Theor Appl Genet (2012) 125:1487–1501
DOI 10.1007/s00122-012-1928-1

ORIGINAL PAPER

Quality control genotyping for assessment of genetic identity and purity in diverse tropical maize inbred lines

Kassa Semagn · Yoseph Beyene · Dan Makumbi · Stephen Mugo · B. M. Prasanna · Cosmos Magorokosho · Gary Atlin

Received: 1 June 2011 / Accepted: 16 June 2012 / Published online: 17 July 2012

Abstract Quality control (QC) genotyping is an important component in breeding, but to our knowledge there are not well established protocols for its implementation in practical breeding programs. The objectives of our study were to (a) ascertain genetic identity among 2–4 seed sources of the same inbred line, (b) evaluate the extent of genetic homogeneity within inbred lines, and (c) identify a subset of highly informative single-nucleotide polymorphism (SNP) markers for routine and low-cost QC genotyping and suggest guidelines for data interpretation. We used a total of 28 maize inbred lines to study genetic identity among different seed sources by genotyping them with 532 and 1,065 SNPs using the KASPar and GoldenGate platforms, respectively. An additional set of 544 inbred lines was used for studying genetic homogeneity. The proportion of alleles that differed between seed sources of the same inbred line varied from 0.1 to 42.3 %. Seed sources exhibiting high levels of genetic distance are mis-labeled, while those with lower levels of difference are contaminated or still segregating. Genetic homogeneity varied from 68.7 to 100 % with 71.3 % of the inbred lines considered to be homogenous. Based on the data sets obtained for a wide range of sample sizes and diverse genetic backgrounds, we recommended a subset of 50–100 SNPs for routine and low-cost QC genotyping, verified them in a different set of double haploid and inbred lines, and outlined a protocol that could be used to minimize errors in genetic analyses and breeding.

Communicated by M. Bohn.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-012-1928-1) contains supplementary material, which is available to authorized users.

K. Semagn (✉) · Y. Beyene · D. Makumbi · S. Mugo · B. M. Prasanna
International Maize and Wheat Improvement Center (CIMMYT), Village Market, P.O. Box 1041, Nairobi 00621, Kenya
e-mail: k.semagn@cgiar.org

C. Magorokosho
International Maize and Wheat Improvement Center (CIMMYT), 12.5 km peg Mazowe Road, Mount Pleasant, P.O. Box MP163, Harare, Zimbabwe

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

Introduction

Maintenance of inbred line genetic purity (homogeneity) and confirmation of the genetic identity of the same inbred line maintained at different locations are important quality control functions in maize breeding programs. These functions have become more critical due to the stringent intellectual property requirements governing plant breeding and variety registration in many countries, and the reductions in the cost of DNA marker technologies that permit highly accurate identification of samples. Microsatellites or simple sequence repeat (SSR) markers are widely used by maize researchers because they are available in large numbers in the public domain (MaizeGDB; http://www.maizegdb.org), co-dominant, multiallelic, highly polymorphic even in closely related individuals, can be exchanged between laboratories, and have uniform distribution in the genome (Gupta et al. 2002; Prasanna et al. 2010). Recent advances in molecular technology, however, have emphasized single-nucleotide polymorphism (SNP) markers (Hamblin et al. 2007). Because of their low genotyping cost per data point, high genomic abundance, locus-specificity, codominance, simple documentation, and potential for high throughput analysis, SNPs have emerged as powerful tools for genetic analyses and molecular marker-assisted breeding for maize improvement, and are emerging as markers of choice for quality control (QC) applications.

The Illumina GoldenGate platform is useful for genotyping with 1,536 SNPs simultaneously in each sample (Low et al. 2006). This platform is suitable for large-scale studies that require genotyping of individual samples with thousands of SNPs. However, due to its high level of multiplexing, total cost per DNA sample, and sometimes lengthy process of initial assay development, the platform becomes unmanageable in studies where only a small to moderate number of SNPs need to be analyzed in a large number of samples, as is the case in breeding program for QC genotyping. In such cases, single-plex SNP genotyping platforms are more suitable (Low et al. 2006). KASPar is a new single-plex SNP genotyping platform that was developed at KBioscience (http://www.kbioscience.co.uk) and is currently used for routine genotyping of at least 24 samples with up to thousands of SNPs. The method uses allele-specific amplification followed by fluorescence detection for genotyping.

Over the past four decades, maize breeders at the International Maize and Wheat Improvement Center (CIMMYT) have developed numerous maize inbred lines, populations, and open-pollinated varieties using germplasm from different backgrounds. The International Maize and Wheat Improvement Center maize inbred lines (CMLs) are widely used by various public and private sector institutions worldwide for different purposes, including hybrid production, pedigree breeding, development of populations for mapping quantitative trait loci (QTL) and molecular breeding, doubled haploid production, and transformation with trangenes, and for studies conducted to understanding molecular evolution (Liu et al. 2003). Therefore, it is important for CIMMYT to maintain these inbred lines as stable and pure entities.

Genetic variability among different sources of maize inbred lines with the same identification and origin has been observed since the beginning of the inbred-hybrid system (Jones 1945). Fleming et al. (1964) evaluated inbred lines maintained by different researchers working in different environments and breeding programs to examine maintenance of genetic purity after several years of sexual propagation by self- or sib-pollinations. They found significant variations in most of the inbred lines. Russell and Vega (1973) evaluated 11 maize inbred lines maintained at different stations for 10 years and found significant differences for several quantitative traits among two inbreds. Gethi et al. (2002) genotyped six inbred lines from 14 sources with SSR markers and reported significant differences among samples of the same inbred lines collected from eight sources. Heckenberger et al. (2002, 2003) genotyped nine inbred and five double haploid (DH) maize lines of different seed sources with amplified fragment length polymorphic (AFLP) and SSR markers, and reported genetic distances of up to 0.120 between different samples of the same line. Yan et al. (2009) genotyped 21 CIMMYT maize lines (CML) maintained at CIMMYT and North Carolina State University for over 30 years with SNP markers and reported SNP mismatch rates that ranged from 0.2 to 19.7 %. In most studies that reported variability in qualitative and quantitative traits of long-time inbred lines, residual heterozygosity and mutation have been cited as the most frequent causes of heritable variations. The major weaknesses of these studies include small sample size or limited number of markers, lack of clear guidelines in data interpretation, and no suggested protocols for routine QC genotyping. Genetic similarity or distance calculated from empirical and simulated molecular marker data have been used for the identification of essentially derived varieties (Heckenberger et al. 2002, 2003, 2005, 2006). However, the genetic distances reported between different seed sources of the same line designation were too small in suggesting guideline in practical breeding programs. The objectives of this study were therefore to (1) develop and test guidelines for verifying genetic identity among different sources for the same inbred line, (2) evaluate the extent of genetic homogeneity within CIMMYT inbred lines, and (3) identify a subset of highly informative SNP markers for routine and low-cost QC genotyping and suggest guidelines for data interpretation.

Materials and methods

Plant materials and DNA preparation

Two sets of maize inbred lines were first used for the genetic identity studies. The first set, hereafter called set 1, consisted of two seed sources of 19 inbred lines from a breeder in Kenya (source A, abbreviated as SA) and the CIMMYT gene bank and/or a breeder in Mexico (SD). The second set of inbred lines (set 2) consisted of two to four seed sources of 22 inbred lines (SA and SD plus SB from the CIMMYT maize breeding program in Kenya and SC from the CIMMYT maize breeding program in Zimbabwe). CIMMYT breeders have used very diverse breeding method and generation of extraction of lines. Pedigree notation used by the breeders was also diverse and idiosyncratic, although based on standard methods. In the early period of pedigree breeding at CIMMYT, lines were extracted by selfing directly out of pools and populations. This practice has been gradually superseded by

Table 1 List of the 19 inbred lines in set 1 and the 22 inbred lines in set 2 used in the present study

