Mol Breeding
DOI 10.1007/s11032-009-9343-2

High-throughput SNP genotyping with the GoldenGate assay in maize

Jianbing Yan · Xiaohong Yang · Trushar Shah · Héctor Sánchez-Villeda · Jiansheng Li · Marilyn Warburton · Yi Zhou · Jonathan H. Crouch · Yunbi Xu

Received: 5 June 2009 / Accepted: 3 October 2009

Abstract Single nucleotide polymorphisms (SNPs) are abundant and evenly distributed throughout the genomes of most plant species. They have become an ideal marker system for genetic research in many crops. Several high throughput platforms have been developed that allow rapid and simultaneous genotyping of up to a million SNP markers. In this study, a custom GoldenGate assay containing 1,536 SNPs was developed based on public SNP information for maize and used to genotype two recombinant inbred line (RIL) populations (Zong3 x 87-1, and B73 x By804) and a panel of 154 diverse inbred lines. Over 90% of the SNPs were successfully scored in the diversity panel and the two RIL populations, with a genotyping error rate of less than 2%. A total of 975 SNP markers detected polymorphism in at least one of the two mapping populations, with a polymorphic rate of 38.5% in Zong3 x 87-1 and 52.6% in B73 x By804. The polymorphic SNPs in B73 x By804 have been integrated with previously mapped simple sequence repeat markers to construct a high-density linkage map containing 662 markers with a total length of 1,673.7 cM and an average of 2.53 cM between two markers. The minor allelic frequency (MAF) was distributed evenly across 10 continued classes from 0.05 to 0.5, and about 16% of the SNP markers had a MAF below 10% in the diversity panel. Polymorphism rates for individual SNP markers in pair-wise comparisons of genotypes tested ranged from 0.3 to 63.8% with an average of 36.3%. Most SNPs used in this GoldenGate assay appear to be equally useful for diversity analysis, marker-trait association studies, and marker-aided breeding.

Keywords Single nucleotide polymorphism · Maize · Goldengate · High-throughput

Abbreviations

BAC	Bacteria artificial chromosomes
DH	Doubled haploid
EST	Expression sequence tag
FPC	Fingerprinted contigs
LD	Linkage disequilibrium
LOD	Logarithm-of-odds

Electronic supplementary material The online version of this article (doi:10.1007/s11032-009-9343-2) contains supplementary material, which is available to authorized users.

J. Yan (✉) · T. Shah · H. Sánchez-Villeda · J. H. Crouch · Y. Xu (✉)
International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 México, D.F., Mexico
e-mail: j.yan@cgiar.org

Y. Xu
e-mail: y.xu@cgiar.org

J. Yan · X. Yang · J. Li · Y. Zhou
National Maize Improvement Center of China, China Agricultural University, 100193 Beijing, China

M. Warburton
USDA-ARS Corn Host Plant Resistance Research Unit, Box 9555, Mississippi State, MS 39762, USA

Published online: 22 October 2009

MARS	Marker-assisted recurrent selection
MAS	Marker-assisted selection
NAM	Nested association mapping
NSF	National science foundation
OPA	Oligo pool assay
PCR	Polymerase chain reaction
QTL	Quantitative trait locus
RFLP	Restriction fragment length polymorphisms
RIL	Recombinant inbred lines
SAM	Sentrix array matrix
SNP	Single nucleotide polymorphism
SSR	Simple sequence repeat
STS	Sequence tagged site

Introduction

Maize is a model plant species for genetic research which has been enhanced over the past two decades by the development and application of various DNA marker technologies. Former generations of widely used markers have been classified as hybridization-based markers such as restriction fragment length polymorphisms (RFLPs) (Helentjaris et al. 1986) and polymerase chain reaction (PCR)-based markers such as simple sequence repeats (SSRs) or microsatellites (Senior et al. 1993). The ideal marker system should be highly polymorphic and evenly distributed across the genome, as well as provide codominant, accurate and reproducible data which can be generated in a high-throughput and cost-effective manner. Although RFLP and SSR marker systems possess several of these attributes, they are not truly low cost or highly scalable. More recently, single nucleotide polymorphisms (SNPs) markers, which are generally developed from sequence information, have become the marker system of choice as they meet all of the above criteria, including the potential for high throughput low cost genotyping.

Single nucleotide polymorphisms 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 (MAS). More than 30 different SNP detection methods have been developed and applied in different species (as reviewed by Gupta et al. 2008). In addition, several high-density platforms are now available that can simultaneously genotype up to 384 DNA samples across 96 to 1 M SNPs (Gupta et al. 2008). The Illumina Company provides two types of genotyping platform; the GoldenGate array for medium-density genotyping that contains 96–1,536 SNPs per array, and the Infinium array for high-density genotyping that contains up to 1 M SNPs per array (Fan et al. 2006a).

