Toward a Cost-Effective Fingerprinting Methodology to Distinguish Maize Open-Pollinated Varieties

Marilyn L. Warburton, Peter Setimela,* Jorge Franco, Hugo Cordova, Kevin Pixley, Marianne Bänziger, Susanne Dreisigacker, Claudia Bedoya, and John MacRobert

ABSTRACT

In Africa, many smallholder farmers grow open-pollinated maize (*Zea mays* L.) varieties (OPVs), which allow seed recycling and outyield traditional unimproved landraces. Seeds of productive OPVs are provided to farmers, often by nongovernmental organizations (NGOs) that help farmers access improved seeds, particularly following disasters in which original seed is lost. However, NGOs often rely on local seed suppliers to provide seed, and in some years the seeds provided to the farmers are suspected not to be of the promised variety. Here we present methodology to prove within a high level of confidence if two samples of seeds are the same genetic population or not, despite the difficulties involved in fingerprinting heterologous populations. In addition to heterogeneity within populations, difficulties can include sampling errors, differences in the fields or years in which the seeds were multiplied, and seed mixing. Despite these confounding sources of variation, we show the possibility to conclusively differentiate each of the populations used in this work. This methodology will allow breeders, seed companies, government agencies, and NGOs to ensure the purity and identity of high-yielding, locally adapted OPVs reach farmers so they can generate the highest yields possible in their fields.

M.L. Warburton, USDA ARS CHPRRU, Box 9555, Mississippi State, MS 39762; P. Setimela and J. MacRobert, Maize Program, CIMMYT, P.O. Box MP 163, Harare, Zimbabwe; J. Franco, Facultad de Agronomía, Univ. de la Republica, Ave. Garzón 780, Montevideo, Uruguay; H. Cordova, K. Pixley, S. Dreisigacker, and C. Bedoya, CIMMYT, Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico; M. Bänziger, Maize Program, CIMMYT, P.O. Box 1041, Village Market-00621, ICRAF House, United Nations Ave., Kenya. Received 20 Feb. 2009. *Corresponding author (p.setimela@cgiar.org).

Abbreviations: AMOVA, analysis of molecular variance; ARDA, Agricultural Rural Development Authority; CBI, Crop Breeding Institute; DUS, distinct, uniform, and stable; NGO, nongovernmental organization; OPV, open-pollinated variety; PCR, polymerase chain reaction; SSR, simple sequence repeat.

In sub-Saharan Africa, improved open-pollinated varieties (OPVs) of maize (*Zea mays* L.) are grown by resource-poor smallholder farmers because they offer the economic advantage of allowing seed recycling for several generations without the yield penalty associated with replanting seeds of hybrid varieties (Pixley and Bänziger, 2004; Setimela et al., 2005) and tend to outyield farmers' unimproved landraces. To improve maize productivity, the International Maize and Wheat Improvement Center (CIMMYT) has developed stress-tolerant and more nutritious OPVs suitable for smallholder farmers' conditions (Bänziger et al., 1999, 2002; Pixley and Bänziger, 2004) that are now grown in more than a million hectares in Africa (Bänziger and de Meyer, 2002; Mwala et al., 2004). Farmers find it a challenge to access quality seeds following drought or natural disaster, as most local seed sources will have been destroyed. Thus, many nongovernmental organizations (NGOs) engage in seed relief programs to help farmers recover, reestablish, and sustain their farming

Published in Crop Sci. 50:1–11 (2010).
doi: 10.2135/cropsci2009.02.0089

677 S. Segoe Rd., Madison, WI 53711 USA

systems. Despite substantial efforts by NGOs to supply quality seed to farmers affected by natural disaster, distribution of quality seed in remote areas is still a major constraint.

Seeds may be purchased from small seed companies, but the cheapest price is usually obtained working with large quantities. Therefore, seeds may be supplied to NGOs in bulk, and repackaged for distribution in smaller amounts to affected farmers, or the NGOs may pay small seed companies to produce and distribute seeds of the chosen OPVs to small farmers for a reduced or no charge (Langyintuo and Setimela, 2007). Seed obtained from local food grain markets is not suitable for planting, as the quality of the plants grown from them can be very poor, especially if they were imported from a distant source where they are adapted to a different environment (Longley et al., 2001).

One of the most popular and best yielding CIMMYT OPVs, ZM521, was released in 2000 and performs particularly well in areas where other maize varieties succumb to diseases that attack maize in Africa. However, NGOs in Nyanga, Masivingo, and Mutare in Zimbabwe have reported that ZM521 distributed in the 2005–2006 cropping season by one seed company was performing far below farmers' expectations. The procurement was part of a seed relief program for vulnerable households. It is suspected that the seeds distributed by this seed company were not, in fact, ZM521. Two methods for determining if two OPVs are the same or not are (i) the comparison of phenotypic attributes of different populations; and (ii) the use of DNA fingerprinting of populations. Current methods for awarding plant breeder's rights and registering a new variety must show that an OPV is distinct, uniform, and stable (known as DUS testing), which is usually done based on morphological traits of field-grown materials for one or more growing seasons. The use of molecular markers for the fingerprinting of lines and populations is a complementary method to identify and distinguish populations at the genetic level.

Open-pollinating populations that are not under strong selection pressure and not being mixed with other seed or pollen sources have stable allele frequencies over generations for all genes in the population (both expressed genes and neutral markers) (Falconer, 1984) and this can be used to determine relationships, purity, and identity. Fingerprinting a population requires sampling sufficient individuals to calculate allele frequencies within the population. However, high levels of within-population genetic diversity typical of maize OPVs call for the analysis of a large and representative sample of individuals for each accession, which makes analyses costly, difficult, and time-consuming. The use of the bulked method of DNA fingerprinting (Dubreuil et al., 2006) allows many populations to be fingerprinted quickly and economically. Past studies of maize populations merely sought to determine relative genetic distances among populations, whereas in this study, we wish to definitively identify a population or subpopulations from the same original population, and distinguish them from other populations in the study. In addition, small changes in allele frequencies in a population may occur following seed regeneration, maintenance of the same population in two different places, subsampling for the fingerprinting itself, and possible contamination of the population with seeds of other populations.

The objectives of this study were to see if the bulked fingerprint method can be used to distinguish (i) genetically different OPVs; (ii) the same OPVs grown for several generations in different locations; (iii) the same OPVs mixed with different percentages of genetically unrelated OPVs; and (iv) two subsamples of the same OPV. In addition, we wished to see how the bulked fingerprint method compares to the more commonly used DUS phenotypic screens when attempting to confirm the identity of a maize OPV.

Table 1. Source of maize seed used for simple sequence repeat (SSR) analysis and field evaluation at Harare, Zimbabwe, 2007–2008 season.