Set	Name	Standard pedigree	Seed source
1	CKL05025	P100-C6-200-1-1-B	SA and SD
1	CML144	P62-C5-FS182-2-1-2-B–B-3-1-B	SA and SD
1	CML159	P63-C2-FS5-1-3-1-B-2-1-1-B	SA and SD
1	CML197	MSR-270-2-B*3-5-1-B	SA and SD
1	CML202	ZSR923-B*4-5-1-B	SA and SD
1	CML204	7794-4-1-B*9-1-4-7-4-5-B	SA and SD
1	CML312	S89500-F2-2-2-1-1-B	SA and SD
1	CML395	90323B-1-B-1-B	SA and SD
1	CML440	G16SEQ-C1-F47-2-1-2-1-B	SA and SD
1	CML442	(M37W/ZM607-#-B-F37SR-2-3SR-6-2-X)-8-2-X-1-B	SA and SD
1	CML444	P43-C9-1-1-1-1-1-B	SA and SD
1	CML445	(TUXPSEQ-C1-F2/P49SR)-F2-45-7-5-1-B	SA and SD
1	CML488	DTPW-C8-F31-4-2-1-5-B	SA and SD
1	CML489	(CML202/LAPOSTASEQ-C3-FS297-2-1-1-2-2-B)-B-3-1-1-8-B	SA and SD
1	CML511	(CML389/CML176)-B-29-2-2-1-B	SA and SD
1	CML78	G32-C19-HS32-1-#-2-B-#*3-3-B	SA and SD
1	CZL00003	DRB-F2-60-1-1-1-B	SA and SD
1	CZL03007	(CML445/ZM621B)-2-1-2-3-1-B	SA and SD
1	CZL03014	MAS(MSR/CML312)-117-2-2-1-B	SA and SD
2	CKL05017	(CML387/CML390)-B-1-1-4-B	SB**
2	CKL05018	(CML387/CML390)-B-1-2-1-B	SB**
2	CKL05022	(CML387/CML390)-B-1-1-5-#-B	SB**
2	CKL05023	(CML388/CML206)-B-4-2-1-B	SB**
2	CML144	P62-C5-FS182-2-1-2-B–B-3-1-B	SA, SB and SC
2	CML158	EV8762SR-2-1-B-1-B	SA and SD
2	CML159	P63-C2-FS5-1-3-1-B-2-1-1-B	SA, SB, SC and SD
2	CML197	MSR-270-2-B*3-5-1-B	SA, SB, SC and SD
2	CML202	ZSR923-B*4-5-1-B	SA, SC and SD
2	CML204	7794-4-1-B*9-1-4-7-4-5-B	SA and SB
2	CML312	S89500-F2-2-2-1-1-B	SB and SC
2	CML334	P590-C3-F374-2-1-2-B-#-3-3-B	SA and SB
2	CML395	90323B-1-B-1-B	SB, SC and SD
2	CML442	(M37W/ZM607-#-B-F37SR-2-3SR-6-2-X)-8-2-X-1-B	SB and SC
2	CML443	(AC8342/IKENNE-1-8149SR//G9A)-C1-F1-500-4-X-1-1-B–B-1-B	SA and SB
2	CML444	P43-C9-1-1-1-1-1-B	SA, SB and SC
2	CML488	DTPW-C8-F31-4-2-1-5-B	SA, SB and SC
2	CML511	(CML389/CML176)-B-29-2-2-1-B	SA and SD
2	CML78	G32-C19-HS32-1-#-2-B-#*3-3-B	SA and SB
2	CZL03014	CML539	SA and SC
2	LaPostaSeqC7	LAPOSTASEQ-C7-F71-1-2-1-1-B	SA and SB
2	P300C5S1B	P300-C5-B–B-2-3-2-#-#-1-1-B	SA and SB

Set 1 was genotyped with 532 SNPs using KASPar while set 2 was genotyped with 1065 SNPs using GoldenGate genotyping platforms

SA source A, *SB* source B, *SC* source C, *SD* source D

** Two seed batches of different generations from source 2B were used

conventional pedigree selection in segregating populations derived from crossing two inbreds. Lines may be derived from F_3, F_4, F_5, or later generations. Table 1 summarizes the list of the inbred lines in set 1 and set 2 and their pedigrees, which are available from the international maize information system (IMIS) database on the CIMMYT website (http://www.imis.cimmyt.org). Thirteen out of the 22 inbred lines in set 2 were the same as those in set 1, but additional seed sources were included in set 2 and/or the seed sources in set 2 were different from set 1. Two seed sources of four inbred lines from set 1 (CML159, CML197, CML202 and CML511) were included in set 2 to serve as positive controls for comparing the SNP data for the same markers from the GoldenGate (Illumina, Inc., USA) and KASPar (KBioscience, UK; http://www.KBioscience.co.uk) genotyping platforms. Altogether, a total of 28 inbred lines were included in both set 1 and set 2. For genetic purity (homogeneity) studies, we included another 544 inbred lines, hereafter referred to as set 3, with details of these lines described in another paper (Semagn et al. 2012). Finally, we also included a different set of 1,306 samples (643 DH lines developed for the water efficient maize for Africa (WEMA) project and 663 inbred lines with various traits of interest), hereafter referred to as set 4, to verify the reliability of a subset of 50–100 SNPs that we recommended for QC genotyping.

Seedlings for all genotypes were raised in plastic seed trays for about 2 weeks, until they reached the 3–4 leaf stage, in a plastic house at the Biosciences Center for Eastern and Central Africa (BecA) hub in Nairobi, Kenya. About equal amounts of leaf tissue were harvested from ten plants, bulked, cut into pieces with scissors, and transferred into 1.2 mL strip tubes that contained two 4-mm stainless steel grinding balls. The tissue was freeze-dried for 3 days using a Labconco freeze dryer (http://www.labconco.com) as described in the user's manual. The lyophilized leaf samples were ground into fine powder using a GenoGrinder at 1,500 strokes/min for 2 min. Genomic DNA was extracted using a modified version of the high-throughput mini-prep cetyl trimethyl ammonium bromide (CTAB) method (Mace et al. 2003). The quality of the isolated DNA was checked after running aliquots of DNA samples on a 0.8 % agarose gel that contained 0.3 µg/mL SYBR safe DNA gel stain (Invitrogen). Deoxyribonucleic acid concentration was measured using NanoDrop ND-1000 spectrophotometer.

Genotyping and statistical analyses

SNP genotyping was carried out using the GoldenGate (Illumina, Inc., USA) and/or KASPar (KBioscience, UK; http://www.KBioscience.co.uk) platforms. The inbred lines in set 2 and set 3 were first genotyped with the CIMMYT 1,536 random SNP chip (Lu et al. 2009) using an Illumina BeadStation 500 G (Illumina, San Diego, CA, USA). Genotyping was done at the Cornell University Life Sciences Core Laboratories Center as described elsewhere (Fan et al. 2006). Alleles were called using the Illumina BeadStudio genotyping software as described in the user manual. Each SNP was checked manually and rescored whenever any error was observed in the clustering of the homozygous and heterozygous genotypes. Of the 1,536 SNPs used for genotyping set 2 and set 3 lines, only 1,065 SNPs (69.3 %) were maintained for statistical analyses. Set 1 was then genotyped with a subset of 540 SNPs using the KASPar system. The 540 SNPs were selected among the 1,065 SNPs using the following criteria: genome coverage, minor allele frequency (Yan et al. 2009), polymorphism information content (Botstein et al. 1980), and agreement between the two genotyping platforms. Allele calling for the 540 SNPs was done by KBioscience. Statistical analyses were performed using the data of 532 SNPs after excluding eight SNPs which were monomorphic across all samples. Set 4 was genotyped with a subset of 50 SNPs recommended by CIMMYT for routine QC genotyping using KASPar assay.

Allele frequency-based Roger's genetic distance (Rogers 1972) was calculated using PowerMarker version 3.25 (Liu and Muse 2005) and used for cluster analysis. A dendrogram was constructed using the neighbor-joining algorithm implemented in PowerMarker and the resulting trees were visualized with MEGA version 5 software (Tamura et al. 2007). Mantel tests (Mantel 1967) were used to compute correlations between (a) different genetic distance matrices derived from different numbers of markers, and (b) cophenetic matrices derived from phenograms produced by each marker number and the original genetic distance matrices. As described elsewhere (Rholf 1993), cophenetic correlation can be used as a measure of goodness of fit for a cluster analysis and it can be interpreted subjectively as follows: $r \geq 0.9$ very good fit; $0.8 \leq r < 0.9$ good fit, $0.7 \leq r < 0.8$ poor fit, and $r < 0.7$ very poor fit. Mantel tests were performed using NTSYS-pc (numerical taxonomy and multivariate analysis system), version 2.11 (Rholf 1993). Line homogeneity was calculated as one minus the proportion of SNPs that were heterogeneous (loci that were not homozygous due either to heterozygosity or bulking of two homozygote genotypes) within the same seed source.

