Original Paper Open Access

Genetic relatedness and population structure within the public Argentinean collection of maize inbred lines

Sofía E Olmos[1]*, Carla Delucchi[1], Melina Ravera[1], María E Negri[1], Cecilia Mandolino[1], Guillermo H Eyhérabide[1]

[1]Instituto Nacional de Tecnología Agropecuaria EEA Pergamino. Ruta 32 Km.4,5 CP 2700. Pergamino Buenos Aires, Argentina
*Corresponding author: E-mail: olmos.sofia@inta.gob.ar

Abstract

Genetic diversity of an Argentinean public temperate inbred maize collection has not been previously assessed. This collection includes mainly locally developed orange flint germplasm and a group of temperate inbred lines introduced from the US or derived from selection of crosses to genetic stocks from other countries, providing representativeness of exotic gene pools. To establish heterotic groups and patterns for breeding purposes and to assess genetic structure and relatedness for association-mapping studies, a public panel of a 103 maize inbreds was characterized using 50 microsatellite markers and pedigree information. By means of clustering-based and model-based methods the flint germplasm collection was split into two subpopulations that were separated from the BSSS-BS13-related lines. Relatedness estimates with coancestry and kinship coefficients provided additional information in the case of structured mixed membership of some germplasm. These three main subpopulations were in agreement with prior pedigree records. Allele diversity was high and sufficient to give major, minor and specific allele profiles to characterize inbred lines. Convenience of the use of minor allele frequency for structure and relatedness assessment is also discussed. In addition, molecular characterization provided useful information to elucidate inbred ancestry origins of germplasm with unknown pedigree records and to group them into known heterotic groups to define heterotic patterns.

Keywords: microsatellites, genetic structure, relatedness, heterotic groups

Introduction

In species that exhibit heterosis, such as maize, information about combining ability with genetically divergent testers is useful to classify inbreds in heterotic groups (Eyhérabide et al, 2006; Melchinger and Gumber, 1998; Delucchi et al, 2012). Such a procedure is based on the positive association observed between grain yield and genetic divergence of the parents of a cross within certain range of diversity (Moll et al, 1965).

The Argentinean breeding program takes advantage of the strength of the Argentine Orange Flint versus the US Yellow Dent germplasm heterotic pattern (Maunder, 1992). Crosses between lines of both heterotic groups often become very highly productive cultivars. Both the US heterotic groups Reid Yellow Dent (RYD) and Lancaster Sure Crop (LSC) and the heterotic group composed of Argentinean germplasm of the Cristalino Colorado (Cateto) race have been defined as the Local Flint versus US Dent heterotic pattern (Eyhérabide et al, 2006). Dent hybrids, either developed in or introduced to Argentina, that follow the RYD versus LSC heterotic pattern exhibit quite good performance, especially in the favorable environments of the Buenos Aires province. These cultivars are more suitable for the wet milling industry. Flint hybrids are appreciated for the hardness of their endosperm, which makes them more suitable for the dry milling industry (Eyhérabide and Gonzalez, 1997) and for their greater tolerance to certain biotic and abiotic stresses.

Germplasm classification based on genetic distances is important because crosses between lines extracted from more divergent sources will probably exhibit greater levels of heterosis (Ordás, 1991). Combining estimates of genetic distance with data on agronomic performance of testcrosses may facilitate the successful use of the backcross-derived lines to increase genetic diversity and improve performance of broadly adapted cultivars (Menkir et al, 2006). Flint-García et al (2009) proved advantages of combining genetic distance estimates between parent lines with other characteristics for improved predictors of yield performance in hybrid combinations. Pedigree information provides a useful guide to place maize inbreds into groups that reflect their degree of genetic similarity (Liu et al, 2003). Thus, pedigree distance calculated for instance as Malécot coefficient of coancestry (Malécot, 1948) allows germplasm classification when pedigree information is available.

Because pedigree information for some inbreds is sometimes incomplete or inaccurate, coancestry coefficients based on molecular markers, also called kinship coefficients, can be used to identify the maize

 RECEIVED 10/29/2013

germplasm pool from which maize inbreds were derived. Kinship coefficients are based on the relative probability of identity of alleles for two homologous genes sampled in some particular way (Hardy and Vekemans, 2002). Previous studies using molecular markers have generally shown a strong correlation between molecular-marker- and pedigree-based distance measures (Bernardo et al, 1997; Smith et al, 1997; Bernardo et al, 2000; Bernardo and Kahler, 2001; Liu et al, 2003). As stated by Liu et al (2003), calculations of relatedness based upon pedigree data are dependent upon the assumptions that both parents contribute an equal number of alleles to the finally selected lines (i.e., no selection, mutation, or genetic drift) and that the pedigree data are accurate. Another assumption is that founder genotypes (genotypes for which no further pedigree information is available) are unrelated by pedigree. This also applies to the case of inbred lines derived from phenotypically selected inbred families from the same population subjected to any procedure of recurrent selection. When any of these assumptions is violated, the correlation is affected (Menkir et al, 2006).

In addition, kinship coefficients and population structure estimated from molecular markers provide useful information to control spurious associations that limit Quantitative Trait Loci (QTLs) association-mapping studies (Stich et al, 2008b). The difficulty with association mapping is that population structure can lead to highly significant associations between a marker and a phenotype, even when the marker is not physically linked to any causative locus (Pritchard, 2001). Here, population structure refers to the presence of subpopulations within the main population. Successful assessment of the genetic structure of maize by Bayesian analysis has been previously reported (Liu et al, 2003; Stich et al, 2005; Camus-Kulandaivelu et al, 2006; Vigouroux et al, 2008; Lia et al, 2009).

Model fit and power of association mapping increase with the inclusion of both population structure (Q) and relative kinship (K) within a sample (Stich et al, 2008b). A Q + K model is able to systematically account for multiple levels of relatedness among individuals. Essentially, the genetic consequence of local adaptation or diversifying selection among different maize populations is accounted for by Q in a gross manner, where relatedness among individuals within and between subpopulations is accounted for by K on a finer scale (Yu et al, 2006).

Simple Sequence Repeat (SSRs) markers are being routinely used for fingerprinting of maize lines. A large proportion of private alleles can be found among maize lines, a factor that is potentially a function of the high mutation rate in maize SSRs (Vigouroux et al, 2002). This feature of SSRs, along with their co-dominance and Mendelian inheritance (Beckmann and Soller, 1990), contributes to their considerable discriminatory power. This allowed Liu et al (2003) to uniquely fingerprint their entire set of 260 lines with as few as 10 SSRs and to Dale et al (2002) to monitor the gene flow between lines. SSRs can also be used to determine pedigrees in maize inbreds and hybrids, although the number of SSRs required to trace a pedigree is larger (e.g., 60 or more SSR loci) than that required for unique line identification, especially when closely related inbreds are considered (Berry et al, 2002).

In this article, we utilized SSRs to analyze the genetic structure and relatedness of a set of inbred lines that represent the diversity available within the current and historical local germplasm of a public Argentinean maize breeding program for association mapping and breeding purposes.

Materials and Methods

Plant material

A set of 103 inbred lines representing a sample of the most important public lines from Argentina, including some reference lines from the United States, were chosen to represent diversity available in the INTA´s (the National Institute of Agriculture Technology - Argentina) maize breeding program. These included public inbred lines mostly adapted to temperate environments. Coding numbers and pedigrees (when records were available) of the lines are listed in Table 1.

SSR genotyping

We used 50 SSR loci that were distributed almost evenly throughout the maize genome, including a set of SSRs previously selected based on their high Polymorphism information Content (PIC) values. For a given number of alleles, PIC reaches the highest value when allele frequencies are equal (Romero-Severson et al, 2001). No prior information about the genomic location of loci in coding or noncoding regions or about locus proximity to genes was used for the selection of loci.

Primer sequences are available at MaizeGDB (http://www.maizegdb.org/). DNA was extracted from 6-day-old seedlings according to Kleinhofs et al (1993). DNA bulks comprising five individual seedlings of each inbred line were analyzed.

PCR reaction mixtures contained approximately 30 ng of DNA, 250 nM each primer, 200 µM each dNTP, 1.5 mM Mg^{2+}, 0.5 unit Taq DNA polymerase (Invitrogen, Cat Num 11615-010), 1x PCR buffer and sterile double-distilled water to a final volume of 13 µl. A touchdown cycling profile (annealing temperature 65–55°C) was used and the PCR products were separated on a 6% denaturing polyacrylamide gel (8 M urea) following standard procedures. Gels were silver-stained. Alleles were identified by comparison with products of known size in the B73 inbred line. Rare alleles were defined as having a frequency lower than 5%. To assess the discriminatory power of rare alleles, a new data set was created wherein rare alleles were replaced by missing data. The two data

Table 1 - List of germplasm used, along with population of origin, pedigree and model-based predicted background.

