

RESEARCH ARTICLE

Open Access

Genetic diversity among INERA maize inbred lines with single nucleotide polymorphism (SNP) markers and their relationship with CIMMYT, IITA, and temperate lines

Abdalla Dao^{1*}, Jacob Sanou¹, Sharon E Mitchell², Vernon Gracen³ and Eric Y Danquah⁴

Abstract

Background: Genetic diversity provides the capacity for plants to meet changing environments. It is fundamentally important in crop improvement. Fifty-nine local maize lines developed at INERA and 41 exotic (temperate and tropical) inbred lines were characterized using 1057 SNP markers to (1) analyse the genetic diversity in a diverse set of maize inbred lines; (2) determine the level of genetic diversity in INERA inbred lines and patterns of relationships of these inbred lines developed from two sources; and (3) examine the genetic differences between local and exotic germplasms.

Results: Roger's genetic distance for about 64% of the pairs of lines fell between 0.300 and 0.400. Sixty one per cent of the pairs of lines also showed relative kinship values of zero. Model-based population structure analysis and principal component analysis revealed the presence of 5 groups that agree, to some extent, with the origin of the germplasm. There was genetic diversity among INERA inbred lines, which were genetically less closely related and showed a low level of heterozygosity. These lines could be divided into 3 major distinct groups and a mixed group consistent with the source population of the lines. Pairwise comparisons between local and exotic germplasms showed that the temperate and some IITA lines were differentiated from INERA lines. There appeared to be substantial levels of genetic variation between local and exotic germplasms as revealed by missing and unique alleles.

Conclusions: Allelic frequency differences observed between the germplasms, together with unique alleles identified within each germplasm, shows the potential for a mutual improvement between the sets of germplasm. The results from this study will be useful to breeders in designing inbred-hybrid breeding programs, association mapping population studies and marker assisted breeding.

Keywords: Maize inbred line, Diversity, Allele frequency

Background

Genetic diversity in agricultural populations provides the capacity to meet changing environments and market requirements [1]. In crop breeding, genetic diversity is very important for an analysis of genetic variability in cultivars [2], the identification of diverse parental combinations to create segregating progenies with maximum

genetic variability for further selection [3], and the introgression of desirable genes from diverse germplasm into the available genetic base [4].

Molecular marker analyses provide an important approach for estimating genetic relationships. Restriction Fragment Length Polymorphism (RFLP), Simple Sequence Repeats (SSR) and Single Nucleotide Polymorphism (SNP) markers have been used to study genetic diversity in maize. Warburton *et al.* [5] characterized 218 elite maize inbred lines from CIMMYT (International Maize and Wheat Improvement Centre) using RFLP markers and

* Correspondence: adao@wacci.edu.gh

¹Institute of Environment and Agricultural Research (INERA), BP 910, Bobo-Dioulasso, Burkina Faso

Full list of author information is available at the end of the article

suggested the use of molecular markers and cross performance information to refine heterotic groups and select representative testers. SSR markers have been used to characterize the genetic structure and diversity of 260 important tropical and temperate maize inbreds [6], and to investigate genetic diversity in CIMMYT lowland tropical [7] and mid-altitude, highland subtropical [8] inbred lines. Population structure and patterns of relationships of 770 inbred lines representing both temperate and tropical/subtropical maize germplasm [9] and 450 maize inbred lines from CIMMYT breeding programs in Zimbabwe and Kenya [10] have been investigated using SNP markers.

A very large number of SNP markers are now available in maize, many of which have been developed from the DNA sequence of known genes. For this reason, SNP markers are now the assay of choice for a variety of tasks in maize improvement including genetic diversity analysis [9].

CIMMYT and IITA (International Institute of Tropical Agriculture) are the source of maize breeding materials for a significant portion of Africa. CIMMYT and IITA inbred lines and OPVs are bred to contain considerable diversity and are then taken by National Agriculture Research Programs and selected for further adaptation in their own particular environment(s). Maize germplasm at the Institute of Environment and Agricultural Research (INERA) in Burkina Faso includes different materials from CIMMYT and IITA. The maize breeding program at INERA was established in the late 1980s but little progress was made because a lot of effort was devoted to Open Pollinated Varieties (OPVs) development, which were more suitable to the majority of farmers. However, INERA initiated a new inbred-hybrid breeding program in 1991. Inbred lines are essentially extracted from Open Pollinated Varieties (OPVs), which have the advantage of being both environmentally adapted and adopted by farmers. Several studies have addressed the relative performance *per se* and in top-cross combination of the newly developed inbred lines but they have not yet been characterized at the molecular level. Knowledge of molecular genetic diversity among the maize inbred lines developed for the national breeding program will provide guidance on how to use the local germplasm more efficiently. A better understanding of how the local lines are related to lines from different sources may be useful for decisions on the incorporation of exotic germplasm in existing breeding program.

The objectives of this study were to (1) analyze the genetic diversity in a diverse set of maize inbred lines; (2) determine the level of genetic diversity in INERA inbred lines and patterns of relationships of these inbred lines developed from two sources; and (3) explore the genetic differences between local and exotic germplasms.

Methods

Plant materials

A total of 100 maize lines representing INERA, CIMMYT, IITA and temperate germplasms were chosen for molecular characterization. CIMMYT and IITA lines were chosen based on their resistance/tolerance to biotic and/or abiotic stresses and INERA lines were chosen to represent a sample of advanced lines selected for morphological characteristics and disease tolerance. The 100 lines included 59 lines from INERA, 16 lines from CIMMYT/Zimbabwe, 15 lines from IITA, and 10 temperate lines obtained from National Institute for Agricultural Research (INRA, France). The temperate lines represented European and U.S inbred lines. The INERA germplasm included 3 OPVs (FBC6, ESPOIR and FBMS1), 34 and 21 lines extracted from FBC6 and ESPOIR, respectively (thereafter called Subset A and B, respectively). FBC6 has a mixed genetic background, It was developed from a mixture of 8 varieties (DMRESR-Y and TZESR-Y C2 from IITA; ROD4, ROD12 and « Révolution précoce » from CIRAD(Agricultural Research Centre for International Development)/IRAT(Institute for Research in Tropical Agriculture)/Réunion and FBC4, Maka and IRAT217 from INERA. ESPOIR was developed from Population 66 SR of CIMMYT/IITA using different cycles of recurrent selection and FBMS1 was developed from a mixture of different sources of sweet corn. FBML10 was derived from IITA line, TZI35, and selected for more uniform grain colour. The list of the inbred lines together with kernel colour, environmental adaptation and reactions to stresses (where available) are listed in Additional file 1: Table S1. The tropical lines included 6 testers from CIMMYT and IITA with known heterotic patterns. The three CIMMYT testers, one from heterotic group A (VL0511298) and two from group B (VL054881 and T02058) were developed from different populations (Additional file 1: Table S1). No known relationship exists between the three testers and the lines except that the tester VL054881 and line VL054794 have CML390 in common in their genetic background. IITA tester, TZEI 17, from heterotic group A and lines TZEI 177 and TZEI 16 were derived from the broad-based *Striga hermonthica* resistant early yellow population, TZE COMP5-Y. The two other IITA testers (TZEI 10 and TZEI 23) which belong to heterotic group B and 7 inbred lines including TZEI 158, TZEI 161, TZEI 124, TZEI 148, TZEI 8, TZEI 149 and TZEI 146, were derived from the broad-based *Striga hermonthica* resistant and drought tolerant early yellow population, TZE-Y Pop DT STR.

SNP genotyping

Genomic DNA for each sample was extracted from seedling leaves (at V3 to V5 stage) using a magnetic bead and Klearcall extraction buffers (protocol <http://www.lgcgenomics>).

com/nucleic-acid-extraction/kits/). They were then genotyped by Kbiosciences (Hoddesdon Herts, UK) using their Kompetitive Allele-Specific PCR (KASP) SNP genotyping system. The KASP assay uses a technique based on allele specific oligo extension and fluorescence resonance energy transfer (FRET) for signal generation. The fluorescent reporting system is comprised of four single-labelled oligonucleotides that hybridize to one another in free solution to form a fluorescent quenched pair which upon introduction of complementary sequences generates a measurable signal. Complete details on the principle and procedure of the assay are available at <http://www.lgcgroup.com/products/kaspgenotyping-chemistry/#.VHhVZlvz0SU>. SNP markers used in this study were chosen to cover all the 10 maize chromosomes and represented all the CIMMYT SNP markers that GCP (Generation Challenge Program) converted to KASP system and made available for maize genotyping. A detailed list of SNPs used can be found in Additional file 2: Table S2.