Causes of genetic variation and threshold scenarios

Several technical reasons related to the molecular marker technologies have been reported as possible sources of variation among different samples of the same line (Heckenberger et al. 2002, 2003). Most of these technical sources of variation may apply to gel-based molecular techniques but are unlikely to apply to SNP markers, which is the genotyping system in the present study. In most breeding programs, the principal causes of large differences in the SNP genotype of samples or stocks originating from the same inbred line but maintained separately are (1) differential drift or fixation of alleles at loci that were heterozygous in the plant from which the line was derived, (2) contamination of the line with pollen or seed of another genotype, and (3) mis-labeling of the seed lot. Mutation is the fourth possible cause of genetic difference among different stocks of the same inbred but the most commonly accepted mutation rate of 10^{-5} (Fleming et al. 1964) is too small to have impact in the time-scale of a typical breeding program. To develop a simple analytical framework for determining the possible reasons for non-homogeneity within sample or genetic identity among seed lots, we assumed, based on our data, that any pair of unrelated CIMMYT inbred lines showed an average rate of 30 % polymorphism (minimum = 20.4 %; maximum = 35.4 %; average = 30.2 %; standard deviation = 3.4 %) for the set of anonymous SNPs used in this study. Hence, any two randomly paired, unrelated inbred lines will differ on average at 30 % of the SNP loci assayed. This provides a threshold for identifying pairs of seed lots that are putatively of the same inbred, but where one of the samples has actually been mis-labeled and contains unrelated material. It should be noted that this frequency will likely differ among marker systems and types of germplasm, and applies only to the germplasm in this study. The expected frequency of polymorphism among random lines is an important parameter to be taken into account when designing a marker-based QC system for a breeding program.

Coupling the assumption of an average polymorphism rate of 30 % for the SNPs used in this study with the standard rate of decay of heterozygosity, we can predict the expected frequency of heterogeneous loci arising from segregation at loci that were expected to be heterozygous in the generation of line derivation, in the absence of major contamination. Most CIMMYT breeding programs now derive lines in the F_4 generation or later, but previously, lines were often derived from earlier generations. In the plant used to establish an F_4-derived line, 12.5 % of the loci that were polymorphic between the parents of the cross are expected to be heterozygous (Table 2), and therefore 12.5 % of these loci are expected to be heterogeneous in advanced self-pollinated generations of the line. We can predict the expected frequency of heterogeneous loci in lines derived from any given generation of self-pollination t as follows: f (heterogeneity) $= 0.30 \times (1 - (1 - 0.5^{t-1}))$. In an average F_4-derived line, we would therefore expect approximately 3.75 % of loci to be heterogeneous due to residual heterozygosity in the founder plant. Samples with substantially more than 3.75 % are likely to have been contaminated by pollen or seed of another genotype. If the sample is simply mis-labeled, it would differ from the reference sample at approximately 30 % of loci (supplementary Table S1), but these loci would not be expected to be heterogeneous. For easy reference and considering that every pollination likely involves a small amount of contamination, we arbitrarily adjusted the thresholds upward somewhat, declaring samples to be contaminated if their level of heterogeneity is greater than 5 %, and we consider two subsamples as differing by an amount greater than expected due to drift if they differ at 3 % or more SNP marker loci. Depending on the tolerance for contamination in the marker or seed system, and the general practice of breeders in the generation of line derivation, these thresholds could be adjusted. Therefore, an inbred line may be considered pure or homogenous if the proportion of heterozygous or heterogeneous loci does not exceed 5 %. An inbred line with 5–15 % heterozygous or heterogeneous loci requires purification by performing ear-to-row selection while one with >15 % heterozygous loci is likely to be contaminated with unrelated genetic material and requires either extensive reselection or should be discarded. We also considered two or more seed sources of the same inbred line as different when genetic distance or the proportion of marker mismatch exceeded 5 %; otherwise, they are considered identical.

Results

Genetic identity in set 1

The 532 SNPs detected a total of 1,064 alleles, with each SNP detecting two alleles as expected. The proportion of alleles that showed differences between two seed sources

Table 2 Expected within-line heterogeneity and between subline polymorphism for CIMMYT's anonymous SNP loci, assuming a 30 % rate of polymorphism among unrelated lines

Generation of line derivation from a bi-parental cross	Mean expected residual heterozygosity in the founding plant at loci that were polymorphic between parents (%)	Expected within-line heterogeneity, assuming 30 % of SNP in random pairs of sublines are polymorphic (%)	Maximum expected polymorphism (percentage of SNP loci assayed) due to drift among separately-maintained sublines (%)
F_2	50.00	15.00	7.50
F_3	25.00	7.50	3.75
F_4	12.50	3.75	1.88
F_5	6.25	1.88	0.94
F_6	3.13	0.94	0.47
F_7	1.56	0.47	0.23

of the same designated line varied from 0.1 to 37.3 % (Table 3). Allelic difference between samples for 11 of the 19 lines was less than 1 %, indicating that the samples were essentially identical. For three line designations (CML395, CML442, and CZL03014), the allelic difference between samples varied from 2.8 to 4.3 %, indicating the samples are derived from the same line but have undergone a minor genetic divergence due to drift. However, there was a high degree of difference between samples for the remaining five lines. In CML312, the allelic difference between samples was 16.7 %, raising concern that the samples may not have been derived from the same source. In CML488, CML489, CZL00003, and CKL05025, the allelic differences ranged from 33.8 to 37.3 %. It thus appears that the samples of these lines were certainly not either derived from a common source or were mis-identified.

Genetic distance between seed sources of the (nominally) same inbred line varied from 0.001 to 0.373. Cluster analyses performed using genetic distance matrices showed clear mis-grouping of the two seed sources of CML488, CML489, CZL00003, and CKL05025 (Fig. 1). The percent allelic differences between samples exceeded 33.8 %, equivalent to levels observed between pairs of unrelated lines. CML488 obtained from SA was found to be basically the same as CML489 from SD, while CML488 from SD appeared to be the same as CML489 from SA. This indicates an error in sampling and archiving seed either in the breeding program or the gene bank. The two seed sources for CML312 grouped together (Fig. 1) despite the relatively large genetic difference between them (Table 3) compared with the expectation within the same line, indicating that these samples may be derived from the same line, but have experienced substantial contamination or drift.

Table 3 Summary of the proportion of alleles that differed between two seed sources of set 1 inbred lines genotyped with 532 SNPs and a subset of 50 SNPs recommended for quality control genotyping using KASPar assay

Line (seed source)*	Percent allele difference (532 SNPs)	Percent allele difference (50 SNPs)	Roger genetic distance (532 SNPs)	Roger genetic distance (50 SNPs)	Genetic identity between seed sources
CML078 (SA vs. SD)	0.2	1.0	0.002	0.010	Same
CML144 (SA vs. SD)	0.6	2.1	0.006	0.021	Same
CML159 (SA vs. SD)	0.2	0.0	0.002	0.000	Same
CML197 (SA vs. SD)	0.4	0.0	0.004	0.000	Same
CML202 (SA vs. SD)	0.1	0.0	0.001	0.000	Same
CML204 (SA vs. SD)	0.2	1.1	0.002	0.011	Same
CML312 (SA vs. SD)	16.7	18.8	0.167	0.188	Different
CML395 (SA vs. SD)	2.8	3.3	0.028	0.033	Same
CML440 (SA vs. SD)	0.9	0.0	0.009	0.000	Same
CML442 (SA vs. SD)	4.3	4.7	0.043	0.047	Same
CML444 (SA vs. SD)	0.7	0.0	0.007	0.000	Same
CML445 (SA vs. SD)	0.8	0.0	0.008	0.000	Same
CML488 (SA vs. SD)	34.9	36.2	0.349	0.362	Different**
CML489 (SA vs. SD)	33.8	39.1	0.338	0.391	Different**
CML511 (SA vs. SD)	0.4	1.1	0.004	0.011	Same
CZL00003 (SA vs. SD)	36.6	43.2	0.366	0.432	Different
CZL03007 (SA vs. SD)	0.2	0.0	0.002	0.000	Same
CZL03014 (SA vs. SD)	3.8	4.2	0.038	0.042	Same
CKL05025 (SA vs. SD)	37.3	52.3	0.373	0.523	Different
CML488-SA vs. CML489-SD	0.8	0.0	0.008	0.000	Same
CML488-SD vs. CML489-SA	0.6	0.0	0.006	0.000	Same