Inbred line code	Inbred register name	Genetic background as determined by pedigree information	Origin	Model-based predicted background[§]
1	P465		Argentine landrace	P465
2	LP611	Fam P465	Recurrent selection in (P465 x D)F2	P465
3	LP662	Fam P465	(P465 x D)F2	P465
4	LP613	Fam P465	Rec Sel in (P465 x D)F2	P465
5	LP168	Fam P465	Rec Sel in (P465 x D)F2	P465
6	LP125-r	Fam LP125r	Synt Colorada Dura	Argentinean x Caribbean Derived Stocks
7	LP317	Fam LP311	Synt Hybrid L100	Mixed
8	LP311	Fam LP311	Synt Hybrid L100	Mixed- Argentinean x Caribbean Derived Stocks
9	LP116	Fam CACaribe	Comp Argentino Caribe	Argentinean x Caribbean Derived Stocks
10	LP122	Fam LP122	Comp Argentino-Caribe	Argentinean x Caribbean Derived Stocks
11	LP1032	Fam Comp I	Compuesto I	Argentinean x Caribbean Derived Stocks
12	LP199	Fam Comp II	Compuesto II	Mixed- Argentinean x Caribbean Derived Stocks
13	LP1044	Fam Comp I	Compuesto I	Argentinean x Caribbean Derived Stocks
14	LP299-2	Fam LP299-2	Synt Hybrid P	Mixed
15	LP197	Fam LP299-2	Synt Hybrid P	Mixed
16	LP223	Fam LP299-2	Synt Hybrid P	Mixed
17	LP304	Fam LP299-2	Synt Hybrid P	Mixed
18	LP2541	Fam BS13	Population BS13	BS13-BSSS
19	LP214	Fam CanArg	Cross Local Flint x Canadian Dent F2	Mixed
20	LP4703	Fam Prolif	Prolific Composite	Argentinean x Caribbean Derived Stocks
21	LP212	Fam CanArg	Cross Local Flint x Canadian Dent F2	Mixed-P465
22	LP236	Fam CanArg	Cross Local Flint x Canadian Dent F2	Mixed
23	LP122-2	Fam LP122	(L3178xL196)F2	Argentinean x Caribbean Derived Stocks
24	LP2542	Fam BS13	Population BS13	BS13-BSSS
25	LP561	Fam CACaribe	Synt R4PC	Mixed- Argentinean x Caribbean Derived Stocks
26	LP29	Fam CCP	Comp Colorado Precoz	Mixed
27	LP179	Fam Suwan	Suwan	Mixed
28	LP612	Fam P465	Rec Sel in (P465 x D)F2	P465
29	LP220	Fam CanArg	Cross Local Flint x Canadian Dent F2	Mixed-BS13-BSSS
30	LP221	Fam CanArg	Cross Local Flint x Canadian Dent F2	Mixed
31	LP605	Fam P465	[(P465 x D)Fn*x ZN6]F2	Mixed-P465
32	LP916	Fam DK	DK752xB73	Mixed
33	LP917	Fam DK	DK752xB73	Mixed-BS13-BSSS
34	LP818	Fam LP299-2	Synt Hybrid P	Mixed-BS13-BSSS
35	LP59		(L10612xB14)F2	Mixed
36	LP124	Fam CCP	Comp Colorado Precoz	Mixed-P465
37	LP438		Comp Semidentado Precoz	Mixed- Argentinean x Caribbean Derived Stocks
38	LP1996	Fam Comp II	Comp II/I	Argentinean x Caribbean Derived Stocks

Table 1 - cont.

39	LP1513	Fam Comp II	Comp II	Argentinean x Caribbean Derived Stocks
40	LP1512	Fam Comp II	Comp GSSS	Argentinean x Caribbean Derived Stocks
41	LP521	Fam LP125r	Synt Colorada Dura	Argentinean x Caribbean Derived Stocks
42	LP126	Fam LP125r	(LP125r x L196)F2	Argentinean x Caribbean Derived Stocks
43	LP453	Fam CACaribe	Comp Argentino Caribe	Argentinean x Caribbean
44	LP5708	Fam CACaribe	Comp Argentino Caribe	Mixed
45	LP1411	Fam LP122	(LP199x L3178)F2	Argentinean x Caribbean Derived Stocks
46	LP153		(A1 x L1571)F2	Argentinean x Caribbean Derived Stocks
47	LP13		Synt Colorada Dura	Argentinean x Caribbean Derived Stocks
48	LP256 r		Rec Sel in (L256 x D)F2	Argentinean x Caribbean Derived Stocks
49	LP509		Comp BSSS x Cuarentín	Mixed
50	LP562		R49022 x Hybrid M370	Mixed- Argentinean x Caribbean Derived Stocks
51	LP563	Fam DK	DK7312 x Landrace Calchaquí	Mixed- Argentinean x Caribbean Derived Stocks
52	LP579		[(5842xLP125)x(28xP1338)]F2	Argentinean x Caribbean Derived Stocks
53	LPB1		L327 (CAC)x Local White	Argentinean x Caribbean Derived Stocks
54	LP2	Caribbean lines 3 Argentine flint synthetic	Compuesto 3:3:B	Mixed
55	LP869		Synt Hybrids	Mixed
56	LPB2		Broad base white endosperm population derived from US germplasm	Mixed
57	LP3830		(B23xB87)F2	Mixed-BS13-BSSS
58	LP580		(Hybrid Titanium F4)F2	P465
59	LP915		[(N28xB73)x(N28x199)]F2	BS13-BSSS
60	CML370014		CML327 (Cimmyt) x BS132	BS13-BSSS
61	A485		(Hybrid ACA 2000)F2	Argentinean x Caribbean Derived Stocks
62	L4674		(Hybrid AX924)F2	Mixed
63	L4637		(LP561 x LP611)F2	Argentinean x Caribbean Derived Stocks
64	B98		Population BS11	Mixed
65	L1445		Rec Sel in [(LP1512xLP199)(LP453xLP58)]F2	Mixed- Argentinean x Caribbean Derived Stocks
66	B100		Developed from B85xH99 The cross was backcrossed to H99, and pedigree selection within the backcross generation used to develop B100	Argentinean x Caribbean Derived Stocks
67	ZN6		Developed from red flint populations	Argentinean x Caribbean Derived Stocks
68	L5665		(P578 x LP116) F2	P465
69	L5605		(P578 x LP116) F2	Mixed-BS13-BSSS
70	L5632(04.5481)		(P578 x LP116) F2	Mixed

Table 1 - cont.

71	LP32	Composite Colorado Precoz	Mixed-BS13-BSSS
72	LP58	Composite Dentado Precoz	Mixed
73	LP923	Hybrid 2F10 F2	Mixed
74	LP178	Suwan	Argentinean x Caribbean Derived Stocks
75	LP598=A485	Hybrid ACA 2000 F2	Argentinean x Caribbean Derived Stocks
76	LP661	(LP662 x LP611)F2	P465
77	LP918	Hybrid AX888 F2	BS13-BSSS
78	08.3326 Fam 2541	Rec selection in BS13 conducted in Argentina using LP612 as tester	Mixed-BS13-BSSS
79	(7310x7266)-1-133	Hybrid C280 F2	P465
80	2915xLP2541-A	(B73 x LP2541)F2	BS13-BSSS
81	2915xLP2541-B	(B73 x LP2541)F2	BS13-BSSS
82	2915xLP2541-C	(B73 x LP2541)F2	BS13-BSSS
83	2915xLP2541-D	(B73 x LP2541)F2	BS13-BSSS
84	AX888IT-A	Hybrid AX888IT F2	Mixed-BS13-BSSS
85	AX888IT-B	Hybrid AX888IT F2	BS13-BSSS
86	AX888IT-C	Hybrid AX888IT F2	Mixed
87	AX888IT-D	Hybrid AX888IT F2	Mixed
88	Z9801-A	Hybrid Z9801 F2	Mixed
89	Z9801-B	Hybrid Z9801 F2	Mixed
90	(LP915x3125-2)-1-10	(DK752xB73)F2	BS13-BSSS
91	(LP915x3125-2)-1-67	(DK752xB73)F2	BS13-BSSS
92	(LP562x3584)-1-39	(M370 x Flint Arg) x Flint Arg	Argentinean x Caribbean Derived Stocks
93	(LP562x3584)-1-53	(M370 x Flint Arg) x Flint Arg	Argentinean x Caribbean Derived Stocks
94	(R4930x3125-2)-1-9	(DK752xB73)F2	Mixed
95	(R4930x3125-2)-1-60	(DK752xB73)F2	Mixed-BS13-BSSS
96	(7310x7266)-1-56	Hybrid C280 F2	P465
97	(7310x7266)-1-84	Hybrid C280 F2	P465
98	(7310x7266)-1-91	Hybrid C280 F2	P465
99	08.3525	High Oleic Acid Population, derived from [(LP1512xLP199)(LP453xLP58)]F2	Argentinean x Caribbean Derived Stocks
100	08.3556	Low Saturated Fatty Acid Population derived from [(LP1512xLP199)(LP453xLP58)]F2	Argentinean x Caribbean Derived Stocks
101	08.3538	High Oleic Acid Population derived from [(LP1512xLP199)(LP453xLP58)]F2	Argentinean x Caribbean Derived Stocks
102	08.3590	High Oleic Acid Population derived from [(LP1512xLP199)(LP453xLP58)]F2	Argentinean x Caribbean Derived Stocks
103	B73	BSSS	BS13-BSSS

[§]Predicted genetic background based on the whole molecular data set STRUCTURE analysis at k=3 allowed the differentiation among P465, Argentinean x Caribbean Derived Stocks, BS13-BSSS, and mixed germplasms. Mixed inbreds with ≥ 60% membership from one subpopulation were also labeled. Mixed inbred with ≥ 0.60 membership were called with the corresponding subpopulation membership.

sets were analyzed separately for each genetic and statistic clustering approach and results were compared.

