP	82.5	46.8	45.5	37.9	40.3	41.1	41.1

Table 3 The correlation coefficients of membership probabilities estimated using different marker sets with various marker number and marker system

Marker sets[a]	SSR10	SSR20	SSR30	SSR40	SSR50	SSR60	SSR70	SSR80	SSR82	SNP100	SNP200	SNP300	SNP400	SNP500	SNP600	SNP700	SNP800	SNP884
SSR20	0.71*																	
SSR30	0.77*	0.96*																
SSR40	0.72*	0.96*	0.96*															
SSR50	0.77*	0.96*	0.97*	0.98*														
SSR60	0.63*	0.95*	0.94*	0.94*	0.95*													
SSR70	0.79*	0.95*	0.96*	0.96*	0.98*	0.95*												
SSR80	0.79*	0.95*	0.96*	0.96*	0.98*	0.96*	1.00*											
SSR82	0.78*	0.94*	0.95*	0.94*	0.97*	0.96*	1.00*	1.00*										
SNP100	0.74*	0.67*	0.70*	0.64*	0.68*	0.60*	0.69*	0.68*	0.67*									
SNP200	0.74*	0.60*	0.65*	0.58*	0.63*	0.53*	0.64*	0.63*	0.62*	0.98*								
SNP300	0.70*	0.49*	0.54*	0.46*	0.52*	0.40*	0.53*	0.53*	0.51*	0.96*	0.98*							
SNP400	0.69*	0.45*	0.50*	0.43*	0.49*	0.36*	0.50*	0.49*	0.48*	0.94*	0.97*	1.00*						
SNP500	0.70*	0.50*	0.55*	0.47*	0.53*	0.42*	0.54*	0.54*	0.53*	0.96*	0.98*	1.00*	1.00*					
SNP600	0.71*	0.51*	0.55*	0.48*	0.54*	0.43*	0.55*	0.54*	0.53*	0.96*	0.99*	1.00*	1.00*	1.00*				
SNP700	0.72*	0.54*	0.58*	0.51*	0.56*	0.46*	0.58*	0.57*	0.56*	0.97*	0.99*	1.00*	0.99*	1.00*	1.00*			
SNP800	0.71*	0.52*	0.56*	0.49*	0.55*	0.44*	0.56*	0.55*	0.54*	0.97*	0.99*	1.00*	0.99*	1.00*	1.00*	1.00*		
SNP884	0.72*	0.53*	0.57*	0.50*	0.56*	0.45*	0.57*	0.56*	0.55*	0.97*	0.99*	1.00*	0.99*	1.00*	1.00*	1.00*	1.00*	
SSR + SNP	0.73*	0.50*	0.55*	0.48*	0.54*	0.42*	0.55*	0.54*	0.53*	0.96*	0.98*	1.00*	1.00*	1.00*	1.00*	1.00*	1.00*	1.00*

* Significant at $P = 0.01$

[a] The number following the marker system is the marker number

Table 4 Assignment of individuals to given populations at $K = 2$ for different marker sets with various marker number and marker system

Marker sets	Percentage of individuals[a]	Individual numbers in P1[b]	Individual numbers in P2[c]
SSR10	29.9	26 (26)	20 (20)
SSR20	59.7	60 (58)	32 (30)
SSR30	59.7	53 (53)	39 (36)
SSR40	66.2	68 (66)	34 (33)
SSR50	72.1	70 (68)	41 (40)
SSR60	64.3	71 (70)	28 (28)
SSR70	78.6	76 (76)	45 (44)
SSR80	77.9	75 (75)	45 (44)
SSR82	79.2	78 (78)	44 (44)
SNP100	31.8	12 (12)	37 (33)
SNP200	46.8	14 (14)	58 (39)
SNP300	79.2	14 (14)	108 (41)
SNP400	82.5	14 (14)	113 (41)
SNP500	76.0	14 (14)	103 (41)
SNP600	77.3	14 (14)	105 (40)
SNP700	70.8	14 (14)	95 (40)
SNP800	76.0	14 (14)	103 (40)
SNP884	73.4	14 (14)	99 (40)
SSR + SNP	82.5	14 (14)	113 (41)

[a] The total percent of individual assignments to given populations at $k = 2$

[b] The number in the bracket indicates the number of common individuals assigned to P1 population using 82 SSRs

[c] The number in the bracket indicates the number of common individuals assigned to P2 population using 82 SSRs