Markers with missing data in one or both seed sources of the same line were excluded from comparison

Two seed sources of an inbred line were considered the same when the difference in percent allele difference or the genetic distance is $\leq$0.050; otherwise, they were considered different

For all comparisons, the correlation between percent allele difference with Roger genetic distance was 1.00

*Seed sources: SA is source A and SD is source D

**The two seed sources for CML488 and CML489 were different due to mis-labeling. CML488 from source A was found to be basically same as CML489 from source D while CML488 from source D appeared to be same as CML489 from source A

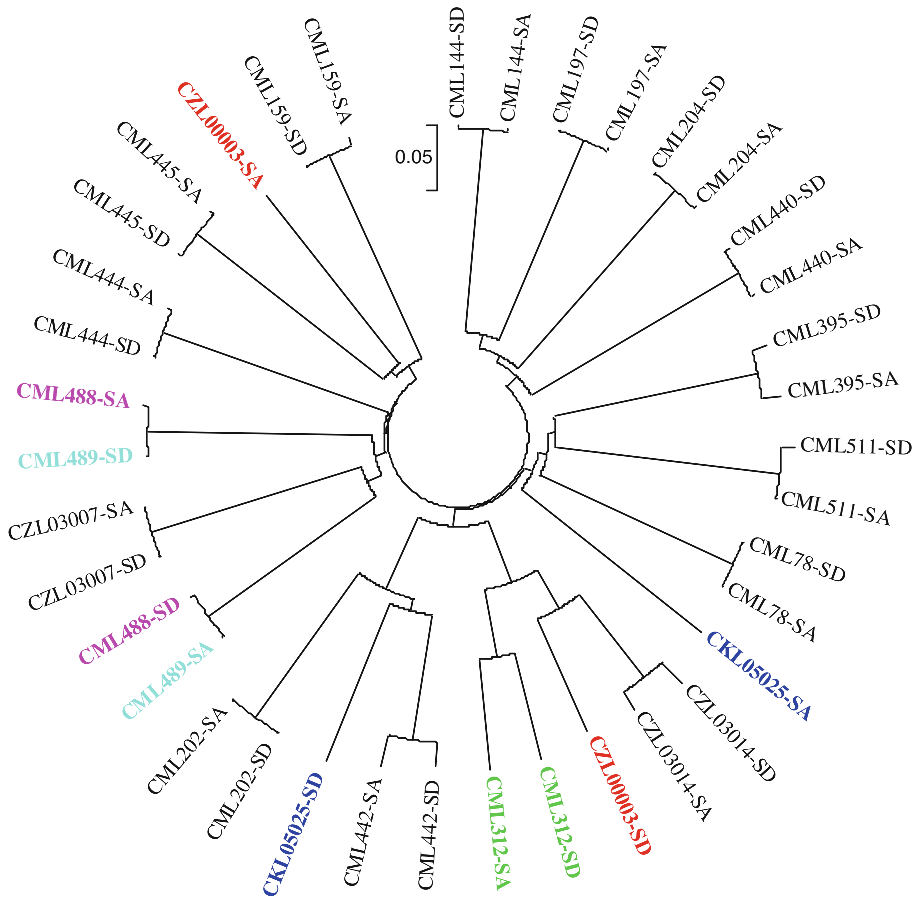

Fig. 1 Neighbor-joining tree for set 1 inbred lines based on Roger's genetic distance computed from 532 SNPs obtained using KASPar platform. Two seed sources of the same line designation that showed >5 % genetic distance are indicated with *same color* in *boldface*. The suffice SA and SD after *line name* indicate seed source A and source D, respectively (color figure online)

Genetic identity in set 2

The proportion of allelic differences of the four positive controls (CML159, CML197, CML202 and CML511) from two sources (eight samples in total) was compared using 532 SNPs. Allelic differences for the same sample on the GoldenGate and KASPar platforms ranged from 0.6 to 0.9 % and 0.1 to 0.4 %, respectively, showing a high congruence of results from the two SNP genotyping platforms. The 1,065 SNPs used with GoldenGate detected a total of 2,130 alleles among the 53 genotypes, with each SNP having two alleles as expected. The proportion of alleles that differed between seed sources of the same line in set 2 varied from 0.30 to 42.30 % (Table 4). Although the overall proportion of allelic differences was higher in set 2 than in set 1, the pattern of their distribution was similar. Among all pairwise comparisons of different seed sources of the 53 lines using the 1,065 SNPs, 28 pairs showed an allelic difference of less than 5 %, indicating the samples are derived from the same line but have experienced a small genetic change. Three pairs showed 5.1 to 14.4 % difference, while the remaining ten pairs showed 25.0 to 33.6 % difference (Table 4). Altogether, six lines (CML159, CML202, CML442, CML443, CML444 and CML488) showed an allelic difference higher than expected for different samples of the same inbred line. Seed of CML159, CML202 and CML442 originating from SC appeared to be very different from those obtained from SA, SB, and/or SD. Genetic distance between different seed sources of the same inbred line varied from 0.008 to 0.423. Cluster analysis performed using the genetic distance matrix of the inbred lines in set 2 revealed the distinct mis-grouping of different seed sources of CML159, CML202, CML442, CML443, and CML488 (Fig. 2). The three seed sources of CML444 grouped together although percent allelic difference and the genetic distance between them were intermediate compared with other samples.

Genetic homogeneity in set 1, set 2, and set 3

In set 1, the proportion of heterogeneity varied from 0.20 to 37.1 % and the average was 3 % (supplementary Table S2). The seed sources for three inbred lines (CML395 and CML442 from SD, and CML511 from SA) showed 5.9 to 13.9 % heterogeneity, while that of CZL00003 from SD showed 37.1 % heterogeneity. In set 2, the proportion of heterogeneity varied from 0.7 to 12.8 % and the average was 3.1 %. The seed sources for eight samples in set 2

Table 4 Summary of the proportion of alleles that differed between seed sources of set 2 inbred lines genotyped with 1,065 SNPs, and a subset of 532 and 50 SNPs using GoldenGate assay