The GoldenGate technology is now being used for genetic analysis in several crop species. For example, a custom oligo pool assay (OPA) containing 1,524 SNPs per assay has been developed to estimate linkage disequilibrium (LD) and perform marker-trait associations in barley (Rostoks et al. 2006). Similarly in soybean, five diverse lines from the parents of three recombinant inbred line (RIL) populations have been used for SNP identification by sequencing the selected genes, expressed sequenced tags (ESTs), bacterial artificial chromosome (BAC)-end sequences and BAC sub-clone sequences. A custom OPA containing 384 SNPs per assay was then designed to genotype the three RIL populations for construction of a high-density consensus linkage map and a 96-line diversity germplasm panel for estimation of allele frequencies (Hyten et al. 2008). In maize, molecular and functional diversity has been studied in the USA National Science Foundation (NSF) funded maize genome project which has developed more than one hundred thousand publicly available SNPs using the re-sequencing and new generation sequencing techniques (www.panzea.org). A maize 1,536 SNP OPA has been developed from this data and used to genotype the Nested Association Mapping (NAM) populations of 5,000 RILs (200 lines from each of 25 families). This has resulted in the development of an integrated linkage map with 1,106 polymorphic SNPs (McMullen et al. 2009).

Maize appears to be an ancestral tetraploid (Helentjaris et al. 1988) with a complex genome structure, containing about 80% repetitive sequences and 32% paralogous sequences (Blanc and Wolfe 2004). Genetic variation in maize is very high to such an extent that the average level of diversity between two maize lines is higher than the level of diversity between humans and chimps (Buckler and Stevens 2005). These factors could limit the utilization of the GoldenGate assay in large-scale analysis of diverse maize germplasm. In soybean and barley, which also

have a complex genome structure but possess considerably less diversity than maize, the rate of successful scoring SNP data from the GoldenGate assay was around 90% (Hyten et al. 2008; Rostoks et al. 2006). The objective of the current study was to characterize a maize GoldenGate assay system by genotyping two RIL populations and a diversity panel of 154 maize inbred lines.

Materials and methods

Plant materials

Two RIL populations were used in this study to generate SNP segregation and map locations: RIL-1 comprising 190 RILs derived from Zong3 x 87-1, a widely used elite hybrid in China, and RIL-2 comprising 174 RILs derived from B73 x BY804, the latter being a line with high oil content (Song et al. 2004). Both populations have been used in previously reported SSR marker linkage maps (Ma et al. 2007; Chander et al. 2008).

In addition, a panel of 154 diverse inbred lines was used to test the performance of SNP genotyping for germplasm analysis. The panel included 91 inbred lines that are parental genotypes of widely used commercial hybrids in Chinese breeding programs (Teng et al. 2004), 34 high-oil lines selected from major high-oil populations of the world, 25 inbred lines selected from Chinese landraces, and four high pro-vitamin A lines introduced from the United States (detailed descriptions of these lines are provided in Supplementary Table 1).

Genomic DNA was extracted from young leaves for all 364 RILs and 154 inbred lines (including the four parental lines of the two RIL populations) following standard protocols (Saghai Maroof et al. 1984).

Development of the OPA

The maize OPA used in this study was developed under the framework of the Molecular and Functional Diversity Team of the USA-NSF Maize Genome Project. This OPA consists of 1,536 well-distributed SNPs and has been used to genotype the NAM populations. A total of 1,106 of these SNPs have been successfully mapped to an integrated linkage map (McMullen et al. 2009). The SNPs used in this OPA together with a further near one million SNPs developed by the Maize Genome Project can be accessed at www.panzea.org. To develop the OPA used in this study, the 1,106 mapped SNPs where combined with 430 SNPs selected from the panzea database on the basis of having a designability score higher than 0.6. Designability scores were provided by the Illumina Company, and a score greater than 0.6 indicates that a SNP has a relatively higher probability of success when used in a GoldenGate assay, the OPA file can be found in Supplementary Table 2.

In silico mapping of SNPs

The 1,536 SNP sequences, of 120 bp or more in length, were used to perform a BlastN (Altschul et al. 1990) search against the maize accessioned golden path (AGPv1) downloaded from the Arizona Genome Institute (http://www2.genome.arizona.edu/genomes/maize). Only the top blast-hits were considered using an *e*-value threshold of e^{-12}. Blast matches to multiple loci, with the same top *e*-value were all selected for further interrogation (Table 1).

SNP genotyping

The genotyping of SNPs was performed using the Illumina BeadStation 500 G (Illumina, San Diego, CA) at the Cornell University Life Sciences Core Laboratories Center using the protocol described by Fan et al. (2006b). The protocol for this assay recommended using a minimum of 50 ng × 5 ul DNA per sample. The DNA quality was checked carefully before genotyping for each sample.