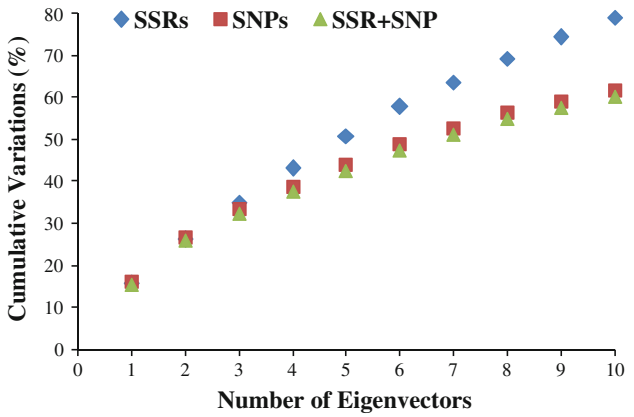

Fig. 3 The cumulative distributions of genetic variations explained by each eigenvector with a range from 1 to 10. The eigenvectors were estimated using three different marker types: 82 SSRs; 884 SNPs; 82 SSRs + 884 SNPs

Referring to the population structure from STRUCTURE analysis, two separate groups with a mixed group were observed by plotting the first two eigenvectors generated with SSR data set (Fig. 4a). The first eigenvector from SSR data set easily differentiate the P1 and P2 groups. The same three groups were seen when using only SNP data set or the full SSR + SNP data set (Fig. 4b, c). Neither of the first two eigenvectors delineated the three groups in this panel. Additionally, the distributions of all individuals were consistent in PCA eigenvector plots from SNP data set and SSR + SNP data set, but had some difference from that in PCA eigenvector plot from SSR data set. This can be further proved by the eigenvector correlation coefficients between marker type pair (Fig. 5). When the eigenvector was 1, the correlation coefficients were 0.67 ($P < 0.001$), 0.72 ($P < 0.001$), 1.00 ($P < 0.001$) for correlations between SSRs and SNPs, SSRs and SSR + SNP, and SNPs and SSR + SNP, respectively. With the increase of eigenvector numbers, the correlation coefficients between SSR and SNP or SSR + SNP decreased sharply from eigenvector 1 to 4 but kept around 0 for the remaining eigenvectors. For the first 10 eigenvectors, the correlation coefficients between SNP and SSR + SNP were significantly high and positive ($r^2 > 0.87$).

Estimation of K matrix

The 82 SSRs, 884 SNPs and 82 SSR + 884 SNP data sets were also used to evaluate kinship in this association panel. For all marker sets, the percentage of individual pairs falling in certain kinship categories showed a similar trend with the increase of kinship coefficient (Fig. 6). The kinship estimated by SNPs greatly agreed with that estimated by SNP + SSR, and the kinship estimated by SSRs and SNPs was moderately correlated ($r^2 = 0.69$, $P < 0.001$). However, the percentages of individual pairs falling in certain kinship categories were slightly different among three marker sets. Furthermore, SSRs had a higher resolution to estimate kinship than SNPs as more individual pairs were classed into kinship categories with high values.

Discussion

The features of SSR and SNP markers are associated with their mutational processes. Due to neutrality, the mutational rates of SSRs (1×10^{-5}) (Kruglyak et al. 1998) are much higher than that of SNPs (1×10^{-9}) (Li et al. 1981; Martinez-Arias et al. 2001). The number of alleles per locus, PIC, polymorphisms between any two lines and the distributions of allelic frequencies in this association panel, estimated by both marker systems, are consistent with previous studies (Vignal et al. 2002; Hamblin et al. 2007). However, the differences between the results from these two markers differed from previous studies. For example, the number of alleles per loci for SSR was 4.1 times of that for SNPs in this association panel while 10.9 times in

