Optimizing automated fingerprinting of maize germplasm using SSR Markers

X.C. Xia[1], M. L. Warburton[1*], D. A. Hoisington[1], M. Bohn[2], M. Frisch[2] and A.E. Melchinger[2]

1. International Maize and Wheat Improvement Center (CIMMYT), Km 45 Carr. Mexico-Veracruz, El Batan, Texcoco, Edo. de Mexico, C.P. 56130.
2. Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, D-70593, Stuttgart, Germany.

*Corresponding author: email: m.warburton@cgiar.org Fax: (52)-58047558

Abstract

The Applied Biotechnology Center of the International Maize and Wheat Improvement Center (CIMMYT) has undertaken molecular marker fingerprinting of maize germplasm in order to better understand the diversity present in breeding lines and populations, and to better classify them into heterotic groups. Fingerprinting of heterogeneous pools and populations such as CIMMYT's open pollinated varieties is more difficult than line fingerprinting, and we have tested methods to accurately but efficiently characterize these populations. We have begun a collaborative study with the University of Hohenheim to optimize high-throughput fingerprinting techniques and analysis using SSR markers multiplexed by size and fluorescent dye color and run on an ABI Prism™ 377 automated DNA sequencer. A pilot study was run where 57 inbred lines and 7 populations of CIMMYT maize were fingerprinted using 38 primers, and data was converted to binary matrices and allele frequencies were calculated for each population. Analyzed data showed that the inbreds clustered according to pedigree and selection history, as expected, and that populations could be uniquely characterized based on allele frequency. Furthermore, when the populations were characterized using simply the presence or absence of the alleles, similar results were found as with analysis using allele frequency. This indicates the possibility of bulking DNA from several individuals in a population to save considerable time and reagents. We will continue to fingerprint additional germplasm using these optimized techniques.

Keywords: SSR, maize, fingerprinting, diversity

Introduction

Knowledge of patterns of diversity of genetic resources is of great importance in maize breeding in order to maximize heterosis in hybrid combinations and to maintain diversity of breeding lines. PCR based SSR markers have been widely used in the fingerprinting of maize germplasm (Smith et al. 1997; Senior et al. 1998), because of their high level of polymorphisms (Saghai Maroof et al. 1994) and their ease of detection via automated systems (Sharon et al. 1997). The CIMMYT Maize Genetic Resources Center and the CIMMYT Maize Breeding Program have over 17,000 inbred lines and populations of maize. The fingerprinting of such a large collection of unique entries will require very high-throughput methodologies in the laboratory and in data collection, storage, and analysis. The objectives of this study are (1) to optimize automated methods for fingerprinting of maize germplasm using SSR markers, and (2) to characterize 57 CIMMYT inbred lines and 7 tropical populations using these methods.

Materials and Methods

Plant Materials and DNA extraction

Fifty-seven CIMMYT inbred lines and 7 populations were employed in this study. Forty-eight individuals were chosen to characterize the diversity present in each population.

DNA was extracted with ‘Sap extractor’ (MEKU Erich Pollaehne GmbH) by a CTAB procedure (Clarke et al. 1989). The nucleic acid preparations were incubated with RNase A and T1 for 1 hour at room temperature, precipitated with cold 70% ethanol, dried, and resuspended in 200µl of 1xTE for storage at 4°C.

Multiplex PCR and amplification conditions

SSR markers were chosen from the MaizeDB database (http:/nucleus.agro.missouri.edu/cgi/bin/ssr_bin.pl) based on bin location (to maximize genomic coverage) and repeat unit. Information on these SSR markers can be found in Table 2. Fluorescent oligonucleotides were

bought from Operon technologies and forward primers were labeled at the 5' end with either 6-carboxyfluorescein (6-FAM), tetrachloro-6-carboxyfluorescein (TET), or hexachloro-6-carboxyfluorescein (HEX). Multiplexed PCR reactions were performed in 10-ul volumes containing 1 µl of template DNA (diluted 5x), 1.2-4.0 pmols of each primer, 1 x PCR buffer, 0.25 mM dNTPs, 1.5-2.5 mM $MgCl_2$ and 0.75 U *Taq* polymerase. The reactions were done on the Peltier Thermal cycler (MJ Research), using the amplification conditions of 94 ^oC for 2 min; followed by 30 cycles of: 94 ^oC for 30 sec, X ^oC for 1 min, and 72 ^oC for 1 min; followed by extension at 72 ^oC for 5 min. X ^oC refers the annealing temperature, which was specific for each primer combination (Table 1).

