Theor Appl Genet (2010) 121:577–588
DOI 10.1007/s00122-010-1331-8

ORIGINAL PAPER

Development of highly polymorphic SNP markers from the complexity reduced portion of maize [*Zea mays* L.] genome for use in marker-assisted breeding

Jafar A. Mammadov · Wei Chen · Ruihua Ren · Reetal Pai · Wesley Marchione · Feyruz Yalçin · Hanneke Witsenboer · Thomas W. Greene · Steven A. Thompson · Siva P. Kumpatla

Received: 18 September 2009 / Accepted: 26 March 2010 / Published online: 18 April 2010

Abstract The duplicated and the highly repetitive nature of the maize genome has historically impeded the development of true single nucleotide polymorphism (SNP) markers in this crop. Recent advances in genome complexity reduction methods coupled with sequencing-by-synthesis technologies permit the implementation of efficient genome-wide SNP discovery in maize. In this study, we have applied Complexity Reduction of Polymorphic Sequences technology (Keygene N.V., Wageningen, The Netherlands) for the identification of informative SNPs between two genetically distinct maize inbred lines of North and South American origins. This approach resulted in the discovery of 1,123 putative SNPs representing low and single copy loci. In silico and experimental (Illumina GoldenGate (GG) assay) validation of putative SNPs resulted in mapping of 604 markers, out of which 188 SNPs represented 43 haplotype blocks distributed across all ten chromosomes. We have determined and clearly stated a specific combination of stringent criteria (>0.3 minor allele frequency, >0.8 GenTrainScore and >0.5 Chi_test100 score) necessary for the identification of highly polymorphic and genetically stable SNP markers. Due to these criteria, we identified a subset of 120 high-quality SNP markers to leverage in GG assay-based marker-assisted selection projects. A total of 32 high-quality SNPs represented 21 haplotypes out of 43 identified in this study. The information on the selection criteria of highly polymorphic SNPs in a complex genome such as maize and the public availability of these SNP assays will be of great value for the maize molecular genetics and breeding community.

Communicated by T. Luebberstedt.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-010-1331-8) contains supplementary material, which is available to authorized users.

J. A. Mammadov (✉) · W. Chen · R. Ren · R. Pai · W. Marchione · T. W. Greene · S. A. Thompson · S. P. Kumpatla
Dow AgroSciences, 9330 Zionsville Rd, Indianapolis, IN 46268-1054, USA
e-mail: jamammadov@dow.com

F. Yalçin · H. Witsenboer
KEYGENE N.V., P.O. Box 216, 6700 AE Wageningen, The Netherlands

Introduction

Single nucleotide polymorphisms (SNPs) are the most abundant forms of genetic variation among individuals within species (Rafalski 2002). SNPs are considered to be versatile tools for many genetic applications such as construction of genetic maps, discovery of genes/quantitative trait loci (QTL), assessment of genetic diversity, pedigree verification, cultivar identification, association analysis and marker-assisted breeding (Zhu et al. 2003). Furthermore, the development of high- to ultrahigh-throughput genotyping methods makes SNPs highly attractive as genetic markers in various commercially important crops such as maize and soybean (Chagné et al. 2007).

Development of SNP markers usually consists of two parts: SNP discovery and SNP assay development. SNP discovery in crops is not an easy task because of genome complexity and often the lack of a reference genome sequence. Even in crops such as maize [*Zea mays*], where a reference genome sequence is available, large-scale SNP discovery efforts are still impeded by the highly repetitive (Meyers et al. 2001) and duplicated (Gaut and Doebley 1997) nature of the genome. To avoid repetitive sequences,

maize researchers have focused on the discovery of SNPs within coding sequences by re-sequencing amplicons derived from unigenes (Wright et al. 2005) or in silico mining of SNPs within ESTs (Batley et al. 2003). The advantage of these approaches is the detection of gene-based SNPs. However, both approaches have some disadvantages: they are low throughput and are unable to detect SNPs located in low-copy non-coding regions and intergenic spaces. Additionally, amplicon re-sequencing is an expensive and labor intensive procedure (Ganal et al. 2009).

The recent emergence of sequencing-by-synthesis (SBS) technologies such as 454 Life Sciences (Roche Applied Science, Indianapolis, IN), Illumina Genome Analyzer/Solexa (Illumina, San Diego, CA) and SOLiD (Life Technologies Corporation, Carlsbad, CA) have elevated expectations toward the rapid genome-wide identification of a large number of SNPs at a much lower price tag (Mardis 2007). However, efficient application of these technologies for SNP discovery in a given crop depends on the availability of a reference genome sequence (Ganal et al. 2009) as well as the level of genome complexity. For instance, in maize, the availability of a reference sequence does not guarantee a painless SNP discovery using SBS technologies. The complexity and existence of re-arrangements in the maize genome complicate the assembly of short-read SBS sequences and their alignment to the reference genome (Morozova and Marra 2008). Thus, the reduction of genome complexity becomes an important prerequisite for the genome-wide discovery of true SNPs in maize using SBS technologies. Several genome complexity reduction techniques have been developed, including High C_0t selection (Yuan et al. 2003), methylation filtering (Palmer et al. 2003; Emberton et al. 2005) and microarray-based genomic selection (Okou et al. 2007). However, the majority of these techniques mainly reduce the number of repetitive sequences and are ineffective in the recognition and elimination of paralogs and homoeologs, which cause the detection of false-positive SNPs. Recently, computational SNP calling methods were developed that could drastically reduce the number of false SNPs, resulting from the alignment of duplicated sequences as well as from re-sequencing errors (Barbazuk et al. 2007; Baird et al. 2008; Gore et al. 2009a). Hence, the availability of reference sequences, the application of genome complexity reduction techniques and SBS technologies coupled with post-re-sequencing computational treatment become important prerequisites for genome-wide detection of true SNPs in complex genomes.

Once true SNPs have been discovered, they can be used for the development of marker assays. In plants, SNP assays are being developed using various chemistries, including TaqMan (Livak et al. 1995), GoldenGate (GG) (Hyten et al. 2008; Jones et al. 2009), Infinium (Illumina technical support, personal communication), iPLEX (Wright et al. 2008), fluorescent polarization detection (FP-TDI) (Chen et al. 1999) and Invader™ (Olivier 2005). The choice of a chemistry strongly depends on the nature of the research. For example, to validate a small subset of SNP markers (<40) for a particular region of the genome, a TaqMan assay implemented using a flexible OpenArray platform (TaqMan OpenArray Genotyping system, Product bulletin) or the Sequenom iPLEX would be an optimal choice. The chemistries with the higher multiplexing abilities such as GG and Infinium assays are suitable for high-throughput genome-wide SNP development within a short period of time. These chemistries allow multiplexing of up to 1,536 [(GG assay) (Fan et al. 2003)] and 200,000 SNPs [(Infinium assay) (Illumina SNP genotyping 2006)], respectively, in one reaction within a 3-day period. Using the GG assay, academic labs (Hyten et al. 2008) and industrial units (Jones et al. 2009) have been validating a large number of SNPs to identify a core set of markers that are highly polymorphic, demonstrate a stable performance and maintain reliable allelic discrimination regardless of genotyped sample size and germplasm.