Statistical analysis

Summary statistics, including the minor allele frequency (MAF), unbiased estimation of gene diversity, observed heterozygosity, and polymorphism information content (PIC) value, were calculated using PowerMarker software [11]. The PIC value, described by Botstein *et al.* [12], was used to refer to the relative value of each marker with respect to the amount of polymorphism revealed. Heterozygosity and unbiased gene diversity were calculated to quantify the genetic variation in the maize lines sampled. Allele frequency was calculated for each locus across each of the four distinct sets of maize germplasm: INERA, CIMMYT, IITA and Temperate. Difference in allele frequency between local germplasm and each of exotic material was calculated and statistical significance of differences in allele frequency was based on the P value from Fisher's exact test [13]. The genetic distance between genotypes was computed using the Roger's genetic distance [14] with PowerMarker software. Genetic distance was calculated between pairwise comparison of all the lines and all the lines bred only in INERA.

SPAGeDi software [15] was used to obtain the kinship matrix between lines using Loiselle method [16]. Loiselle's estimator is expected to be unbiased with respect to allelic frequencies [15]. The relative kinship reflects the approximate degree of identity between two given individuals. Negative values between two individuals (indicating less relationship than expected between two random individuals) were changed to zero. The relative kinship coefficients were estimated between pairs of the entire germplasm set and INERA germplasm set.

Three multivariate analysis including cluster analysis, principal component analysis and model-based population structure analysis were employed to subdivide inbred lines into genetic subgroups. A dendrogram was constructed, in

cluster analysis, from the Roger's genetic distance matrix using the neighbor-joining algorithm [17] with PowerMarker and the resulting trees were visualized using MEGA version 5.2.2 [18]. Principal component analyses (PCA) were conducted using TASSEL software [19]. To infer the structure of the population, the software STRUCTURE [20] was used with 1057 informative SNPs. The dataset was tested for a number of subpopulations ranging from $k = 1-12$. Three runs for each k value were performed using the admixture model and correlated allele frequencies [21]. The burn-in length and iterations were all set to 500,000. To infer the most likely number of groups within the population, the Evanno transformation method [22] was used on the STRUCTURE outputs. In the model-based method, membership coefficients (Q values) for each inbred line were estimated to have its memberships in multiple subgroups. Inbred lines with membership probabilities ≥ 0.60 were assigned to the corresponding subgroup and lines with membership probabilities < 0.60 assigned to a mixed subgroup.

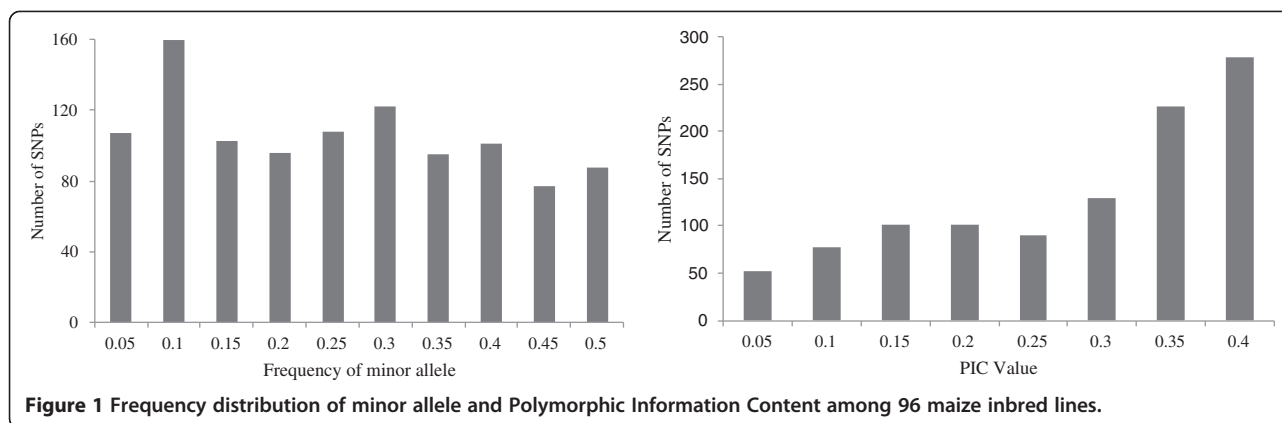
Results

Descriptive summary and statistics of 1 237 SNPs in the assay

Of the 1237 SNPs in the KASP assay, 1151 (93%) were successfully called in the 100 lines. SNP markers that were monomorphic (75 SNPs) or had missing data points more than 20% (19 SNPs) in the diversity panel of inbred lines were removed from further analysis. As a result, a total of 1057 SNPs (91.83%) was called successfully with high quality. Of the 100 maize lines, 4 lines were deleted from the next statistical analysis due to missing data called $\geq 20\%$. A detail list of these informative SNP loci including chromosome, base change, minor allele frequency (MAF), heterozygosity, gene diversity, and PIC is presented in Additional file 2: Table S2.

Of the 1057 SNPs, 55.91% (591 of 1057) had a MAF > 0.2 and were selected as markers with normal allele frequencies. Approximately 25.26% SNPs (267 of 1057) had a MAF ≤ 0.1 , and 10.12% (107 of 1057) had a MAF ≤ 0.05 . In addition, 88 (8.33%) showed almost equal allele frequencies (with MAF close to 0.5) for two alternative alleles (Figure 1). The average PIC was 0.256, ranging from 0.02 to 0.375 with a peak distribution between 0.350 and 0.375 (Figure 1). The PIC value (0.256) was consistent with the highest reported value of 0.259 using 1034 informative SNPs and 770 maize inbred lines [9] and was higher than the value of 0.239 found by Hao *et al.* [23] using 1006 informative SNPs and 80 maize inbred lines.

The average heterozygosity of each line was 3.8%, this is within expected ranges of normal level of residual heterozygosity in inbred lines of maize. The heterozygosity rate of the 96 inbred lines is provided in Additional file 1: Table S1.



As the quality of markers have an effect on diversity estimation, we identified 580 high quality SNPs out of 1057 based on normal MAF (>0.2) and high PIC values (>0.25). Lu et al. [9] recommended 449 out of 1034 SNPs that were found to be the best for the detection of genetic diversity in temperate, subtropical and tropical maize germplasm and Semagn et al. [10] proposed 644 out of 1065 SNPs for routine genetic diversity and mapping studies in tropical and subtropical CIMMYT maize germplasm. Out of the 580 high quality SNPs identified in the present study, 327 (56.4%) and 278 (47.9%) SNPs were common between the present study and, Lu et al. [9] and Semagn et al. [10] studies, respectively.

To understand the effect of SNP subset and the influence of the two subsets in INERA germplasm on genetic diversity, the parameters of genetic diversity were estimated for each group with the 1057 SNPs and the 580 SNPs. Using all the 1057 informative SNPs and 96 inbred lines, temperate germplasm was found to show the highest average PIC value and gene diversity, followed by INERA germplasm for gene diversity and by CIMMYT germplasm for PIC value, whereas the IITA germplasm showed the lowest PIC value and gene diversity (Table 1). On the other hand, using 580 high-quality markers, INERA germplasm was found to show the highest average PIC value and gene diversity, followed by CIMMYT and temperate germplasm respectively, while the IITA germplasm still showed the lowest (Table 1). The order of genetic diversity in different germplasm sets changed between the 1057 SNPs and the 580 high quality markers selected, contrary to what has been reported in the study of Lu et al. [9], which showed that different subsets of SNPs did not change the order of genetic diversity in different germplasm collections. The sample size could explain this difference. However there was a significant increase in the estimates of PIC and gene diversity for all germplasm collections, which was consistent with Lu et al. [9] finding. Compared to the entire set of INERA germplasm, the estimates for PIC and gene diversity decreased in the two subsets. In subset A (FBC6 derived

lines), the order of genetic diversity in different germplasm collections did not change whereas with subset B (ESPOIR derived lines) the order changed, indicating that the level of genetic diversity in subset A was slightly higher than that in subset B. However, the PIC value and gene diversity in each subset of INERA germplasm were not significantly different from the entire INERA germplasm set. This could be attributed to the fact that the two subsets share a common genetic background. In all cases, the IITA maize lines tested in this study appear to have relatively narrow genetic base as revealed by their estimates for both PIC and gene diversity. The genetic diversity in the temperate germplasm with 1057 SNPs is much higher than that in the tropical germplasm, in agreement with a previous study [9] but opposite results was found with studies of Yan et al. [24] and Liu et al. [6] using SNP and SSR markers, respectively. These temperate lines are an important resource to find new functional alleles of desired traits to improve tropical lines.