We also examined the precision of our estimates of band sizes by comparing the estimates of SSR alleles of the B73 bulk local source line and the corresponding SSR amplicon size predicted in the AGI's B73 RefGen_v2 reference sequence, for all cases in which the SSR loci were physically mapped.

Allelic richness, diversity and genotype display

The PowerMarker software (Liu, 2002) was used to calculate major allele frequency, residual heterozygosity (observed heterozygosity) and average gene diversity indices. Graphical genotyping and visualization of similarities was achieved by the FlapJack program (Milne et al, 2010) which allowed the genotype profiling by chromosomes and the assessment of inbred identification.

Analysis of relatedness

The relative kinship (K) matrix was calculated on the basis of the 50 SSR loci, using the method of Loiselle et al (1995) implemented in SPAGeDI (Hardy and Vekemans, 2002). This method is adapted to heterozygote diploid individuals in the case of multiallele and multilocus data sets. Negative values between individuals were set to 0, as this indicates that they are less related than random individuals. Essentially, the degree of genetic covariance caused by polygenic effects was defined as 0 for a pair of individuals that are not related and as positive for a pair of individuals that are related. This threshold is similar to the pedigree-based coancestry matrix in which individuals with unknown relationship are set to 0 (Yu et al, 2006).

Additionally, coancestry coefficients were previously obtained in a set of 60 inbred lines with known pedigree information by using the Malécot (1948) analysis (Eyhérabide, personal communication). Inbred line pairs with unknown pedigree were considered unrelated and assigned a coancestry coefficient equal to 0.

To compare the genetic relationships obtained from the kinship coefficient matrices calculated from molecular data with those obtained from pedigree information, the cophenetic correlation coefficient was calculated and a Mantel test between matrices was carried out with Infostat software (Di Rienzo et al, 2010).

Analysis of the genetic structure

To compare two approaches to estimate genetic structure of our maize inbred lines, a similarity-based and a model-based approach were applied.

Firstly, for the similarity-based clustering method, every microsatellite allele was scored for the presence (1) or absence (0) in each of the 103 Operational Taxonomic Unit (OTU's). The resulting OTU × OTU matrix was then the input to calculate the Simple Matching coefficient in order to construct a similarity matrix and build a phenogram by the unweighted pair-group method using arithmetic averages (UPGMA). The distortion of the phenogram was measured by computing the cophenetic correlation coefficient (r). This computational work was done using the Info-Gen software package (Balzarini and Di Rienzo, 2012).

Secondly, lines were subdivided into genetic clusters using a Bayesian model-based approach implemented with the software package STRUCTURE 2.3.3 (Pritchard et al, 2000a). Given a value for the number of subpopulations (clusters), this method assigns lines from the entire sample to clusters in such a way that Hardy-Weinberg disequilibrium and linkage disequilibrium (LD) were minimized. Two independent runs of STRUCTURE were performed when setting the number of subpopulation (K) simulations from 1 to 12. No prior information regarding the pedigree origin of the inbred line was used to infer subpopulations. As recommended by Pritchard et al (2010), the admixture model was used as a starting point for data analysis. Under this model, each individual draws some fraction of its genome from each of the K subpopulations and conditional on the ancestry vector, q(i), the origin of each allele is independent. That is, this model assumes that all markers are unlinked and provides independent information on an individual's ancestry. For each run, burn-in time and replication number were both set to 1,000,000. The program CLUMPP (Jakobsson and Rosenberg, 2007) was utilized to line up the cluster labels across the two different runs prior to data plotting. The average output matrix from CLUMPP was plotted by using the STRUCTURE program in which the STRUCTURE input file of a certain simulation run was replaced by the corresponding CLUMPP output matrix file fitted with the STRUCTURE column format.

Besides the known pedigree records, graphical results, maximum likelihood and the rate of change in the log probability of data between successive K values (ΔK) were used to infer the correct value of K, which is usually the one with highest posterior probability. To assign inbred lines into clusters, lines with membership probabilities ≥0.80 were considered to belong to discrete clusters; whereas inbred lines with membership probabilities <0.80 were assigned to the "mixed" subpopulation.

Nei´s genetic distances (Nei, 1972) between clusters resulting after selecting the right K value were then calculated with the software Info-Gen (Balzarini and Di Rienzo, 2012). Correlation coefficients were calculated by Infostat (Di Rienzo et al, 2010).

Results

SSR genotyping

SSR target sequence features and amplicon repeat motifs were studied (Table 2). Fourteen out of the 50 SSR loci selected were found to be contained in known genes. Thirty-eight out 50 SSRs mapped

Table 2 - Microsatellite and sequence target site features.

Locus	Chr Bin	Sequence motif	Allele number	AGI's B73 RefGen_v2 sequence region position (http://maizesequence.org)	Sequence features	Amplicon coding feature	Gene target by the SSR loci
phi056	1.01	CCG	6	1:2037854:2038473:1	GRMZM2G164696	translated	
bnlg1429	1.02	AG	12	gap			
bnlg439	1.03	AG	8	1:43691148:43691776:1	...		
umc2025	1.05	AGCT	5	1:91247418:91248041:1	GRMZM2G105863	translated	
bnlg504	1.11	...	5	gap			
umc2246	2.00	CCT	8	2:961241:961866:1	GRMZM2G165535	translated	
phi96100	2.01	ACCT	6	2:2816572:2817190:-1	GRMZM2G043932	translated	
phi083	2.04	AGCT	6	2:40588030:40588654:1	GRMZM2G102356	translated	prp2 pathogenesis-related protein2
dupssr21	2.05	AG	11	2:63373371:63373994:-1	GRMZM2G087059	translated	
umc1749	2.06	GA	10	2:181683132:181683755:1	GRMZM2G336456	translated	
phi127	2.08	AGAC	3	2:189737223:189737848:1	...		
phi104127	3.01	ACCG	4	3:3478660:3479281:-1	GRMZM2G589470	translated	
bnlg420	3.05	...	12	gap			
phi053	3.05	ATAC	5	3:126236290:126236908:1	...flanked by MIPs and transposons		
umc2050	3.07	CGC	5	3:195963252:195963875:1	GRMZM2G105863	translated	
phi047	3.09	ATC	4	3:223681409:223682032:1	GRMZM2G071977	untranslated	
phi072	4.01	AAAC	6	4:1075707:1076334:1	GRMZM2G164229	translated	mtl1 metallothionein1
nc004	4.03	AG	5	4:13398344:13398963:-1	GRMZM2G098346	translated	adh2 alcohol dehydrogenase2
bnlg1217	4.05	AG	13	4:41850169:41850788:1	...Transposons		
umc1299	4.06	AAG	4	4:159667635:159668258:1	GRMZM2G126505	translated	abh2 abscisic acid 8'-hydroxylase2
bnlg1137	4.06	AG	11	4:169740006:169740628:1	...MIPs		
phi019	4.11	ATT	6	4:240106649:240107274:1	GRMZM2G079348	translated	cat3 catalase3
umc1240	5.00	TTG	2	5:533858:534482:1	AC220970.4_FG002, flanked by transposons	translated	
phi113	5.03	GTCT	5	5:12290892-12291009:1	GRMZM2G102926	translated	ole3 oleosin3
umc1752	5.06	CGG	3	5:195453047:195453671:-1	GRMZM2G481755, flanked by transposons	translated	
phi128	5.07	AAGCG	3	5:208741547:208742175:1	GRMZM5G801076	untranslated	
umc1792	5.08	CGG	6	5:212613385:212614008:1	GRMZM5G852886 with MIPs, flanked by transposons	translated	
phi075	6.00	CT	4	6:1339993:1340621:1	GRMZM2G122337, flanked by transposons	translated	fdx1 ferredoxin1
umc1887	6.03	CGA	4	6:102720521:102721144:1	...		
umc1979	6.04	GCG	4	6:106067155:106067775:1	GRMZM2G148460	translated	
umc2318	6.05	GAC	3	6:123777161:123777784:1	GRMZM2G064096	translated	
umc2059	6.08	CAG	5	6:167981902:167982525:-1	GRMZM2G456570	translated	
bnlg1367	7.00	AG	10	7:2133458:2134076:-1	AC205122.4_FG003, flanked by MIPs and transposons	untranslated	
umc1583	7.00	GAA	2	7:59865977:59866598:-1	...flanked by MIPs and transposons	untranslated	
phi057	7.01	GCC	3	7:10795406:10796026:1	GRMZM2G015534, flanked by MIPs and transposons	translated	o2 opaque endosperm2
phi034	7.02	CCT	5	gap			cyp6 cytochrome P450
bnlg1070	7.03	AG	9	7:133139523:133140142:1	...		
umc2190	7.06	CCT	5	7:173817722:173818345:-1	AC155434.2_FG005	translated	
phi115	8.03	AT/ATAC	3	8:100396776:100397078:-1	GRMZM2G126010	translated	act1 actin1
phi014	8.04	GGC	3	8:109163425:109164050:1	GRMZM2G063536, flanked by MIPs and transposons	translated	rip1 ribosome-inactivating protein1
phi080	8.08	AGGAG	5	8:173117572:173118193:1	GRMZM2G116273	translated	gst1 glutathione-S-transferase1
bnlg1131	8.09	AG	12	8:175698379:175699000:1	GRMZM2G111354	untranslated	
umc2093	9.01	ACAT	3	9:11749306:11749933:1	GRMZM2G177098	untranslated	stc1 sesquiterpene cyclase1
phi065	9.03	CACTT	5	9:61300782:61301408:-1	GRMZM2G083841	untranslated	pep1 phosphoenol-pyruvate carboxylase1
umc1078	9.05	GT	11	9:128473364:128473987:-1	GRMZM2G323479	untranslated	
bnlg1270	9.06	AG	17	9:129511397:129512022:-1	....		
umc1380	10.0	CTG	5	10:2255441:2256063:-1	GRMZM2G138659	translated	
phi041	10.0	AGCC	3	10:2646451:2647072:-1	... next to GRMZM2G172596	untranslated	
umc1938	10.03	TGC	2	10:63816004:63816627:-1	GRMZM2G478370	translated	
phi084	10.04	GAA	3	10: 87270300:87270925:-1	GRMZM2G015605	translated	