In this study, we describe the development process of over 1,000 maize SNP markers. These markers were discovered by Complexity Reduction of Polymorphic Sequences (CRoPS) technology (KeyGene N.V., Wageningen, The Netherlands) (van Orsouw et al. 2007). CRoPS reduces genome complexity using AFLP technology followed by re-sequencing of AFLP fragments with Genome Sequencer (GS) 20/GS FLX SBS technology (454 Life Sciences). We used Keygene's proprietary automated SNP mining tool to reduce the number of false-positive SNPs and Illumina's GoldenGate assay for the validation of informative SNPs. Finally, we provide information on stringent criteria that were developed and applied for the selection of highly polymorphic and genetically stable SNPs. These SNPs were made public and recommended for application in marker-assisted selection projects in maize.

Materials and methods

Genetic material

Genomic DNA samples were isolated from the root tips of two genetically distant Dow AgroSciences (DAS) proprietary maize inbred lines, 'DAS-NSS-22' and 'DAS-SS-25', using the CTAB method (Stuart and Via 1993). Root tips were chosen for DNA extraction to reduce the occurrence of chloroplast DNA. Three F_2 mapping populations of ~300 individuals each, DAS-NSS-22 × DAS-SS-25, DAS-NSS-24 × DAS-SS-12 and DAS-MISC-5 × DAS-SS-30, hereafter referred to as

Pop_1, Pop_2 and Pop_3, respectively, were used to map developed SNP markers, named as DZm SNPs. DNA for SNP genotyping was extracted from lyophilized leaf tissue of all F_2 individuals using a Qiagen DNA extraction kit (Qiagen, Hilden, Germany).

A panel of 86 genetically diverse public and proprietary maize inbred lines was developed to calculate the minor allele frequency (MAF) estimates. The inbred lines represent North and South American geographies, stiff, non-stiff stalk and miscellaneous heterotic groups. Another panel of ~6,000 DAS maize samples from various internal marker-assisted breeding projects was developed to inspect the stability of GG genotypic clusters in Illumina's GenomeStudio software (Illumina, San Diego, CA) in case of a dramatic sample increase. DNA from inbred lines of both panels was isolated from eight leaf punches per sample using a Qiagen DNA extraction kit.

CRoPS analysis

Complexity reduction using the AFLP technology

Complexity reduction using AFLP was performed as described by Vos et al. (1995). Briefly, AFLP templates were generated using a *Pst*I/*Mse*I enzyme combination. Pre-amplification reactions were performed using one selective nucleotide on each primer (A for the *Pst*I primer and C for the *Mse*I primer). This material was used as the input for the subsequent CRoPS library preparation (van Orsouw et al. 2007).

Quality control of the CRoPS fragment library

Quality control of the CRoPS fragment library was performed as suggested by van Orsouw et al. (2007). Briefly, a sample of the fragments from the CRoPS library was cloned and sequenced using the Sanger sequencing method to quantify the chloroplast (cp) and mitochondrial (mt) sequences. To assess the optimal amount of DNA to be used in the emulsion PCR (em-PCR), a titration run was performed (van Orsouw et al. 2007). Sequences obtained from the titration run were checked for duplication with cp- and mt-DNA sequences. To assess the level of repetitiveness among the total sequences, an internal clustering was performed.

GS FLX sequencing run, data preprocessing and SNP mining

The CRoPS run was performed as described by van Orsouw et al. (2007). SNPs were mined using Keygene's proprietary automated SNP mining tool. To minimize the complication of downstream GG assay design, additional quality criterion was applied to SNPs: 12 base-minimal interval of flanking sequence that must be devoid of additional SNPs. Additionally, SNPs in homopolymer sequence stretches were discarded, as the error rate of base calling is relatively large in those regions.

In silico and experimental validation of SNPs

In silico validation of SNPs included three steps. Step 1 covered the identification of sequences representing mobilome of the maize genome using RepeatMasker software (http://www.repeatmasker.org). Step 2 involved screening for paralogous and homoeologous sequences. SNP-containing sequences were BLASTed against maize high-throughput genomic sequence database (dbhtgs) at NCBI. If a sequence retrieved hits from two or more BAC clones representing different chromosomes, it was declared homoeologous. On the other hand, a sequence was declared a paralogue if a query sequence hit only one BAC clone, but had several copies within the same BAC. Single copy SNPs were passed onto the final step of in silico validation, i.e., assessment of assay designability using Illumina's preliminary Assay Design Tool (ADT) at http://www.illumina.com. The assessment was conducted with Illumina's proprietary criteria where SNPs were assigned a designability score of 1 (highly designable), 0.5 (moderately designable) or zero (low designability). SNPs with zero designability scores were discarded. The remaining SNPs were sent to Illumina for the synthesis of 1536-plex Oligo Pool All (OPA). Experimental validation and genotyping of SNPs were performed using Illumina's BeadArray Technology and GG assay (Illumina, San Diego CA) as per the methodology described by Fan et al. (2003) and Hyten et al. (2008). Polymorphic SNPs were subject to further analysis using two parameters of GenomeStudio software, the Gen_Train_Score (GTS) and Chi_test100 score (CT100). A GTS parameter was used to determine SNPs with well-separated and tight clusters exhibiting high-intensity signals. GTS was used instead of a Cluster_Separation parameter because the latter measures the degree of separation between three genotype clusters in the theta dimension only, without considering the texture of the clusters (tight vs. diffused) or intensity of the calls. The value of a GTS parameter ranges from zero to one, with one being the highest score. If GenomeStudio detects three well-separated tight clusters with the correct positions of a normalized theta value and high intensity, the marker gains a higher GTS. The correct positions of clusters regarding the normalized theta value are: "zero" and "1" for homozygous clusters and "0.5" for heterozygous clusters. The value of MAF > 0.1 (Hyten et al. 2008) was set as an initial default to select SNPs potentially useful in marker-assisted breeding.

Data analysis

Genotypic data generated from the project were analyzed using Illumina's GenomeStudio 3.0. JoinMap 4.0 (van Ooijen and Voorrips 2001) was used to create genetic linkage maps.