Table 1 PIC and gene diversity as revealed by 580 SNPs selected compared to their entire counterparts

	MAF	Gene diversity	Heterozygosity	PIC
1057 SNPs				
Temperate	0.256	0.302	0.008	0.268
IITA	0.202	0.252	0.038	0.218
CIMMYT	0.223	0.278	0.039	0.238
INERA	0.211	0.282	0.048	0.232
Subset A	0.199	0.263	0.044	0.219
Subset B	0.205	0.263	0.055	0.222
580 SNPs				
Temperate	0.306	0.348	0.01	0.305
IITA	0.287	0.348	0.053	0.295
CIMMYT	0.314	0.372	0.054	0.311
INERA	0.321	0.408	0.071	0.326
Subset A	0.305	0.383	0.067	0.312
Subset B	0.302	0.37	0.079	0.307

Population structure and relative kinship

An admixture model-based clustering method in STRUCTURE was implemented to infer population structure for all 96 tested lines and it was run for the number of fixed subgroups k from 1 to 12. The likelihood (L_n) value of this analysis is shown in Figure 2. Likelihood increases continuously with no obvious inflection point. This could imply that the lines included in the analysis were very diverse as well as highly mixed, however the L_n value for each given Pritchard's K (the supposed number of subpopulations based on the model) increased sharply when $K < 5$, and the increasing trend became more moderate for $K > 5$. In addition, the Evanno criterion supported the choice of $k = 5$ as the highest level of structure, so that five genetically distinct subgroups can be claimed. Three of these groups (Groups 3, 4, and 5) included all INERA maize inbred lines and the other two groups (1 and 2) included maize lines predominantly from temperate in group 1 and exclusively from IITA for group 2. The five groups (1,2,3,4 and 5) are named as Temperate, IITA, INERA-1, INERA-2 and INERA-3 respectively (Figure 3).

Group 1 (named temperate) consisted of 10 temperate lines, 7 CIMMYT lines and 1 IITA line (Additional file 3: Table S3). The temperate lines belong to different heterotic groups (BSSS, Lancaster, European group) and had high membership Q value between 93 to 100%. Line VL05616 from CIMMYT, included in this group, was also classified with temperate lines particularly with Lancaster heterotic group in a previous study [9]; a temperate line (FR812) constitutes 50% of its pedigree. Group 2, IITA, contained only 5 IITA inbred lines. Group 4, also named INERA-2, was the largest group containing 41 INERA inbred lines, of which 23 were extracted from FBC6, 17 from ESPOIR, and 1 line (FBML10) selected from TZI35 in INERA. In addition, 3 other IITA inbred lines belong to this group. The group 3, INERA-1, and the group 5, INERA-3, contained 4 and 5 INERA inbred lines, respectively, that are closely related in pedigree. In addition to

the inbred lines that were clearly assigned with probability ≥ 0.60 to a single group (population), 19 inbred lines (19.8% of the total) could not be clearly assigned to any of these groups. These lines, were placed in a mixed group and, include 8 lines from CIMMYT, 6 from IITA and 5 from INERA. Out of the six testers used in this study belonging to heterotic groups A and B, only TZEI 17 (heterotic group A) was assigned to a group (INERA-2). The remaining testers (one from heterotic group A and 4 from group B) were included in the mixed group. Inbred lines with proportional memberships in the model-based groups are presented in Additional file 3: Table S3.

Relative kinship

The relative kinship reflects the approximate degree of identity between two given individuals. Relative kinship coefficients between pairs of lines varied from 0 to 1.55, with an overall average of 0.039, and 61.3% of the pairwise kinship estimates had a value of zero (Figure 4a), indicating that the lines were unrelated. Kinship analysis of INERA germplasm showed the lack of redundant lines among the germplasm since kinship coefficients for approximately 64% of the pairs of lines had a value of zero.

Cluster analysis

In order to gain further insight into the genetic diversity among different groups of the maize inbreds, a neighbor-joining tree based on Roger's genetic distance was constructed. The 96 inbred lines were classified into three major groups (Figure 5). In the first major group, composed largely of INERA lines, FBC6 derived lines were grouped together with 2 lines from CIMMYT and 3 lines from IITA. The second major group was divided into two subgroups: 21 INERA lines, representing ESPOIR derived lines, and 1 CIMMYT line in the first subgroup and the 10 temperate lines plus 2 CIMMYT and 10 IITA lines in the second subgroup. The third major group, represented



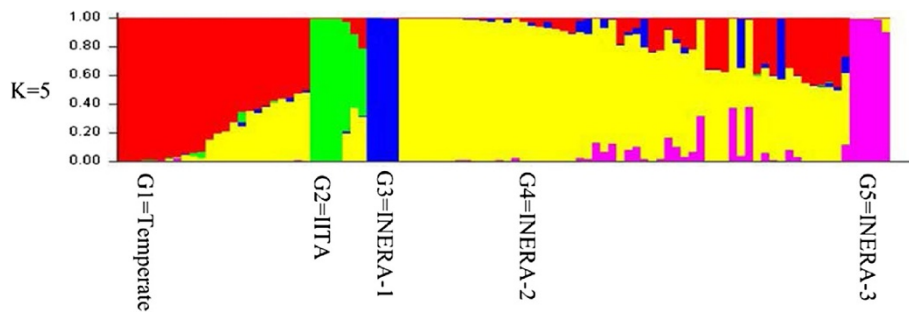


Figure 3 Population structure of the 96 maize lines shown as membership coefficients (Q values). Each vertical bar represents one maize line, which is partitioned into up to k coloured segments. Colour codes are as follows: Temperate, red; IITA, green; INERA-1, blue; INERA-2, yellow; INERA-3, purple.

mostly CIMMYT germplasm and contained 11 lines from CIMMYT and 2 lines from IITA. Each major group included 2 testers with well-known heterotic grouping. Testers VL0511298 and TZEI17 (Group A) were clustered in the first major group containing mostly lines from INERA (FBC6 derived). The testers, TZEI10 and TZEI23, belonging to heterotic group B, were placed in the second major

group while testers VL054881 and T02058 from heterotic group B, were included in the third major group. All the groups identified by the STRUCTURE analysis were also identified by the cluster analysis except the group 3 (named as INERA-2). Furthermore, some inbred lines assigned to some defined groups by structure analysis were placed in different groups by the cluster analysis. In addition, the third cluster group was mainly composed of lines assigned to the mixed group by structure analysis.

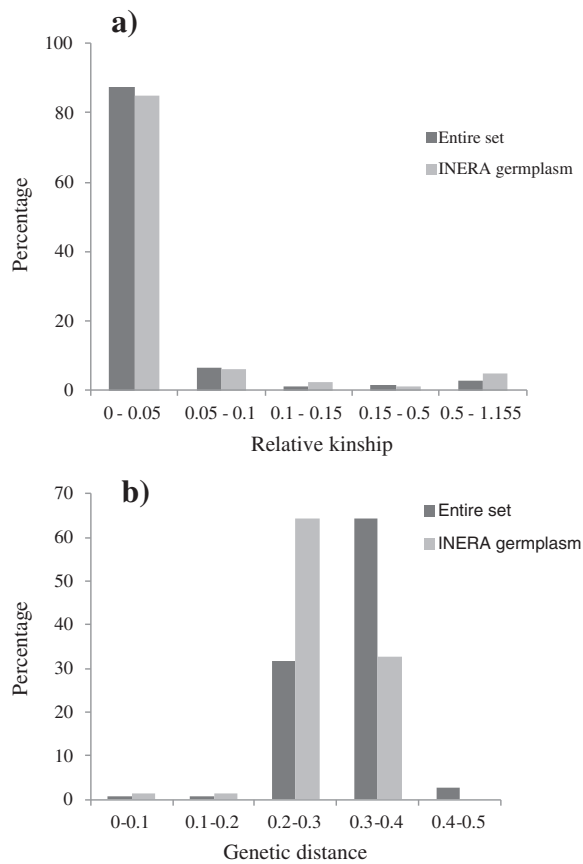


Figure 4 Distribution of pairwise relative kinship values (a) and Roger's genetic distance (b) for 96 (entire set) and 54 (INERA) maize inbred lines.