The samples (364 RILs and 154 inbred lines) were divided into six groups and analyzed using separate Sentrix Array Matrices (SAMs), which accommodate 96 samples per SAM. Two inbred lines (A619 and RY737) were included in duplicate to verify the genotype reproducibility. All SNP data were analyzed using the Illumina BeadStudio genotyping software that can cluster and call the data automatically, thereby allowing visualization of the data directly for further analysis (Fig. 1). Each SNP was re-checked manually and re-scored if any error was

Table 1 Summary of SNPs used in this study and the comparison of the in silico mapping and linkage map results

Chr.	Number[a]	Multiple hits[b]			Matching ratio with in silico mapping results (%)[d]	Conserved order of SNPs (%)
		Within chr.	Different chr.	Mapped[c]		
1	264	1	6	115 (21/94)	96.8	96.7
2	172	3	3	72 (16/56)	92.9	86.5
3	169	1	2	56 (11/45)	95.6	93.0
4	163	3	5	61 (8/53)	100.0	92.5
5	194	4	6	72 (7/65)	96.9	95.2
6	130	2	3	59 (18/41)	100.0	78.0
7	101	0	2	60 (11/49)	91.8	86.7
8	142	1	1	66 (20/46)	95.7	100.0
9	115	2	1	56 (19/37)	91.9	94.1
10	98	1	0	45 (14/31)	100.0	87.1
Total	1,548	18	29	662 (145/517)	96.2	91.0

[a] In silico mapping results by blasting the maize genome sequence at $E < 10^{-12}$ level

[b] One SNP with more than one hit at the same *e*-level was counted one time if the hits were identified in same chromosome, and was counted multiple times if the hits were identified in different chromosomes (in this study, the same SNP was identified two hits in two chromosomes mostly)

[c] Map results based on the B73*BY804 RIL population used in this study. In brackets are the numbers of SSRs and SNPs, respectively

[d] Matching ratio in this case the results in brackets are when excluding results that are not mapped onto the in silico map

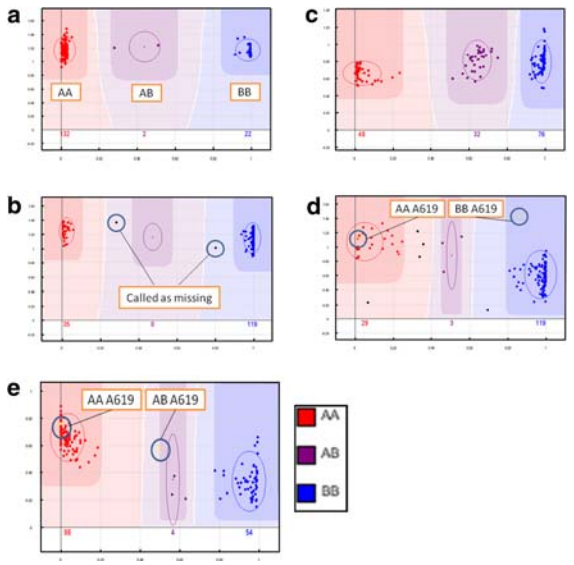

Fig. 1 Scoring of SNP genotyping data using the BeadStudio genotyping software. **a** Typical score by Goldengate (SNP PHM13183.12); **b** plots located in the border of cluster were excluded (SNP PZB00054.3); **c** typical score with obvious 3 clusters (*AA*, *AB*, *BB*) that AB cannot be re-clustered to any homozygous cluster manually (SNP PZA02186.1); **d** one type of genotyping error (AA was called as BB) (PZA03058.22); **e** another type of genotyping error (AA was called as AB) (SNP PZB00008.1)

observed in the clustering of homozygous and heterozygous groups.

Linkage map construction

Data from all SNP markers that detected polymorphisms between B73 and BY804 were used to construct the linkage map. SNPs with the same genotype as B73 were scored as "1"; SNPs with the same genotype as BY804 were scored as "2"; heterozygous SNPs were scored as missing data. All data from polymorphic SNPs and previously mapped SSRs or sequence tagged site (STS) markers were combined to construct an integrated genetic linkage map using Mapmaker 3.0 (Lincoln et al. 1992). The NAM map and previous maps from B73 x BY804 were regarded as reference maps for constructing the new map. The threshold of logarithm-of-odds (LOD) score for the test of independence of marker pairs was set at 3.0, and the marker order with the highest LOD score was then selected. The Kosambi mapping function was used for calculating map distances.

Results

SNP performance and quality

Scoring of SNP genotyping data using the BeadStudio genotyping software generally produced three clear clusters denoting the AA homozygote, BB homozygote, and AB heterozygote (see Fig. 1a). Data points ambiguously located between these clusters were scored as missing data (see Fig. 1b). In this study, only fixed RILs or elite inbred breeding lines were used for genotyping which contained none or only a few heterozygotes. In a few other cases SNP markers with a high ratio of heterozygotes could not be re-clustered manually (Fig. 1c). These SNPs were excluded from the analysis in the present study.