Linkage disequilibrium and haplotype block assessment

The linkage disequilibrium (LD) between every pair of SNP markers used in the study was computed using the |D′| measure as described in Devlin and Risch (1995). A haplotype block was defined as three or more markers (Phillips et al. 2003) for which all pairwise |D′| values exceeded 0.8. If two or more blocks overlapped, the longest of the blocks was chosen as the final haplotype block.

Results

CRoPS fragment library and CRoPS run

To assess the quality of the fragment library, 84 AFLP fragments were cloned and sequenced. Sequence comparison revealed that four AFLP fragments (4.8%) showed significant homology to cp-DNA while ten fragments (12%) showed significant sequence similarity to mt-DNA. Though the percentage of mitochondrial sequences was higher than expected, it was still acceptable and we proceeded with the CRoPS titration run step, which was necessary to assess the optimal amount of DNA required in the emulsion PCR (em-PCR). The sequences obtained from this titration run (6,751 in total) were used as an additional quality control and were also checked for mt- and cp-DNA. Also, an internal clustering was performed to assess the level of repetitiveness among these sequences. The percentage of mt-DNA reads was 13.7 and the percentage of cp-DNA reads was 5.1. The biggest cluster contained 3.1% of the total reads. Based on these results, we proceeded with the final GS FLX sequencing run.

GS FLX sequencing run and data preprocessing

A total of 650,330 reads were obtained from the genome sequencer. After the first filtering, the number of good reads was 635,719, which included 274,244 DAS-NSS-22 sequences and 361,475 DAS-NSS-25 sequences. The sequence reads were analyzed, restriction sites restored and sequences assembled into contigs. One cluster that was identified in this process contained 264,995 reads. This cluster consisted mainly of repetitive sequences and was discarded. This means that the SNP mining was actually based on 370,724 (635,719–264,995) sequence reads. In total, 2,248 SNPs representing 1,034 loci were identified. However, not all SNP sequences were suitable for GG assay design because of linked polymorphisms within 1–12 nucleotides flanking the target SNP. All SNPs were scanned for sequences that did not have any neighboring polymorphisms within 12 nucleotides of a target SNP. As a result, 1,123 SNPs representing 761 loci were selected for further in silico validation.

In silico validation of SNPs

All 1,123 SNPs were scanned for homoeologous and paralogous sequences. As much as 59 sequences (~5%) were found to be either homoeologous or paralogous and were removed from further genotyping. Twenty sequences retrieved multiple hits in all ten chromosomes suggesting their repetitive nature and were eliminated from further study as well. The remaining SNPs (1,044) were single copy and submitted to the Assay Design Tool (ADT) at http://www.illumina.com (verification date: 9/17/2008). About 8% (85) of the submitted sequences were discarded, as their SNP scores were below 0.5. An SNP score represents the probability that a particular assay will be successfully designed. For instance, 0.5 score suggests that an assay is predicted to have 50% likelihood of success. For the synthesis of a custom OPA, 959 SNPs with scores equal to or more than 0.5 were selected. These 959 SNPs were part of a 1536-plex OPA that contained 577 additional SNPs, which were not part of this study. This OPA was used to genotype 288 F_2 individuals of Pop_1 to generate a genetic linkage map.

Redundancy checks and sequence annotation

To identify non-redundant SNPs for genetic mapping, in silico validated SNPs were further checked for redundancy with respect to existing SNPs in the public domain, namely PanZea SNPs (http://www.panzea.org) and Maize Assembled Genomic Island (MAGI) SNPs (Fu et al. 2005) by sequence alignment. Only 2 out of 959 identified sequences were identical with Panzea SNPs and none with MAGI SNPs. Comparison of SNPs to dbEST at NCBI revealed that 394 out of 959 (49%) were from coding regions/exons, while the remaining 51% of SNPs either represented introns or non-coding areas of the maize genome.

SNP validation and mapping using GoldenGate assay

GG genotyping yielded 660 (69%) polymorphic SNPs and 90 (9%) monomorphic SNPs, while the remaining 207 (22%) assays failed. Thus, the success rate of functional GG assays (polymorphic and monomorphic SNPs) was 78%. Fifty-six polymorphic markers were further excluded

Table 1 Distribution of SNPs in maize genome

Chromosome	Length of the chromosome (cM)	Number of mapped SNPs	Density of DZm SNPs (SNP/cM)	Number of >20 cM gaps
1	132	105	1.26	1
2	148	81	1.82	0
3	90	64	1.42	0
4	92	63	1.46	0
5	134	67	2.00	0
6	95	33	2.88	0
7	116	50	2.32	0
8	135	49	3.86	0
9	103	35	2.94	1
10	109	57	1.91	0
Total	1,154	604		

from mapping because of bad clustering patterns of genotypes and distorted segregation ratios. As a result, 604 SNPs out of 959 were selected for mapping. These SNPs represent 61,004 bp in maize genome. Hence, the overall success rate of the GG assay was ~63%. The markers were found to be evenly distributed across the ten linkage groups of the maize genome in Pop_1 (Table 1). The number of SNPs varies from chromosome to chromosome and ranges from 33 (chr. 6) to 105 (chr. 1) SNPs per linkage group. The average density of SNPs within chromosome 1 was one SNP every 1.26 cM, the highest density among all ten maize chromosomes. Chromosome 8 had the lowest density of SNPs. In the genetic linkage map, gaps over 20 cM were observed only in two chromosomes, 1 and 9, which had 21 and 22 cM gaps, respectively (Table 1).

Identification of haplotype blocks represented by SNP markers

Among 86 maize inbred lines analyzed with 604 SNP markers, 43 haplotype blocks were identified. Distribution of blocks among the chromosomes was uneven. Chromosome 1 was represented by 11 haplotype blocks, while one haplotype block was identified in each of chromosomes 6 and 9 (Table 1S). Almost 31% of all mapped SNPs (188/604) were involved in the formation of haplotype blocks. The physical length of haplotypes ranged from 111 to several million bases. Haplotype blocks represented by SNPs from this study make up ~13% (313 of 2,500 Mb) of the entire maize genome.

Potential usefulness of developed SNP markers in GG assay-based marker-assisted selection projects

To determine the potential value of developed SNP markers in GG assay-based marker-assisted selection (MAS) projects, three major criteria were taken into consideration: (1) the value of the MAF estimate, (2) behavior of an SNP in a segregating population (CT100 parameter) and (3) quality of genotypic clusters and their degree of separation (GTS parameter).