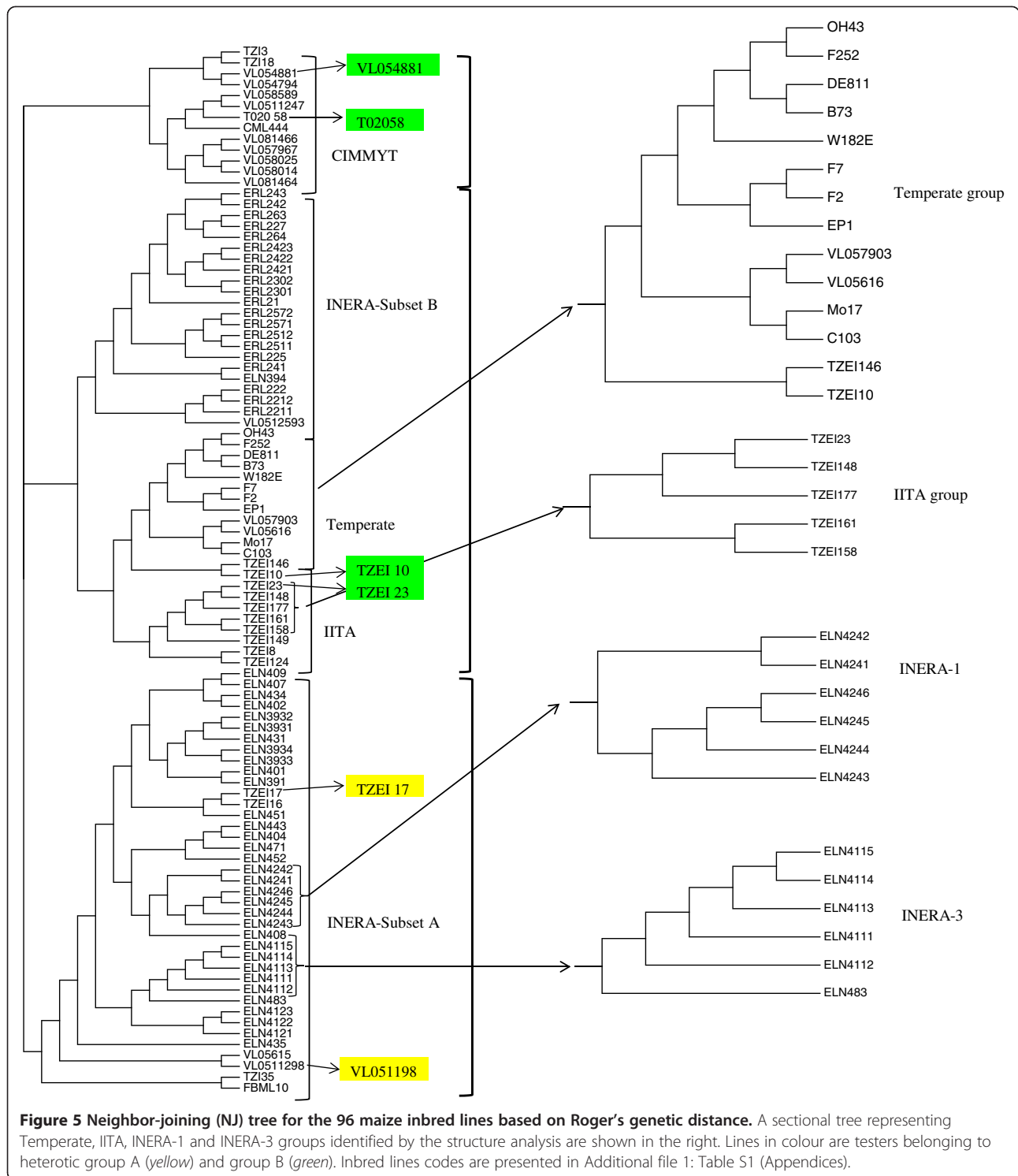
Principal component analysis

Principal Component Analysis (PCA) has been proposed as an alternative to Structure analysis for studying population structure of genotypic data [25]. Principal Component Analysis results were consistent with those of the structure analysis. PCA on the entire set of 96 inbred lines (Figure 6) showed a clear separation of the 5 groups identified in structure analysis. The second PC separated group 1 (temperate) from the 4 other groups. Group 2 (IITA) was well separated from the 3 groups (INERA-1, INERA-2 and INERA-3) by the first PC.

Principal component analysis (PCA) classified the INERA inbred lines into four distinct groups (Figure 7) which included the two groups (INERA-1, INERA-3) that were consistent with structure analysis, plus two other groups obtained by the separation of the lines in group 3 (INERA-2) from structure analysis. PCA identified a subgroup that consisted of ESPOIR derived lines from INERA-2 group comprised of FBC6 derived and ESPOIR derived lines.

In order to gain further insight into genetic differentiation between INERA germplasm and CIMMYT, IITA and temperate germplasms, a principal component analysis was performed on the following sets of inbred lines: set 1 consisted of INERA and temperate lines, set 2 : INERA and CIMMYT lines and set 3 : INERA and IITA lines.

PCA on set 1 showed a clear separation of INERA and temperate germplasms (Figure 8a). However, the PCA graph of set 2 did not separate the inbred lines on the basis of the breeding program. The first PC divided the



lines into two groups, a group consisted of a subset of FBC6 derived lines while the other group contained a mixture of CIMMYT and INERA (FBC6 derived and ESPOIR derived) lines (Figure 8b). Inbred lines of set 3 were separated into three groups with the first PC separating IITA and INERA (FBC6 derived) lines while the

second PC separated the two groups from the mixed group (Figure 8c).

Genetic distance

Roger's genetic distance of the 96 lines ranged from 0.0205 to 0.448, with the overall average distance of 0.314. However,

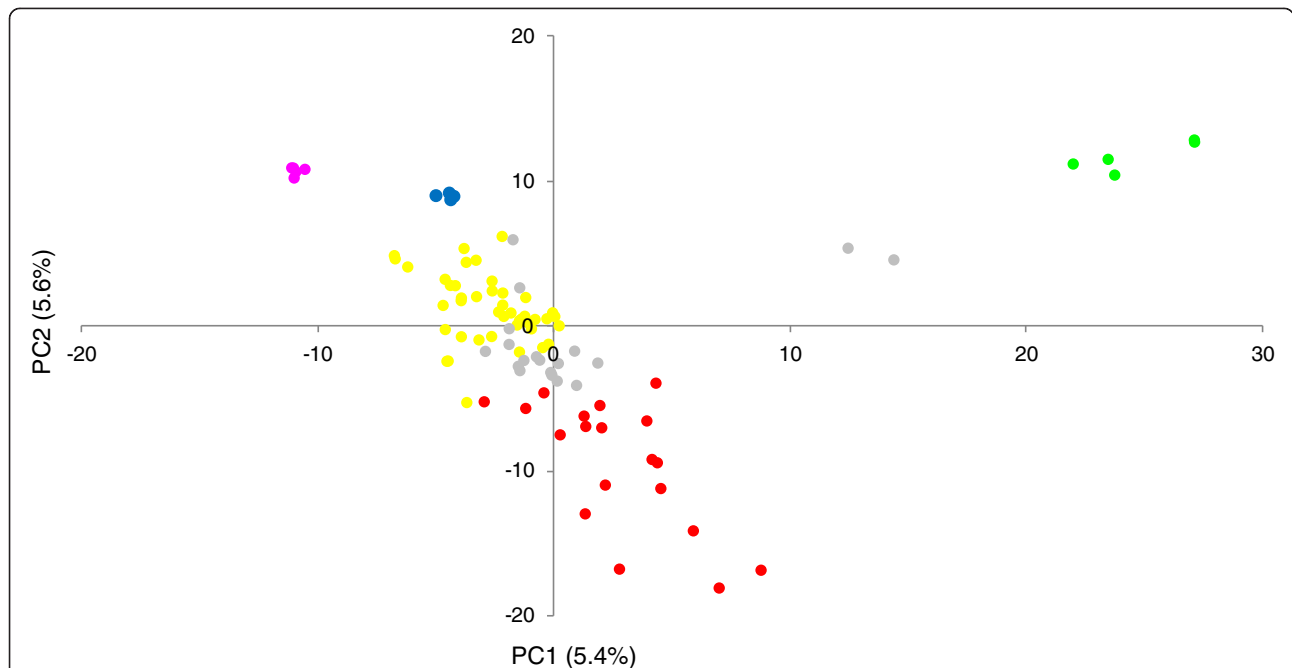


Figure 6 Principal component analysis for the entire set of maize lines. Groups identified by the structure analysis are shown in colours, the codes are as follows: Temperate (G1), red; IITA (G2), green; INERA-1 (G3), blue; INERA-2 (G4), yellow; INERA-3 (G5), purple; mixed group (GM), gray.

the majority (64.2%) of the inbred lines fell between 0.300 and 0.400 (Figure 4b). Genetic distance based on pairwise comparisons of the 54 INERA lines ranged from 0.029 to 0.348, with the overall average distance of 0.283 while 64.4% of the inbred lines fell between 0.200 and 0.300 (Figure 4b).