Table 1. Multiplexes at the PCR reaction and gel level for the 38 SSR markers used in this study. Annealing temperature and concentration of $MgCl_2$ are included.

PCR Group	Gel Group	FAM	TET	HEX	Anneal. temp.	$MgCl_2$
1	A	phi051	phi033	phi015	56 ^oC	1.5 mM
2	B	phi085	phi093	phi024	60 ^oC	1.5 mM
3	B	phi006, phi014	phi127		52 ^oC	1.5 mM
4	C	phi072	phi083	phi090	52 ^oC	2.5 mM
5	D	phi121	phi053	phi034	56 ^oC	1.5 mM
6	D	phi078	phi032	phi064	56 ^oC	1.5 mM
7	E	phi073	phi050	phi056	56 ^oC	1.5 mM
8	E	phi059	phi096		60 ^oC	1.5 mM
9	F		phi031	phi115	56 ^oC	1.5 mM
10	G		phi029	phi062	56 ^oC	1.5 mM
11	H	phi112	phi079	phi076	60 ^oC	2.5 mM
12	C		zcaa 391	phi041	56 ^oC	1.5 mM
13	G	phi011			60 ^oC	2.5 mM
14	G	phi022			56 ^oC	2.5 mM
15	F	phi116			56 ^oC	2.0 mM
16	A		phi070		56 ^oC	1.5 mM
17	F			phi002	60 ^oC	2.5 mM
18	H			zct118	60 ^oC	2.5 mM

Electrophoresis

Samples containing two PCR reactions (0.5 ul / each), 0.3 ul GeneScan 350 or 500 internal lane standard labeled with N, N, N, N-tetramethyl-6-

carboxyrhodamine (TAMRA), and 30% formamide were heated at 95^{o}C for 5 min, placed on ice, then loaded on 4.5% denaturing (6 M urea) acrylamide:bisacrylamide (29:1) gels (36 cm well-to-read). DNA samples were electrophoresed in 1 x TBE buffer (PH 8.3) at constant voltage (3.00 KV) for 2.5 hours on an automatic DNA sequencer (Perkin Elmer/ABI Prism^{TM} 377 DNA Sequencer).

Table 2. Number of alleles, allele size range, repeat type and PIC (Polymorphic Information Content) value for SSR loci used in this study. PIC value is calculated using the 57 CIMMYT maize inbred lines genotyped in this study.