Minor allele frequency estimates

The first criterion that was taken into consideration while assessing markers for their usefulness in MAS projects was the MAF estimate. The higher the MAF value, the more powerful the discriminating ability of an SNP in germplasm analysis and, consequently, the more useful the SNP can be in MAS. To calculate the MAF estimate, a panel of 86 diverse maize inbred lines was designed and genotyped with 604 SNPs. Figure 1 demonstrates the uniform MAF distribution among the five classes. About 20% (122/604) of all SNPs fell into the "0–0.1" class, which made them (by default) not suitable for MAS. The availability of markers with <0.1 MAF was expected, because SNPs developed in this study originated from two genetically different inbred lines, DAS-NSS-22 and DAS-SS-25. Presumably, when parents are too distinct, there is always a risk of discovering a large portion of unique SNPs with <0.1 MAF, which are polymorphic only in this particular cross. To verify this, all SNPs were attempted to be mapped in two more mapping populations, namely Pop_2 and Pop_3. The parents of these two populations were distantly related to each other as well as parents of Pop_1, except the line DAS-NSS-24 (Pop_2), which was genetically close to DAS-NSS-25 (Pop_1) (data not shown).

Mapping efforts confirmed that the "0–0.1" class of SNPs were polymorphic only between DAS-NSS-22 and DAS-SS-25 and were mapped in Pop_1 only (Fig. 1). The "0.11–0.2" class was also largely (60%) represented by SNPs polymorphic only in the paternal cross (Pop_1). Interestingly, SNPs that were mapped in all three

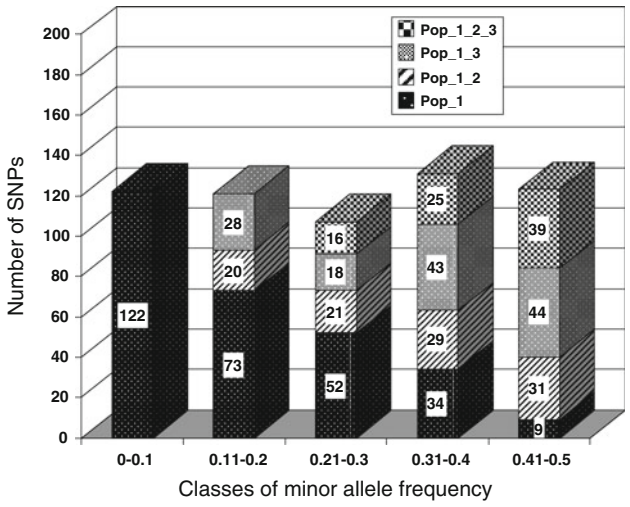

Fig. 1 Minor allele frequencies of SNPs in 86 diverse maize inbred lines. Pop_1 stands for the number of SNPs mapped in Pop_1 only, Pop_1_2, Pop_1_3 and Pop_1_2_3 stand for the number of SNPs mapped between Pop_1 and Pop_2, Pop_1 and Pop_3 and among all three populations, respectively

populations first appear only in the "0.21–0.3" class. The number of SNPs mapped in more than one population accounts for 51% (55) in this class. The "0.31–0.4" and "0.41–0.5" classes include 74% (97) and 92% (114) SNPs shared between two or more populations. Based on these results, a preliminary conclusion was made that SNPs with >0.3 MAF estimates are highly polymorphic and subsequently could be useful in MAS.

GenTrainScore

The GTS of polymorphic SNPs ranged from 0.2642 (DZm-2597774) to 0.9536 (DZm-2612320) (Fig. 2). The graph indicates that a majority of SNPs (61%) fell into the "0.81–0.9" class. In total, 534 SNPs with good and excellent cluster separation patterns had >0.8 GTS. Those SNPs did not need any manual editing of clusters. The remaining SNPs with 0.2< GTS <0.7 were in need of manual editing prior to genotype calling. Although SNPs with 0.2< GTS <0.7 represent a valuable source of information, the texture

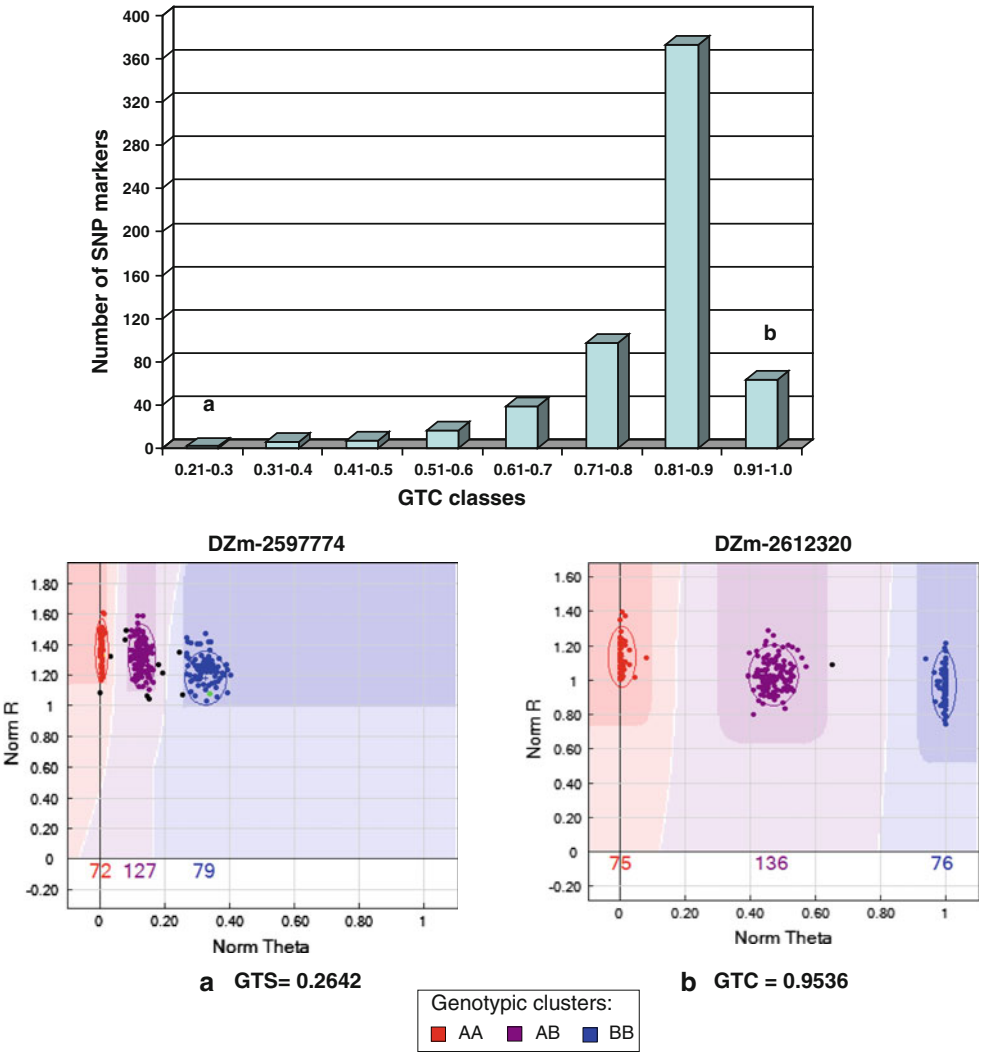

Fig. 2 Classes of SNPs based on Gen_Train_Score in Pop_1. The *lower case letters* over the bars on the chart and beneath the genotypic clusters of SNPs correspond to each other

of their genotypic clusters and the degree of their separation are not adequate for leveraging them in MAS projects. As a result, a GTS value of >0.8 was established as a cutoff value in the selection of SNP markers for GG assay-based MAS projects.