Differences in allele frequencies

To reveal genetic differences among local and exotic maize inbred lines, comparative analysis of allele frequencies was performed for three pairwise comparisons: INERA versus CIMMYT lines, INERA versus IITA, and INERA versus

Temperate. Of the 1057 SNPs, a significant difference in allele frequency ($P < 0.01$) was observed for 331 (31.3%) SNPs in INERA versus Temperate, 263 (24.6%) SNPs in INERA versus CIMMYT lines, and 252 (23.8%) SNPs in INERA versus IITA. A distribution of allele frequency difference observed in the three pairwise comparisons, ranging from 0.1% to 71% is shown in Figure 9. The minimum difference was found in INERA versus IITA comparison whereas the maximum was between INERA versus temperate. In INERA versus temperate comparison, the highest difference was 71% for PZA01352.5 and PZA00643.13 for alleles A/G

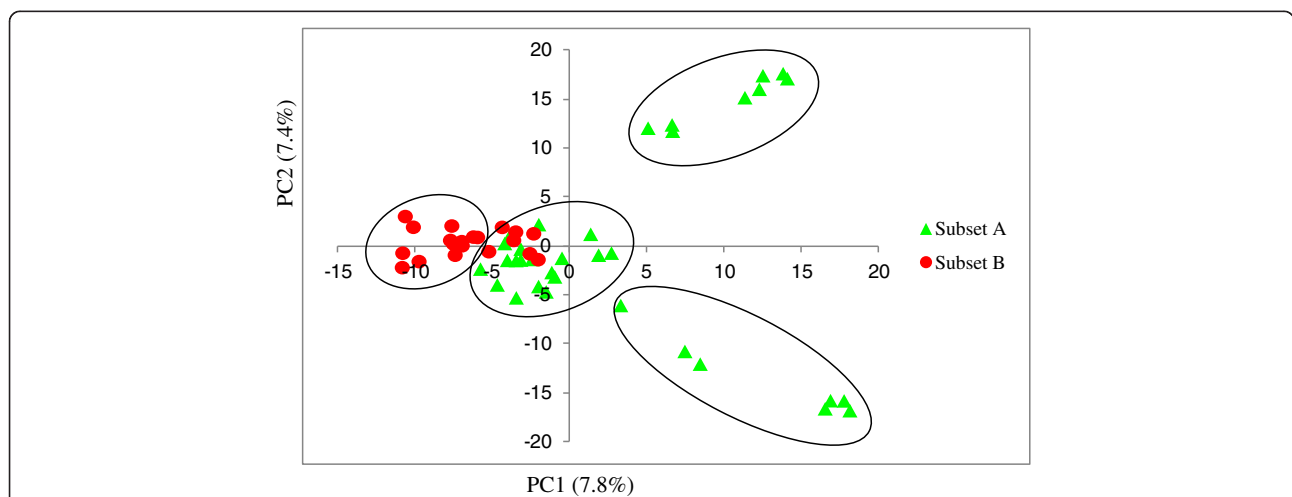


Figure 7 Principal component analysis for INERA inbred lines.

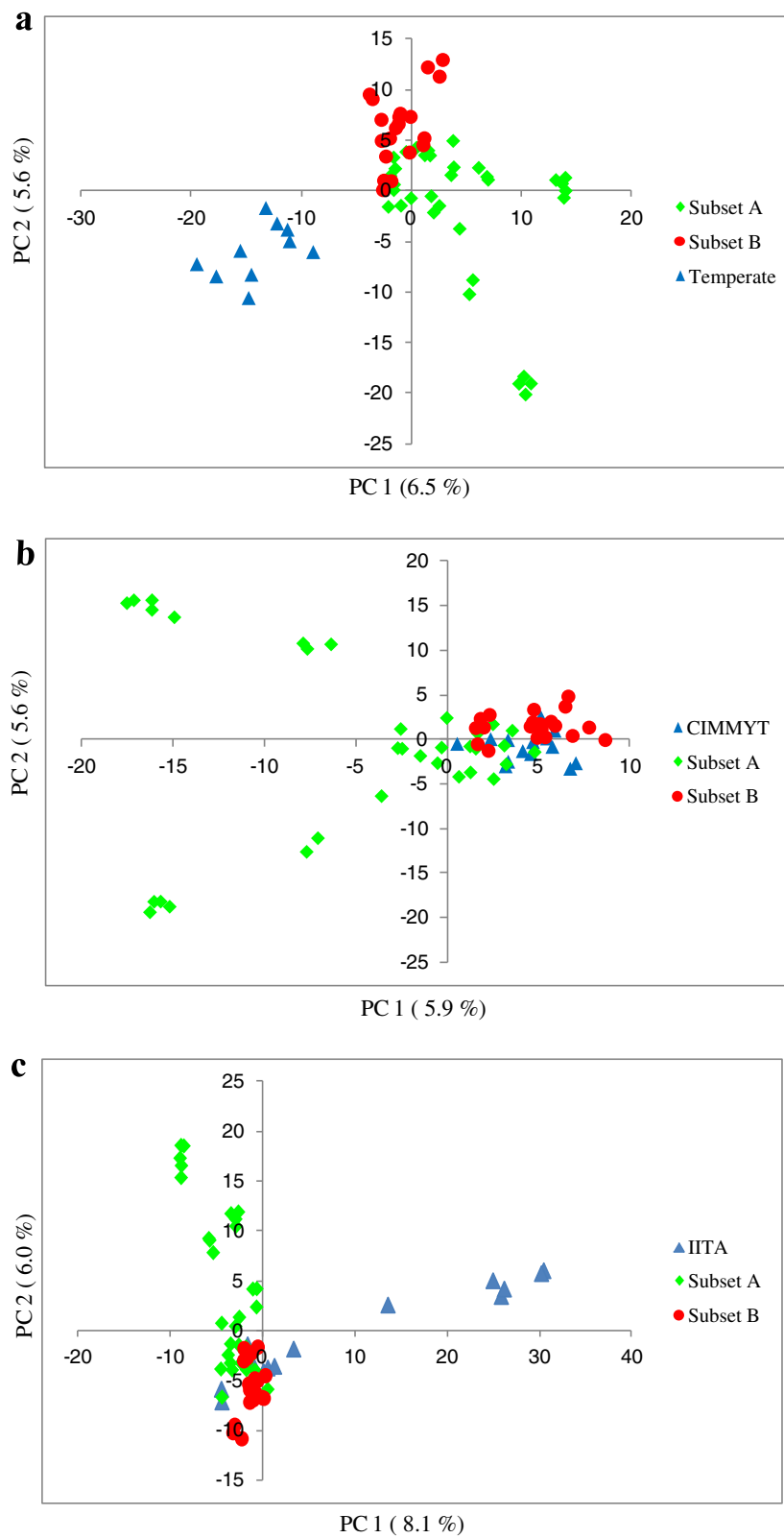


Figure 8 Principal component analysis for three germplasm sets. Set 1(INERA-Temperate lines) (a), set 2 (INERA-CIMMYT lines) (b) and set 3 (INERA-IITA lines) (c). Subset A and Subset B representing INERA FBC6 and ESPOIR derived lines respectively.

and G/T, respectively, the highest difference in INERA versus IITA was 67% for PZA02398.2 for alleles A/G and the least highest difference among the three pairwise comparison was 57% for PZA01073.1 for alleles A/G in INERA versus CIMMYT comparison (Table 2).