in silico to predicted genes, with GRMZM2G and AC prefix, at the www.maizeseqence.org data base. Twenty-nine of these were located in exons (translated sequences) and nine in introns (untranslated sequences). Eight out of 50 SSRs were assigned to non-coding sequences. The final four SSRs matched gap genomic regions of the AGPv2, 2009-03-20 assembly version, which means these sequences were not found in the B73 RefGen_v2 sequence. Thus, most of the SSR loci used in this study (76%) were contained in gene sequences. In addition, two SSRs were found in transposable elements sequences and nine SSRs were mapped to sequences surrounding transposable elements. SSR repeat motifs varied from two to five nucleotides. Among the 13 dinucleotide SSR motifs, AG was the most common repeat. No trend was

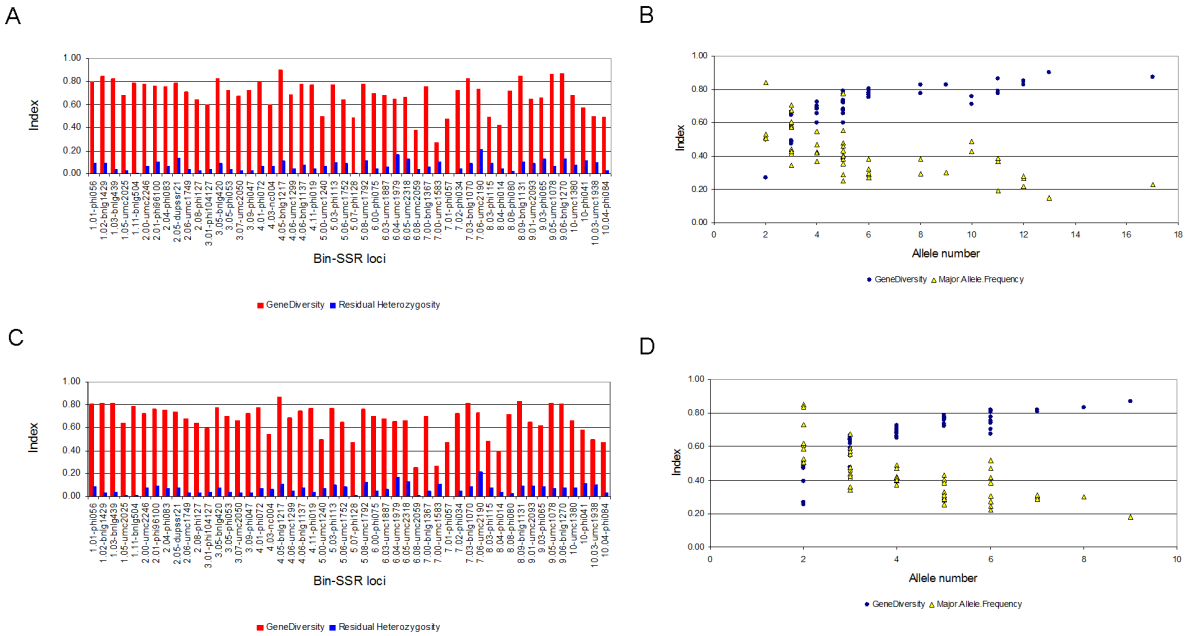

Figure 1 - Variation in diversity statistics within and among chromosomes and allele data sets. A) Diversity indexes according to chromosomes and B) diversity indexes according to allele number per locus, respectively, for the entire data set. C) Diversity indexes according to chromosomes and D) Diversity indexes variation according to allele number per locus, respectively, for data set excluding ≤ 5 % allele frequency.

found between the SSR locus sequence features and the hypervariability of loci. However, the correlation between the number of sequence motif repetitions and the variation in allele number per locus was moderate and negative (r=-0.49, P< 0.001).

Allelic richness and diversity

Among the inbred panel, all 50 microsatellite loci were found to be polymorphic. A total of 300 alleles were detected (Supplementary Table1). The number of alleles per locus was variable, ranging from 2 (umc1240, umc1583, umc1938) to 17 (bnlg1270), with an average of 6 alleles per locus (Supplementary Table1).

Within the sample of one hundred three inbred lines and by taking into account the whole molecular data assayed, we found that 76% of the 300 alleles occurred at a frequency of 0.25 or less, predicting high gene diversity, often referred to as expected heterozygosity. Average gene diversity and residual heterozygosity were 0.69 and 0.07, respectively. The average residual heterozygosity was low, as would expected for inbred materials, however, all residual heterozygosity was detected in all 50 SSR loci. Among these loci, bnlg504 detected the minimum (0.01) and umc2190 the maximum residual heterozygosity (0.21). In addition, residual heterozygosity was not correlated with the allele number per locus. Bulk heterozygosity indicates either the presence of individual heterozygous plants or the presence of two homozygote alleles fixed among individual plants composing the bulk. In either case, a bulk was considered to be composed of more than one allele when the minor allele represented ≥ 25 % of the total genotypes obtained per locus. Thus, only 15 out of 103 inbred lines lacked residual heterozygosity and were completely stabilized (#11,18, 20, 21, 25, 37, 56, 59, 60, 66, 82, 85, 88, 89, and 103, Supplementary Table 1).

When analyzing diversity statistics according to chromosomes, we found variation within and among chromosomes. For instance, greater and less variable average gene diversity indexes within chromosomes were obtained in chromosomes 1, 2, 3, 4, and 9 (Figure 1A). This indicates some chromosomes might provide differential genotype patterns for inbred line fingerprinting.

Gene diversity increased when the number of alleles per locus increased (Figure 1B) and the number of major alleles decreased (Figure 1B). Highly informative SSRs, for instance locus bnlg1217 from chromosome 4, can be utilized to fingerprint a bigger population and the less informative loci, such as the biallelic locus umc1583 from chromosome 7, can be removed from analysis.

Within the 103 inbred lines, 28% of the alleles (83 out of 300) occurred at a frequency of 0.05 or less (minor alleles are indicated in red letters in Supplementary Table 1). Reanalysis of data excluding alleles at ≤0.05 frequency resulted in an average number of 4.4 alleles per locus (Figure 1B). Average gene diversity and residual heterozygosity remained similar to those found in the entire data set (0.67 and 0.07, respectively). Variation in gene diversity and major allele frequency values according to chromosome

and variation in allele number variation per locus also showed trends similar to those observed in the entire data set (Figure 1 C, D).

Microsatellites on chromosome 1 showed high gene diversity and the five SSRs loci mapped to this chromosome provided unique genotypes for 99 out of the 103 inbred lines as visualized with FlapJack (Milne et al, 2010). The complete locus set of chromosomes 2, 4, and 7 also allowed inbred fingerprinting greater than 90 %. Chromosomes 8 and 10 were less discriminatory and showed a high number of genotypes shared among inbreds.