Chi_test100 score

Since GTS does not take into consideration the segregation patterns of an SNP within a population, all selected SNPs were scanned for CT100 parameter, which implements a "goodness of fit" test. The goal was to discard markers with segregation ratios, which strongly deviated from the expected 1(AA): 2(AB): 1(BB) ratio eventhough they had high GTS value. SNPs with <0.5 CT100 parameter demonstrated certain degrees of deviation from the expected ratio, while SNPs with <0.2 CT100 were impossible to map. Thus, >0.5 CT100 value was determined to be used as an additional criterion in choosing SNPs for GG assay-based MAS projects.

Minor allele frequency and SNP assay stability

Since most of the marker-assisted breeding projects involve genotyping a large number of samples, it was important to investigate whether there was any correlation between the value of an MAF estimate of a particular SNP and the stability of SNP genotyping clusters in the GG assay. A total of 436 SNPs with >0.1 MAF, >0.8 GTS and >0.5 CT100, were selected from the study above and examined for the stability and consistency of clustering patterns in the GG assay in the case of a dramatic sample increase. Using the GG assay, a ~6,000 sample panel was genotyped with 426 SNPs. Genotype clustering patterns of 426 SNPs were compared on ~6,000 and 288 (Pop_1) sample size projects. The initial assumption was that if a SNP assay had >0.8 GTC value and demonstrated good cluster separation in a population of 288 samples, it would behave similarly in any sample size project. However, such a consistency was not observed: 249 out of 426 SNP assays failed to maintain clustering stability, which was mainly manifested by shifts of heterozygous clusters toward one of the homozygous clusters that made scoring challenging and unreliable. The example of an unstable SNP is shown in Fig. 3a. The remaining 177 SNPs preserved their excellent clustering abilities regardless of sample number increase. An example of a stable SNP is depicted in Fig. 3b.

It was observed that 76% [188(112 + 76/249)] of all unstable SNPs had <0.3 MAF (Table 2). The ratio of stable/unstable SNPs is 0.29 and 0.43 in the "0.11–0.2" and "0.21–0.3" MAF classes, respectively. However, starting with the "0.31–0.4" MAF class, the number of stable SNPs dramatically increases. The ratio of stable/unstable increases too and is equal to 1.97 and 2.94 in the "0.31–0.4" and "0.41 and 0.5" classes, respectively (Table 2). This finding shows that there is a correlation between the MAF estimate and the stability of the GG SNP assay: the higher the value of MAF estimates, the more stable is the GG assay.

In summary, stringent criteria were established for the selection of highly polymorphic and stable GG SNPs toward their use in marker-assisted selection projects. Based on this study, 120 SNPs (67 + 53) (Table 2) with >0.3 MAF, >0.8 GTC and >0.5 CT100 scores were selected for utilization in GG assay-based MAS projects (Table 2S).

Discussion

SNP discovery in the complex maize genome

The duplicated and the highly repetitive nature of the maize genome has historically been a major impediment to the discovery of true SNPs. In the last 7 years, the maize scientific community has been developing experimental and in silico methods of targeted SNP discovery mainly within the transcriptome (Edwards et al. 2008). Undoubtedly, the transcriptome is a valuable source to mine SNPs. However, in organisms with a duplicated genome such as maize, the sources and methods of SNP discovery have been significantly debilitated by the presence of homoeologs and paralogs. Moreover, use of expressed sequences as the only source of SNP discovery deprives us of the repertoire of valid polymorphisms within introns or intergenic spaces. Recent advances in genome complexity reduction methods coupled with SBS technologies allow for genome-wide SNP discovery targeting both coding and non-coding single and low-copy regions of a genome (van Orsouw et al. 2007; Baird et al. 2008; Gore et al. 2009a).

In this study, we have applied CRoPS technology for genome-wide identification of polymorphic SNPs between two genetically distinct maize inbred lines of North and South American origins. Our efforts resulted in the discovery of over 1,123 polymorphic SNPs with a high probability of the absence of any other polymorphisms within at least 12 nucleotides flanking the target SNP. The number of discovered allelic variations in this study is much lower compared to 126,683 identified by Gore et al. (2009a) using the HMPR technique. Although the HMPR technique and CRoPS use methylation-sensitive enzymes to reduce genome complexity, the latter applies an AFLP component as a second step of complexity reduction, which drastically reduces the number of false variations. Our