Missing and unique alleles in different germplasm collections

The temperate, CIMMYT and IITA inbred lines were included in this study for comparison purpose. They are not representative of all temperate, CIMMYT, and IITA germplasms. Thus, we only identified missing alleles present in INERA lines. In total, there were one hundred missing alleles identified in INERA germplasm which were present in other germplasm. The counterpart allele frequency in temperate, CIMMYT and IITA germplasm ranged from 3% to 70%. There are 28, 7 and 3 of such alleles that were frequent only in temperate, CIMMYT and IITA germplasm, respectively, but completely lacking in INERA germplasm (Figure 10a; Additional file 4: Table S4). In addition, 26, 11 and 3 of the missing alleles were present in both CIMMYT-temperate, temperate-IITA and CIMMYT-IITA germplasms respectively and 22 were frequent in the three germplasms (temperate, CIMMYT and IITA) (Additional file 4: Table S4). For instance, the allele C at PHM537.22 (chromosome 10) was missing in INERA germplasm but present in 70%, 50% and 13.3% of IITA, temperate and CIMMYT germplasm, respectively.

By comparing allele frequencies of a specific germplasm collection with those in the entire germplasm set (96 lines), 55 unique alleles were identified that only existed in that specific germplasm collection but not in others (Figure 10b, Additional file 2: Table S2). The highest number (28) of unique alleles was found in the temperate lines. All the temperate lines had at least one unique allele and 9 of the 10 lines had more than 2 unique alleles. Unique allele frequencies in temperate, CIMMYT and IITA collection varied from 0.01 to 0.05 while it varied from 0.01 to 0.1 in INERA germplasm.

Discussion

Population structure and genetic relationship

The extent of genetic differentiation, population structure, and patterns of relationship among a set of 96 maize inbred lines was investigated using 1057 SNP markers. The model-based population structure analysis, NJ-cluster analysis, and principal component analysis were used to explore whether the population of the 96 maize lines from diverse origins (temperate, CIMMYT, IITA and INERA) was homogeneous or contained genetically distinct subgroups. All the different multivariate methods supported the presence of genetically distinct groups. The model-based population structure and principal component analysis showed, to some extent, a separation by origin of the lines with related lines tending to cluster together. It has been reported that the clustering observed in the tropical

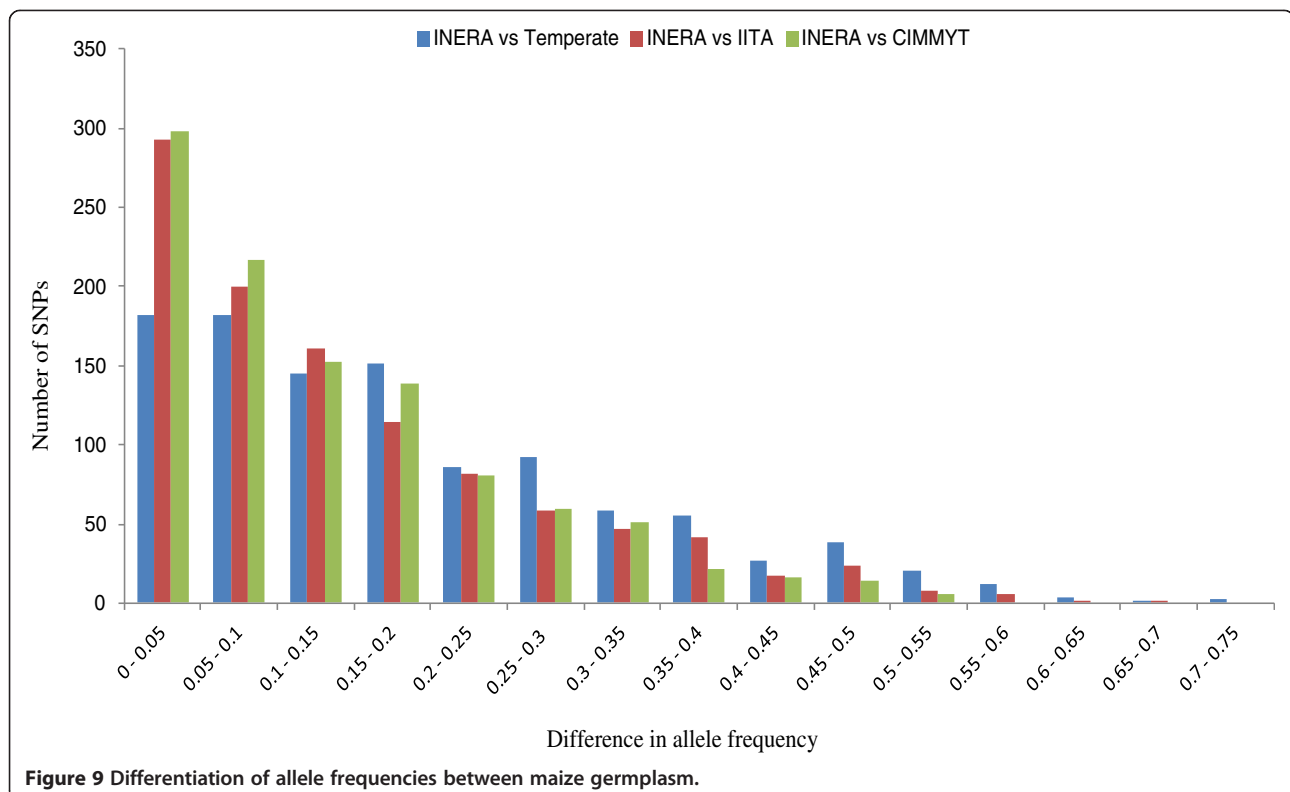


Figure 9 Differentiation of allele frequencies between maize germplasm.

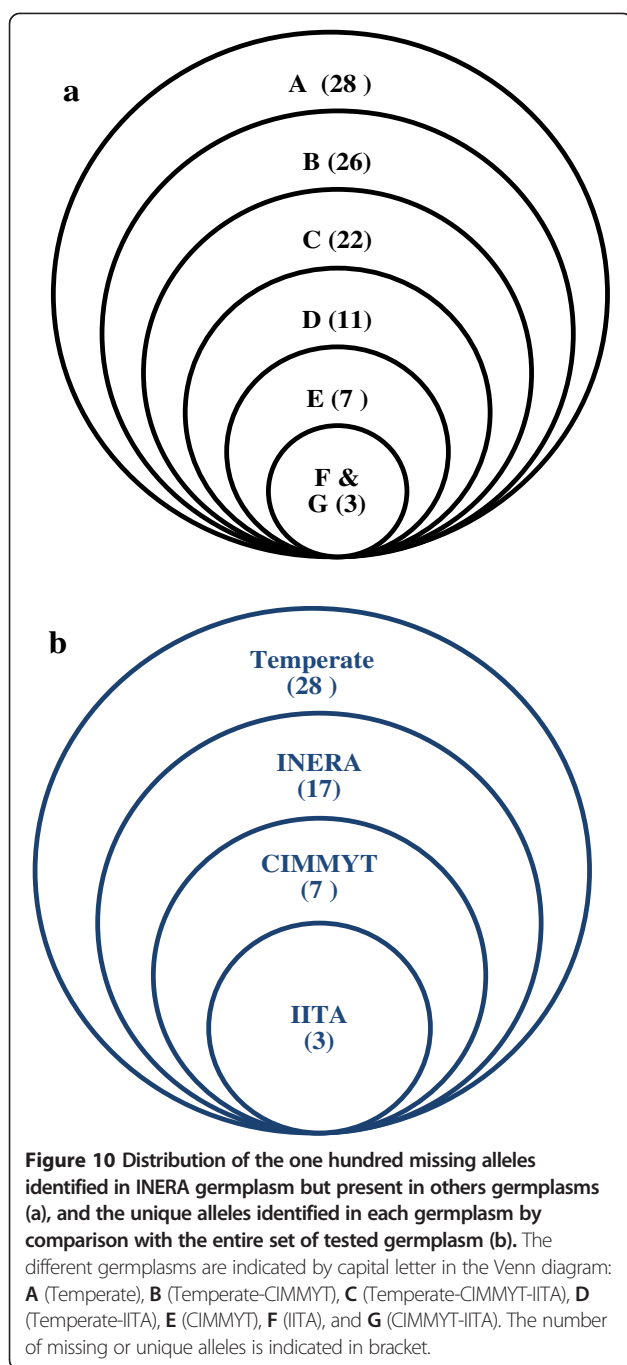
Table 2 Top ten SNPs with significantly different allele frequencies among different germplasm origins