SSR locus	Repeat type	Bin no.	# alleles	Size range (bp)	PIC value
phi002	AAGG	1.08	4	69-81	0.36
phi006	CCT	4.11	12	70-109	0.78
phi011	GCT	1.09	4	215-230	0.59
phi014	GGC	8.04	5	419-434	0.67
phi015	TTTG	8.09	8	82-114	0.69
phi022	GTGC	9.03	7	368-412	0.73
phi024	CCT	5.01	5	360-372	0.54
phi029	CCCT-CT	3.04	4	148-162	0.54
phi031	GTAC	6.04	6	185-225	0.74
phi032	TTTC	9.04	3	233-241	0.50
phi033	CTT	9.01	7	245-263	0.69
phi034	CCT	7.02	7	122-146	0.81
phi041	AGCC	10.00	3	196-204	0.40
phi050	AAGC	10.03	5	77-92	0.64
phi051	AGG-AAAG	7.06	3	139-147	0.57
phi053	ATGT	3.05	6	160-196	0.72
phi056	GCC	1.01	8	237-258	0.73
phi059	CCA	10.02	6	147-165	0.65
phi062	GAC	10.04	3	161-176	0.31
phi064	ATCC	1.11	12	69-113	0.89
phi070	GAGCT	6.07	8	70-105	0.77
phi072	AAAC	4.01	7	139-163	0.64
phi073	CAG	3.05	4	184-193	0.65
phi076	GAGCGG	4.11	3	161-173	0.66
phi078	AAAG	6.05	3	300-308	0.55
phi079	CATCT	4.05	5	180-200	0.47
phi083	CTAG	2.04	6	117-137	0.78
phi085	GCGTT	5.07	5	237-267	0.68
phi090	ATATC	2.08	3	141-151	0.10
phi093	CTAG	4.08	5	282-298	0.65
phi096	GAGGT	4.04	1	238	0
phi112	AG	7.01	13	136-175	0.83
phi115	TA-ATAC	8.03	3	242-258	0.54
phi116	TGAC-GAC	7.06	6	152-173	0.79
phi121	CCG	8.04	7	94-112	0.63
phi127	AGAC	2.08	9	96-128	0.84
bnlg391	CAA	6.01	14	68-110	0.86
bnlg118	CT	5.07	8	105-121	0.78

Data analyses

Fragment sizes were automatically calculated with GeneScan 3.1 (Perkin Elmer/Applied Biosystems) using the Local Southern sizing method. The GeneScan data were appended to a table with Genotyper 2.1, and then converted to binary matrices that was saved in an Excel file. A Hypercard routine was written to convert the data to the proper configuration for subsequent analysis and is available upon request. Binary data for inbred lines were converted to a simple matching similarity coefficient matrix and this was used to create a dendrogram using the UPGMA function. All multivariate analyses were performed using NTSYSpc 2.01. For population analysis, binary data from the 48 individuals in each population were converted to allele frequency data and Nei's genetic distance was calculated between each pair of populations in the study. This matrix was then used to create a dendrogram using UPGMA.

Results and Discussion

Genetic Diversity of Inbred Lines

The dendrogram of the analysis of the maize inbred lines is shown in Figure 1. Lines do not cluster clearly on pedigree, except for the highly related sister lines (TS lines and LP lines). Nor do they cluster based on mega-environment where grown (tropical, subtropical, highland); nor kernel color or type. This is not entirely unexpected, because CIMMYT inbred lines are generally drawn from a pool, population, or mixture of pools and populations. Pools and populations contain a very broad range of diversity, and may contain more variation within a pool or population than between them. Thus, two lines drawn at random from any given pool or population may not actually contain many alleles in common. Furthermore, lines that have been selected for each environment may have a similar initial pedigree; thus, looking for correlations in allele diversity and pedigree or environment may prove difficult. We have found, however, that the performance of many hybrids made from specific pairs of inbred lines do correlate to the genetic similarity between those parental inbred lines (data not shown). We must do much more

work to see if this descriptive trend can be used as a predictive tool for hybrid breeding, but the implications are highly favorable.