Source of ZM521	Source company or institute	Source of seed	Year of production
ZM521-CIMMYT†	CIMMYT	Harare	2006
ZM521-CBI	Crop Breeding Institute	Harare	2005
ZM521-ARDA	Crop Breeding Institute	ARDA‡	2005
ZM521-VR-grain§	VR Grain	Nganga	2005
ZM521-green	Seed Co Ltd. (Zimbabwe)	Seed Co	2004
ZM521-CBI (Check1)¶	Crop Breeding Institute	Gwebi	2005
ZM521-CBI (Check2)¶	Crop Breeding Institute	Chisumbanje	2005

†Standard reference source of ZM521.

‡Agricultural Rural Development Authority.

§Included in the SSR analysis, but not the DUS (distinct, uniform, and stable) study.

¶Check: Included in the DUS study, but not included in the SSR analysis.

MATERIALS AND METHODS

Source of Seed for Farmers' Tests

Farmers planted two seed lots that were both procured by the NGO Concern World Wide and labeled ZM521, the first from one private seed company for the 2004–2005 growing season, and the second from a different seed company in South Africa for the 2005–2006 growing season. Farmers were given 5 kg of seed in the 2004–2005 and 2005–2006 seasons, enough for a 0.5- to 1-ha plot. Because of poor rainfall in 2004–2005, farmers only planted part of their seeds, and saved the rest, which were planted side by side with the second seed lot from 2005–2006, allowing direct comparison. The differences that farmers observed between the two seed sources sparked the debate on the poor performance of ZM521 from the 2005–2006 season. To address these concerns, CIMMYT and Concern World Wide visited fifteen randomly chosen farmers in the area to investigate their observations between the two seed lots of ZM521.

DUS Phenotypic Tests

Five different sources of ZM521 were collected from companies and institutions that maintain breeder's and foundation seed of ZM521 (Table 1), the main known sources of ZM521 in the region. The CIMMYT source of ZM521 is considered the reference sample in this study. Because the disputed seeds of the 2005–2006 season had

all been used by the farmers, these were not included in the phenotypic or genotypic tests below. The five sources of the ZM521 were planted at the CIMMYT Harare maize research station in the 2007–2008 planting season. For each source, 10 × 10-m rows were planted for DUS testing, conducted according to procedures and guidelines outlined by the International Union for the Protection of New Varieties of Plants (http://www.upov.int/en/publications/tg-rom/tg002/tg_2_6.pdf [verified 23 Nov. 2009]) (Table 2). The field data were transformed using the natural logarithm of each ordinal variable as response and analyzed for significant differences among the different seed sources using a General Linear Model in SAS V9.1 software (SAS Institute, 2004).

Bulked SSR Marker Fingerprinting Tests

Two simple sequence repeat (SSR) marker fingerprinting tests were conducted for the objectives of this study. The first looked at the relationship among the different sources of ZM521 included in the phenotypic DUS tests (Table 1). The second was run using nine different, unrelated OPVs, as a test of the methodology. This test compared two independent bulks from the same OPV, different contamination levels to simulate the mixing of seeds, and OPVs with the same name from different sources (institutions, companies, fields, or years). Contaminated bulks of DNA were created by taking seeds of one population and mixing them with seeds from an unrelated population in proportions of 5, 10,

Table 2. Table of characteristics measured on the different sources of ZM521 maize for conducting DUS (distinct, uniform, and stable), according to guidelines from Union for the Protection of New Varieties of Plants.

No.	Characteristic	Scale	ZM521 CIMMYT†	ZM521 CBI†	ZM521 ARDA†	ZM521 CBI Chisumbanje (check‡)	ZM521 CBI Gwebi (check‡)	In
1	First leaf: anthocyanin coloration of sheath	1–9	3.0	1.5	1.0	1.5	1.0	NS
2	Leaf: angle between blade and stem (on leaf just above upper ear)	1–9	3.5	3.5	3.5	3.0	3.5	NS
3	Leaf: attitude of blade (on leaf just above upper ear)	1–9	4.0	3.0	4.0	4.5	4.0	NS
4	Stem: degree of zigzag	1, 3	1.5	1.3	1.8	1.5	1.5	NS
5	Stem: anthocyanin coloration of brace roots	1–9	2.0	5.5	5.5	2.5	3.5	***
6	Tassel: time of anthesis (on middle third of main axis, 50% of plants)	1–9	3.0	3.3	3.8	2.3	2.0	***
7	Tassel: anthocyanin coloration at base of glume (in middle third of main axis)	1–9	2.5	2.0	3.5	3.0	2.0	*
8	Tassel: anthocyanin coloration of glumes excluding base (in middle third of main axis)	1–9	2.5	4.0	5.0	2.5	4.0	*
9	Tassel: anthocyanin coloration of anthers (in middle third of main axis on fresh anthers)	1–9	1.5	1.5	2.5	1.5	3.5	*
10	Tassel: density of spikelets (in middle third of main axis)	1–9	4.5	4.0	3.5	4.0	4.0	NS
11	Tassel: angle between main axis and lateral branches (in lower third of tassel)	1–9	4.0	4.0	3.5	4.0	4.5	NS
12	Tassel: attitude of lateral branches (in lower third of tassel)	1–9	4.5	4.0	6.5	4	3.5	*
13	Tassel: number of primary and lateral branches	1–9	5.5	6.0	6.5	6	6	NS
14	Ear: time of silk emergence (50% plants)	1–9	2.25	3.8	3.75	2.75	2.5	*
15	Ear: anthocyanin coloration of silks	1, 9	3	7.0	9	5	9	*
16	Leaf: anthocyanin coloration of sheath (in middle of plant)	1–9	1.5	1.0	2	1	1.5	NS
17	Tassel: length of main axis above lowest side branch	1–9	6	4.0	6.5	5.5	6.5	NS
18	Tassel: length of main axis above upper side branch	1–9	2.5	6.0	6.5	6	6.5	**
20	Plant: length (up to flag leaf)	1–9	4.5	4.5	6.5	4.5	5.5	*
22	Plant: ratio between height of insertion of upper ear to plant length	1–9	4.5	4.5	5.5	6	6	NS
23	Leaf: width of blade (leaf of upper ear)	1–9	5.5	4.0	5	6	5	NS
24	Ear: length of peduncle	1–9	3.5	3.7	6.5	4.5	6	NS
25	Ear: length without husk	1–9	4	3.5	3.5	3	3	NS
26	Ear: diameter without husk (in middle)	1–9	4.5	4	5.5	4	4	NS
27	Ear: shape	1–9	1.5	1.5	2.25	1.75	1.75	NS
28	Ear: number of rows of grains	1–9	5.5	6	6.5	6.5	6	NS
29	Ear: type of grain (in middle third of ear)	1, 7	2.75	1.75	1.75	2	3	NS
30	Ear: color of top of grain	1–9	1	1	1	1	1	NS
31	Ear: color of dorsal side of grain	1–9	1	3	1.25	1.5	1.75	NS
32	Ear: anthocyanin coloration of glumes of cob	1, 9	1	1	1	1	1	NS
33	Kernel: row arrangement	1–9	2.5	1.5	1.75	1.75	2.25	NS
34	Grain shape	1–9	2.5	1.75	2	2.5	2.5	NS
35	Grain size (1000-grain weight)	1–9	4	4.5	5.5	6	5	NS

*Significant at the $P < 0.5$ probability level.