SNP number	SNP name	chr	Position	Allele	Allele frequency		Allele frequency difference
					I	II	
<i>INERA (I) versus Temperate (II)</i>							
54	PZB01062.3	1	56846728	A	0.25	0.8	0.55
70	PHM12633.15	1	1.03E + 08	A	0.99	0.4	0.59
187	PZA00172.12	2	4177515	A	0.14	0.8	0.66
245	PZA00637.6	2	1.7E + 08	A	0.35	0.95	0.6
282	PZA01352.5	2	2.26E + 08	A	0.86	0.15	0.71
310	PZA00297.2	3	39992968	C	0.24	0.78	0.54
594	PZA00643.13	5	91096945	G	0.91	0.2	0.71
779	PZA02854.13	7	1.38E + 08	A	0.32	0.9	0.58
902	PHM11946.19	9	9886093	A	0.26	0.9	0.64
<i>INERA (I) versus CIMMYT (II)</i>							
234	PZA01537.2	2	1.51E + 08	A	0.78	0.27	0.51
254	PZA00824.2	2	1.94E + 08	A	0.92	0.43	0.5
498	PZA02479.1	4	2.18E + 08	A	0.57	0.09	0.48
628	PHM532.23	5	1.93E + 08	A	0.98	0.47	0.51
671	PHM2551.31	6	85125455	A	0.68	0.2	0.48
689	PZA01591.1	6	1.25E + 08	A	0.08	0.57	0.49
871	PHM12749.13	8	1.55E + 08	C	0.12	0.59	0.48
876	PZA00838.2	8	1.59E + 08	A	0.15	0.67	0.51
1034	PZA01073.1	10	1.45E + 08	A	0.36	0.93	0.57
1046	PZA00311.5	-	-	A	0.08	0.59	0.51
<i>INERA (I) versus IITA (II)</i>							
344	PZB02179.1	3	1.58E + 08	A	0.8	0.27	0.53
404	PHM2423.33	3	2.28E + 08	A	0.25	0.85	0.6
469	PZA03116.1	4	1.66E + 08	A	0.71	0.17	0.54
522	PZA01887.1	5	656148	A	0.05	0.6	0.55
865	PZB01454.1	8	1.46E + 08	A	0.19	0.75	0.56
998	PZB00409.1	10	84002430	A	0.13	0.69	0.56
1002	PZA02398.2	10	99471436	A	0.06	0.73	0.67
1033	PHM5435.25	10	1.44E + 08	A	0.25	0.82	0.57
1036	PZA01001.2	10	1.47E + 08	A	0.33	0.86	0.53
1055	PZA02474.1	-	-	A	0.69	0.07	0.62

Chr = chromosome.

populations is largely consistent with the pedigree information [5,10]. Comparisons of the different multivariate analyses revealed high consistency among the PCA and model-based population methods in terms of the number of groups and members of each group. However, cluster analysis showed low concordance with the other methods in assigning genotypes into their respective groups, similar results were reported by Semagn et al. [10]. This could be explained by the fact that, in cluster analysis, different combinations of genetic distance/similarity matrix and clustering algorithms can give rise to different groups.

Even a single distance matrix and a single clustering algorithm may produce several alternative clusters that often create ambiguity in selecting the best one. PCA produces 2 or 3-dimensional scatter plots of the samples in which geometrical distances among samples in the plot reflect the genetic distances among them with a minimum of distortion and ambiguity compared to cluster analysis [26]. Therefore, our population differentiation was based on PCA and population structure analysis which are more reliable than the NJ-clustering. Population structure grouping and different pairwise PCAs between different sets of



inbred lines led to the identification of 4 distinct groups. All the temperate lines, 19% of INERA and 13% of IITA lines were well differentiated, CIMMYT lines could not be divided into groups with significant genetic differences. This corroborates the results of previous studies that showed no clear grouping in CIMMYT germplasm. Although 394 maize lines from CIMMYT's global maize breeding programs were tested in the study of Lu et al. [9], the lines were not differentiated into groups thus supporting the previous reports that no clear clustering or

heterotic patterns could be identified in either the CIMMYT lowland tropical maize [7] or subtropical, tropical mid-altitude and highland maize populations [8]. A set of INERA, CIMMYT and IITA lines, representing 61% of the 96 lines included in this study, were not separated based on origin. All the INERA maize lines derived from ESPOIR were grouped with 88% and 53% of CIMMYT and IITA lines respectively, in PCA. These maize lines might share a common genetic background. The population source of ESPOIR was population 66 developed by CIMMYT in collaboration with IITA. Although the three breeding programs share common germplasm, results of this study identified subgroups between INERA and IITA germplasm with a large genetic differentiation which was not observed in the INERA and CIMMYT germplasms. The genetic distance observed between maize inbred lines from IITA and a national breeding program (the Cameroon Institute of Agronomic Research) has been reported [27]. The same authors suggested that maize breeding programs isolated in space can play a significant role in generating divergent inbred lines. A clear separation between temperate and INERA lines was observed in this study thus confirming the results of a previous study [9] on genetic difference between temperate and tropical germplasm. Genetic distance and kinship analysis showed that the lines tested in this study are distantly related, with only 0.61% the pairwise comparison of the 96 lines falling within a genetic distance less than 0.1. In addition, 61.3% of the pairwise kinship estimates had a value of zero. The result on kinship coefficient estimation is a little lower than that of Hao et al. [23] who reported pairwise kinship values close to zero for about 66.6% of 80 maize inbred lines. The lack of redundant lines among the germplasm suggests that each line is probably contributing new alleles to a breeding program [10]. This is supported by the identification of missing alleles or unique alleles and significant allelic frequency differences among the germplasm collections studied. The higher rate of alleles in temperate germplasm but missing in a tropical germplasm in the present study has been previously reported [9]. As heterotic group assignment is made based on combining ability from diallel or line by tester experiments, several authors have suggested the use of molecular markers in heterotic grouping [28-30]. In this study, cluster analysis separated the testers of well-known heterotic groups into different clusters. In addition, population structure analysis separated some testers into the groups. However, assigning lines into heterotic groups based on this result might not be consistent with field experiments. Semagn et al. [10] used three multivariate analysis (population structure, NJ-clustering and PCA) to separate 220 CIMMYT lines into heterotic groups A and B, but the SNP makers did not reveal clear population structure and genetic differentiation for most inbred lines in heterotic groups A and B, as defined by CIMMYT breeders.

Genetic diversity in maize germplasm in Burkina Faso

There are few studies on molecular characterization of maize germplasm from Burkina Faso. Previous studies (unpublished) characterized a collection of landraces and improved varieties using enzymatic and SSR markers and showed the influence of flux of genes in maize diversity of Burkina Faso.

Hybrid maize programs were initiated at INERA but little progress has been made. The current hybrid breeding program uses improved OPVs, adapted to local environments and adopted by farmers as source populations for extraction of inbred lines. FBC6 and ESPOIR are the main sources from which the available advanced inbred lines were extracted. The present study identified genetic variation and subgroups among INERA inbred lines. Model-based population structure and PCA of the 54 INERA maize lines, extracted from the two different sources, included in this study showed a separation of the lines into three distinct subgroups and a mixed group, which is believed to include recombinant lines since the two parental sources had a common parent. The extraction of inbred lines from FBC6, which was developed by mixing many different varieties, might have resulted in the separation of the lines into genetically distinct subgroups. In contrast, the lines extracted from ESPOIR did not show any major differentiation suggesting that the separation of lines from FBC6 could be explained by the presence in its genetic background of genes from diverse geographic origins. Lines extracted from FBC6 and tested in this study consisted of three groups. Two of the three groups included lines that are closely related in terms of the pedigree. In other studies of tropical maize lines considered extremely diverse, SSR marker variation did not provide any evidence of population structure other than among individuals closely related by pedigree [5,31]. It has been suggested that relatedness among highly diverse maize lines is difficult to measure accurately regardless of the marker system [32].

The present study showed that INERA inbred lines are fixed (lower level of heterozygosity) and exhibited an amount of genetic diversity between different lines. This makes them a valuable source for association mapping studies. Allelic frequency differences observed between INERA and temperate, CIMMYT, and IITA lines, together with unique alleles identified within each germplasm set, suggests that a mutual improvement between INERA and each of temperate, CIMMYT and IITA sets of germplasm is possible.

Conclusions

The present study investigated the genetic diversity among maize inbred lines developed at INERA and the relationship between these lines and temperate lines, and CIMMYT and IITA lines. The 1057 informative SNP markers revealed genetic variation among the inbred lines from different sources

as well as between the INERA germplasm set. Two subsets of INERA germplasm included in this study were divergent but there was also a mixed group that presumably share a common genetic background. There appeared to be substantial progress in national program inbred line development as revealed by the low level of heterozygosity and the uniqueness of the majority of the lines. Principal component analysis showed a genetic differentiation between INERA and temperate and IITA lines but not with CIMMYT lines. However, the unique alleles identified within each set of lines suggest that the exotic lines can provide new desirable alleles for local lines. This study has confirmed a set of SNPs previously reported by Lu et al. [9] and Semagn et al. [10] which can provide good results at low cost in genetic characterization of tropical maize germplasm.