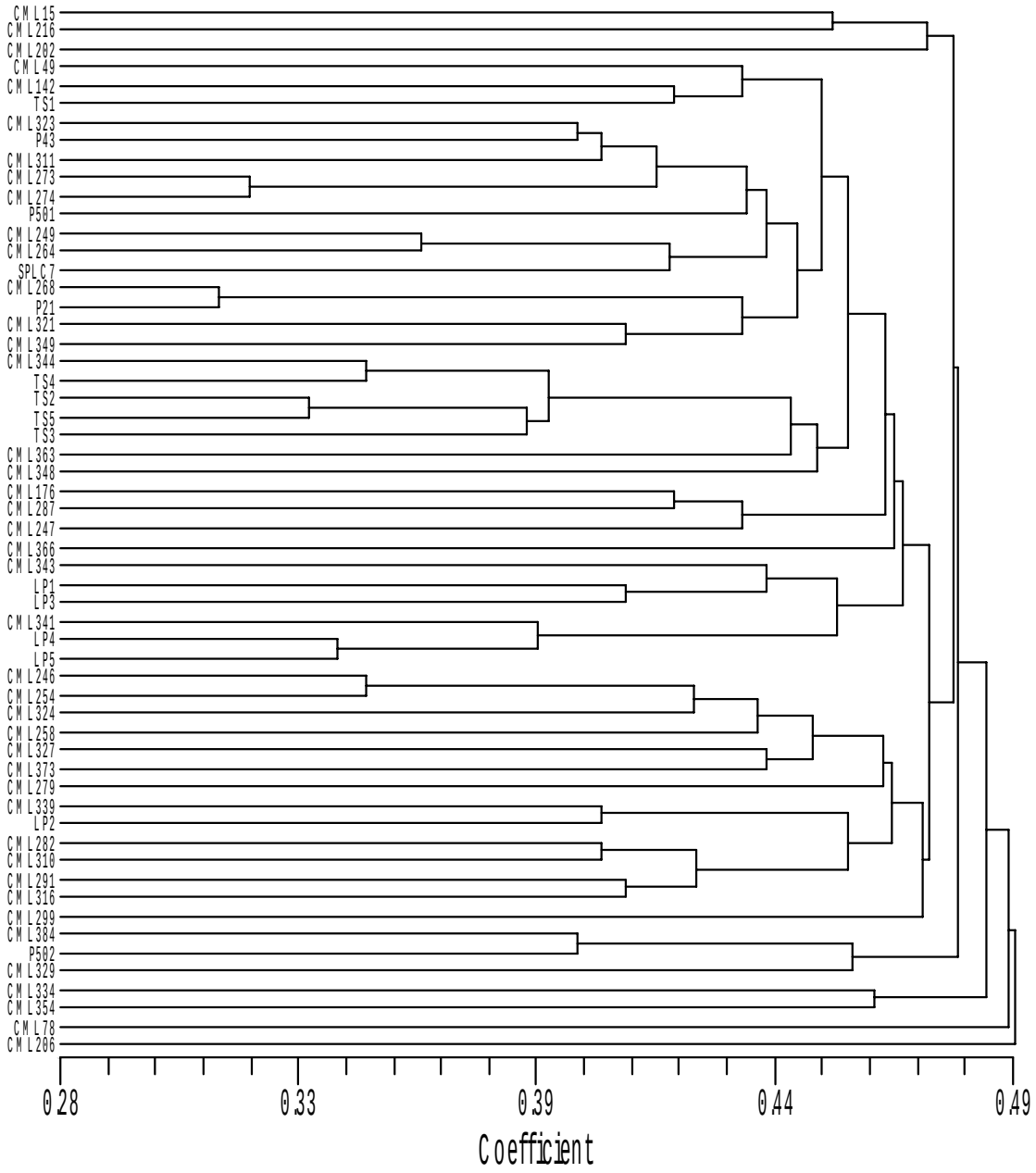

Figure 1. Dendrogram constructed with a Unweighted Paired Group Method Using Arithmetic Averages (UPGMA) clustering algorithm from the pairwise matrix of genetic similarity among 57 maize inbred lines.

Genetic Diversity of Populations

The dendrogram of the analysis of the maize populations using allele frequency of 48 individuals per population is shown in Figure 2. Populations clearly cluster according to pedigree and heterotic group. In

order to test the feasibility of bulking individuals from a population and using simply presence or absence of alleles, rather than frequency, to determine genetic relationships, all frequencies greater than zero were converted to presence of the allele. This produced a matrix of similarity coefficients which was somewhat similar to the matrix produced using allele frequency. The correlation coefficient was –0.811 (it is negative because the allele frequencies were analyzed using Nei's genetic distance, and the presence/absence data were analyzed using the simple matching similarity coefficient. Similarity = 1 – Distance). This correlation was not as high as hoped, however, we believe that these results show promise, and more bulking experiments will test if this can become a faster, more efficient method for analyzing genetic diversity in heterogenous populations.