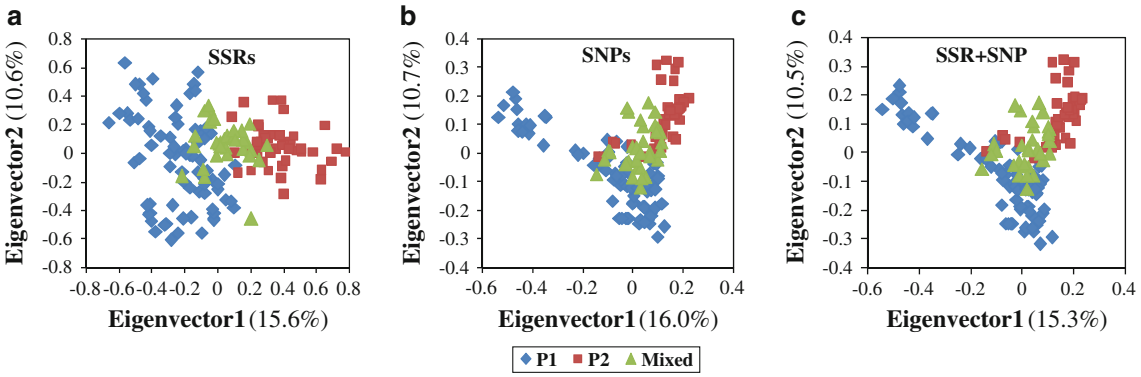

Fig. 4 PCA eigenvector plot of three groups in this association *panel*. The two eigenvectors were estimated from 82 SSRs (**a**), 884 SNPs (**b**) and 82 SSR + 884 SNP (**c**)

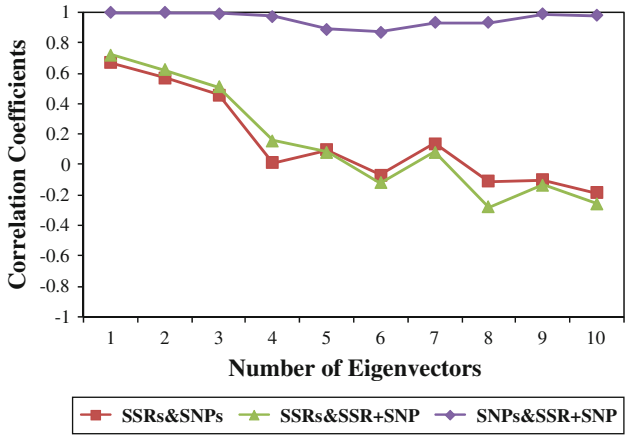

Fig. 5 The distributions of eigenvector correlation coefficients for marker type pair. The eigenvector correlation coefficients were calculated between SSRs and SNPs, SSRs and SSR + SNP, and SNPs and SSR + SNP

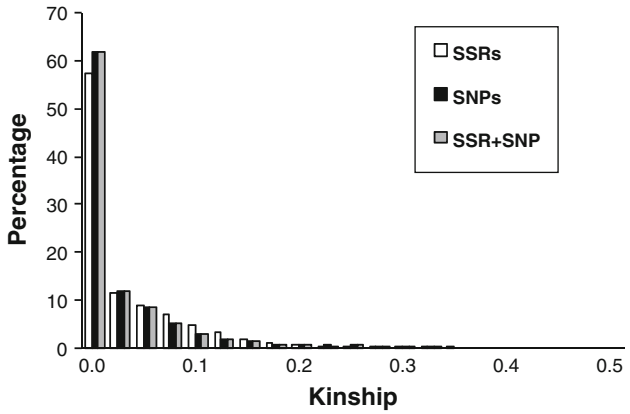

Fig. 6 Distribution of pairwise relative kinship estimates in 155 maize samples for different marker sets: 82 SSRs, 884 SNPs, and 82 SSR + 884 SNP. For simplicity, only percentages of relative kinship estimates ranging from 0 to 0.50 are shown

another association panel consisting of 260 inbred lines (Liu et al. 2003; Hamblin et al. 2007). This main reason for this deviation can be attributed to the composition of germplasm present in these two maize association panels (Yang et al. 2010).

All 154 lines were inferred into two groups using 82 SSRs based on both LnP(D) and Δk values (Yang et al. 2010). The same number of groups ($k = 2$) were obtained using 884 SNPs and SSR + SNP although the Δk value was not the highest for SSR + SNP at the true k. The lower Δk value at the true k may be due to mixed marker system as the Δk value is sensitive to marker type (Evanno et al. 2005). When k was set to 2, 20.6, 26.4 and 17.5% lines were assigned into mixed groups using SSRs, SNPs and SSR + SNP, respectively. Furthermore, only a few lines clustered in P1 population inferred using 82 SSRs separated from the whole maize panel using SNP markers and SSR + SNP markers. This was different from the previous studies presented by Hamblin et al. (2007), Inghelandt et al. (2010). The possible reasons may be due to the complex relationship of present maize lines and/or the ascertainment bias of SNP markers in this study though we have deleted the SNPs with serious bias by cutting off the SNPs with MAF less than 0.20. SSRs performed better at assessing the genetic similarity among lines due to informativeness as multiallelic markers, which can explain why there were relatively fewer lines classified into mixed groups using SSRs (Table 2) and more lines clustered into P1 population separated from the maize panel (Table 3).