Cluster separation scores provided by the BeadStudio software were used as an indicator to describe the separation of the three classes. However, this may not be a perfect indicator, since cluster separation scores are calculated based on the degree of separation of the two homozygous clusters versus the heterozygous cluster rather than between the two homozygous clusters (Hyten et al. 2008). Nevertheless, this measure still provides some general information about SNP quality. As the materials tested in this study are RIL populations or inbred lines with very few heterozygotes, SNPs with a cluster separation score as low as 0.3 can still be successfully used. Figure 2 shows the distribution of cluster separation scores between 0.3 and 1.0. More than half of the successfully scored SNPs were well separated with a cluster separation score of 1. The detailed cluster separation score of the 1,362 SNPs can be found in Supplementary Table 3.

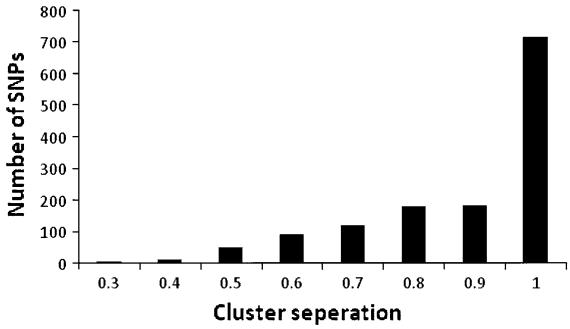

Fig. 2 Cluster separation distribution in the 154 diversity lines based on 1,362 SNPs

Around 92% of the SNPs tested (1,414 of 1,536) were called successfully with a cluster separation score of 0.3 or greater and less than 20% missing data for the inbred lines. Among the 1,414 reliably scored SNPs, a further 52 had three genotypic classes (as shown in Fig. 1c), so data from these markers was excluded from further analysis. The proportion of heterozygote individuals indentified per marker ranged from 0 to 8.33% with an average of 1.01%. The level of missing data per marker ranged from 0 to 19.5% with an average of 1.6%. About 90% of the SNPs (1,256 of the 1,414), the level of missing data was less than 5%. In the RIL populations, 288 SNPs were excluded because they had separation scores less than 0.3 or more than 20% missing data on this basis, data from a total of 1,248 SNPs were used for further analysis.

Each of the two control samples (A619 and RY737) was put onto two different plates as technical repeats. Two types of genotyping errors were observed from the data collected from these control samples: (1) Class A Errors, a SNP was scored as different homozygotes in different plates (Fig. 1d), and (2) Class B Errors, a SNP was scored as a homozygote in one plate but as a heterozygote in another plate (Fig. 1e). A total of ten Class A errors (1.23%) and 23 Class B errors (1.75%) were observed across the two control genotypes (SNPs called successfully in one plate but failed the other were excluded). Thus, the accuracy (repeatability) of the GoldenGate assay in this study was more than 98%.

SNP in silico mapping and distribution

The original sequences used to develop these SNPs were used to carry out BLAST comparisons with the maize accessioned golden path (AGPv1). Of the 1,536 sequences used in this study, only 22 did not have a BLAST match below the *e*-value threshold of e^{-12}.Thus, we were able to generate in silico map positions for 1,514 unique SNPs used in this study. The number of SNPs per chromosome ranged from 98 SNPs on chromosome 10 and 264 SNPs on chromosome 1. A total of 18 SNPs were found to have more than one BLAST match with the same *e*-value on the same chromosome, while 29 SNPs had two significant BLAST matches on different chromosomes (Table 1).

Allele frequencies in diversity inbred lines

The SNPs used in this study were selected based on the genetic polymorphisms detected in 27 diverse inbred maize lines that were used as parental genotypes of the NAM population (Yu et al. 2008; McMullen et al. 2009). Particular emphasis was placed on SNP markers that detected polymorphisms between a common parent (B73) and the 26 other parents of NAM population. Information on the allelic frequencies of these SNPs in other maize germplasm should help determine the usefulness of this OPA for analysis of a broader range of maize germplasm. Most of the 154 inbred lines used in this study would be classified as temperate germplasm, more than half of which are widely used in Chinese maize breeding programs. Although it might be expected that this germplasm would contain limited allelic diversity, in fact only 20 of the 1,414 SNP markers were monomorphic across all 154 lines. An even distribution of minor allelic frequency (MAF) was observed (Fig. 3) with 10 continued classes from 0.05 to 0.5 with a similar number of SNPs in each MAF class. Only 7.3% (100/1,362) of the SNPs had a MAF of less than 0.05, while 16.3% (222/1,372) of the SNPs had a MAF of less than 0.1. Polymorphic ratios for pair-wise comparisons of the genotypes tested ranged from 0.3 to 63.8% with an average of 36.3%. The highest level polymorphism was observed between B73 and WMR (a line derived from a Chinese landrace), while the lowest level of polymorphism was observed between two Chinese commercial lines (CY72 and 4F1). The average polymorphic ratio between any given line compared with all other lines tested ranged from 33.3 (HZS) to 52.6% (B73).