Line	Seed origins*	Percent allele difference (1,065 SNPs)	Percent allele difference (532 SNPs)	Percent allele difference (50 SNPs)	Roger genetic distance (1,065 SNPs)	Roger genetic distance (532 SNPs)	Roger genetic distance (50 SNPs)	Genetic identity of two seed sources
CML159	SB vs. SC	30.2	37.8	52.0	0.302	0.378	0.520	Different
	SC vs. SD	29.8	37.5	52.0	0.298	0.375	0.520	Different
	SA vs. SC	29.8	37.2	53.0	0.298	0.372	0.530	Different
	SB vs. SD	1.2	0.7	0.0	0.012	0.007	0.000	Same
	SA vs. SB	1.8	1.3	1.0	0.018	0.013	0.010	Same
	SA vs. SD	1.7	0.9	1.0	0.017	0.009	0.010	Same
CML202	SC vs. SD	33.6	42.3	54.1	0.336	0.423	0.541	Different
	SA vs. SC	33.6	42.2	53.1	0.336	0.422	0.531	Different
	SA vs. SD	1.2	0.6	1.0	0.012	0.006	0.010	Same
CML442	SB vs. SC	25.0	30.2	43.0	0.250	0.302	0.430	Different
CML443	SA vs. SB	32.6	39.9	49.0	0.326	0.399	0.490	Different
CML444	SB vs. SC	7.6	9.4	10.0	0.076	0.094	0.100	Different
	SA vs. SC	14.4	17.3	17.0	0.144	0.173	0.170	Different
	SA vs. SB	11.9	14.3	11.0	0.118	0.143	0.110	Same, but original line was highly heterogeneous
CML488	SB vs. SC	30.8	38.1	35.0	0.308	0.381	0.350	Different
	SA vs. SC	32.0	40.3	44.0	0.320	0.403	0.440	Different
	SA vs. SB	30.2	35.3	39.0	0.302	0.353	0.390	Different
CKL05017**	SB-a vs. SB-b	0.9	0.5	1.0	0.010	0.005	0.010	Same
CKL05018**	SB-a vs. SB-b	1.6	1.2	0.0	0.016	0.012	0.000	Same
CKL05022**	SB-a vs. SB-b	2.2	1.6	0.0	0.022	0.016	0.000	Same
CKL05023**	SB-a vs. SB-b	3.8	3.3	4.0	0.038	0.033	0.040	Same
CML144	SB vs. SC	4.5	4.7	4.7	0.045	0.047	0.047	Same
	SA vs. SC	4.3	3.7	4.2	0.043	0.037	0.042	Same
	SA vs. SB	4.0	4.1	3.1	0.040	0.041	0.031	Same
CML158	SA vs. SD	1.0	0.8	1.0	0.010	0.008	0.010	Same
CML197	SB vs. SC	5.1	4.5	6.1	0.051	0.045	0.061	Same
	SC vs. SD	4.4	3.7	5.0	0.044	0.037	0.050	Same
	SA vs. SC	3.7	3.4	5.0	0.036	0.034	0.050	Same
	SB vs. SD	2.9	1.5	5.0	0.029	0.015	0.050	Same
	SA vs. SB	2.5	1.8	2.0	0.025	0.018	0.020	Same
	SA vs. SD	1.2	0.8	3.0	0.012	0.008	0.030	Same
CML204	SA vs. SB	0.5	0.3	0.0	0.050	0.003	0.000	Same
CML312	SB vs. SC	0.8	0.4	0.0	0.008	0.004	0.000	Same
CML334	SA vs. SB	0.8	1.0	1.0	0.008	0.010	0.010	Same
CML395	SB vs. SC	2.8	2.1	3.1	0.028	0.021	0.031	Same
	SC vs. SD	2.5	1.7	4.0	0.025	0.017	0.040	Same
	SB vs. SD	3.0	2.1	3.1	0.030	0.021	0.031	Same
CML511	SA vs. SD	4.9	4.3	1.0	0.049	0.043	0.010	Same
CML78	SA vs. SB	2.0	0.6	1.0	0.020	0.006	0.010	Same
CZL03014	SA vs. SC	4.3	4.6	4.9	0.043	0.046	0.049	Same
LaPostaSeqC7	SA vs. SB	1.4	1.5	0.0	0.014	0.015	0.000	Same
P300C5S1B	SA vs. SB	1.4	0.8	3.0	0.014	0.008	0.030	Same

Markers with missing data in one or both seed sources of the same line were excluded from comparison

Two seed sources of an inbred line were considered the same when the differences in percent allele difference or the genetic distance is $\leq$0.050; otherwise, they were considered different

* Seed sources: *SA* source A; *SB* source B; *SC* source C; *SD* source D

** a and b in column 2 refer to different generations for the same line obtained from source 2B

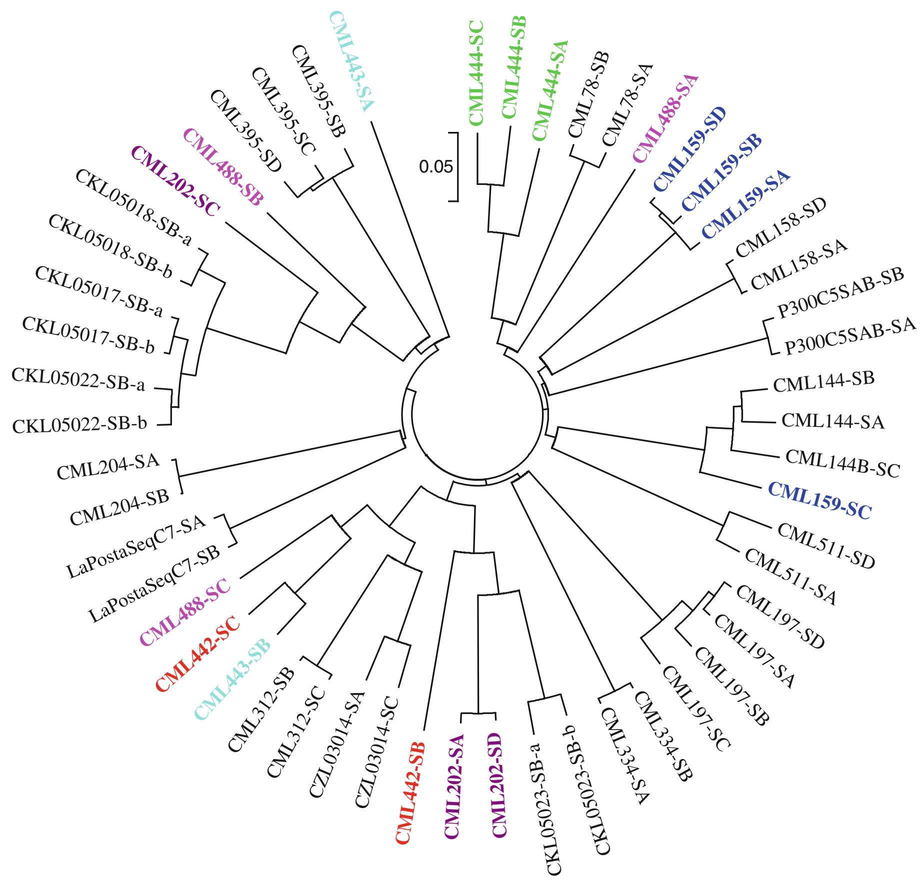

Fig. 2 Neighbor-joining tree for set 2 inbred lines based on Roger's genetic distance computed from 1,065 SNPs obtained using GoldenGate platform. Seed sources of the same line designation that showed >5 % genetic distance are indicated with *same color* in *boldface*. The suffice SA, SB, SC, and SD after *line name* indicate seed source A, source B, source C, and source D, respectively (color figure online)

(CML444-SA, CML444-SC, CML159-SC, CML144-SB, CML144-SC, CML442-SB, CML443-SB, and CKL05023-S2b) showed heterogeneity that varied from 5.2 to 12.8 %. In set 3, heterogeneity for the 544 inbred lines varied from 0.4 to 31.3 % and the average was 5.3 %. As shown in Fig. 3 for the 1,065 SNPs, about 71 % of the 544 inbred lines were considered homogenous with <5 % heterogeneity. The remaining 21 % and 8 % had 5–15 % and >15 % heterogeneity, respectively.

Selection of a subset of SNPs for QC genotyping

In order to identify a subset of informative markers that could be used for routine genotyping QC, we attempted to determine the minimum number SNPs needed to obtain results comparable with those obtainable with the entire datasets (532 SNPs for set 1, 1,065 SNPs for set 2 and set 3). Subsets of 50 SNPs (Table 5) provided comparable results to the entire datasets as far as genetic identity and homogeneity is concerned. Cophenetic correlation coefficients for the different numbers of markers varied from 0.81 (good fit) to 0.97 (very good fit), with the highest being for set 1 and set 2 genotyped with 532 SNPs and 1,065 SNPs, respectively. The correlations between the genetic distance matrices derived from 50 SNPs and 532

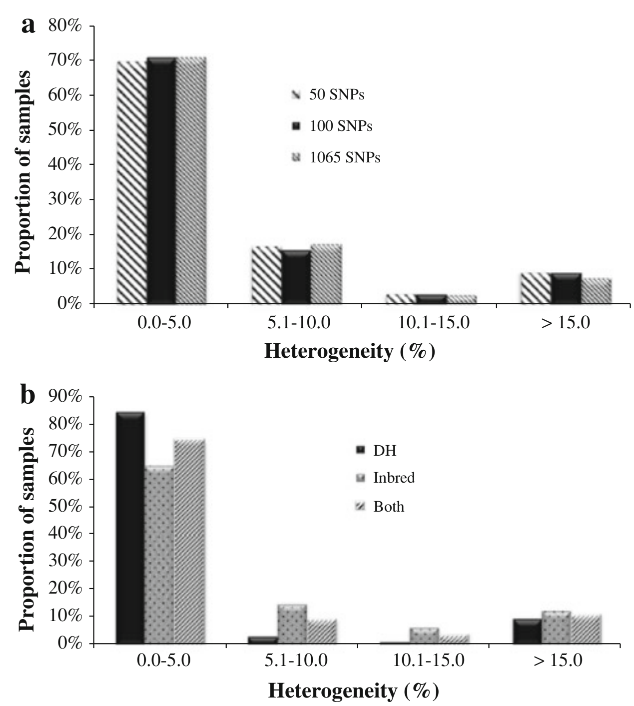

Fig. 3 Heterogeneity summary for set 3 and set 4 samples: **a** 544 inbred lines genotyped with different set of SNPs using the GoldenGate platform; **b** 1,306 samples (643 DH lines and 663 inbred lines) genotyped with 50 SNPs recommended for quality control genotyping using KASPar platform