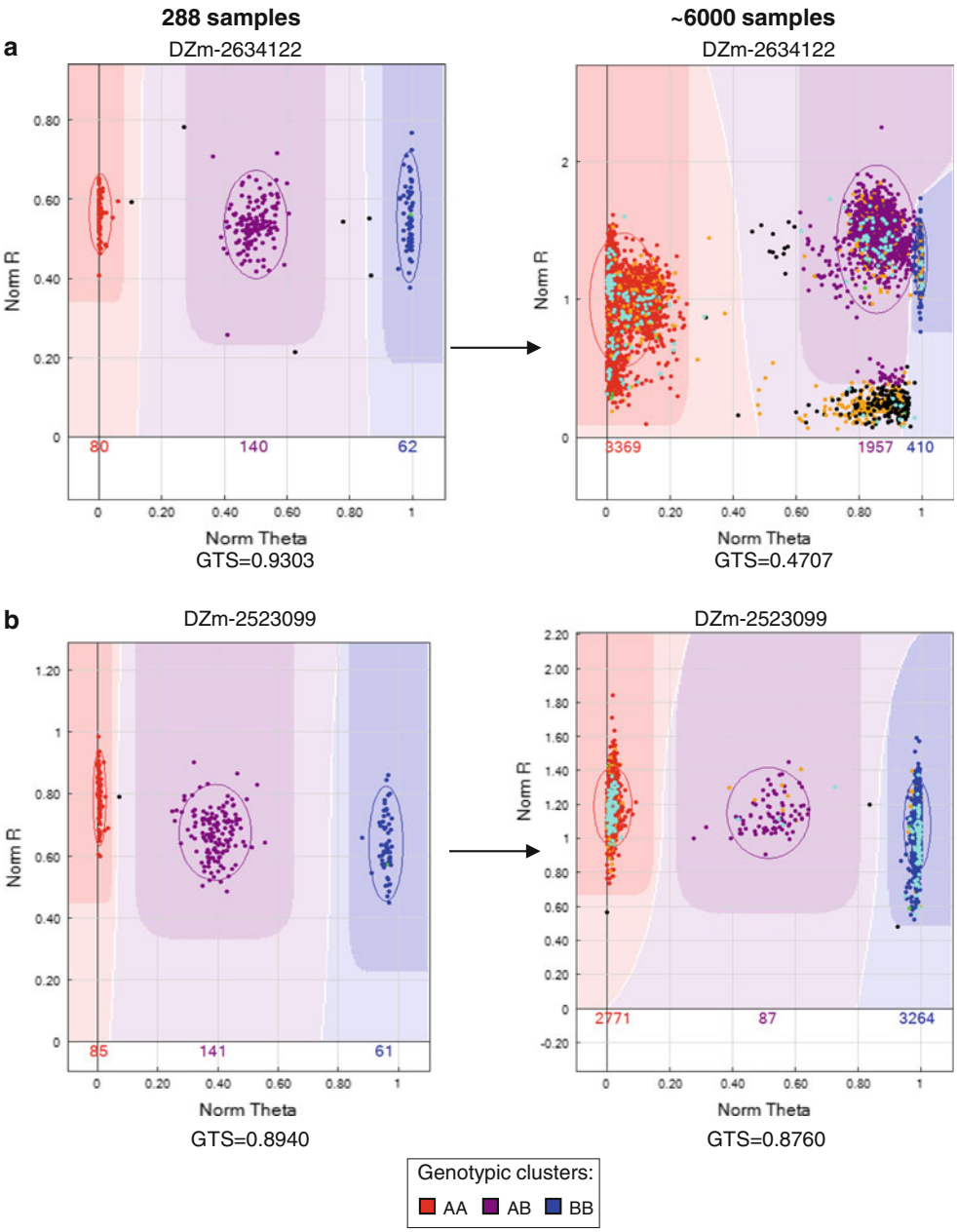

Fig. 3 Examples of performance of SNPs with >0.1 minor allele frequency based on the genotyping of 288 (*left panel*) and ~6,000 (*right panel*) samples. **a** DZm-2634122 is an example of an unstable SNP assay: with 288 samples it exhibits perfect cluster separation (GTS = 0.9303), while significantly worsens its performance with increased number of samples (~6,000). **b** DZm-2523099 SNP demonstrates stable performance regardless of the number of samples of genotype

Table 2 Correlation between minor allele frequency estimate and stability of GoldenGate SNP assay

Status of GG SNP assay	Classes of minor allele frequencies			
	0.11–0.2	0.21–0.3	0.31–0.4	0.41–0.5
Stable (*S*)	33	33	67	53
Unstable (*U*)	112	76	34	18
S/*U* ratio	0.27	0.43	1.97	2.94

validation experiments showed that 63% of all discovered SNPs were true representatives of the overall genome. Thus, despite the relatively lower throughput of SNPs discovered by the AFLP component of CRoPS, the detection efficiency of valid variations is still high. The AFLP component might increase the bias toward variations detected by nucleotides at the end of the +1 selective primer; therefore, one may consider this a limitation of the technology. However, this bias can be eliminated by varying the numbers of selective nucleotides. The number of discovered SNPs could also be increased by using several enzyme combinations. With these two aspects of genome complexity reduction, namely choice of different enzyme combinations and varying the number of selective nucleotides, the method can be modified for any target species, including polyploid organisms. The discrimination

power of CRoPS has been achieved due to the two-step complexity reduction component of the technology: gene copies that have SNPs either in the restriction sites or selective nucleotides have a small chance to end up in the same pre-amplification reactions and will not be sequenced simultaneously. In our experiments, CRoPS demonstrated the superiority in eliminating repetitive and duplicated sequences. We scanned all putative SNPs for the presence of homoeologous, paralogous and repetitive sequences and observed that only 7% of them were represented by high copy sequences.

Conversion of discovered SNPs into GoldenGate assay

It has been previously reported that the conversion rate of CRoPS-discovered SNPs to a successful assay such as SNPWave (Keygene N.V., Wageningen, The Netherlands) was over 75% (van Orsouw et al. 2007). However, the experiment was based on a small subset of 30 loci. In our experiments, GG assay was developed for nearly 1,000 SNPs representing ~700 loci. The GG assay is a highly multiplexed chemistry and allows parallel genotyping of up to 1,536 SNPs in one reaction (Fan et al. 2003). This feature of the GG assay is very attractive for a rapid validation of a large number of SNP markers. However, the chemistry also applies stringent requirements to the quality of SNP sequences and is very sensitive to variations from repetitive and duplicated regions. For example, availability of oligos designed based on repetitive sequences in the OPA will consume tremendous amount of enzymatic resources during implementation of GG genotyping and might result in complete failure of a whole experiment. Whereas, GG assay designed from duplicated sequences will result in amplification of both target and paralogous loci resulting in distorted segregation ratios or severely compressed clusters, where heterozygous clusters will not be easily distinguishable from one or both of the homozygous classes (personal experience, data now shown). Consequently, the use of the GG assay for the validation of SNPs discovered in highly duplicated genomes might seem disadvantageous. Fortunately, recent experiments in soybean (Hyten et al. 2008) and maize (Jones et al. 2009) demonstrated that the GG assay can be successfully applied for SNP genotyping in highly duplicated organisms. For instance, success rates of GG assays in soybean and maize were reported to be 89% (Hyten et al. 2008) and 88–93% (Jones et al. 2009), respectively. In barley, an organism with a lower number of paralogous genes, the success rate of the GG assay (90%) was similar to soybean and maize (Rostoks et al. 2006).

It must be mentioned that a majority of assayed SNPs from the above-mentioned studies were discovered within genes. Whether a robust GG assay can be developed from non-coding intergenic sequences remained unknown. Intergenic regions in maize are usually highly variable (Fu and Dooner 2002), which complicates marker design due to linked polymorphisms. Additionally, higher GC content is also characteristic for those regions. The Illumina ADT is very sensitive to both features (linked polymorphisms and higher GC content) of non-genic regions, which can hinder the assay design. In our study, the success rate of the functional GG assay (polymorphic + monomorphic assays) was about 78%. The relatively low conversion rate compared to the one previously reported by Jones et al. (2009) can be explained by the fact that a little over half of the discovered sequences (59%) represent non-genic regions. Our experiments have demonstrated that robust SNP assays from intergenic regions can be successfully developed. Although gene-based SNP markers are considered ideal markers, SNP markers from non-genic regions can play an important role in filling the gaps in genetic maps and serve as "bridges" connecting gene-rich islands in the maize genome.