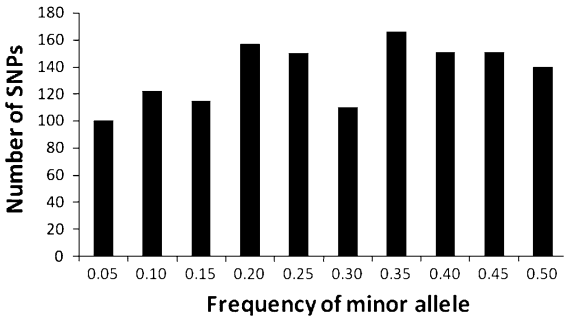

Fig. 3 Minor allelic frequency distribution in the 154 diversity lines based on 1,362 SNPs

Genetic mapping of SNPs

A total of 190 families from RIL-1 and 174 families from RIL-2 were genotyped using the tailored 1,536-SNP OPA. Of the 1,393 successfully scored SNPs, 975 SNPs were polymorphic in one or both of the RIL populations: 536 SNPs (38.5%) were polymorphic between the parents of RIL-1, and 733 SNPs (52.6%) were polymorphic between the parents of RIL-2, while 294 SNPs (21.1%) were polymorphic in both populations. RIL-2 was used as an example to build a linkage map by integrating the SSRs and SNPs. Finally, a linkage map including 662 markers (145 SSRs and 517 SNPs) was constructed which covered the ten chromosomes with a total length of 1,673.7 cM and an average of 2.53 cM between two markers (supplemental Fig. 1). Totally, the locations of 96.2% linkage mapped SNPs were same with the results of in silico mapping. The detailed comparison of the linkage and physical map can be found in the Supplementary Table 4.

Discussion

Data scoring and management

About 50% of the SNPs screened in this study had cluster separation scores near 1 (Fig. 2) and can be considered of sufficient quality to be correctly scored by the Illumina BeadStudio genotyping software without manual intervention. Conversely, about 10% of the SNPs showed different cluster separation scores in different groups of germplasm. These markers cannot be reliably scored using the automated software. For example, in Fig. 1c for SNP PZA02186.1, the diverse panel of inbred lines is clustered into three distinct groups, with a much higher number of heterozygotes than should be observed in fixed lines. This may be caused by the complex genome structure of maize germplasm, including many paralogous or homologous copies of the target locus, or by mismatch in the primer region. The clustering patterns of about 4% SNP markers (52 of the total 1,414 analyzed here) were too distinct to be reliably rescored (see Fig. 1c). Thus, data from these SNPs were excluded in this study but the putative heterozygotes could be considered as a third genotypic class in marker-trait

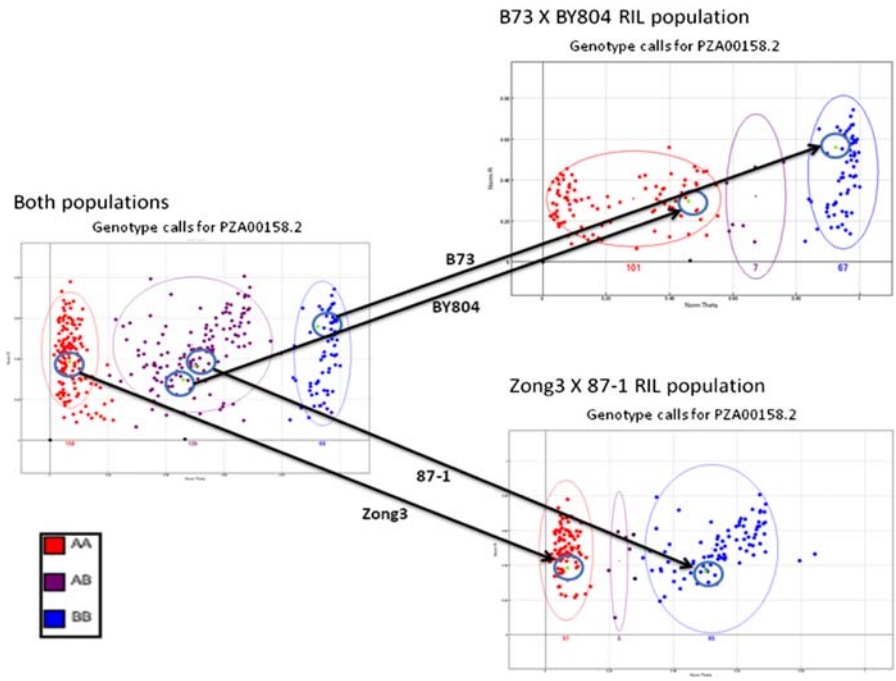

Fig. 4 Example of cluster compression in two RIL populations with GoldenGate assay (*X* axis is normalized theta and *Y* axis is normalized *R*)

