GENOMIC SIGNATURE OF ADAPTATION IN MAIZE LANDRACES

Laura Natalia González García¹, Juan Castro, Giovanna Danies¹, Silvia Restrepo¹, Sarah Hearne²
and Charles Chen³*

¹Department of Biological Sciences, Universidad de los Andes, Bogotá, Colombia
²International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico
³Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK, USA

*Corresponding author. Mailing address: 246B Noble Research Center, Stillwater, OK, USA. Email: charles.chen@okstate.edu
The tremendous amount of genetic diversity in landraces has been the building blocks for the success in breeding maize as one the most important food resources. The extremely broad adaptability shown during maize domestication, growing in areas ranging from almost sea level to ~2,800 meter highland fields, is the result of the interplay between genomic variation and natural selection. Genetic characterization of maize landraces does not only provide the gateway to understand important genetic factors in adaptation, but also help to uncover these under-used, yet desirable genetic variants for the anticipated agricultural and climate challenge. Using next-generation sequencing technologies, this study is aimed to understand the genomic composition of maize landraces, identify regions in the genome responsible for genetic adaptation during domestication; and, finally, to mine pre-adapted, breeding potential genes according to mega-environment classification of maize landrace accessions. Admixture analyses and relatedness networks were performed to assess the genomic relationship of landraces. Maize landraces population structure was correlated with geographical and environmental variables. Highland landraces in Mexico were evidently differentiated from lowland accessions; Caribbean accessions and South American collection both demonstrated its unique genomic composition. Though genomically diverse, large proportion of landrace accessions can be related with as little as four landraces based on the networks constructed from shared identical-by-descent (IBD). In addition to their agronomical significance, the results of this research suggest future conservation efforts and as well advocate the importance of uncovering these under-utilized, breed-able genetic variations from germplasm banks. Genetic loci responsible for population differentiation were related with flowering time variation, response to UV-light; development, response to pathogens, and water deprivation genes were also related with Highland versus lowland differentiation. Comparisons
involving South American landraces, were significantly enriched in genes related with circadian cycle and responses to bacteria. In addition to flowering time variation, differentiation of SNP allele frequency between lowland races and US derived accession highlighted the genes related with root hair development and lipid storage. And finally, from GO term enrichment analysis, seed dormancy and epigenetic development play an important role in genomic differentiation between dry lowland and wet lowland maize landraces.

Key words: maize landraces, genetic adaptation, genotype-by-sequencing, admixture analysis, IBD and genetic networks.
Maize (*Zea mays* L.), one of the most important crop in the world, is the main source of human food, animal feed, and raw material for some industrial processes (FAOSTAT, 2014). Furthermore, maize is a significant model plant for the scientific community to study phenomena such as hybrid vigor, genome evolution, and many other important biological processes. Maize has been one of the most-studied plants and great amounts of genomic and genetic information has been generated and made publicly available (Bonavia 2008). The complexity of the maize genome and the substantial amount of genetic diversity in landraces have been the building blocks for the success in breeding (Romay et al., 2013). Genetic variation and selection through domestication have played an important role in the broad adaptability of maize, allowing it to grow in areas ranging from almost sea level (lowland fields) to ~ 2,800 meters above sea level (highland fields) (Bracco et al. 2012).

The origin of maize has given rise to a large number of controversies. However, all specialists agree that maize originated in Mesoamerica (Bonavia 2008). Two hypotheses to explain domestication have arisen. The first one suggests a single domestication event, which took place in Southern Mexico from Teosinte (*Z. mays* subsp *parviglumis* or *Z. mays* subsp *mexicana*) between 6000 and 9000 years ago (Matsuoka et al., 2002) and then taken to South America. The second hypothesis, suggests several independent centers of domestication including the Mesoamerican highlands and the highlands of Peru. This hypothesis further implies the translocation of wild maize from Mexico to Peru before the domestication process (Mangelsdorf 1974). Most ancient corn from Mexico was pod corn and popcorn. There were four original races in Mexico called Palomero toluqueño, Nal-tel, Arrocillo Amarillo, and Chapalote. The introgression of Teosinte genes and the selection process originated the pre-historic races, such as Conico, Tehua, Zapalote, etc. The posterior breeding of those races with Teosinte germplasm and the influx of pre-colombian exotic varieties from the South originated the modern incipient races such as Celaya and Conico Norteño.
The process of artificial selection was carried out in two phases, domestication and improvement (Yamasaki et al. 2007). The first phase involved the domestication of traits such as the sugar pith, starch, oil, and protein content, the kernel size, number of lines and rows, etc. selected by farmers (indigenous). This gave rise to the appearance of landraces (Smalley and Blake, 2003; Jaenicke-Deprés and Smith, 2006). The second phase involved the improvement of maize by selecting for traits such as provitamin A, through plant breeding programs (Owens et al. 2014). These traits confer an additional value to maize kernels contributing with food security policies.