**Significant at the $P < 0.01$ probability level.

***Significant at the $P < 0.001$ probability level.

†Included in the SSR analysis. ARDA, Agricultural Rural Development Authority; CBI, Crop Breeding Institute.

‡Check: Various sources of ZM521 grown for DUS study but not included in SSR analysis.

Table 3. List of the maize populations, sources of seeds, and the ratio and identity of the contaminating sources, used in the study comparing sources and significance of variation.

Bulk ID	Variety name	% Contamination	Seed source (field and year)
01_1	Across 0025 + 15% Across 0045	15	AF04B-5051-24
02_1	Turipana 0030	0	AF02B-5022
02_2	Turipana 0030	0	
03_1	Turipana 0030	0	AF02B-5022
03_2	Turipana 0030	0	
04_1	Across 0025	0	AF02B-5037
04_2	Across 0025	0	
05_1	S97 TLW GH "A" + 10% Across 0045	10	PR99A-448
05_2	S97 TLW GH "A" + 10% Across 0045	10	
06_1	Turipana 0030	0	AF04B-5051-1
06_2	Turipana 0030	0	
07_1	Agua Fria 0021	0	AF02B-5027
07_2	Agua Fria 0021	0	
08_1	S97 TLW GH "B" + 20% Across 0045	20	PR99A-449
08_2	S97 TLW GH "B" + 20% Across 0045	20	
09_1	Agua Fria 0021 + 15% Across 0045	15	AF04B-5051-13
09_2	Agua Fria 0021 + 15% Across 0045	15	
10_1	Across 0025 + 10% Across 0045	10	AF04B-5051-24
10_2	Across 0025 + 10% Across 0045	10	
11_1	S97 TLW GH "A"	0	PR99A-448
11_2	S97 TLW GH "A"	0	
12_1	Across 0025	0	AF04B-5051-24
12_2	Across 0025	0	
13_1	Omonita 9243	0	AF03B-5440-20
13_2	Omonita 9243	0	
14_1	S98 TLY-1B	0	AF03B-5440-31
14_2	S98 TLY-1B	0	
15_1	Across 0025 + 20% Across 0045	20	AF04B-5051-24
15_2	Across 0025 + 20% Across 0045	20	
16_1	S97 TLW GH "A&B" (2)	0	PR99A-451
16_2	S97 TLW GH "A&B" (2)	0	
17_1	S97 TLW GH "B" + 5% Across 0045	5	PR99A-449
17_2	S97 TLW GH "B" + 5% Across 0045	5	
18_1	S97 TLW GH "A"	0	PR99A-448
18_2	S97 TLW GH "A"	0	
19_1	S97 TLW GH "A" + 15% Across 0045	15	PR99A-448
19_2	S97 TLW GH "A" + 15% Across 0045	15	
20_1	Across 0025	0	AF02B-5037
20_2	Across 0025	0	
21_1	S97 TLW GH "A&B" (1)	0	PR99A-450
21_2	S97 TLW GH "A&B" (1)	0	
22_1	S97 TLW GH "B" + 50% Across 0045	50	PR99A-449
22_2	S97 TLW GH "B" + 50% Across 0045	50	
23_1	S97 TLW GH "A" + 20% Across 0045	20	PR99A-448
23_2	S97 TLW GH "A" + 20% Across 0045	20	
24_1	S97 TLW GH "B"	0	PR99A-449
24_2	S97 TLW GH "B"	0	
25_1	Agua Fria 0021 + 50% Across 0045	50	AF04B-5051-13
25_2	Agua Fria 0021 + 50% Across 0045	50	
26_1	S99 TLW BNSEQ(1)	0	TL00A-1427
27_1	S97 TLW GH "A&B" (2)	0	AF04B-5051-32
27_2	S97 TLW GH "A&B" (2)	0	
28_1	Omonita 9243	0	AF03B-5440-20

(cont'd)

20, and 50% mixtures. All populations tested were white populations, and the contaminants were always yellow, for ease of seed handling. Table 3 includes a list of all populations, seed sources, and mixtures (contamination levels) tested in the study. The following possible sources of differences between any two given seed samples were tested, using different subsets from the populations described in Table 3: (i) differences caused by sampling different bulks from the same OPV (differences between two random bulks of 15 seeds per population from the same source were tested); (ii) differences caused by possible contaminations, and the level of contamination needed before a difference was registered by the methodology (5, 10, 20, 50% levels, using an unrelated OPV to "contaminate" the population being tested by mixing of DNA, were tested); (iii) differences caused by the seed source of the same named OPV (where the source is the field and growing season where the current generation of seed has been grown, and differences between two or three sources of seed per population were tested); and (iv) true genetic differences between populations (nine different [unrelated] populations were tested).

To generate populations with different levels of contamination from other populations, we created a sample of 100 seeds; for the 5% contamination level we took 95 yellow seeds and 5 white, etc. From this sample, we took a random subsample of 15 seeds (regardless of color) to form the bulk. In all fingerprinting tests, each population was fingerprinted using bulks of DNA from 15 individual plants, all from the same population (or mixed sample, in the case of the contamination study). One or two bulks of 15 plants each are routinely characterized per population using the bulking technique (Dubreuil and Charcosset, 1999; Dubreuil et al., 2006); however, Test 1, above, will rigorously test if one bulk is sufficient. DNA was extracted from individual plants and mixed after quantification to form the bulk. Genomic DNA was extracted using the CTAB method from lyophilized leaf tissue according to CIMMYT protocols (http://www.cimmyt.org/english/docs/manual/protocols/abc_amgl.pdf [verified 23 Nov. 2009]). Two bulks per population were used in all but two cases (due to low seed germination, listed in Table 3).