We were able to map 604 SNPs (63% of all GG-converted markers), which were evenly distributed across all ten maize chromosomes with the marker density of one SNP per ~2 cM in a population of 288 F_2 individuals (Table 1). Thus, one run of CRoPS per bi-parental cross generated a very balanced collection of true SNPs from both genic and non-genic components of the maize genome. Moreover, the number of true SNPs generated from this method was sufficient to create a high-density genetic linkage map, which could be used for the initial gene/QTL mapping (Schőn et al. 2003). Importantly, most of the SNPs developed in this study appeared to be unique when compared to existing public SNPs including Panzea and MAGI SNPs. Both Panzea and MAGI SNPs were detected by re-sequencing of coding sequences, including unigenes (http://www.panzea.org) and transcriptome sequences (Barbazuk et al. 2007), respectively.

We used 604 SNPs to score a diverse panel of 86 maize inbred lines to investigate LD in a genome-wide fashion. It was previously reported that in maize, LD decay distance was on average less than 2,000 bp (Remington et al. 2001). Later studies suggested that in commercial inbred lines, LD decay may span more than 100–500 Kb (Ching et al. 2002; Tian et al. 2009). The recently developed first-generation haplotype map of maize possesses evidence of much longer haplotypes spanning several million bases (Gore et al. 2009b). In our study, we identified 43 haplotype blocks conserved among 86 maize inbred lines. Interestingly, out of 43 haplotype blocks, 6 decayed within the previously reported 2,000 bp. The remaining haplotype blocks extended over 65 Kb–40 Mb, which is in correspondence with the above-mentioned reports.

Development of criteria for selection of highly polymorphic SNP markers for marker-assisted selection projects

Several methods have been developed to determine the level of polymorphism of a particular marker. These include the effective number of alleles (Hedrick 2000), expected heterozygosity value (Hedrick 2000) and polymorphism information content (Ott 2001). Although those methods were mainly designed for population genetics research (Aminafshar et al. 2008), their applicability in molecular breeding have been demonstrated in Zhu et al. (2003), Hyten et al. (2008) and Jones et al. (2009). In this report, we calculated minor allele frequencies to determine the level of polymorphism of each SNP marker. We created a panel of 86 genetically diverse proprietary maize inbred lines of different geographical origins and heterotic patterns. Using the GG assay, these lines were genotyped and the MAF was calculated for each SNP. Hyten et al. (2008) suggested that SNPs with a >0.1 MAF could potentially be useful in marker-assisted selection, QTL mapping and association analysis. In our study, about 80% of the 604 mapped SNPs had a >0.1 MAF value (Fig. 1). To determine the minimum threshold of the MAF value in which the selection of highly polymorphic SNPs would begin, two more F_2 populations (Pop_2 and Pop_3) were genotyped using the same set of SNPs. The parents of Pop_2 and Pop_3 were genetically distant from the parents of Pop_1 from which the SNPs were developed. We postulated that highly polymorphic SNPs could be mapped in all three mapping populations. The mapping efforts demonstrated that the first two MAF classes ($0 < MAF < 0.2$) were heavily populated with SNPs unique to Pop_1 (247/350). In contrast, SNPs mapped in two or more populations have a >0.2 MAF value (Fig. 1). Thus, SNPs with $0.1 < MAF < 0.2$ do not appear to be highly polymorphic. Consequently, they are inadequate in their ability to discriminate diverse maize germplasms and to be informative in molecular breeding projects. Furthermore, we determined the correlation between the MAF value and the stability of the GG SNP assay regardless of the sample size of a project and genetic distance between samples. We developed a new panel of ~6,000 genetically diverse samples and GG-genotyped them with a subset of 426 SNPs (MAF >0.2). We compared the genotype clustering patterns of the 426 SNPs between both the 288 and ~6,000 sample size projects. Surprisingly, a large portion of SNPs that demonstrated excellent clustering in a smaller sample size project did not show adequate clustering in the larger sample size project. The majority of SNP assays that demonstrated stable performance regardless of the project size had a >0.3 MAF (Table 2). Despite the evident correlation between SNP assay stability and the MAF value, this association is not absolute. About 28% [52/182 (Table 2)] of all SNPs with a MAF >0.3 were still unstable. Based on these findings, we conclude that SNPs with >0.3 MAF are more informative and genetically more stable than SNPs with $0.1 < MAF < 0.3$.

The success of SNP marker application in a MAS project using the GG assay depends not only on the level of polymorphism of a particular marker, but also on the clustering patterns of genotypes. Scoring a large number of SNP genotypes becomes increasingly difficult if the genotypic clusters of the SNPs are close to each other (theta compression) or the heterozygous cluster is shifted toward either of the homozygous clusters. To address this, in addition to the MAF value, a GTS parameter of Illumina's GenomeStudio software was used to identify SNPs with excellent clustering patterns. We experimentally proved that SNPs with a GTS >0.8 had well-defined clusters adequate to successfully conduct marker-assisted selection. Finally, due to duplicated sequences, certain SNPs with excellent clustering patterns demonstrated distorted segregation ratios, which severely affected our mapping efforts. To avoid an SNP assay, which amplifies both target SNP and its paralog, we employed a CT100 parameter of the GenomeStudio. We established that SNPs with CT100 >0.5 segregated in Mendelian fashion.

The novelty of our study is that we have determined and clearly stated a specific combination of stringent criteria (>0.3 MAF, >0.8 GTS and >0.5 CT100) necessary for the identification of highly polymorphic and genetically stable SNP markers. To the best of our knowledge, a combination of these criteria has not been employed in the available scientific literature. Due to these criteria, we identified a subset of 120 highly polymorphic and genetically stable SNP markers to leverage in GG assay-based marker-assisted selection projects. More than 25% of these SNPs represented 21 haplotypes out of 43 identified in this study. The sequence information of 120 assays and sequences from which they were developed have been made available to the public (Table 2S). The information on the development of SNPs from complexity reduced genomic DNA and the criteria used for the identification of stable and highly polymorphic SNPs in a complex genome such as maize will be of great value for the molecular genetics and breeding community.

Acknowledgments We thank Sarah Bohl and Jan Eric Backlund of Trait Genetics and Technologies (TG&T) Department of Dow AgroSciences for providing us with the DNA of 86 maize inbred lines. We also thank Rebecca Aus of TG&T for genotyping the 6000-sample marker application project. Our special thanks go to Ryan Gibson of TG&T for proofreading and editing the manuscript.