Agriculture is currently facing several challenges, including the rapid human population growth, climate change, and the need to increase food production while reducing the environmental impact (Romay et al., 2013). Jones & Thornton (2003) through weather simulations showed the potential impact of climate change in maize production in the year 2055 in Africa and Latin America. They indicated an overall reduction of 10% in maize production and furthermore, they showed an increase in yield production, mainly in highland and mesic environments. The genetic diversity of landraces preserved in germplasm banks as well as the understanding of gene flow are useful resources for both geneticists and breeders (Flint-Garcia et al., 2005; Liu et al., 2003). However, the genetic differences among landraces, the genetic components driving the domestication process, and the historical pattern of maize dispersion and adaptation are still unclear.

Several markers have been used to study the populations of maize. Among these are: chromosome knobs (Brown 1949); quantitative trait loci (QTLs); single genes including the *tunicate* (*Tu*) and the *tunicate inhibitor* (*Ti*) (Mongelsdorf 1974) genes, the *alcohol dehydrogenase* (*adh2*) (Goloubinoff 1993), *teosinte branched 1* (*tbl1*) (Dobley 2004), and *teosinte glume architecture 1* (*tga1*) (Wang et al. 2005) genes; and single sequence repeats (SSRs) (Vigoroux et al. 2002; Vigoroux et al. 2008). These markers have provided insights into the origin and domestication
process of maize. Nowadays, large-scale genomic data enable researchers to elucidate the evolutionary processes and optimize the characterization, discovery, and use of functional genetic variation (Romay et al., 2013). Technologies such as genotyping-by-sequencing (GBS) and the Diversity Array Technology (DArT), have been developed to reduce genome complexity for SNP genotyping analyses of large samples (Elshire et al., 2011; Wenzl et al., 2004). These technologies allow the analyses and identification of genomic regions and polymorphisms associated to adaptation.

The aim of this study is to understand the genomic composition of the maize breeder’s core collection. More specifically we aimed to identify genetic regions subject to selection during the maize domestication process and genomic differentiation between geographical and environmental defined groups of landraces. This knowledge will help worldwide maize breeders and CIMMYT to improve maize breeding programs including traits that confer adaptability to the future climatic changes.

Materials and Methods

Germplasm composition for this study

The “Breeder’s Core Collection”, composed of 4,293 maize landraces from the CIMMYT maize seed bank accessions, was initiated with the goal of improving and monitoring agricultural performance of maize landraces (Taba 2005). Accessions of this collection were selected based on a multivariate clustering analysis performed on agro-morphological data. The top 20% of the most representative, non-overlapping agro-morphological clusters were selected according to their combining ability with both heterotic and non-heterotic elite accessions (Taba 2005).

Information, including phenotypes, SNP genotypes, as well as geographic coordinates of collection sites for the Breeder’s Core Collection can be obtained via a research agreement with
In this study, in total 133 maize landraces were included, with the majority collected from Mexico (2,594 accessions, 60%). These Mexican accessions corresponded to 41 different landraces (for more details refer to Table S1). Non-Mexican accessions (Figure 1) were collected from 35 countries. In detail, a total of 273 accessions were collected from the Caribbean Islands (Antigua and Barbuda, Barbados, British Virgin Islands, Cuba, Dominican Republic, Grenada, Guadalupe, Haiti, Jamaica, Martinique, Puerto Rico, Saint Vincent, Trinidad and Tobago, and Virgin Islands), 1,139 from South America (Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, French Guiana, Guyana, Paraguay, Peru, Suriname, Uruguay, and Venezuela), and 322 from Central America (Costa Rica, Guatemala, Honduras, Nicaragua, Panama, and El Salvador) (Figure 1). Relatively larger in proportion, 569 accessions came from Brazil, 185 from Uruguay, 137 from Argentina, 122 from Guatemala, 118 from Venezuela and 75 from Cuba. No accessions from the United States were included in this analysis.