The classifications of population structure in maize association panel based on PCA were consistent with those based on STRUCTURE for all marker systems. However, some differences were observed among three marker data sets. In sorghum, the same clusters were observed when using only SSRs or only SNPs or both SSRs and SNPs, but some individuals shifted the given groups (Murray et al. 2009). In this study, fewer eigenvectors from SSRs inferred

the population differentiation and explained 80% of genetic variation in this association panel, which may be explained by SSRs having more number of alleles at a given locus than SNPs . This result demonstrated that the performance of SSRs with PCA was better than that of SNPs and SSR + SNP.

The ability of marker systems in evaluating relative kinship was similar to that in population structure inferred by both STRUCTURE and PCA approaches. 57.2, 61.9 and 61.9% pairwise kinship were detected to be zero using SSRs, SNPs and SSR + SNP, respectively. Yu et al. (2009) found that kinship estimation was more sensitive to the number of markers used than population structure estimation, and the kinship estimated using 1,000 SNPs was consistent with that estimated using 100 SSRs. In agreement with Yu et al (2009), in this present study using 154 inbred lines, we found that the ability of 82 SSRs to estimate kinship was similar to that of 884 SNPs. This might be the major reason for attaining the similar distributions of kinship with three marker systems.

In summary, Q, PC and *K* were the main parameters in estimating the genetic relatedness. They can be estimated by random markers and then be used as covariances in models to control false positives in association studies (Yu et al. 2006; Price et al. 2006). All these three parameters inferred by either SSRs or SNPs were similar, although SSRs performed better than SNPs. It could be improved by increasing SNP marker density, which is easily conducted as SNP detection methods are high-throughput, cost effective, and a great number of SNPs in maize are available although the limited increasing number of SNPs in our study did not improve the inference of population structure. However, Yu et al. (2009) suggested over 10 times more SNPs than SSRs should be used to estimated relative kinship and Inghelandt et al. (2010) proposed between 7 and 11 times should be used to infer population structure in maize association analysis. Therefore, SNPs will be widely applied in maize and other species including genetic diversity analysis that will provide useful information for crop improvement in the future.

Acknowledgments This research was supported by the National High Technology Research and Development Program of China.

References

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322:881–888

Aranzana MJ, Kim S, Zhao KY, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang CL, Toomajian C, Traw B, Zheng HG, Bergelson J, Dean C, Marjoram P, Nordborg M (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genet 1(5):e60

Beló A, Zheng P, Luck S, Shen B, Meyer DJ, Li BL, Tingey S, Rafalski A (2008) Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. Mol Genet Genomics 279:1–10

Breseghello F, Sorrells ME (2006) Association analysis as a strategy for improvement of quantitative traits in plants. Crop Sci 46:1323–1330

Buckler ES, Gore M (2007) An Arabidopsis haplotype map takes root. Nat Genet 39(9):1056–1057

Chan EKF, Rowe HC, Kliebenstein DJ (2009) Understanding the evolution of defense metabolites in *Arabidopsis thaliana* using genome-wide association mapping. Genetics 109:108522

Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, Chen H, Frazer KA, Huson DH, Schölkopf B, Nordborg M, Rätsch G, Ecker JR, Weige D (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. Science 317:338–342

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Fan JB, Gunderson KL, Bibikova M, Yeakley JM, Chen J, Garcia EW, Lebruska LL, Laurent M, Shen R, Barker D (2006) Illumina universal bead arrays. Methods Enzymol 410:57–73

Gore M, Chia JM, Elshire R, Ersoz E, Hurwitz B, Grills G, Ware D, Buckler ES (2009) A first generation haplotype map of the maize genome. In: The 51st maize genetics conference abstracts, p 39

Gupta PK, Rustgi S, Mir RR (2008) Array-based high-throughput DNA markers for crop improvement. Heredity 101:5–18

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 12:e1367

Hardy OJ, Vekemans X (2002) Spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2:618–620

Harjes CE, Rocheford TR, Bai L, Brutnell TP, Kandianis CB, Sowinski SG, Stapleton AE, Vallabhaneni R, Williams M, Wurtzel ET, Yan JB, Buckler ES (2008) Natural genetic variation in *lycopene epsilon cyclase* tapped for maize biofortification. Science 319:330–333