References

Aminafshar M, Amirina C, Vaez Torshizi R (2008) Genetic diversity in buffalo population of Guilan using microsatellite markers. J Anim Vet Adv 7:1499–1502

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3:e3376

Barbazuk WB, Emrich SJ, Chen HD, Schnable P (2007) SNP discovery via 454 transcriptome sequencing. Plant J 51:910–918

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

Chagné D, Batley J, Edwards D, Forster JW (2007) Single nucleotide polymorphisms genotyping in plants. In: Oraguzie NC, Rikkerink EHA, Susan E, Gardiner SE, De Silva HN (eds) Association mapping in plants. Springer, New York, pp 77–94

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

Chen X, Levine L, Kwok P-Y (1999) Fluorescence polarization in homogeneous nucleic acid analysis. Genome Res 9:492–498

Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine scale mapping. Genomics 29:311–322

Edwards KJ, Poole RL, Barker GLA (2008) SNP discovery in plants. In: Henry RJ (ed) Plant genotyping II: SNP technology. CABI, Wallingford, Oxfordshire, pp 1–29

Emberton J, Ma J, Yuan Y, SanMiguel P, Bennetzen JL (2005) Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries. Genome Res 15:1441–1446

Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P, Galver L, Hunt S, McBride C, Bibikova M, Rubano T, Chen J, Wickham E, Doucet D, Chang W, Campbell D, Zhang B, Kruglyak S, Bentley D, Haas J, Rigault P, Zhou L, Stuelpnagel J, Chee MS (2003) Highly parallel SNP genotyping. Cold Spring Harb Symp Quant Biol 68:69–78

Fu H, Dooner HK (2002) Intraspecific violation of genetic colinearity and its implications in maize. Proc Natl Acad Sci USA 99:9573–9578

Fu Y, Emrich SJ, Guo L, Wen TJ, Ashlock DA, Aluru S, Schnable PS (2005) Quality assessment of maize assembled genomic islands (MAGIs) and large-scale experimental verification of predicted genes. Proc Natl Acad Sci USA 102:12282–12287

Ganal M, Altmann T, Röder MS (2009) SNP identification in crop plants. Curr Opin Biotechnol 12:211–217

Gaut BS, Doebley JF (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. Proc Natl Acad Sci USA 96:6809–6814

Gore MA, Wright MH, Ersoz ES, Bouffard P, Szekeres ES, Jarvie TP, Hurwitz BL, Narechania A, Harkins TT, Grills GS, Ware DH, Buckler ES (2009a) Large-scale discovery of gene-enriched SNPs. The Plant Genome 2:121–133

Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH, Buckler ES (2009b) A first-generation haplotype map of maize. Science 326:1115

Hedrick PW (2000) Genetics of populations. Jones and Barlett, Sudbury, MA

Hyten DL, Song Q, Choi I-Y, Yoon M-S, Specht JE, Matukumalli LK, Nelson RL, Shoemaker RC, Young ND, Cregan PB (2008) High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor Appl Genet 116:945–952

Illumina SNP genotyping (2006) Infinium Assay II Workflow. URL:http://www.illumina.com/Documents/products/workflows/workflow_infinium_ii.pdf

Jones E, Chu WC, Ayele M, Ho J, Bruggeman E, Yourstone K, Rafalksi A, Smith OS, McMullen MD, Bezawada C, Warren L, Babayev J, Basu S, Smith S (2009) Development of single nucleotide polymorphism (SNP) markers for use in commercial maize (*Zea mais* L.) germplasm. Mol Breeding 24:165–176

Livak KJ, Flood SJ, Marmaro J, Giusti W, Deetz K (1995) Oligonucleotide with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. Genome Res 4:357–362

Mardis ER (2007) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141

Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution and transcriptional activity of repetitive elements in the maize genome. Genome Res 11:1660–1676

Morozova O, Marra MA (2008) Application of next-generation sequencing technologies in functional genomics. Genomics 92:255–264

Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME (2007) Microarray-based genomic selection for high-throughput resequencing. Nat Methods 4:907–909

Olivier M (2005) The Invader® assay for SNP genotyping. Mutat Res 573:103–110

Ott J (2001) Program HET Version 1.8, utility programs for genetic analysis of linkage. Rockfeller University, NY

Palmer LE, Rabinowicz PD, O'Shaughnessy A, Balija V, Nascimento L, Dike S, de la Bastide M, Martienssen RA, McCombie WR (2003) Maize genome sequencing by methylation filtration. Science 302:2115–2117

Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF, Ankener WM, Alfisi SV, Kuo F-S, Camisa AL, Pazorov V, Scott KE, Carey BJ, Faith J, Katari G, Bhatti HA, Cyr JM, Derohannessian V, Elosua C, Forman AM, Grecco NM, Hock CR, Kuebler JM, Lathrop JA, Mockler MA, Nachtman EP, Restine SL, Varde SA, Hozza MJ, Gelfand CA, Broxholme J, Abecasis GR, Boyce-Jacino MT, Cardon LR (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet 33:382–387

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

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci USA 98:11479–11484

Rostoks N, Ramsay L, Mackenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, Graner A, Close TJ, Waugh R (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proc Natl Acad Sci USA 103:18656–18661

Schön CC, Utz HF, Grob S, Truberg B, Openshaw S, Melchinger AE (2003) Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics 167:485–498

Stuart CN, Via LE (1993) A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR application. Biotechniques 14:748–750

Tian F, Stevensa NM, Buckler ES (2009) Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. Proc Natl Acad Sci USA 106(Suppl 1):9979–9986

van Ooijen JW, Voorrips RE (2001) JoinMap® 3.0, software for the calculation of genetic linkage maps. Plant Research International, Wageningen, The Netherlands

van Orsouw N, Hogers RC, Janssen A, Yalcin F, Snoeijers S, Verstege E, Scheiders H, van der Poel H, van Oeveren J, Verstegen H, van Eijk MJT (2007) Complexity reduction of polymorphic sequences (CRoPS^{TM}): a novel approach for large-scale polymorphism discovery in complex genomes. PLoS One 11:1–10

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

Yuan Y, SanMiguel PJ, Bennetzen JL (2003) High-C_0t sequence analysis of the maize genome. Plant J 34:249–255

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

Wright WT, Heggarty SV, Young IS, Nicholls DP, Whittall R, Humphries SE, Graham CA (2008) Multiplex MassARRAY spectrometry (iPLEX) produces a fast and economical test for 56 familial hypercholesterolaemia-causing mutations. Clin Genet 74:463–468

Zhu YL, Song QL, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003) Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134