Table 5 Summary of the 50 highly informative SNPs in set 1 recommended for quality control genotyping

SNP name	Chromosome	Physical map position (bp)	MAF ($n = 38$)	MAF ($n = 53$)	MAF ($n = 544$)	MAF ($n = 1{,}306$)	PIC ($n = 38$)	PIC ($n = 53$)	PIC ($n = 544$)	PIC ($n = 1{,}306$)	Agreement between GoldenGate and KASPar (%)	Extent of polymorphism out of 15 crosses (%)
PZA00175.2	1	8,510,027	0.242	0.377	0.386	0.304	0.300	0.360	0.362	0.334	100.0	33.3
PZA03561.1	1	60,212,051	0.414	0.472	0.419	0.411	0.368	0.374	0.368	0.367	100.0	66.7
PHM1932.51	1	118,875,639	0.486	0.434	0.415	0.254	0.375	0.371	0.368	0.307	100.0	40.0
PZA02741.1	1	161,072,169	0.286	0.302	0.252	0.404	0.325	0.333	0.306	0.366	97.4	73.3
PHM12706.14	1	212,356,401	0.458	0.443	0.495	0.419	0.373	0.372	0.375	0.368	100.0	40.0
PZA02269.3	1	252,722,026	0.263	0.208	0.423	0.383	0.313	0.275	0.369	0.361	100.0	46.7
PHM4752.14	1	298,874,066	0.382	0.434	0.394	0.348	0.361	0.371	0.364	0.351	95.0	46.7
PHM13440.13	2	2,527,344	0.471	0.377	0.422	0.459	0.374	0.360	0.369	0.373	92.1	26.7
PZA02378.7	2	35,040,818	0.276	0.462	0.414	0.327	0.320	0.374	0.367	0.343	95.2	40.0
PHM3457.6	2	62,804,122	0.371	0.302	0.497	0.359	0.358	0.333	0.375	0.354	97.6	53.3
PHM3626.3	2	125,642,617	0.426	0.406	0.475	0.437	0.375	0.366	0.374	0.371	95.2	26.7
PZA01885.2	2	206,881,202	0.289	0.396	0.422	0.388	0.327	0.364	0.369	0.362	100.0	40.0
PZA02090.1	3	4,138,512	0.365	0.340	0.388	0.384	0.356	0.348	0.362	0.361	97.5	53.3
PHM5502.31	3	67,284,067	0.486	0.378	0.402	0.483	0.375	0.360	0.365	0.375	97.5	53.3
PZA00413.20	3	125,192,432	0.303	0.330	0.432	0.488	0.333	0.345	0.370	0.375	100.0	53.3
PZA00667.2	3	161,516,227	0.329	0.311	0.417	0.269	0.344	0.337	0.368	0.316	100.0	53.3
PHM15964.16	3	221,986,592	0.405	0.302	0.484	0.360	0.366	0.333	0.375	0.354	100.0	40.0
PZA02358.1	4	11,329,241	0.472	0.472	0.476	0.399	0.374	0.374	0.374	0.365	97.4	26.7
PZA00726.10	4	60,768,063	0.368	0.255	0.335	0.378	0.357	0.308	0.346	0.360	97.6	73.3
PZA03409.1	4	128,632,208	0.446	0.434	0.463	0.435	0.372	0.371	0.374	0.371	100.0	53.3
PZA01477.3	4	172,301,064	0.500	0.434	0.370	0.496	0.375	0.371	0.357	0.375	94.9	66.7
PZA00399.11	4	229,644,826	0.417	0.387	0.388	0.436	0.368	0.362	0.362	0.371	100.0	53.3
PZA02462.1	5	6,820,571	0.314	0.434	0.448	0.404	0.338	0.371	0.372	0.366	97.6	40.0
PZA00981.3	5	37,030,384	0.351	0.462	0.484	0.272	0.352	0.374	0.375	0.318	94.6	46.7
PZA02164.16	5	112,179,855	0.408	0.423	0.425	0.431	0.366	0.369	0.369	0.370	100.0	80.0
ae1.7	5	167,873,309	0.500	0.387	0.389	0.467	0.375	0.362	0.362	0.374	97.5	26.7
PHM3512.186	5	203,434,263	0.386	0.330	0.472	0.486	0.362	0.345	0.374	0.375	100.0	73.3
PZA00440.1	6	22,404,308	0.474	0.472	0.478	0.455	0.374	0.374	0.375	0.373	97.6	53.3
PZA00355.2	6	78,756,133	0.447	0.385	0.461	0.387	0.372	0.362	0.373	0.362	100.0	66.7
PZB01658.1	6	102,953,833	0.306	0.358	0.380	0.413	0.334	0.354	0.360	0.367	100.0	40.0
PZA02187.1	6	139,106,115	0.443	0.425	0.497	0.315	0.372	0.369	0.375	0.338	97.5	26.7
PHM3466.69	6	167,148,384	0.314	0.283	0.380	0.448	0.338	0.323	0.360	0.372	100.0	53.3
PHM3078.12	7	5,963,009	0.361	0.321	0.347	0.286	0.355	0.341	0.351	0.325	100.0	46.7
PZA00084.2	7	43,948,264	0.306	0.302	0.368	0.407	0.334	0.333	0.357	0.366	100.0	66.7

Table 5 continued

SNP name	Chromosome	Physical map position (bp)	MAF (n = 38)	MAF (n = 53)	MAF (n = 544)	MAF (n = 1,306)	PIC (n = 38)	PIC (n = 53)	PIC (n = 544)	PIC (n = 1,306)	Agreement between GoldenGate and KASPar (%)	Extent of polymorphism out of 15 crosses (%)
PZA03645.1	7	73,892,322	0.338	0.415	0.407	0.232	0.347	0.368	0.366	0.293	100.0	33.3
PZB00752.1	7	131,103,240	0.300	0.288	0.346	0.337	0.332	0.326	0.350	0.347	97.3	40.0
PZA01533.2	7	162,381,818	0.368	0.311	0.458	0.405	0.357	0.337	0.373	0.366	94.6	33.3
PZA02174.2	8	4,101,256	0.433	0.321	0.438	0.460	0.371	0.341	0.371	0.373	92.7	46.7
PZA00793.2	8	64,421,988	0.485	0.358	0.494	0.393	0.375	0.354	0.375	0.363	95.2	53.3
PZA03135.1	8	100,564,485	0.375	0.349	0.402	0.337	0.359	0.351	0.365	0.347	100.0	40.0
PHM5805.19	8	120,875,248	0.375	0.443	0.353	0.457	0.359	0.372	0.352	0.373	100.0	40.0
PZA02746.2	8	163,067,200	0.391	0.406	0.464	0.338	0.363	0.366	0.374	0.348	95.0	46.7
sh1.12	9	11,340,882	0.386	0.481	0.449	0.411	0.362	0.375	0.372	0.367	97.6	73.3
PHM229.15	9	30,003,189	0.365	0.481	0.451	0.469	0.356	0.375	0.373	0.374	100.0	53.3
PZA01062.1	9	88,057,320	0.472	0.423	0.381	0.468	0.374	0.369	0.361	0.374	97.4	53.3
PHM7916.4	9	132,762,904	0.458	0.311	0.465	0.441	0.373	0.337	0.374	0.372	97.5	53.3
PHM1752.36	10	9,746,552	0.395	0.274	0.307	0.239	0.364	0.318	0.335	0.298	100.0	26.7
PHM2770.19	10	72,565,410	0.334	0.349	0.277	0.383	0.361	0.351	0.320	0.361	91.9	66.7
PZA01919.2	10	111,260,278	0.382	0.255	0.431	0.204	0.361	0.308	0.370	0.272	100.0	40.0
PZA03605.1	10	141,830,532	0.319	0.500	0.447	0.369	0.340	0.375	0.372	0.357	91.9	33.3