association analyses. Another SNP marker (PZA00158.2) was observed to be polymorphic between the parental genotypes of both RIL populations (Fig. 4). Three clusters were observed for both RIL populations with a high ratio of heterozygotes. Zong3 (a parent of RIL-1) and B73 (a parent of RIL-2) were correctly located in the AA and BB homozygous clusters with normalized thetas of 0.1 and 0.9, respectively. However, 87-1 (the second parent of RIL-1) and BY804 (the second parent of RIL-2) were located in the AB cluster with a similar normalized theta of 0.5. When the two populations were analyzed separately, the AA homozygous cluster had a normalized theta of 0.1 in RIL-1 but varied between 0.1 and 0.5 in RIL-2 whereas the BB homozygous cluster had the normalized thetas of 0.5 in RIL-1 and 0.9 in RIL-1. This result implies that there may be an additional locus in the genome with sequence to the target locus used for designing PZA00158.2 (In silico mapping results show that the original sequence of this SNP has two hits with same *e*-value in the maize genome). In RIL-1, the identical sequence of PZA00158.2 affected the BB cluster by increasing the background signal and causing the normalized thetas to vary from 0.6 to 0.9. In RIL-2, the identical sequence might only affect a part of the AA cluster that made the normalized theta vary between 0.1 and 0.5. The results from these SNP markers are unusual and clearly difficult to interpret, and since they may affect overall results they need to be managed carefully. In cases such as these, where the results are not clearly unambiguous, we prefer to remove the data from the analysis. When analyzing material with a highly homozygous background [RIL/doubled haploid (DH) populations or inbred lines], it is simple to make a judgement on most clustering patterns where the cluster separation scores are high. However, if the analyzed materials contain many heterozygotes (such as segregating genetic and breeding populations, or open pollinated cultivars and landraces of maize), it may be difficult to distinguish a true AB cluster from a cluster caused by paralogs, homologs, or other repeated gene sequences as discussed above. During genotyping of heterologous or heterozygous populations with SNP markers, experimental designs should include known, fixed genotypes as controls for aiding scoring, as also suggested by Hyten et al. (2008).

Validating the wide potential utility of the custom OPA

In this study, we have shown that the GoldenGate array can be used successfully for genotyping of

diverse inbred maize germplasm. A total of 1,414 SNP markers were observed to produce clear separation allelic classes with less than 20% missing data across the 154 genotypes tested. Our successful average calling rate of 92% is in line with that of 90% reported for barely (Rostoks et al. 2006) and 89% for soybean (Hyten et al. 2008). However, it should be noted that this OPA was developed using 1,106 SNPs that had already been selected for genotyping of the NAM population and only a further 430 SNPs had not been previously screened. Thus, this success rate is higher than may be expected from an OPA based entirely on SNPs not previously screened. However, we have developed a parallel OPA in maize based on selected candidate genes that had not been previously tested. This second custom OPA was used to genotype more than 600 diverse maize inbred lines providing a calling success rate of 85% (Yan JB et al. unpublished data). Therefore, it appears that the GoldenGate array can be successfully and efficiently applied to a diverse range of genetic analyses in maize. The OPA used in the current study was designed for linkage mapping of the NAM population; i.e., SNPs were selected to maximize polymorphisms between B73 and 26 other inbred parental genotypes. Consequently, the highest polymorphic ratio was seen between B73 and other lines, with an average polymorphic rate of 52.6% for pairwise genotype comparisons. However, a uniform distribution of allelic frequencies in different classes was also observed amongst the Chinese temperate and commercial lines tested. Only 16.3% of the SNPs showed MAFs below 10% (Fig. 3). Alleles present at very low frequencies generally have very little impact on large-scale diversity studies and have a low probability of being polymorphic in mapping studies. However, markers with low MAF scores may be highly valuable in allele mining projects. Meanwhile, markers with higher MAF scores should be valuable for screening diverse sources of maize germplasm in genetic diversity analysis projects.

Future development of the OPA

Although the current OPA has widespread applicability, we are now focusing on improvements in two areas towards the development of a universal OPA for multiple research objectives: (1) ease of scoring, preferably the scoring of all SNPs should be fully automated, and (2) selection of SNPs that detect a substantial amount of polymorphism in any panel of germplasm. For the currently available OPA, we have had to manually score some of the SNPs, which would present a significant bottleneck for large-scale genotyping projects. For a few SNPs (e.g. PZA00158.2; Fig. 4), the data were too ambiguous to be reliably called even through manual scoring. It is likely that the complex genome structure in maize will always cause a small percentage of markers to be rejected based on this criteria. SNP markers with low MAF scores below 0.05 may not be informative for most diversity analysis, linkage mapping studies, and MAS application programs. Removing SNP markers that cannot be easily scored or that represent very rare alleles will be necessary to further optimize the value of this OPA. For routine QTL mapping studies and marker-assisted recurrent selection (MARS) applications (Bernardo 2008), fewer polymorphic markers ($\sim$200) are required. These applications would then only required screening with around 600 SNPs from this OPA, as around one-third have been shown to be polymorphic between any given pair of genotypes. In the current study, we report the identification of nearly 600 SNPs that are highly polymorphic across the material tested as well as easy to score. Thus, we propose to establish a universal 384-SNP OPA based on the data reported here, and validate that more than 40% of the SNPs detect polymorphisms in any given segregating population.