In addition, 28 of the 83 rare alleles were exclusive to 20 inbreds, although certain lines were much better discriminated. Thus, inbreds named 9, 43 and 56 had three unique alleles, inbreds 38 and 39 had two alleles, and the rest of the 20 inbreds (3, 10, 11, 13, 31, 41, 57, 60, 63, 64, 66, 67, 89, 94, and 98) carried only one specific allele.

Allele detection based on acrylamide gels and silver staining resolved size fragments differing by 1 bp (Supplementary Table 1). Comparison of our visual allele size estimation based on the amplicon fragment obtained from our local B73 source line (inbred line named 103) with the corresponding SSR amplicon size predicted on the AGI's B73 RefGen_v2 reference sequence revealed differences in length that varied from 1 to 5 bp. Ten out of 44 comparisons resulted in exact size matches, while 14 gave differences of 1 bp. The remaining 20 SSRs loci differed in size primarily by 2 and 3 bp. However, the SSRs umc1749 and bnlg1367 showed differences in length that varied from 4 to 5 bp with a smaller size than expected based on the B73 reference sequence.

Analysis of relatedness

Fifty-eight percent of the pairwise kinship coefficients obtained from the entire data set were zero and 40.0% fell in the 0.25-0.5 range (Supplementary Table 2). When minor alleles were excluded similar results were obtained (Supplementary Table 3). Three pairs of lines (32-33, 62-75, 80-81) consistently showed (Supplementary Figure 1) higher kinship coefficients (0.75-1). In addition, both data sets had similar counts fall within the kinship coefficient ranges (Table 3).

The Malécot coefficient of coancestry matrix (Supplementary Table 4) obtained by using pedigree information resulted in 96.9% of pairwise comparisons equal to zero. The remaining coefficients ranged from 0.03 to 0.5. Only nine inbred pairs shared the highest possible coancestry score of 0.5 (1-2, 1-3, 1-4, 1-5, 1-28, 6-42, 10-23, 10-45, and 13-45).

The cophenetic correlation between matrices obtained from the entire allele data set and the data set excluding minor alleles was high (r=0.99, P < 0.001) (Supplementary Table 5). However, cophenetic correlations between Malécot and Loiselle kinship coefficient matrices obtained from these data set were low but highly significant (r=0.36 and r=0.37, P=0.001

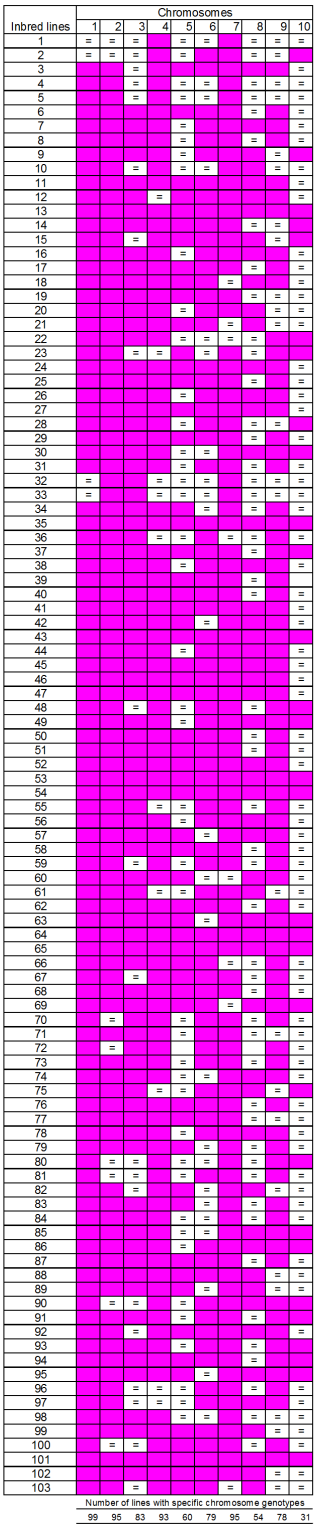

Figure 2 - Genotype profiling by chromosomes that allowed the assessment of inbred identification using the FlapJack program. Filled purple squares indicates that a line can be uniquely fingerprinted using the marker arrangements for a given chromosome. Filled equal symbol square indicates that a line share the same genotype for a given chromosome.

Table 3 - Kinship coefficients frequency range distribution for pairwise relationships between inbred lines.

	Entire data set		Data set excluding ≤ 5% allele frequency	
	Cell count	Percentage (%)	Cell count	Percentage (%)
≤ 0	3044	57.9	3012	57.3
> 0 ≤ 0.25	2103	40.0	2120	40.4
> 0.25 ≤ 0.5	89	1.7	102	1.9
> 0.5 ≤ 0.75	14	0.3	16	0.3
> 0.75 ≤ 1	3	0.1	3	0.1

respectively) (Supplementary Table 5). Because the pedigree-based matrix did not include all inbred lines, only partial comparisons between known pedigree lines could be effectively obtained. Pairwise comparisons between P465 (1) and P465-derived lines (2, 3, 4, 5, and 28) showed coefficients of coancestry of 0.5. Pairwise kinship coefficients between the inbred line pairs 1-2, 1-4, and 1-28 were in the range of 0.5-0.75 whereas the 1-3 and 1-5 pairs had coefficients in the 0.25-0.50 range. Coancestry coefficients between inbred line 6 (LP 125-r) and its derived lines 41 (LP521) and 42 (LP126) were 0.1 and 0.5, respectively. Similar results were obtained for these pairwise comparisons using both data sets. Thus, for the 6-42 and 6-41 pairs, kinship coefficients were 0.1 and 0.37, respectively. Other related inbred pairs such as 10-23, 10-45 and 12-45 resulted in pairwise coancestry coefficients of 0.5 and both data sets provided positive but low kinship coefficients.

Analysis of genetic relationships

Genetic relationships were revealed by cluster analysis of SSR data. The clustering obtained was then compared with the known pedigree of the inbred line panel (Table 1) to verify or predict phylogenetic relationships.

Simple matching coefficients varied from 0.73 to 0.98 for both data sets, with all alleles (Supplementary Figure 2) and without ≤ 5% allele frequency (Supplementary Figure 3). Whole data set analysis grouped P465 and P465-derived lines and related inbreds (2, 3, 4, 5, 28, and 76) with a group of local inbreds (31, 73, 36, 55, 49, 51, 74, 56, 27, and 44), which were differentiated from the rest. Minor allele exclusion split inbred lines 27 and 44 from the cluster mentioned above and grouped both at a lower level. Both data sets showed that line B73 (103) was consistently the closest to inbred CML370014 (60).

The model-based assignment method implemented in the program STRUCTURE was used for the two data sets with the parameter k ranging from 2 to 12. The graphical outputs obtained from both data sets are presented in Supplementary Figures 4 and 5, respectively. Using a burnin and run length of 1,000,000 consistent results were obtained between replicate runs. The observation of the Q matrices across repetitions showed no substantial label switching among subpopulations, consequently multimodality was not observed and reproducibility was verified. The CLUMPP program was mainly used to average membership coefficients between runs and preparing an average Q matrix for the graphical layout. According to Pritchard et al, (2009) reproducibility of results between repetitions was also verified by comparing likelihoods (Figure 3A) and likelihood variance (Figure 3B) estimations across runs of a given k. Our results showed similar Ln P(D) and Var[LnP(D)] estimations across run repetitions.

Evaluation of the STRUCTURE analysis simulation summary statistics showed a constant increase in ln(X/K) when k parameters increased (Figure 3A). However, the rate of change in the log probability between successive k values (Δk) was greater for the k=2 and k=3 analyses (Figure 3B). Based on this result and considering the biological information from the genetic background of the inbreds mentioned above, the k=3 parameter was chosen as the value best capturing the major structure of the data.

Analysis of the full data set (Supplementary Figure 4) showed that at k=2, 84% of the inbred lines were discretely assigned to one subpopulation. The two clusters clearly separated local derived germplasms from the BS13-BSSS derived inbred lines.

At k=3 (Supplementary Figure 4), we obtained three discrete subpopulations and a mixed cluster. We named the three discrete subpopulations P465, Argentinean x Caribbean Derived Stocks and BS13-BSSS (here in after P465, ACDS, and BSSS, respectively). Inbreds with and without known pedigree records were clustered into these four groups (Table 1) providing valuable information for characterizing these lines. We additionally lowered the membership assignment criterion from ≥0.80 to ≥ 0.60 to reduce the number of lines in the mixed group. Some mixed inbred with ≥ 0.60 membership were then called with the corresponding subpopulation membership.

Inbred lines derived from commercial hybrids or their crosses were mostly assigned to the mixed cluster. For inbreds in the mixed group, the kinship coefficients provide additional information to clarify ancestry.