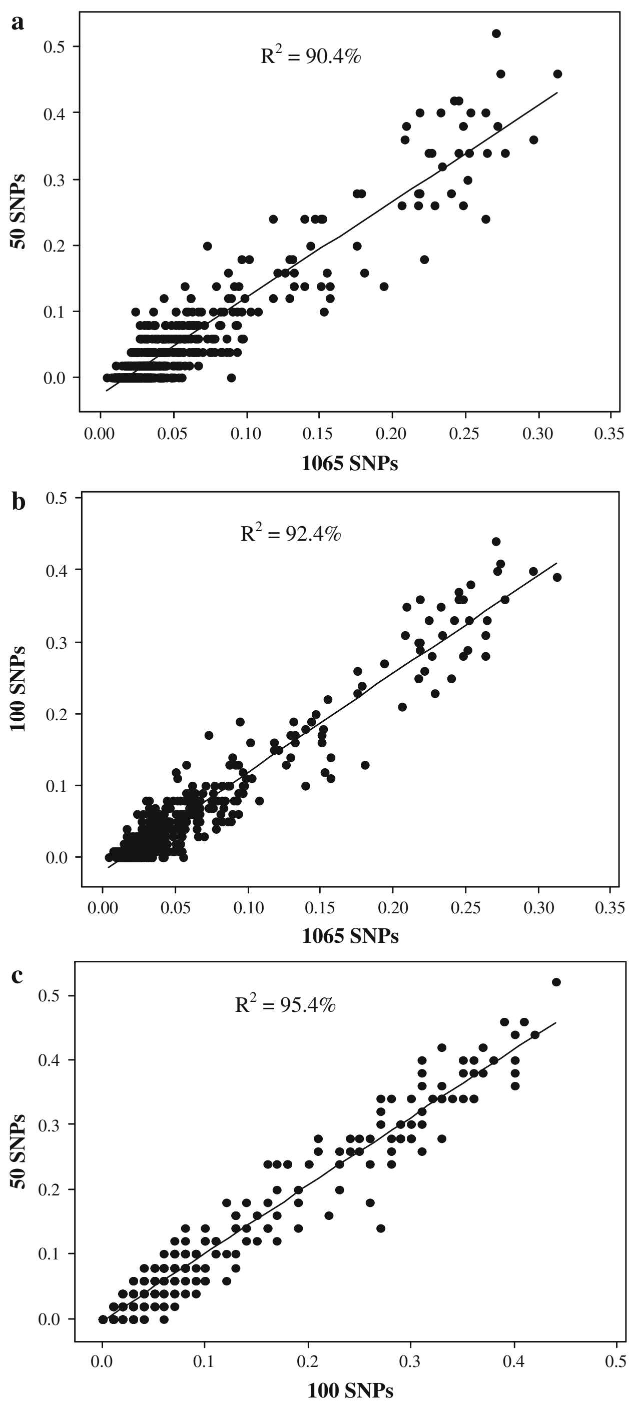

Fig. 4 Correlations between genetic heterogeneity values calculated from 1,065 SNPs and a subset of 50 and 100 SNPs that were used to genotype the 544 inbred lines in set 3. *Each point* represents heterogeneity between a pair of samples: **a** 50 SNPs versus 1,065 SNPs ($r = 0.951$); **b** 100 SNPs versus 1,065 SNPs ($r = 0.961$); and **c** 50 SNPs versus 100 SNPs ($r = 0.977$)

SNPs in set 1, and between 50 SNPs and 1,065 SNPs both in set 2 and set 3 (Fig. 4) remained the same as the cophenetic correlations. The SNPs that we recommend for quality control genotyping were selected based on the following criteria from the different datasets in this study: (1) ease of scoring with unambiguous separation of the two homozygous and heterozygous genotypes both in GoldenGate and KASPar genotyping platforms; (2) SNPs with minor allele frequency (MAF) and polymorphism information content (PIC) of at least 0.20 and 0.25, respectively (Table 5); (3) good distribution across chromosomes based on the physical map (supplementary Fig. S1); (4) at least 90 % agreement between the GoldenGate and KASPar SNP genotyping platforms; and (5) polymorphism in at least 25 % of the parents used for developing 15 mapping populations. Table 5 shows details of the informative SNPs selected for quality control purposes, including chromosomal position, MAF, and PIC for sample sizes that varied from 38 to 1,306. In set 4 lines genotyped with the 50 SNPs recommended for QC genotyping, heterogeneity varied from 0 to 42.9 % in the DH and 0–62 % in the inbred lines. The average heterogeneity for the DH and inbred lines in set 4 was 4.7 and 5.3 %, respectively, with 85 % of the DH and 65 % of the inbred lines considered to be homogenous with <5 % heterogeneity. Since QC genotyping on the DH lines was made after one generation of seed multiplication, the observed higher (>5 %) than expected heterogeneity for the remaining 15 % of these lines is likely to be due to pollen contamination during hand pollination at the nurseries. Heterogeneity was observed in some of the DH lines irrespective of the source population used for developing them. However, Fig. 5 illustrates the possibility of making wrong conclusions using a subset of 50 SNPs rather than 1,065 SNPs. For example, heterogeneity for entry 2153 and 2486 were 3.1–3.8 % and 8.2–9.1 % when the data of the 1,065 and 50 SNPs were considered, respectively, indicating overestimation of heterogeneity using fewer markers. Therefore, users need to be aware of the possibility of arriving at incorrect conclusions by using the subset of 50 SNPs that we recommend for QC, compared to decisions based on larger sets of SNPs. For this purpose, we increased the number of SNPs for QC genotyping from 50 to 100 by adding the next 50 best SNPs (supplementary Table S3) and compared the results with the entire data of the different types of samples (Figs. 3 and 5). The increase in the number of SNPs from 50 to 100 increased the correlation between 0.03 and 0.14 depending on the purpose of the QC and reduced the error in classifying genotypes into different genetic purity groups up to 4.0 %.

Discussion

Genetic purity and identity

Inbred lines should be genetically pure and possess all the genetic qualities that the breeder has selected for. Small

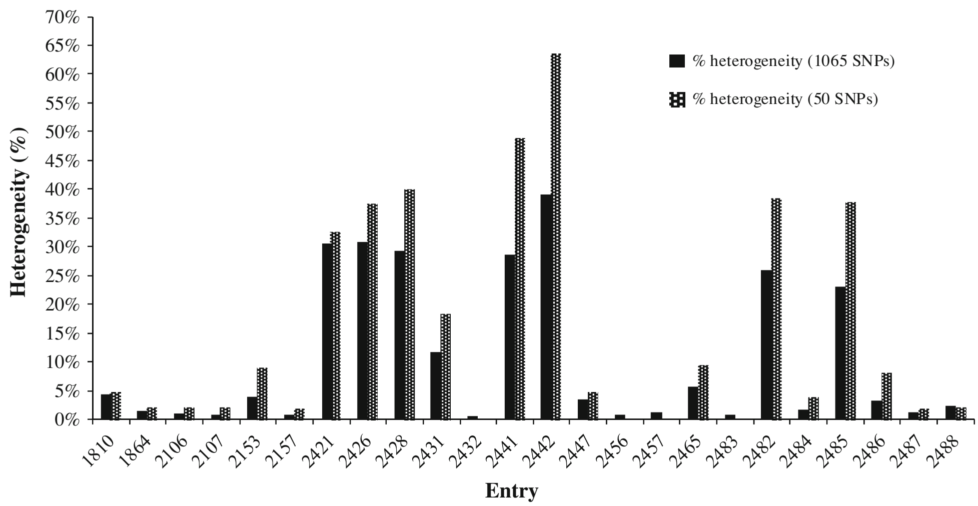

Fig. 5 Heterogeneity comparisons for selected double haploid lines (entries 1810–2465) and inbred lines (entries 2482–2488) that were genotyped with 1,065 SNPs and a subset of 50 SNPs recommended for quality control genotyping. The correlation between the 50 and 1,065 SNPs was 0.986