Using of the custom GoldenGate assay for QTL mapping and genome-wide selection

Quantitative trait locus linkage mapping continues to be widely used for identifying and locating genes affecting complex traits. Construction of a genetic map remains the first essential step towards QTL mapping. Before the establishment of high throughput SNP genotyping systems, linkage maps in maize were generally constructed using RFLP or SSR markers that were time-consuming and expensive even with the advent of capillary electrophoresis for data capture from PCR-based markers. Although it is difficult to compare cost and time efficiency of marker analyses across laboratories (especially in across countries), we provide some general indication. In the present study,

we constructed linkage maps based on screening two RIL populations with 263 and 237 PCR-based markers (SSR/STS) using gel electrophoresis. The genomic data generation for this project took two full-time well-trained students for more than 1 year, with each student able to map an average of just one marker in one population per day. Using the GoldenGate assay, all genotyping was completed within 1 week, which is 100-fold faster than gel-based methods. In the CIMMYT Applied Biotechnology Center, the genotyping cost per SSR marker is about US$ 1 per sample per SSR (excluding the cost of DNA extraction and data management). Thus, the cost of generating the genotype data for these two linkage maps is more than US$ 200 per individual. The cost of generating SNP data for the same populations using the GoldenGate assay would be less than US$ 100 per individual. In addition, the density of the resultant map would be 2–3 times higher (assuming one-third to one half of the 1,536 SNP markers on the OPA would be polymorphic and easily scored in any given mapping population). This would result in a cost saving of about 75% for SNP genotyping versus SSR-based methods.

Large-scale MARS applications have been widely incorporated into commercial maize breeding programs to increase the speed and genetic gain of the breeding process. More recently, many private sector breeding programs are implementing genome-wide selection systems that do not require conventional QTL mapping (Bernardo 2008). The cost of obtaining reliable phenotypic data from replicated trails is now estimated to be at least US$75 per entry in the USA (Bernardo 2008). Marker-based technologies including genome-wide selection allow breeders to perform selection without phenotyping (or with less phenotyping). The annual genetic gain that can be achieved with genome-wide selection for complex traits such as yield in maize (using up to 3 cycles per year) is also significantly higher than for phenotypic selection (0.5 cycle per year) (Bernardo 2008). A SNP array containing 384 SNPs can provide enough polymorphic markers for genome-wide selection. The cost of genotyping using such an array is only about US$ 20 per sample (Jeff Ehlers, personal communication). Thus, for complex traits, genotyping has become more cost effective than phenotyping. However, for genome-wide selection to be effective, the genotyping results must be generated every cycle prior to pollination. If this window of timeline is missed, the benefits of genotyping over phenotyping will be largely lost. Achieving 3 cycles per year in maize is also a substantial logistical challenge, requiring the use of off-season nurseries. Thus, the whole process of DNA extraction, genotyping and data analysis must be completed within 4–6 weeks for each cycle. The recent development of a single seed-based DNA extraction system (Gao et al. 2008), should greatly assist large-scale genome-wide screening systems to routinely complete 3-cycles of selection per year. The CIMMYT global maize program is currently introducing genome-wide selection to assist drought tolerance breeding. For this purpose, a 384-SNP OPA is currently being developed based on the results presented here. This will then allow dozens of segregating populations to be annually subjected to high throughput cost effective genome-wide selection.

Enhancing the information generated from SNPs for improved diversity analysis

The genetic information provided by SSRs and SNPs during linkage mapping and MAS activities is broadly similar when the same numbers of markers is used and a small number of parental genotypes are involved. The use of high throughput genotyping systems for SNPs then provides an efficiency advantage over SSR in mapping and molecular breeding applications. However, SNPs are bi-allelic markers that generally detect only two alleles per SNP marker, while SSRs are multi-allelic markers capable of detecting a very large number of alleles per locus (Lu and Bernardo 2001; Liu et al. 2003), only limited by the nature and extent of the germplasm being tested. Most maize SNPs have been developed through re-sequencing of known genomic regions across a limited number of diverse lines. This approach may lead to an even distribution for SNP allele frequencies (Fig. 2). In contrast, since all possible alleles can be detected by appropriate SSR markers, the allelic frequencies detected are often skewed towards rare alleles. Thus, SSR markers are likely to be more informative than SNPs when performing diversity and relatedness analyses (Hamblin et al. 2007). Consequently, in these situations a larger number of SNPs will be required in order to obtain the same level of information as provided by currently

available SSR markers. For analysis of highly divergent maize germplasm, it may be difficult to accurately estimate relatedness using SNP markers.