Twenty-seven SSR markers were used to distinguish the same populations of ZM521 as were used in the DUS study, and 45 SSR markers (including 11 overlapping with the 27) were run on the nine populations to test the bulked methodology first reported in Dubreuil et al. (2006). Not all SSR loci are suitable for bulked amplification, as stuttering, preferential amplification, or complicated banding patterns cannot be resolved in a bulk. The SSRs published by Dubreuil et al. (2006) and additional markers optimized for this study can be found along with standard polymerase chain reaction (PCR) amplification protocols at http://www.cimmyt.org/english/docs/manual/protocols/abc_amgl.pdf (verified 23 Nov. 2009). Fluorescently labeled PCR products were separated by capillary electrophoresis in an ABI 3100 automatic DNA sequencer (Applied Biosystems, Foster City, CA). Genescan v3.0 was used to generate input files for the Freqs-R program (Franco et al., 2005), which removes background noise and PCR artifacts, and calculates allele frequencies for bulked pools. It can be downloaded free of charge from http://www.generationcp.org/bioinformatics.php [verified 23 Nov. 2009].

Once allele frequencies were calculated with the Freqs-R program, the FtoL-R (frequencies to lengths) program (http://www.generationcp.org/bioinformatics.php [verified 23 Nov. 2009]) was used to simulate the alleles (reported as length in base pairs) for 15 individuals that would satisfy the bulked allele frequencies and expected heterozygosity of each sample. This was done because other software packages used in this study do not accept population frequencies as input files. The program DARwin 5.0 (Perrier and Jacquemoud-Collet, 2006) was used to calculate Euclidean distances between bulks to create a neighbor-joining dendrogram for both the ZM521 seed source tests and the tests of the factors contributing to the differences between populations. Bootstrap values were generated using 1000 iterations of the clustering procedure for the dendrogram of the ZM521 bulks. A neighbor-joining phylogram of the ZM521 seed sources plus two unrelated populations was also generated as a reference as to the significance of the distances between the ZM521 bulks. Finally, the significance of each of the factors contributing to differences between the populations was studied using the analysis of molecular variance (AMOVA) according to Weir (1996) with Arlequin V3.01 (Excoffier et al., 2005). The significance of the differences between populations was calculated using resampling (10,000 repetitions) of the F_{ST} parameter, per Berg and Hamrick (1997).

RESULTS AND DISCUSSION

Farmers' Tests

The characteristics of the two sources of ZM521 (2004–2005 and 2005–2006) are described in Table 4. Farmers preferred the ZM521 from the 2004–2005 season, based on the earlier, taller plants, and larger cob size (Table 4). Early-maturing varieties are able to escape drought and are thus more suitable for the short growing season than late-maturing varieties. Larger cob size is associated with higher yielding varieties (Setimela et al., 2004). Many farmers were familiar with the characteristics of the ZM521, as they have planted them before and expected a better performance in 2005–2006.

Tests of Different Sources of ZM521

Some of the DUS characteristics were significantly different among the sources of ZM521, while for other traits there were no significant differences (Table 2). The seed of ZM521 from Crop Breeding Institute (CBI) and Agricultural Rural Development Authority (ARDA) in Harare had higher scores than the reference ZM521 for time of silk emergence (50% plants), attitude of lateral branches in the lower third of the tassel, time to anthesis, and plant height to the flag leaf. Although some traits may appear the same between different (unrelated) populations, plants from the same population must appear the same for every trait measured. Open-pollinated varieties do have a heterogeneous genetic base; however, for important agronomic traits, and certainly those used for DUS studies, these populations must be fixed and stable and display very low variation between individual plants. The phenotypic differences of CBI and ARDA from the other sources of ZM521 indicate low genetic similarities among CBI, ARDA, and the reference ZM521 populations in this study (Table 2).

In the dendrogram of the five different seed sources of ZM521 presented in Fig. 1, the two bulks of each seed source (labeled "a" and "b") always cluster together

Table 3. Continued.

Bulk ID	Variety name	% Contamination	Seed source (field and year)
28_2	Omonita 9243	0	
29_1	S97 TLW GH "B" + 15% Across 0045	15	PR99A-449
29_2	S97 TLW GH "B" + 15% Across 0045	15	
30_1	Across 0025 + 50% Across 0045	50	AF04B-5051-24
30_2	Across 0025 + 50% Across 0045	50	
31_1	S99 TLW BNSEQ(1)	0	AF04B-5051-34
31_2	S99 TLW BNSEQ(1)	0	
32_1	S98 TLY-1B	0	AF03B-5440-31
32_2	S98 TLY-1B	0	
33_1	Across 0025 + 5% Across 0045	5	AF04B-5051-24
33_2	Across 0025 + 5% Across 0045	5	
34_1	Agua Fria 0021 + 10% Across0045	10	AF04B-5051-13
34_2	Agua Fria 0021 + 10% Across0045	10	
35_1	S97 TLW GH "A" + 50% Across 0045	50	PR99A-448
35_2	S97 TLW GH "A" + 50% Across 0045	50	
36_1	S97 TLW GH "A&B" (2)	0	AF04B-5051-32
36_2	S97 TLW GH "A&B" (2)	0	
37_1	S97 TLW GH "A&B" (1)	0	PR99A-450
37_2	S97 TLW GH "A&B" (1)	0	
38_1	S97 TLW GH "B"	0	PR99A-449
38_2	S97 TLW GH "B"	0	
39_1	Agua Fria 0021 + 5% Across 9745	5	AF04B-5051-13
39_2	Agua Fria 0021 + 5% Across 9745	5	
40_1	Agua Fria 0021	0	AF04B-5051-13
40_2	Agua Fria 0021	0	
41_1	Agua Fria 0021 + 20% Across 0045	20	AF04B-5051-13
41_2	Agua Fria 0021 + 20% Across 0045	20	
42_1	S97 TLW GH "A" + 5% Across 0045	5	PR99A-448
42_2	S97 TLW GH "A" + 5% Across 0045	5	
43_1	S97 TLW GH "B" + 10% Across 0045	10	PR99A-449
43_2	S97 TLW GH "B" + 10% Across 0045	10	
44_1	S99 TLW BNSEQ(1)	0	AF04B-5051-34
44_2	S99 TLW BNSEQ(1)	0	
45_1	Agua Fria 0021	0	AF02B-5027
45_2	Agua Fria 0021	0	

Table 4. Farmers' comparison of two maize seed sources planted in 2005–2006.