In addition, at k=3 parameter the Argentinean Orange Flint germplasm was separated into clusters that we defined as the P465 and ACDS representative subpopulations. For instance, local line 67 (ZN6), which was developed from red flint populations and released in 1959, shared membership with other ACDS backgrounds and shared almost no descendent relationship with the line 1 (P465) family based on kinship coefficients. Moreover, local line 31, which resulted from a cross between P465 and line 67 (ZN6), showed different genetic relationship by means of STRUCTURE and kinship coefficients: at k=3 using the full set of alleles inbred 31 was clustered, but not discretely, into the P465 subpopulation. Whereas kinship coefficients (Supplementary Table 2) showed closer relationships of inbred 31 to 73 (L882), 36 (LP124) and then to 67 (ZN6), the two former belonging to the P465 subpopulation and the

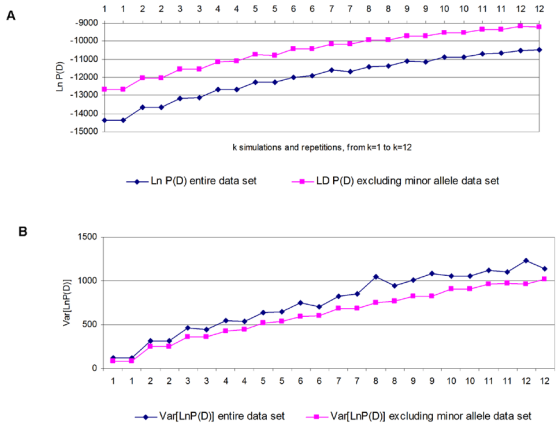

Figure 3 - STRUCTURE simulation summary statistics. A) Ln(X/K) variation according to the k parameters. B) Rate of change in the log probability (Δk) between successive k values. Two independent run repetitions for each simulated k are shown.

latter to the ACDS subpopulation.

Across subsequent values of k (from k=1 to k=12), two of the three discrete subpopulations obtained at k=3 remained conserved. One of these two subpopulations included the Argentinean P465-derived lines (1, 2, 4, 5, and 28), whereas the other cluster comprised the BS13-BSSS composite-derived lines from Iowa State University, USA, and local inbred lines derived from planned crosses and recurrent selection involving that genetic background (18, 60, 77, 80, 81, 82, 83, 84, 85, and 103=B73). The remainder of lines were admixed between these two subpopulations that were sequentially assigned to a specific cluster as the k parameter increased. However, increasing k from k=11 to k=12 did not improve clustering assignment because the k=12 analysis only differentiated 11 fully discrete subpopulations.

Different assignment of inbred lines was obtained using the full data set and the data set excluding minor alleles (Supplementary Figure 5). For instance, after minor allele exclusion, a more homogenous grouping of the Argentinean P465-derived lines was obtained (Supplementary Figures 4 and 5 at k=3). Additionally, inbreds 58, 68, 79, 96, 97 and 98 were removed from the P465-derived families subpopulation and moved to the BSSS subpopulation. Minor allele exclusion also resulted in inbred membership switches in subpopulations of inbreds that completely lacked minor alleles, for example inbreds 58 and 68. In addition, minor allele exclusion reduced membership within subpopulations but not switching among subpopulations of certain inbred lines that had unique minor alleles (e.g., 9, 43, 13, 63, and 66). The exception was inbred 56 which had three unique alleles but did not separate from its subpopulation membership.

Further genetic relationships among subpopulations defined at k=3 were analyzed based on Nei´s 1972 genetic distance (Supplementary Table 6). As a result of minor allele exclusion, the genetic distance between the Argentinean P465-derived subpopulation and the other subpopulations increased.

Similarity-based and model-based clustering results were compared. When analyzing the entire data set, a simple matching-based phenogram discriminated four main clusters in the inbred population and a small cluster composed of three lines (38, 43 and 86) that joint with the others approximately at a similarity coefficient of 0.74. For this data set, STRUCTURE results at k=4 were compared (Supplementary Figure 4). At k=4, four discrete subpopulations were obtained, three of them were those previously identified at k=3. The fourth subpopulation included lines that were mixed at k=3. The clustering by the similarity-based phenogram (Supplementary Figure 2) clustered the P465, ACDS and BSSS discrete subpopulations identified at k=3. Consequently, the simple matching-based phenogram depicted genetic relationships closer to those obtained in the STRUCTURE k=3 analysis.

Minor allele exclusion resulted in a simple matching-based phenogram with six main clusters that were grouped at a similarity of 0.74. The Argentinean P465-derived and the B73-related lines remained separated into two distinct clusters. For instance, inbred line 86, which showed a high level of heterozygosity but carried only one minor allele (and was classified as a mixed inbred in the STRUCTURE k=3 analysis of both data sets), was clustered by this phenogram alongside the Argentinean P465-derived lines. Inspection of the inbred 86 in Supplementary Table 1 revealed that the residual heterozygosity was higher, particularly on chromosomes 2, 3, and 10. Thus, it seems that the higher residual heterozygosity along with a reduced total allele number after minor allele exclusion might have biased the genetic similarity of inbred 86 and P465-derived lines in the Simple Matching Similarity analysis. In addition, the k=3 analysis with minor allele exclusion (Supplementary Figure 5) showed increased membership of inbred 31 in the ACDS subpopulation that includes its parental inbred 67. This may have resulted from narrowing of the P465 subpopulation membership. In contrast, kinship relationships were not affected after minor allele exclusion.

Discussion

A set of 103 inbred lines was selected for this study. This set included three US dent lines (B73, B98, and B100) and 100 inbred lines developed by INTA from different sources. Sources covered a wide range of synthetics, composites, and planned crosses. However, whenever sources were commercial hybrids, the exact parental genetic background from 38 out of the 100 INTA-derived inbreds remained unavailable but were estimated by the 50 SSR loci assayed and the statistical approaches implemented to infer population genetic relationships.

SSRs had a very high discrimination because they showed multiple allele variation in our diverse inbred collection. The extent of this variation depended on the molecular feature of the SSR loci and chromosomal location.

The average amount of genetic diversity revealed by SSRs compared to other maize studies showed high values. The average allele number per locus of 6 and the average gene diversity of 0.64 were similar to those found by Camus-Kulandaivelu et al, (2006), who conducted genetic characterization of a much more diverse collection of maize inbred lines with 55 SSRs. The low allele frequency and the consequent high average gene diversity observed, has also been reported in several studies of SSR characterization in maize (Chin et al, 1996; Senior et al, 1998; Stich et al, 2005).

Information deposited at the maize genomic data bases was helpful to interpret SSRs readability and to conduct allele size estimations. Also, the developed and selected for high-throughput genotyping SSRs (Register et al, 2001) provided useful characteristics (such as robustness, informative character, easy scorability etc). The MaizeGDB data base contains SSR primer sequences which allowed retrieving loci coordinates and features from the B73 sequence projects at the maizesequence.org data base.

Reports indicate that dinucleotide repeats are the most frequent and polymorphic classes of microsatellites among plant species but often produce stutter bands, which differ by 2-bp increases, hampering precise band size estimation. The inclusion of dinucleotide SSRs with high mutation rates (Vigouroux et al, 2002) in our study might have resulted in an upward bias of allele richness and diversity compared to this former study. Our randomly selected SSR dinucleotide repeats had a prevalence of AG/CT repeats as found by Chin et al (1996).

SSR hypervariability of dinucleotide SSRs is attributed to the slippage during DNA replication (Levinson and Gutman, 1987). Since this process is a neutral process, SSRs are expected to be randomly distributed throughout the euchromatic portion of the genomes of species, including maize. We found that almost half of the selected dinucleotide SSRs (6 out 13) were from coding regions mostly showing high allelic variation. Holton (2001) stated that stutter bands would not complicate scoring of well separated fragments and that selection of trinucleotide higher-order repeats for mapping purposes eliminates the problem. Stuttering, however, was also present when we employed trinucleotide SSRs. The inclusion of a B73 bulk in our study also helped band size estimation through comparison to the expected B73 amplicon size from the maizesequence.org data base.

Feature sequences of mostly mapped SSRs were possible by searching SSR positions at the B73 reference sequences. A few loci were located in gap regions and the loci characterization was not possible. We also found that some loci mapped transposable elements. In plant genomes such as that of barley, SSRs are often found in proximity to long terminal repeats of retrotransposons (Ramsay et al, 1999). Avoiding flanking sequences corresponding to known repetitive DNA has become a routine procedure during the development of SSR markers for mammalian genomes (Steen et al, 1999) because positioning PCR primers in repetitive regions generates spurious or nonspecific products. In rice, association of (AT)n dinucleotide repeats with dispersed repetitive elements seems to explain the poor amplification of these repeats, as primers from their flanking sequences recognize many targets and do not amplify cleanly from a unique site (Temnykh et al, 2001).