Inghelandt DV, Melchinger AE, Lebreton C, Stich B (2010) Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor Appl Genet 120(7):1289–1299

Jones ES, Sullivan H, Bhattramakki D, Smith JSC (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

Kruglyak S, Durrett RT, Schug MD, Aquadro CF (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci USA 95:10774–10778

Li WH, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237–239

Liu KJ, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21(9):2128–2129

Liu KJ, Goodman M, Muse S, Smith JS, Buckler ES, Doebley J (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165:2117–2128

Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (*Rubiaceae*). Am J Bot 82:1420–1425

Mackay TF (2001) The genetic architecture of quantitative traits. Annu Rev Genet 35:303–339

Malosetti M, van der Linden CG, Vosman B, van Eeuwijk FA (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. Genetics 175:879–889

Martinez-Arias R, Calafell F, Mateu E, Comas D, Andrés A, Bertranpetit J (2001) Sequence variability of a human pseudogene. Genome Res 11:1071–1085

Murray SC, Rooney WL, Hamblin MT, Mitchell SE, Kresovich S (2009) Sweet sorghum genetic diversity and association mapping for brix and height. Plant Genome 2:48–62

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang ZW, Costich DE, Buckler ED (2009) Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell (http://www.plantcell.org/cgi/doi/10.1105/tpc.109.068437)

Nei M (1972) Genetic distance between populations. Am Nat 106:283–292

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904–909

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

Reif JC, Warburton ML, Xia XC, Hoisington DA, Crossa J, Taba S, Muminovic J, Bohn M, Frisch M, Melchinger AE (2006) Grouping of accessions of Mexican races of maize revisited with SSR markers. Theor Appl Genet 113:177–185

Rohlf FJ (2000) NTSYS-pc. Numerical taxonomy and multivariate analysis system, version 2.1. Exeter Soft ware, New York

Smith JSC, Chin ECL, Shu H, Smith S, 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

Song TM, Chen SJ (2004) Long term selection for oil concentration in five maize populations. Maydica 49:9–14

Teng WT, Can JS, Chen YH, Liu XH, Jing XQ, Zhang FJ, Li JS (2004) Analysis of maize heterotic groups and patterns during past decade in China. Sci Agric Sin 37:1804–1811

Thomson MJ, Septiningsih EM, Suwardjo F, Santoso TJ, Silitonga TS, McCouch SR (2007) Genetic diversity analysis of traditional and improved Indonesian rice (*Oryza sativa* L.) germplasm using microsatellite markers. Theor Appl Genet 114:559–568

Vignal A, Milan D, Sancristobal M, Eggen A (2002) A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol 34:275–305

Wang SC, Basten CJ, Zeng ZB (2005) Windows QTL cartographer 2.5 user manual. North Carolina State University, Raleigh

Yan JB, Yang XH, Hector S, Sánchez H, Li JS, Warburton M, Zhou Y, Crouch JH, Xu YB (2010) High-throughput SNP genotyping with the golden gate assay in maize. Mol Breed 25:441–451

Yang XH, Yan JB, Zheng YP, Yu JM, Li JS (2007) Reviews of association analysis for quantitative traits in plant. Acta Agron Sin 33:523–530

Yang XH, Yan JB, Shah T, Warburton ML, Li Q, Li L, Gao YF, Chai YC, Fu ZY, Zhou Y, Xu XT, Bai GH, Meng YJ, Zheng YP, Li JS (2010) Genetic analysis and characterization of a new maize association mapping panel for quantitative trait loci dissection. Theor Appl Genet 121:417–431

Yu JM, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17:1–6

Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Yu JM, Zhang ZW, Zhu CS, Tabanao DA, Pressoir G, Tuinstra MR, Kresovich S, Todhunter RJ, Buckler ES (2009) Simulation appraisal of the adequacy of number of background markers for relationship estimation in association mapping. Plant Genome 2:63–77

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2007) An Arabidopsis example of association mapping in structured samples. PLoS Genet 3:e4

Zheng G, Freidlin B, Li ZH, Gastwirth JL (2005) Genomic control for association studies under various genetic models. Biometrics 61:186–192

Zhu CS, Yu JM (2009) Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. Genetics 182:875–888

Zhu CS, Gore M, Buckler ES, Yu JM (2008) Status and prospects of association mapping in plants. Plant Genome 1:5–20