changes in allele frequencies may occur during seed regeneration, maintenance at two different places, bulking during maintenance breeding, and possible contamination with seeds or pollen of other samples (Heckenberger et al. 2002; Warburton et al. 2010). Significant changes in the genetic makeup of a germplasm may affect performance, and in the worst case result in distribution of wrong hybrids or varieties. Our study clearly shows the presence of high genetic heterogeneity within some of the inbred lines (Fig. 3, supplementary Table S2). Results from the present study also demonstrate the presence of high genetic difference between different seed sources of five lines in set 1 and six lines in set 2. Altogether, ten of the 28 lines showed genetic differences higher than expected. Several authors (Fleming et al. 1964; Gethi et al. 2002; Heckenberger et al. 2002, 2003; Jones 1945; Revilla et al. 2005; Russell et al. 1963; Russell and Vega 1973; Schuler 1954; Yan et al. 2009) have also reported the presence of a wide range of genetic differences among different sources of seed of the same line designation. Mis-labeling is clearly observed in set 1, where the two samples of CML488 have an average allelic difference of about 33.3 % and the two samples of CML489 have an average allelic difference of 34.5 %. On the other hand, the SA sample for CML488 was a near-perfect match for the SD sample of CML489, while the SA sample for CML489 was a near-perfect match for the SD sample of CML488. This indicates that the two lines have been mis-labeled either in the breeder's working collection or in the gene-bank. Contamination is the most likely cause for the observed difference between the two seed sources of CZL00003. This is supported by the presence of 30.9–37.1 % heterogeneity in the SD sample compared with the 3.5–11.6 % heterogeneity in the SA sample. CML159 from SC, CML442 and CML443 from SB, and CML444 from both SA and SC showed fairly high levels of heterogeneity that could be due to residual heterozygosity.

Why quality control?

Breeding programs must monitor the quality of seed increase and line maintenance processes to ensure the genetic homogeneity and identity of their products. Our results identified some mis-labeled seed samples used by CIMMYT breeders. Monitoring differences among samples of the same line maintained separately, and heterogeneity within lines is an important element in assessing the quality of seed stock maintenance in a breeding program. Quality control genotyping is essential for maintaining the genetic identity of the inbred lines and minimizing errors at different levels. Based on results of the present study, we recommend a subset of 50 highly informative SNPs (Table 5) described in this paper for monitoring seeds of fixed tropical maize inbred lines. The correlations between the entire dataset and the 50 SNPs selected for QC varied from 0.81 to 0.97. As the composition of the germplasm under study may influence the number of markers that will be used for genotyping, an additional 50 SNPs (supplementary Table S3) were identified to obtain reliable results. However, increasing the number of SNPs from 50 to 100 had limited impact in QC in the present study. The selected SNPs in this study can easily be used either with the Illumina or KBioscience SNP genotyping platforms and are likely to be transferable to other platforms. These SNP markers also allow breeders to outsource genotyping to commercial agencies that provide fairly quick and cheap genotyping service. Some commercial genotyping service providers return allele calls in Excel spreadsheet format within 4–8 weeks at a cost of 0.05 to $ 0.22 per data point, equivalent to 2.5 to $ 11.0 per DNA sample.

Quality control genotyping can be done at different stages in the breeding program, particularly when advanced lines are first bulked, before or at the time of release of inbred lines, before using inbred lines as parents for making crosses, etc. If resources are limiting, QC genotyping may be delayed until fixed lines are developed. Mis-labeling can be avoided if the parents of all new pedigree breeding projects are genotyped at high density, so that derived lines can be compared with parental genotypes to confirm their provenance. The informative markers identified in the present study, along with the suggested protocol, can potentially aid in undertaking such a task in a time- and cost-effective manner for monitoring the extent of heterozygosity or heterogeneity in the present inbred lines. CIMMYT is collaborating with Cornell University and USDA Agricultural Research Service in developing an integrated platform and analytical tools for genotyping via next-generation sequencing for breeding, reducing the genotyping cost below a row charge for field evaluation. Genotyping-by-sequencing (Elshire et al. 2011) is likely to bring the genotyping cost as low as $20 per DNA sample (http://www.maizegenetics.net/gbs-overview) in generating about half a million SNPs, which may replace the use of a subset of SNPs for QC genotyping.

Acknowledgments We thank Veronica Ogugo for sample preparation and DNA extraction. This work was carried out under the Drought Tolerant Maize for Africa (DTMA) and Water Efficient Maize for Africa (WEMA) projects undertaken by CIMMYT and national partners in Africa, and funded by the Bill and Melinda Gates Foundation.

References

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A Robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379

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

Fleming AA, Kozelnicky GM, Browne EB (1964) Variations between stocks within long-time inbred lines of maize (*Zea mays* L.) 1. Crop Sci 4:291–295

Gethi JG, Labate JA, Lamkey KR, Smith ME, Kresovich S (2002) SSR variation in important US maize inbred lines. Crop Sci 42:951–957

Gupta K, Balyan S, Edwards J, Isaac P, Korzun V, Rodër M, Gautier MF, Joudrier P, Schlatter R, Dubcovsky J, De La Pena C, Khairallah M, Penner G, Hayden J, Sharp P, Keller B, Wang C, Hardouin P, Jack P, Leroy P (2002) Genetic mapping of 66 new microsatellite (SSR) loci in bread wheat. Theor Appl Genet 105:413–422

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

Heckenberger M, Bohn M, Ziegle JS, Joe LK, Hauser JD, Hutton M, Melchinger AE (2002) Variation of DNA fingerprints among accessions within maize inbred lines and implications for identification of essentially derived varieties. I. Genetic and technical sources of variation in SSR data. Mol Breed 10:181–191

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

Heckenberger M, Bohn M, Frisch M, Maurer HP, Melchinger AE (2005) Identification of essentially derived varieties with molecular markers: an approach based on statistical test theory and computer simulations. Theor Appl Genet 111:598–608

Heckenberger M, MuminovicÂ′ J, Voort JR, Peleman J, Bohn M, Melchinger AE (2006) Identification of essentially derived varieties obtained from biparental crosses of homozygous lines. III. AFLP data from maize inbreds and comparison with SSR data. Mol Breed 17:111–125

Jones DF (1945) Heterosis resulting from degenerative changes. Genetics 30:527–542

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

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

Low YL, Wedrén S, Liu J (2006) High-throughput genomic technology in research and clinical management of breast cancer. Evolving landscape of genetic epidemiological studies. Breast Cancer Res 8:209

Lu Y, Yan J, Guimarães CT, Taba S, Hao Z, Gao S, Chen S, Li J, Zhang S, Bindiganavile SV et al (2009) Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. Theor Appl Genet 120: 93–115.

Mace EM, Buhariwalla HK, Crouch JH (2003) A high-throughput DNA extraction protocol for tropical molecular breeding programs. Plant Mol Biol Report 21:459a–459h

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

Prasanna B, Pixley K, Warburton M, Xie CX (2010) Molecular marker-assisted breeding options for maize improvement in Asia. Mol Breed 26:339–356

Revilla P, AbuÃn MC, Malvar RA, Soengas P, Ordas B, Ordas A (2005) Genetic variation between Spanish and American versions of sweet corn inbred lines. Plant Breed 124:268–271

Rholf FJ (1993) NTSYS-pc, numerical taxonomy and multivariate analysis system. Exeter software, New York

Rogers JS (1972) Measures of genetic similarity and genetic distance. Stud Genet VII Univ Texas Publ 7213:145–153

Russell WA, Vega UA (1973) Genetic stability of quantitative characters in successive generations in maize inbred lines. Euphytica 22:172–180

Russell WA, Sprague GF, Penny LH (1963) Mutations affecting quantitative characters in long-time inbred lines of maize1. Crop Sci 3:175–178

Schuler JF (1954) Natural mutations in inbred lines of maize and their heterotic effect. I. Comparison of parent, mutant and their F_1 hybrid in a highly inbred background. Genetics 39:908–922

Semagn K, Magorokosho C, Vivek BS, Makumbi D, Beyene Y, Mugo S, Prasanna BM, Warburton ML (2012) Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single nucleotide polymorphic markers. BMC Genomics 13:113. doi:10.1186/1471-2164-13-113

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599

Warburton ML, Setimela P, Franco J, Cordova H, Pixley K, Banziger M, Dreisigacker S, Bedoya C, MacRobert J (2010) Toward a cost-effective fingerprinting methodology to distinguish maize open-pollinated varieties. Crop Sci 50:467–477

Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PLoS One 4:e8451