Trait	Scale	ZM521 2004–2005	ZM521 2005–2006
Time to maturity	Early, medium, and late maturing	Early	Late
Cob size	Small, medium, and large	Large	Small
Plant height	Short, medium, and tall	Tall	Short
Plant stand	Good, poor, average	Good	Poor
Drought tolerance	Very good, average, poor	Very good	Poor

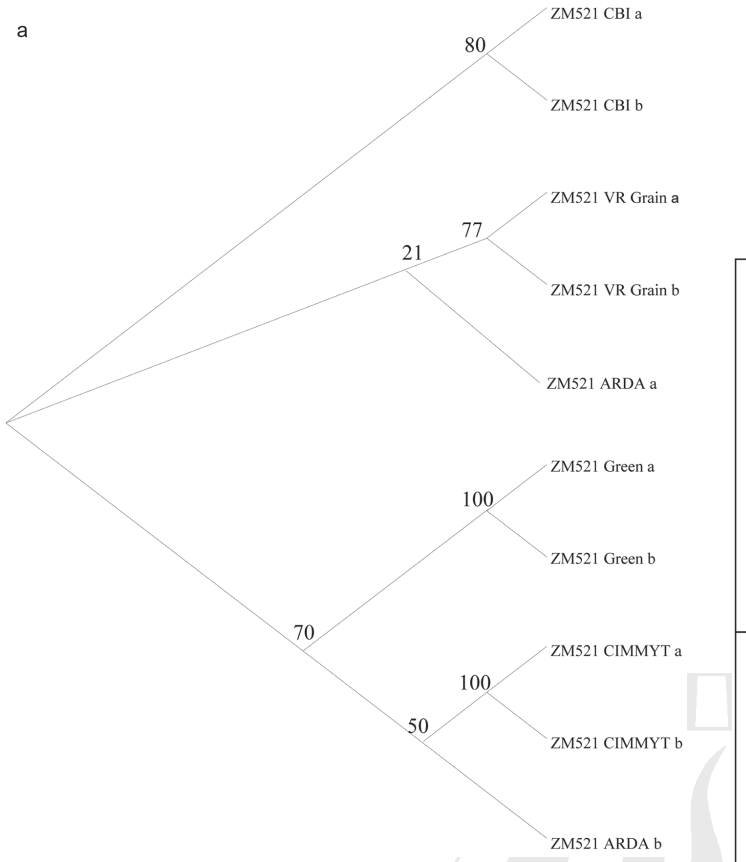

Figure 1. Unpaired group method for arithmetic means dendrogram of the five different seed sources of ZM521 maize used in this study and described in Table 1, based on the shared allele genetic similarity between pairs of populations calculated using 27 simple sequence repeat markers. Numbers at the junctions of clusters are bootstrap confidence intervals based on 10,000 repetitions.

except the ARDA source, which had much missing data in bulk "a" for the 27 markers, so results must be interpreted with caution for this bulk. There is a high level of diversity between these populations, belying the hypothesis that they are all drawn from the same original source of ZM521. The average Euclidian distance between all bulks is 0.21 (data not shown). The reference population (CIMMYT) bulks, ARDA bulk "b," and Green bulks cluster together with an average distance of 0.19, and the AMOVA analysis indicates no difference between these populations at the $P = 0.05$ level (data not shown). The ARDA source, bulk "a," clusters with the VR Grain bulks, but with only a 21% confidence level according to the bootstrap analysis. The AMOVA confirms that these three bulks are not different at the $P = 0.05$ level of significance, and the average Euclidian distance between these bulks to all other bulks in the analysis is 0.24. The CBI bulks cluster together and show no difference at the $P = 0.05$ confidence level, but they have an average Euclidian distance of 0.26 to the other bulks in the study. The AMOVA cannot conclude that the VR Grain and especially the CBI sources are ZM521.

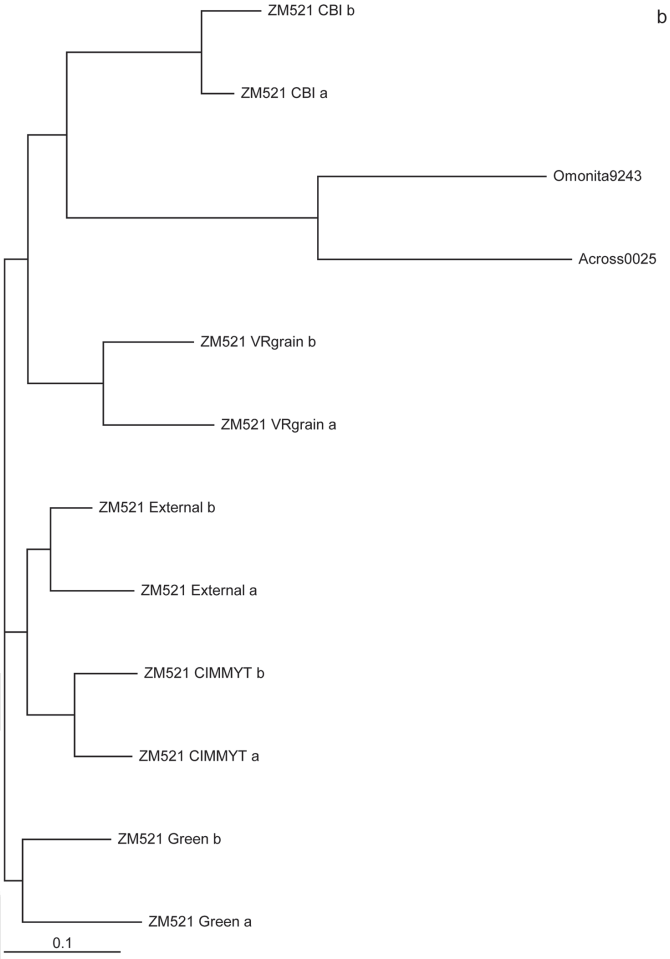

Figure 2. Neighbor-joining phylogram of the five different seed sources of ZM521 maize and two additional populations unrelated by pedigree based on the shared allele genetic similarity between pairs of populations calculated using 11 simple sequence repeat (SSR) markers. Shared allele genetic similarity is measured on a scale of 0 (indicating no alleles shared in common) to 1 (indicating exact identity), and the scale at the bottom indicates 1/10th of this range.

The neighbor-joining phylogram of the ZM521 populations including two additional populations, unrelated by pedigree, is shown in Fig. 2. The same patterns as were seen in Fig. 1 are still evident: the reference and both "ARDA" and "Green" bulks cluster together and far from the unrelated populations; and the VR Grain and CBI sources of the ZM521 population are far distant from the other ZM521. In fact, the CBI source looks more similar to the two unrelated populations than to the other ZM521.