Considerations in Large-Scale Fingerprinting

PCR conditions first had to be optimized for each SSR marker separately. Based on these conditions, and on fragment size range and dye color of each SSR, 12 PCR reactions, which were multiplexed in the PCR tube, were developed (Table 1). In each multiplexed PCR, the concentration of each primer was adjusted to have fragments of the same intensity on the gel. Multiple PCR reactions could be further multiplexed when the gel was loaded (Table 1). A total of 8 multiplexes was necessary to include all 38 SSR markers. It was found that the fragment sizes of different loci were best distinguished if there was not an exact overlap in sizes of different SSRs, even if they were of different colors, because the spectrum of each color may overlap a bit into the spectrum of the next color. This would create artifacts and introduce error into the data). We found that the dye-labeled primers are not highly stable, especially in diluted working solution, and that the fluorescent dye may be sensitive to repeated freezing and thawing. It is therefore recommended not to make more working solution than can be used in a period of a few days, and to make several aliquots of the concentrated stock solution.

GeneScan data contains all possible peaks, including contaminants, SSR stuttering, and background noise. Furthermore, because of the error in

size calling, (which usually occurs due to imperfections in the PCR reaction and electrophoresis, and which can be slightly greater than 1 base pair on the ABI prism 377), you can see a range of peaks that may be separated by less than one base pair, rather than by the expected SSR repeat unit. The Genotyper program is used to define the expected size range of the locus, the repeat unit, and the error in size calling, to convert a series of peaks into expected alleles at an SSR locus. In order to calculate the error, a series of tests are run with diverse maize lines, and the frequency of each peak is counted. These frequencies are used to create a histogram, which should show an oscillating pattern with the highest frequencies corresponding to the “true” alleles, and the distance between peaks corresponding to the repeat unit of the SSR. The peaks that deviate from the “true” sizes would be included in the category under allowable tolerance (error). We have written a program in the ICIS database (ICIS 2000) which calculates the histograms in order to aid in the definition of the categories (allele sizes and error) and this is available upon request. The ranges of tolerances calculated in this study were 0.85 bp, 1.25 bp, 1.50 bp, 2.00 bp, and 2.50 bp for the SSRs with di-, tri-, tetra-, penta- and hexa-nucleotide repeats, respectively. Any peaks occurring outside of the range of tolerance would not be scored. However, because SSRs with dinucleotide repeats have an error tolerance (0.85bp) which is smaller than the actual observed error on the sequencer (1 bp), it is very difficult to reliably assign these peaks to alleles. It is therefore suggested that only SSRs with a trinucleotide repeat or greater be used for fingerprinting studies. The error in size calling on manual gels is generally greater than on the sequencer, which would only exacerbate the problem of using dinucleotide repeats.

References

Clarke, BC, Moran LB, Appels R (1989) DNA analyses in wheat breeding. Genome 32: 334-339

Saghai Maroof, MA, Biyashev RM, Yang CP, Zhang Q, Allard RW (1994) Extraordinarily polymorphic microsatellite DNA in barley: species diversity, chromosomal locations, and population dynamics. Proc. Natl. Acad. Sci. (USA) 91:5466-5470

Senior ML, Murphy JP, Goodman MM, Stuber CW (1998) Utility of SSRs for determining genetic similarities and relationships in maize using an agarose gel system. Crop Sci. 38: 1088-1098

Sharon E M, Kresovich S, Jester CA, Hernandez CJ, Szewc-McFadden AK (1997) Application of multiplex PCR and fluorescence-based, semi-automated allele sizing technology for genotyping plant genetic resources. Crop Sci. 37:617-624

Smith JSC, Chin ECL, Shu H, Smith OS, Wall SJ, Senior ML, Mitchell SE, Kresovich S,and Ziegle J (1997) An evaluation of utility of SSR loci as molecular markers in maize (*Zea mays L*.): comparisons with data from RFLPs and pedigree. Theor Appl Genet 95: 163-173