Haplotypes that combine information from several SNPs within the same gene or locus may provide a partial solution to the disadvantage that SNP markers have when used in diversity analyses. For homozygous lines, one SNP (as A/T) can produce two alleles (A and T), while two SNPs from one locus (as A/T, G/C) can produce four allele combinations or haplotypes (AG, AC, TG and TC). In theory, n SNPs from one locus can produce 2^n haplotypes. For 1,536 unique SNPs (with just one SNP per locus), a maximum total of 3,072 alleles can be detected with an allelic frequency ranging from 0 to 0.5. However, retaining the same total number of SNPs but reducing the number of loci to 512 (i.e., three SNPs per locus) or to 384 loci (i.e., four SNPs/loci), can result in up to 4,096 or 6,144 detectable haplotypes. In these cases the haplotype frequencies will range from 0 to 1 with a large accumulation in the low frequency class. Not all theoretically possible haplotypes can be observed due to linkage disequilibrium in most loci. It is probably that at least 3,000 haplotypes would be detected by four SNP markers in each of 384 genes. Even at this hypothetical level of 50% redundancy, SNP marker haplotypes would be as informative as screening with 140 SSRs detecting an average of 22 alleles (Liu et al. 2003). Hamblin et al. (2007) have reported that "SNP haplotypes" are slightly more informative than standard SNP data when determining population structure. However, this study used two and three SNP haplotype combinations per locus. Screening more SNPs from within each locus should improve the power of diversity analyses even further. Although this needs to be proven, combined use of both genotypes and haplotypes makes SNP markers more powerful than using genotypes alone and should be functional well in genetic diversity analysis.

Acknowledgments We highly appreciate the molecular and functional diversity team of the NSF Maize Genome Project for making available all the SNPs information used in this study. We thank Dr. Ortiz Rodomiro (CIMMYT) for his critical review this manuscript. This research was supported by the National Hi-Tech Research and Development Program of China and the Bill & Melinda Gates Foundation through the Drought Tolerant Maize for Africa project (http://dtma.cimmyt.org/).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410

Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci 48:1649–1664

Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1667–1678

Buckler ES, Stevens NM (2005) Maize Origins, Domestication, and selection. In: Motley TJ, Zerega N, Cross H (eds) Darwin's harvest. Columbia University Press, New York, pp 67–90

Chander S, Guo YQ, Yang XH et al (2008) Using molecular markers to identify two major loci controlling carotenoid contents in maize grain. Theor Appl Genet 116:223–233

Fan JB, Chee MS, Gunderson KL (2006a) Highly parallel genomic assays. Nat Rev Genet 7:632–644

Fan JB, Gunderson KL, Bibikova M et al (2006b) Illumina universal bead arrays. Methods Enzymol 410:57–73

Gao SB, Martinez C, Skinner DJ et al (2008) Development of a seed DNA-based genotyping system for marker-assisted selection in maize. Mol Breeding 22:477–494

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 2:e1367

Helentjaris T, Slocum M, Wright S et al (1986) Construction of genetic linkage maps in maize and tomato using restriction fragment length polymorphisms. Theor Appl Genet 72:761–769

Helentjaris T, Weber D, Wright S (1988) Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms. Genetics 118:353–363

Hyten DL, Song Q, Choi IY et al (2008) High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor Appl Genet 116:945–952

Lincoln S, Daly M, Lander E (1992) Constructing genetics maps with MAPMAKER/EXP 3.0. Whitehead Institute Technical Report, Whitehead Institute, Cambridge

Liu K, Goodman MM, Muse S et al (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165:2117–2128

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

Ma XQ, Tang JH, Teng WT et al (2007) Epistatic interaction is an important genetic basis of grain yield and its components in maize. Mol Breeding 20:41–51

McMullen MM, Kresovich S, Villeda HS et al (2009) Genetic properties of the maize nested association mapping population. Science 325:737–740

Rostoks N, Ramsay L, MacKenzie K et al (2006) Recent history of artificial outcrossing facilitates whole-genome

association mapping in elite inbred crop varieties. Proc Natl Acad Sci USA 103:18656–18661

Saghai Maroof MA, Soliman KM, Jorgensen RA et al (1984) Ribosomal DNA spacer length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc Natl Acad Sci USA 81:8014–8018

Senior L, Lynn M, Heun M (1993) Mapping maize microsatellites and polymerase chain reaction confirmation of the targeted repeats using a CT primer. Genome 36:884–889

Song XF, Song TM, Dai JR et al (2004) QTL mapping of kernel oil concentration with high-oil maize by SSR markers. Maydica 49:41–48

Teng WT, Can JS, Chen YH et al (2004) Analysis of maize heterotic groups and patterns during past decade in China. Sci Agric Sin 37:1804–1811

Yu JM, Holland JB, McMullen MD et al (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551