SSR Tests of the Mixed Populations

Effect of Sampling in the Bulked Procedure

The two bulks of each population clustered most closely together in 36 out of 43 pairs of bulks. This indicates that there is a small difference caused by the subsampling of populations when creating the bulks, or in errors when scoring the bulks using the bulked method. When tested with the F_{ST} parameter, six of these seven pairs were significantly different at the $P = 0.05$ level (data not shown), indicating

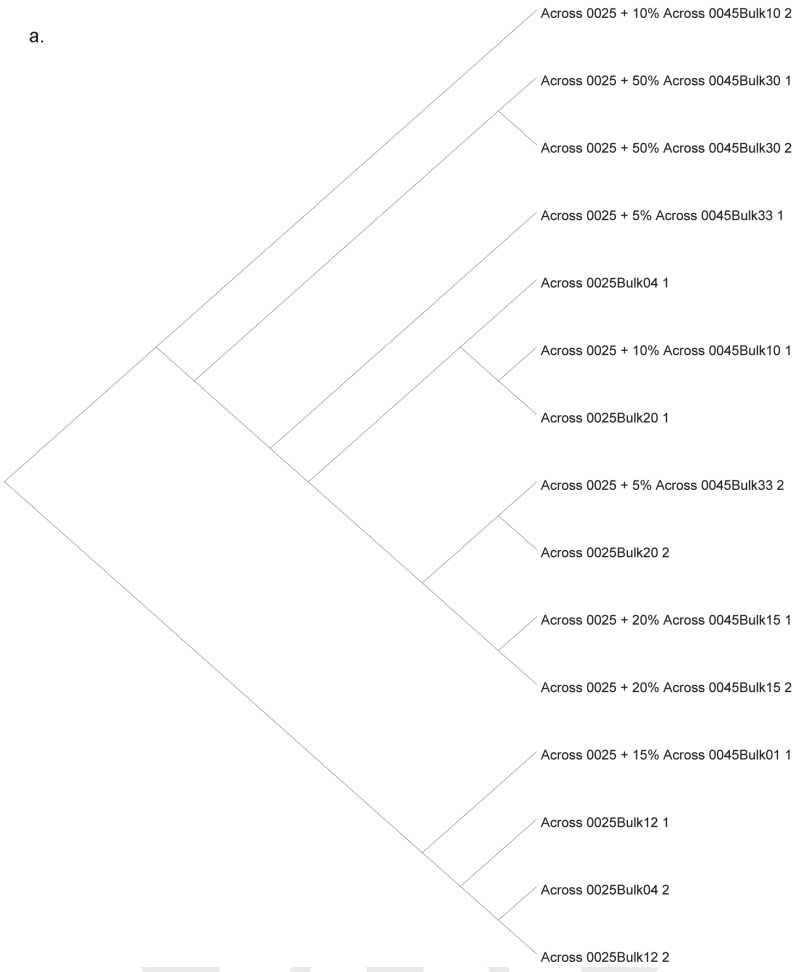

Figure 3. Unpaired group method for arithmetic means dendrogram of each of four named maize populations, including only the different sources of seeds and the contaminated samples of the same populations (described in Table 3), based on 45 simple sequence repeat markers. (a) Open-pollinated variety (OPV) Across 0025.

that the sampling used in the bulks is causing a small but significant source of variation in the analyses. Past studies of maize populations usually included one or a few (at most 12) individuals per population. Due to the heterogeneous nature of maize populations, sampling with such a low number will not be representative of the population from which the sample was drawn. This study found that 30 individuals is more satisfactory than 15. If following the stricter guidelines for DUS testing, which require 80 individuals to be characterized for OPVs (http://www.upov.int/en/publications/tg-rom/tg002/tg_2_6.pdf [verified 23 Nov. 2009], six bulks of 15 individuals each per population could be fingerprinted to have marker information for 90 individuals at a fraction of the cost of running 80 individuals one at a time.

Effect of Contaminating Populations on the Bulked Procedure

Analyzing each named population with the mixed (contaminated) populations of the same name tended to form one or two clusters of the pure populations (on rare occasions including one of the lower percentage mixtures); one or two clusters based on the most heavily mixed populations; and occasionally one intermediate cluster with the slightly mixed and some of the pure populations (Fig. 3a–d). Clustering of the pure selections of populations from different seed sources separately indicates a difficulty in keeping seed sources pure (as discussed in the section below). When looking at the F_{ST} statistics for each named population, the pure sample is always significantly different from the contaminated samples, except with the Agua Fria population, in which the 15% contaminated sample was not significantly different than the pure sample, and the S97 TLW GH "A" population, in which the 20% contaminated sample was not significantly different than the pure sample (Table 5). This analysis indicates that populations contaminated by moderate levels seed mixing (>20%) will be consistently differentiated from the pure populations, and even low levels (5– 10%) can usually be identified (unless the contaminating population happens to be very closely related to the pure sample, a condition we did not test in this study). Pollen flow from neighboring fields may also be identified using this technique, although exact quantification of pollen flow may be underestimated.

Figure 3. Continued. Unpaired group method for arithmetic means dendrogram of each of four named maize populations, including only the different sources of seeds and the contaminated samples of the same populations (described in Table 3), based on 45 simple sequence repeat markers. (b) OPV Agua Fria 0021.

Effect of Different Seed Sources on the Bulked Procedure

The F_{ST} statistic used to test the significance of differences between the same named populations grown in different field sites or years found significant differences in 13 of 18 possible comparisons (data not shown). Differences due to seed source depend on the care taken by each field manager when increasing seed for each population, a problem already noted in the ZM521 comparison. It is apparently quite difficult to ensure seed production with absolutely no pollen or seed flow from other populations and, in addition, genetic differences can be caused by unintended selection during seed increase, genetic drift from small sample sizes, or genetic substructure from possible assortative or disassortative mating (crossing most similar or dissimilar plants with each other), which often happens if all plants do not shed pollen on the same day. Genetic differences have been seen between different sources of the same cultivar, including inbred lines and doubled haploids, in past marker studies (Smith et al., 1991; Heckenberger et al., 2002).

Effect of Different Populations on the Bulked Procedure

In every case, populations with a different name were found to be significantly different, according to the F_{ST} values (Table 6). Although some of the bulks drawn from the same named variety are also significantly different (as discussed in the above sections), the average F_{ST} for comparisons from within the same named population are always much lower than the F_{ST} among varieties (0.027 vs. 0.14).