Availability of supporting data

The data sets supporting the results of this article are included within the article and its additional files.

Additional files

Additional file 1: Table S1. Summary of the 100 maize inbred lines.

Additional file 2: Table S2. Summary statistics for the 1057 informative SNP markers identified from the 1151-SNPs.

Additional file 3: Table S3. Inbred lines with their proportional memberships in the model-based subgroups determined by structure.

Additional file 4: Table S4. Markers with missing alleles identified in INERA collection compared with CIMMYT, IITA and Temperate germplasm collections.

Abbreviations

CIMMYT: International Maize and Wheat Improvement Centre; CIRAD: Agricultural Research Centre for International Development; GCP: Generation Challenge Program; IITA: International Institute of Tropical Agriculture; INERA: Institute of Environment and Agricultural Research; INRA: National Institute for Agricultural Research; IRAT: Institute for Research in Tropical Agriculture; KASP: Kompetitive Allele Specific PCR; MAF: Minor Allele Frequency; OPVs: Open Pollinated Varieties; PCA: Principal Component Analysis; PIC: Polymorphism Information Content; RFLP: Restriction fragment length polymorphism; SNP: Single Nucleotide Polymorphism; SSR: Simple Sequence Repeats.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AD selected the inbred lines and was responsible for data analysis and writing the manuscript; JS assisted in selection of the inbred lines and conception of the study. SEM contributed to the data analysis. VG and EYD participated in the design of the study and interpretation of data. All authors have made their contribution in editing the manuscript and approved the final version.

Acknowledgements

The provision of maize germplasm from CIMMYT/Zimbabwe, IITA and INERA is gratefully acknowledged. The authors are grateful to Generation Challenges Program (GCP) for funding the genotyping service. This work was carried as part of a Ph.D research that has been supported by Alliance for Green Revolution in Africa (AGRA) and West Africa Centre for Crop Improvement (WACCI).

Author details

¹Institute of Environment and Agricultural Research (INERA), BP 910, Bobo-Dioulasso, Burkina Faso. ²Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853, USA. ³Department of Plant Breeding and Genetics, 520 Bradfield Hall, Cornell University, Ithaca, NY 14850, USA. ⁴West Africa Centre for Crop Improvement (WACCI), University of Ghana, BMP 30, Accra, Legon, Ghana.

Received: 30 April 2014 Accepted: 5 November 2014

Published online: 25 November 2014

References

- Nassiry MR, Javanmard A, Tohidi R: Application of statistical procedures for analysis of genetic diversity in domestic animal populations. *AJAVS* 2009, **4**:136.
- Smith JSC: Genetic variability within US hybrid maize: multi-variate analysis of isozyme data. *Crop Sci* 1984, **24**:1041–1046.
- Barrett BA, Kidwell KK: AFLP-based genetic diversity assessment among wheat cultivars from the Pacific Northwest. *Crop Sci* 1998, **38**:1261–1271.
- Thompson JA, Nelson RL: Utilization of diverse germplasm for soybean yield improvement. *Crop Sci* 1998, **38**:1362–1368.
- Warburton ML, Ribaut JM, Franco J, Crossa J, Dubreuil P, Betran FJ: Genetic characterization of 218 elite CIMMYT inbred maize lines using RFLP markers. *Euphytica* 2005, **142**:97–106.
- Liu KJ, Goodman M, Muse S, Smith JS, Buckler E, Doebley J: Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* 2003, **165**:2117–2128.
- Xia XC, Reif JC, Hoisington DA, Melchinger AE, Frisch M, Warburton ML: Genetic diversity among CIMMYT maize inbred lines investigated with SSR markers: I. Lowland tropical maize. *Crop Sci* 2004, **44**:2230–2237.
- Xia XC, Reif JC, Melchinger AE, Frisch M, Hoisington DA, Beck D, Pixley K, Warburton ML: Genetic diversity among CIMMYT maize inbred lines investigated with SSR markers: II. Subtropical, tropical midaltitude, and highland maize inbred lines and their relationships with elite US and European maize. *Crop Sci* 2005, **45**:2573–2582.
- Lu Y, Yan J, Guimaraes GT, Taba S, Hao Z, Gao S, Chen S, Li J, Zhang S, Vivek BS, Magorokosho C, Mugo S, Makumbi D, Parentoni SN, Shah T, Rong T, Crouch JH, Xu Y: Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. *Theor Appl Genet* 2009, **120**:93–115.
- Semagn K, Magorokosho C, Bindiganavile SV, Makumbi D, Beyene Y, Mugo S, Prasanna BM, Warburton ML: Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single nucleotide polymorphic markers. *BMC Genomics* 2012, **13**(1):113.
- Liu K, Muse SV: PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 2005, **21**:2128–2129.
- Botstein DR, White RL, Skolnick M, Davis RW: Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 1980, **32**:314–333.
- Fisher RA: On the interpretation of χ^2 from contingency tables, and the calculation of P. *J R Stat Soc* 1922, **85**:87–94.
- Rogers JS: Measures of genetic similarity and genetic distance. In *Studies in Genetics*, Volume 7213. Texas: Univ. Texas Publ; 1972:145–153.
- Hardy OJ, Vekemans X: SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol Ecol Notes* 2002, **2**(4):618–620.
- Loiselle BA, Sork VL, Nason J, Graham C: Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *Am J Bot* 1995, **82**(11):1420–1425.
- Saitou N, Nei M: The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987, **4**:406–425.
- Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011, **28**:2731–2739.
- Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES: TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 2007, **23**:2633–2635.
- Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. *Genetics* 2000, **155**:945–959.
- Falush D, Stephens M, Pritchard JK: Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 2003, **164**:1567–1587.
- Evanno G, Regnaut S, Goudet J: Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 2005, **14**(8):2611–2620.
- Hao Z, Li X, Xie C, Weng J, Li M, Zhang D, Liang X, Liu L, Liu S, Zhang S: Identification of functional genetic variations underlying drought tolerance in maize using SNP markers. *J Int Plant Biol* 2011, **53**(8):641–652.
- Yan JB, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J: Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS One* 2009, **4**:e8451.
- Patterson NJ, Price AL, Reich D: Population structure and eigenanalysis. *PLoS Genet* 2006, **2**:e190.
- Karp A, Kresovich S, Bhat KV, Ayad WG, Hodgkin T: *Molecular Tools in Plant Genetic Resources Conservation: A Guide to the Technologies*. Rome: IPGRI Technical Bulletin No. 2; 1997.
- Menkir A, Melake-Berhan A, The C, Ingelbrecht I, Adepoju A: Grouping of tropical mid-altitude maize inbred lines on the basis of yield data and molecular markers. *Theor Appl Genet* 2004, **108**:1582–1590.
- Flint-Garcia SA, Buckler ES, Tiffin P, Ersoz E, Springer NM: Heterosis is prevalent for multiple traits in diverse maize germplasm. *PLoS One* 2009, **4**:e7433.
- Reif JC, Melchinger AE, Xia XC, Warburton ML, Hoisington DA, Vasal SK, Srinivasanb G, Bohna M, Frisch M: Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. *Crop Sci* 2003, **43**:1275–1282.
- Reif JC, Melchinger AE, Xia XC, Warburton ML, Hoisington DA, Vasal SK, Srinivasanb G, Bohna M, Frisch M: Use of SSRs for establishing heterotic groups in subtropical maize. *Theor Appl Genet* 2003, **107**:947–957.
- Warburton ML, Xia XC, Crossa J, Franco J, Melchinger AE, Frisch M, Bohn M, Hoisington DA: Genetic characterization of CIMMYT maize inbred lines and open pollinated populations using large scale fingerprinting methods. *Crop Sci* 2002, **42**:1832–1840.
- Hamblin MT, Warburton ML, Buckler ES: Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. *PLoS One* 2007, **12**:e1367.

doi:10.1186/s12863-014-0127-2

Cite this article as: Dao et al.: Genetic diversity among INERA maize inbred lines with single nucleotide polymorphism (SNP) markers and their relationship with CIMMYT, IITA, and temperate lines. *BMC Genetics* 2014 **15**:127.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