In our experiments, the two SSRs that mapped to transposable elements had a difficult pattern to score but were finally readable. The detection in the INTA maize panel of 13 and 11 allele variations for these (bnlg1217 and bnlg1137, respectively), may indicate that high numbers of mutation and recombination events during the inbred line selection process could have occurred. However, one of the two primers of these markers colocalized to a segment of the transposable element sequence. Thus, the high allele variations observed may also be the consequence of non-specific priming. BLAST, the alignment program that determines sequence identity between the SSR primer sequences and B73 genome sequences, usually displayed multiple aligned genomic regions. However, highest scores retrieved by BLAST correspond to alignments that include both flanking primer sequences. In all cases we only described SSR sequence features of the highest score alignment retrieved which contained primer sequences on both sides.

Residual heterozygosity was present in 85% of the INTA inbred panel, indicating that the lines are not completely stabilized. As stated by Hallauer and Miranda (1988), the effect of inbreeding in maize usually produces undesirable phenotypes and inbreeding depression. It is necessary to maximize additive and non-additive effects for favorable alleles to produce uniform and superior hybrids. Complete or nearly complete homozygosity of parental inbred lines allow production with genetic fidelity of elite hybrids. Although some pedigree-selection against unfavorable traits was usually performed, some small degree of heterozygosity remains after several inbreeding generations. Our results showed that residual heterozygosity does not contribute much to the presence of minor or unique alleles. Novel alleles might arise either from incidental pollination with closely related sister lines or de novo mutations in the SSR alleles that subsequently become fixed by genetic drift (Romero-Severson et al, 2001). The utility of minor and specific alleles for line identification needs to be analyzed and verified by testing different seed sources of the same genotype to check for rare alleles fixed

by genetic drift or artificial selection. Several reports have filtered minor alleles from the data set to reduce spurious linkage disequilibrium (Rostoks et al, 2006) and false positive associations between marker and phenotypes (Kang et al, 2008; Brachi et al, 2010). However, as stated by Gorlov et al (2008), excluding low minor allele frequency Single Nucleotide Polymorphisms (SNPs) in association studies may hamper the ability to detect rare human disease-causing polymorphisms. We found twenty maize inbred lines carrying specific alleles. The stability of such alleles through generations of seed production and its association with specific phenotypes needs to be further studied.

Results obtained from the inbred panel show the effect of breeding decisions and selection on the population structure during the last half century. The development of new second cycle lines from breeding populations obtained by crossing existing Argentinean inbred lines and BSSS-related genetic stocks has led to a subpopulation of highly related families of inbred lines that were discriminated as a discrete subpopulation when running the STRUCTURE. Thus, local inbred lines related to BSSS or BS13 clustered together, whereas B73 became the representative inbred within Stiff Stalk materials in our population. From the local-derived lines, two distinct subpopulations, the P465 subpopulation and the ACDS subpopulation, were separated. Origin and genetic background of Argentinean flint modern germplasm is still a controversial subject (Luna et al, 1964; Ferrer, 2012). One hypothesis claims that it derived from the introduction of adapted flint gene pools by immigrants from Italy. In such a case, genetic background would have relationship with germplasms from Central America and Caribbean, previously introduced in Italy. A second hypothesis asserts that modern flint germplasm of Argentina is derived from indigenous populations of the Pampean Plains and neighbor countries. Considering the greater level of genetic variability and adaptability of local landraces compared with Italian varieties introduced by immigrants, Luna et al (1964) proposed the latter hypothesis seems most likely. In addition, Camus-Kulandaivelu et al (2006) who studied local Argentinean inbred line ZN6 found that this line is admixted between Andean and Italian Flint groups. This finding might suggest a linkage between both hypotheses. An intriguing finding is that the US line B100 clustered in the ACDS subpopulation. This may be due to the heterogeneous origin of this group of lines. By reviewing the public Argentinean maize breeding history, it is known that, at least in the public sector, the breeding strategy generalized in the 1950s through late 1980s was primarily based on developing broad-base composites or pools followed by population improvement methods such as recurrent selection, with minor attention paid to development of heterotic patterns. As an example, genetic sources of the breeding population named Composite II (Table 1) include Argentinean landraces, Caribbean germplasm and US Corn Belt dent germplasm. Lines derived from Composite II clustered in the ACDS subpopulation. Since no detailed documentation is available to us, we speculate that the relationship between B100 and local lines from the ACDS cluster could result from common US dent germplasm incorporated into Composite II.

It is expected that these three subpopulations would serve as sources of different alleles and desirable phenotypes for planning breeding crosses to exploit heterotic patterns between the US Yellow Dents and our local germplasm. In this study, we generated information that allowed clustering of some particular inbred lines into the three main subpopulations mentioned above, which is in agreement with the definition of heterotic patterns based on agronomic traits, of the Argentinean flints into the A and B composite groups proposed by Delucchi et al (2012). In cases where genotypes were classified with mixed membership, additional approaches to clarify genetic relationships, such as kinship coefficients between inbred pairs, will help in the prediction of heterotic patterns.

In the present study, we compared coancestry coefficients based on pedigree with those obtained from molecular data-based kinship coefficients. Values between inbred coefficient pairs were positive and in the same order of magnitude. However, as shown by Menkir et al (2006), fixation of a high proportion of a particular donor´s alleles through deliberate selection on favorable disease resistance traits during line conversion can cause the proportion of the genome from that donor parent retained in backcross-derived lines to be significantly higher than expected. By consequence we expect coancestry coefficients might not accurately reflect the true genome contribution of parents and relatedness after line conversion.

Among several types of kinship coefficients listed by Hardy and Vekemans (2002), the estimator of Loiselle et al (1995) weights allele contribution in a manner least subject to bias caused by low-frequency alleles. As stated by Romero-Severson et al (2001), in maize germplasm development the pedigree of interest involves recent ancestry rather that ancient relationships. Thus, genetic-based distance measures can reveal descent from common progenitors regardless of multiple generations of intermating and introgression, and rare shared haplotypes can allow detection of essential derivation, a circumstance in which inbred lines are extracted directly from the population produced by selfing a single hybrid. Consequently, the estimators used in our study can be used for different purposes. The kinship coefficient described by Loiselle et al (1995), which is defined as ratios of differences of probabilities of identity in state and in which the coefficient is computed as a correlation coefficient between allelic states, cannot

be used to estimate probability of identity by descent because alleles come from an arbitrary sample rather than from a population (Rousset, 2002). For a given SSR allele, identity in state may not result from identity by descent (Romero-Severson et al, 2001). Lia et al (2009) showed that all ten SSRs assayed in maize landraces showed homoplasy and that evolutionary forces such as divergence, rather than convergence, were driving size homoplasy. For instance, SSR locus phi127 assayed in their report carried an INDEL 2-bp long at the 3´ position of the tetranucleotide repeat. In our work, we detected a similar pattern of variation (tetra- and dinucleotide) in this locus. Mogg et al (2002) found that some of the allele length polymorphisms seen with SSRs could be due to the presence of INDELs within the flanking regions rather than changes in the number of repeats at the primary SSRs motifs, leading to SSR homoplasy, whereby different (sequence-based) SSR alleles have evolved to be of identical size. Another important feature of the relative kinship coefficient is that it defines the degree of covariance between a pair of individuals. Thus, relatedness estimations for association mapping purposes, among individuals within and among subpopulations are accounted for by a relative kinship on a finer scale than the population structure estimator given by STRUCTURE analysis (Yu et al, 2006).

We conclude that the number and the distribution of SSRs assayed were adequate to clearly infer by a similarity model-based approach three subpopulations of inbreds with different ecogeographic distribution and ancestry origin. Also, most inbreds with undisclosed pedigrees were clustered by similarity to one of the subpopulations mentioned above. The additional information provided by kinship coefficients constitutes an additional tool for predicting heterotic patterns for the maize inbred breeding program. Further studies are needed to study the extent of pairwise linkage disequilibrium between adjacent, linked and unlinked SSR markers and its implication for marker/trait associating studies within our breeding population.

Acknowledgements

This research was conducted at the National Institute of Agricultural Technology (INTA) Pergamino station and financed by INTA projects. We thank all staff members and students that participated during laboratory analysis. The authors would like to thank to the reviewers for their kind remarks on the earlier version of the manuscript.

References

Balzarini MG, Di Rienzo JA, 2012. Info-Gen: FCA, Universidad Nacional de Córdoba, Argentina. URL http://www.info-gen.com.ar

Beckmann JS, Soller M, 1990. Toward a unified approach to genetic mapping of eukaryotes based on sequence tagged microsatellite sites. Bio/Technology 8: 930-932

Bernardo R, Kahler A, 2001. North American study on essential derivation in maize: inbreds developed without and with selection from F_2 populations. Theor Appl Genet 102: 986-992

Bernardo R, Murigneux A, Maisonneuve JP, Johnsson C, Karaman Z, 1997. RFLP-based estimates of parental contribution to F_2- and BC_1-derived maize inbreds. Theor Appl Genet 94: 652-656

Bernardo R, Romero-Severson J, Ziegle J, Hauser J, Joe L, Hookstra G, Doerge RW, 2000. Parental contribution and coefficient of coancestry among maize inbreds: pedigree, RFLP, and SSR data. Theor Appl Genet 100: 552-556

Berry DA, Seltzer JD, Xie C, Wright DL, Smith JSC, 2002. Assessing probability of ancestry using simple sequence repeat profiles: Applications to maize hybrids and inbreds. Genetics 161: 813-824.

Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, Nordborg M, Bergelson J, Cuguen J, Roux F, 2010. Linkage and association mapping of Arabidopsis thaliana flowering time in nature. PLoS Genetics, 6 (5): e1000940

Camus-Kulandaivelu L, Veyrieras JB, Madur D, Combes V, Fourmann M, Nordborg M, Bergelson J, Cuguen J, Roux F, 2006. Maize adaptation to temperate climate: relationship between population structure and polymorphism in the *Dwarf8* gene. Genetics 172: 2449-2463

Chin EC, Senior ML, Shu H, Smith JS, 1996. Maize simple repetitive DNA sequences: abundance and allele variation. Genome 39: 866–873

Dale PJ, Clarke B, Fontes EMG, 2002. Potential for the environmental impact of transgenic crops. Nat Biotech 20: 567–574

Delucchi C, Eyhérabide GH, Lorea RD, Presello DA, Otegui ME et al. 2012. Classification of argentine maize landraces in heterotic groups. Maydica 57: 26-33

Di Rienzo JA, Casanoves F, Balzarini MG, Gonzalez L, Tablada M, Robledo CW, 2010. InfoStat, versión 2010, Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina

Eyhérabide GH, Gonzalez AS, 1997. Interactions between testers and argentine maize landraces. Maydica 42: 29-38

Eyhérabide GH, Nestares G, Hourquescos MJ, 2006. Development of a heterotic pattern in orange flint maize, pp. 368-379. In: Plant breeding: the Arnel R Hallauer Internacional Symposium. Lamkey KR, Lee M, eds. Blackwell Publishing. Ames, Iowa

Ferrer M, 2012. Los recursos genéticos del maíz, pp. 107-124. In: Bases para el manejo del cultivo de maíz. Eyhérabide G ed. Ediciones INTA. Buenos Aires

Flint-Garcia SA, Buckler ES, Tiffin P, Ersoz E, Springer NM, 2009. Heterosis Is Prevalent for Multiple Traits in Diverse Maize Germplasm. PLoS ONE

4(10): e7433. doi:10.1371/journal.pone.0007433
Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI, 2008. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82: 100-112
Hallauer AR, Miranda Filho, JB, 1988. Quantitative genetics in maize breeding. Ames. Iowa State University
Hardy OJ, Vekemans X, 2002. SPAGeDi: a versatile computer program to analyze spatial genetic structure at the individual or population levels. Molecular Ecology Notes 2: 618-620
Holton TA, 2001. Plant Genotyping by Analysis of Microsatellite. In: Plant Genotyping the DNA Fingerprinting of Plants, United Kingdom. CAB International Wallingford
Jakobsson M, Rosenberg NA, 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23: 1801-1806
Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E, 2008. Efficient control of population structure in model organism association mapping. Genetics 178: 1709-1723
Kleinhofs A, Kilian A, Saghai Maroof M, Biyashev RM, Hayes P, Chen FQ, Lapitan N, Fenwick A, Blake TK, Kanazin V, Ananiev E, Dahleen L, Kudrna D, Bollinger J, Knapp SJ, Liu B, Sorrels M, Heun M, Franckowiak JD, Hoffman D, Skadsen R, Steffenson BJ, 1993. A Molecular isozyme and morphological map of the barley (*Hordeum vulgare*) genome. Theor Appl Genet 86: 705-712
Levinson G, Gutman GA, 1987. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. Nucleic Acids Research 15: 5323-5338
Lia VV, Poggio L, Confalonieri VA, 2009. Microsatellite variation in maize landraces from Northwestern Argentina: genetic diversity, population structure and racial affiliations. Theor Appl Genet 119: 1053-1067
Liu J, 2002. Powermarker - A Powerful Software for Marker Data Analysis. North Carolina State University Bioinformatics Research Center, Raleigh, NC (www.powermarker.net).
Liu K, Goodman MM, Muse S, Smith JSC, Buckler ES, Doebley J, 2003. Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165: 2117-2128
Loiselle BA, Sork VL, Nason J, Graham C, 1995. Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (*Rubiaceae*). Am J Bot 82: 1420-1425
Luna JT, Kugler WF, Godoy EF, Mazzoni L, 1964. Maíz, pp. 553-589. In : Enciclopedia Argentina de Agricultura y jardinería, vol 111, primera parte. Parodi L ed. Buenos Aires
Malécot G, 1948. Les Mathématiques de l'hérédité. Paris: Masson et Cie
Maunder AB, 1992. Identification of useful germplasm for practical plant breeding programs, pp.147-169. In: Plant Breeding in the 1990s. Stalker HT, Murphy JP, eds. Proceedings of the Symposium on Plant Breeding in the 1990s. CAB
Melchinger AE, Gumber RK, 1998. Overview of Heterosis and Heterotic Groups in Agronomic Crops. In: Concepts and Breeding of Heterosis in Crop Plants. Special Publication 25. CSSA: Crop Sci Society of America
Menkir A, Olowolafe MO, Ingelbrecht I, Fawole I, Badu-Apraku B, Vroh BI, 2006. Assessment of testcross performance and genetic diversity of yellow endosperm maize lines derived from adapted x exotic backcrosses. Theor Appl Genet 113: 90–99
Milne I, Shaw P, Stephen G, Bayer M, Cardle L, Stephen G, Marshall D, 2010. Flapjack – graphical genotype visualization. Bioinformatics 26: 3133-3134
Mogg R, Batley J, Hanley S, Edwards D, O'Sullivan H, Edwards KJ, 2002. Characterization of the flanking regions of Zea mays microsatellites reveals a large number of useful sequence polymorphisms. Theor Appl Genet 105:532–543
Moll RH, Longquist JH, Fortuna JV, Johnson EC, 1965. The relation of heterosis and genetic divergence in maize. Genetics 52: 139–144
Nei M, 1972 Genetic distance between populations. Am Nat 106: 283-291
Ordás A, 1991. Heterosis in Crosses between American and Spanish Populations of Maize. Crop Science 31: 931-935
Pritchard JK, 2001. Deconstructing maize population structure. Nat Genet 28: 203-204
Pritchard JK, Stephens M, Donnelly P, 2000a. Inference of population structure using multilocus genotype data. Genetics, 155: 945-959
Pritchard JK, Wen X, Falush D, 2009. Documentation for STRUCTURE software: Version 2.3. The University of Chicago Press
Ramsay L, Macaulay M, Cardle L, Morgante M, Degli Ivanissevich S, Maestri E, Powell W, Waugh R, 1999. Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. Plant J : 415-425
Register JC, Sullivan HR, Yun Y, Cook D, Vaske DA, 2001. A set of microsatellite markers of general utility in maize. Available: http://www.agron.missouri.edu/mnl/75/02register.html. accessed 14 January 2010
Romero-Severson J, Smith JSC, Ziegle J, Hauser JL, Hookstra G, 2001. Pedigree analysis and haplotype sharing within diverse groups of *Zea mays* L inbreds. Theor Appl Genet 103: 567–574
Rostocks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, Graner A, Close TJ, Waugh R, 2006. Recent history of artificial outcrossing fa-

cilitates whole genome association mapping in elite crop varieties. Proc Natl Acad Sci USA 103: 18656-18661

Rousset F, 2002. Inbreeding and relatedness coefficients: what do they measure? Heredity 88: 371-380

Senior ML, Murphy JP, Goodman MM, Stuber CW, 1998. Utility of SSRs for Determining Genetic Similarities an Relationships in Maize Using an Agarose Gel System. Crop Science 38(4): 1088-1098

Smith JSC, Chin ECL, Shu H, Smith OS, Wall SJ, et al. 1997. An evaluation of the 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

Steen RG, Kwitek-Black AE, Glenn C, Gullings-Hindley J, Van Etten W, et al, 1999. A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. Genome Res 9: 1–9

Stich B, Melchinger AE, Frisch M, Maurer HP, Heckenberger M, et al, 2005. Linkage disequilibrium in European elite maize germplasm investigated with SSRs. Theor Appl Genet 111: 723–730

Stich B, Mohring J, Piepho HP, Heckenberger M, Buckler ES, Melchinger AE, 2008b. Comparison of mixed-model approaches for association mapping. Genetics 178: 1745-1754

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S, 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L): Frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11: 1441-1452

Vigouroux Y, Glaubitz JC, Matsuoka Y, Goodman MM, Sánchez GJ, 2008. Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. Am J Bot 95: 1240–1253

Vigouroux Y, Jaqueth JS, Matsuoka Y, Smith OS, Beavis WD, Stephen J, Smith C, Doebley J, 2002. Rate and pattern of mutation at microsatellite loci in maize. Mol Biol Evol 19(8): 1251-1260

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES, 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203–208