Significance of Sources of Differences between Subsamples

The AMOVA used to test the significance of each factor that could make two subsamples of the same population look different is shown in Table 7, and shows that the majority of the variation occurs between individuals within populations in the study, as to be expected with an out-breeding crop like maize (Warburton et al., 2002, 2008). However, in agreement with all the F_{ST} tests described above, significant differences are seen among different named populations, as when contaminants are added to the populations. Much smaller but still

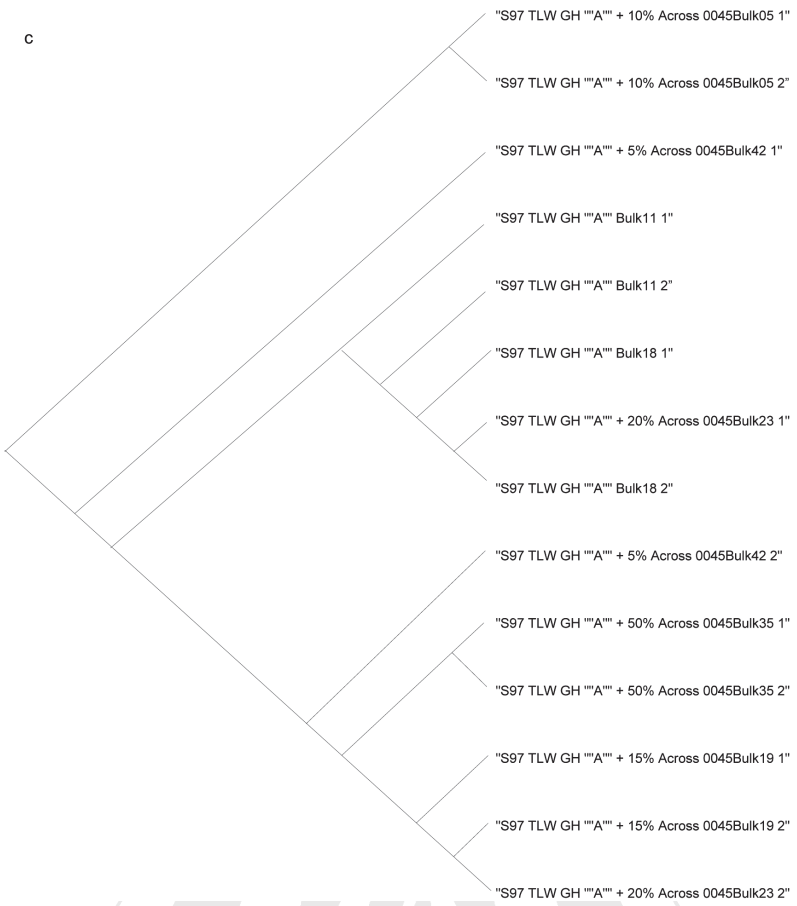

Figure 3. Continued. Unpaired group method for arithmetic means dendrogram of each of four named maize populations, including only the different sources of seeds and the contaminated samples of the same populations (described in Table 3), based on 45 simple sequence repeat markers. (c) OPV S97TLWGHA.

significant differences can be seen between different sources of seed of the same named populations, and due to differences between the two bulks sampled from the same source. This indicates that different subsamples of the same OPV may look slightly different, either due to sampling error, as 15 is apparently too few individuals for a true representation of the diversity within a population of maize, or due to error in the bulked analysis technique. We would therefore recommend that when the identity of a population is being established (rather than the degree of relationship between two populations), no fewer than two bulks of 15 individuals each be sampled and the average allele frequencies for both bulks used. In addition, the bulked assay should be used following training and practice to avoid additional error.

Variation caused by different sources of seed is much lower than the other sources of variation (except the sampling caused by the repeated bulks), but is a significant source of variation among samples. This methodology can be used to help keep different stocks and sources of an OPV pure and not drifting due to sampling, selection, or gene flow. Variation caused by different levels of contaminating gene flow will complicate identification, as Fig. 3 shows how mixed populations greatly confuse the relationships between similar populations. This method can distinguish some of the contaminated populations from the pure source, but low levels of contamination, or contamination from related seed sources, may be undetectable by either the markers or phenotypic screens.

CONCLUSIONS

The seed lot from the 2004–2005 season performed better than 2005–2006 seed source and farmers preferred it. The genetic purity of ZM521 from the 2005–2006 season was demonstrated by SSR markers and DUS testing to be variable, depending on seed source. The SSRs were able to distinguish unrelated OPVs and can be used to investigate the claims of seed companies as to population identity, and distinguish potential causes of differences among the groups, including subsamples (including different seed sources) of the same population and contaminated subpopulations vs. the original source. This can be used to set guidelines to use SSRs for declaring two samples to belong to the same population, or distinguish them definitively, especially as laboratories analyze seeds of dubious identity. This may provide additional information in the DUS registration of new varieties and can aid seed companies, governmental agencies, and NGOs to ensure a pure seed supply to farmers, free of inadvertent or purposeful seed mixing or substitution.

Figure 3. Continued. Unpaired group method for arithmetic means dendrogram of each of four named maize populations, including only the different sources of seeds and the contaminated samples of the same populations (described in Table 3), based on 45 simple sequence repeat markers. (d) S97TLWGHB.

Table 5. F_{ST} values for pairwise comparisons of "pure" (0%) vs. "contaminated" (5%, 10%, etc.) subsamples from four different maize populations.

Contamination level	Population			
	Across 0025	Agua Fria 021	S97 TLW GH "A"	S97 TLW GH "B"
	0%	0%	0%	0%
5%	0.0620**	0.0555**	0.0297*	0.0626**
10%	0.0629**	0.0537**	0.1999**	0.0357**
15%	0.0781**	0.0126 NS†	0.0217*	0.0316**
20%	0.0794**	0.1556**	−0.0775 NS	0.0832**
50%	0.0584**	0.0695**	0.0338*	0.0238**

*$P \leq 0.05$, F_{ST} values showing differences (rejecting the hypothesis of nondifference) in 10,000 bootstrap repetitions.

**$P \leq 0.01$, F_{ST} values showing differences (rejecting the hypothesis of nondifference) in 10,000 bootstrap repetitions.

†NS, nonsignificant ($P > 0.05$).

References

Bänziger, M., and J. de Meyer. 2002. Collaborative maize variety development for stress-prone environments in southern Africa. p. 269–296. *In* D.A. Cleveland and D. Soleri (ed.) Farmers, scientists, and plant breeding: Integrating knowledge and practice. CABI, Oxon, UK.

Bänziger, M., J. de Meyer, M.S. Mwala, M.A.R. Phiri, B. Vivek, and K.V. Pixley. 2002. Progress in delivering stress tolerant maize varieties to farmers in southern Africa. p. 67–71. *In* J. DeVries et al. (ed.) Biotechnology, breeding and crop systems for African crops: Research and product development that reaches farmers: Proc. of an Int. Workshop, Los Baños, Philippines.

Bänziger, M., G.O. Edmeades, and H.R. Lafitte. 1999. Selection for drought tolerance increases maize yields over a range of nitrogen levels. Crop Sci. 39:1035–1040.

Berg, E.E., and J.L. Hamrick. 1997. Quantification of genetic diversity at allozyme loci. Can. J. For. Res. 27:415–424.

Dubreuil, P., and A. Charcosset. 1999. Relationships among maize inbred lines and populations from European and North-American origins as estimated using RFLP markers. Theor. Appl. Genet. 99:473–480.

Dubreuil, P., M. Warburton, M. Chastanet, D. Hoisington, and

Table 6. F_{ST} values between differently named varieties (Across, AguaFria, Omonita, S97 TLW GH, S97 TL AB(1), S97 TL AB(2), S97 TLW GHB, S98 TLY B, and S99 SEQ), and average F_{ST} values between bulks within the same named varieties (Average within). All of the differences between varieties are significant using 10,000 bootstrap repetitions.

Variety	Across	AguaFria	Omonita	S97 TLW GH	S97 TL AB(1)	S97 TL AB(2)	S97 TLW GHB	S98 TLY B	S99 SEQ	Average within
Across	–									0.064
AguaFria	0.153	–								0.084
Omonita	0.098	0.134	–							0.041
S97 TLW GHA	0.127	0.153	0.048	–						0.015
S97 TL A&B(1)	0.104	0.179	0.053	0.089	–					0.000
S97 TL A&B(2)	0.135	0.164	0.070	0.070	0.041	–				0.015
S97 TLW GHB	0.158	0.241	0.126	0.153	0.051	0.088	–			0.036
S98 TLY B	0.135	0.183	0.111	0.120	0.096	0.082	0.172	–		0.000
S99 SEQ(1)	0.134	0.291	0.169	0.192	0.122	0.141	0.215	0.100	–	0.005
Turpiana	0.144	0.239	0.117	0.176	0.155	0.165	0.149	0.199	0.250	0.064
Mean									0.139	0.028

Table 7. Analysis of molecular variance of the simple sequence repeat differences measured on the populations listed in Table 3. % Variation is the percentage of the total variance explained by each variance component.

Test 1†			Test 2			Test 3		
Source of variation	df	% Variation	Source of variation	df	% Variation	Source of variation	df	% Variation
Among bulks	42	15.07**	Among populations	4	7.24**	Among populations	4	20.81**
Among repetitions within bulks	43	2.22**	Among contamination levels within populations	20	8.14**	Among seed sources within populations	5	1.78**
Between individuals within repetitions	2494	82.71**	Between individuals within levels	1746	84.63**	Between individuals within seed sources	860	77.41**

***Sources of variation are significant at the $P = 0.001$ level.

†Test 1 tests the effect of the variation due to sampling error in the bulking procedure (two independent bulks of 15 individuals are chosen from the same open-pollinated variety [OPV]). Test 2 tests the effect of gene flow from contaminating populations, either via seed or pollen mixing. Test 3 tests the effect of different sources (more than one field or field season where the same named OPV has been grown for seed increase).

A. Charcosset. 2006. More on the introduction of temperate maize into Europe: Large-scale bulk SSR genotyping and new historical elements. Maydica 51:281–291.

Excoffier, L., G. Laval, and S. Schneider. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evol. Bioinformatics Online 1:47–50.

Falconer, D.S. 1984. An introduction to quantitative genetics. Eugen Ulmer, Stuttgart.

Franco, J., M. Warburton, P. Dubreuil, and S. Dreisigacker. 2005. User's manual for the FREQS-R Program for estimating allele frequencies for fingerprinting and genetic diversity studies using bulked heterogeneous populations. CIMMYT, Mexico, D.F.

Heckenberger, M., M. Bohn, J.S. Ziegle, L.K. Joe, J.D. Hauser, M. Hutton, and A.E. Melchinger. 2002. Variation of DNA fingerprints among accessions within maize inbred lines and implications for identification of essentially derived varieties. Mol. Breed. 10:181–191.

Langyintuo, A., and P.S. Setimela. 2007. Assessment of effectiveness of maize seed assistance to vulnerable farmers in Zimbabwe. CIMMYT, Mexico, D.F.

Longley, C., G. Kayobyo, and R. Tripp. 2001. Guidelines for seed production and the dissemination of improved varieties. ODI, London.

Mwala, M.S., J. de Meyer, P.S. Setimela, and M. Bänziger. 2004. Participatory maize variety evaluation for increased adoption. *In* CIMMYT (ed.) Resilient crops for water-limited environments: Abstracts from an international symposium, Cocoyoc, Mexico.

Perrier, X., and J.P. Jacquemoud-Collet. 2006. DARwin software. Available at http://darwin.cirad.fr/darwin (verified 18 Nov. 2009).

Pixley, K.V., and M. Bänziger. 2004. Open-pollinated maize varieties: A backward step or valuable option for farmers? p. 22–29. *In* D.K. Friesen and A.F.E. Palmer (ed.) Integrated approaches to higher maize productivity in the new millennium. Proc. of the Eastern and Southern Africa Regional Maize Conf., 7th, Nairobi, Kenya. 5–11 Feb. 2002.

SAS Institute. 2004. SAS OnlineDoc 9.1.3. SAS Inst., Cary, NC.

Setimela, P.S., M. Bänziger, and M.S. Mwala. 2004. Choosing the crop and variety. p. 23–26. *In* P.S. Setimela et al. (ed.) Successful community-based seed production strategies/Produção de sementes de culturas alimentares na região da SADC. CIMMYT, Mexico, D.F.

Setimela, P.S., X. Mhike, J.F. MacRobert, and D. Muungani. 2005. Maize hybrid and open-pollinated varieties. CIMMYT, Mexico, D.F.

Smith, J.S.C., O.S. Smith, S.L. Bowen, R.A. Tenborg, and S.J. Wall. 1991. The description and assessment of distances between inbred lines of maize. III. A revised scheme for the testing of distinctiveness between inbred lines utilizing DNA RFLPs. Maydica 36:213–226.

Warburton, M.L., J.C. Reif, M. Frisch, M. Bohn, C. Bedoya, X.C. Xia, J. Crossa, J. Franco, D. Hoisington, K. Pixley, S. Taba, and A.E. Melchinger. 2008. Trends in genetic diversity in CIMMYT non-temperate maize germplasm. Crop Sci. 48:617–624.

Warburton, M.L., X. Xianchun, J. Crossa, J. Franco, A.E. Melchinger, M. Frisch, M. Bohn, and D. Hoisington. 2002. Genetic characterization of CIMMYT inbred maize lines and open pollinated populations using large scale fingerprinting methods. Crop Sci. 42:1832–1840.

Weir, B.S. 1996. Genetic data analysis II: Methods for discrete population genetic data. Sinauer Assoc., Sunderland, MA.