Using 19 bioclimatic variables, Mexican and Central American accessions were classified into mega-environments such as wet lowland, dry lowland, lowland, wet lower mid-altitude, wet upper mid-altitude, dry mid-altitude, mid-altitude, and highland. Accessions collected between 0 and 1,100 meters above sea level (m.a.s.l) from either wet or dry lowland mega-environments were further classified as lowland. Accessions collected between 900 and 1,900 m.a.s.l from wet or dry, upper or lower mid-altitude were classified as middle altitude. Finally, accessions collected above 1,700 m.a.s.l from highland mega-environments were classified as highland.

Due to the lack of complete information of geographical coordinates and corresponding altitudes, Caribbean and South American accessions were only classified into mega-environments based solely on the bioclimatic variables. According to their collection sites, we categorized the geographic origin of South American landraces as Caribbean accessions, northern South American
(Colombia, Ecuador, French Guiana, Guyana, Peru, Suriname, and Venezuela), and middle and southern South American accessions (Argentina, Bolivia, Brazil, Chile, Paraguay, and Uruguay).

**Genotypic information: GBS and DArT SNP markers**

The genotypic values of each accession in the Breeder’s Core Collection were derived from a single plant. For each accession, kernels were collected, DNA was extracted and sequenced by both GBS (Genotyping-by-Sequencing, Elshire et al 2011) and DArT (Diversity Array Technology) using Illumina HiSeq 2000.

The global build of GBSv2.7 accommodates maize genomic single nucleotide polymorphisms generated from an international collection of over 40,000 maize accessions. Short read sequences of GBSv2,7 were aligned to the current maize B73 reference genome available in the Maize Genetics and Genomics Database (http://www.maizegdb.org/); SNP profile was generated for each of the individuals based on the comparison with B73 reference genotypes as well as the rest of maize diversity. Currently GBSv2.7 has three different versions depending on the interpretation of missing genotypes: (i) original, un-imputed SNP profile with 955,121 SNPs that on average have 65% of missing data without mitochondria, chloroplast, and unassigned SNPs (Hearne et al., 2014a), and (ii) two imputed GBSv2.7 SNP tables, GBSv2.7FILLIN and GBSv2.7Beagle4 that were completed using the FILLIN and the Beagle 4 imputation algorithms, respectively (Hearn, et al. 2014b and 2014c)

DArT SNP markers were produced using two different calling pipelines: referenced with B73 (1st pipeline) and the non-referenced (2nd pipeline). In DArT’s 1st pipeline, short reads were aligned and SNPs were determined considering information that is only present in the maize B73 reference assembly. On the other hand, Illumina short reads which could not be mapped on B73 reference genome were aligned against each other and SNPs were determined when a single
mutation can be defined by directly comparing the aligned sequence tags. DArT 2nd pipeline, non-referenced SNP markers are considered less biased towards B73 genotypes.

Before the cleaning and filtering steps there were in total 43,265 SNPs from the 1st pipeline (referenced pipeline and 31,713 removing unassigned SNPs) and 45,800 SNPs from the 2nd pipeline (DArT non-referenced pipeline).

**Principal component analysis and Admixture analysis**

In order to obtain quality SNP genotypic values, SNPs from DArT’s 1st pipeline and unimputed SNPs from GBS2.7 were merged and then filtered based on the percentage of missing data (all resultant SNPs that had less than 20% missing data). A genomic similarity matrix was then computed using the filtered SNP genotypes as done by Van Randen (2001). A principal component analysis (PCA) of the resultant genomic similarity matrix was done using the FactoMineR package (Lê et al 2008).

SNPs from DArT’s 1st pipeline and unimputed SNPs from GBS2.7 were merged and then filtered based on the percentage of missing data (< 25%). The Hapmap files were converted into plink files (.map, .ped) using both TASSEL (Bradbury et al., 2007) and a custom script. Plink files were then used as the input files for the Admixture analyses (a maximum likelihood multi-locus estimation of individual ancestries) (Alexander, Novembre, & Lange, 2009). Input data were iterated through the Admixture analysis from K=2 to K=15, to obtain the most informative K for further analyses. Genomic admixture between mega-environment classification and landrace groups were visualized with K=3 and K=6.

**Identical-by-Descent analysis**
The IBD (identical by descent) algorithm by PLINK (Purcell et al 2007) package was used to analyze genomic ancestry of maize landrace accessions. In a population, a DNA segment is identical by descent if it has the same ancestral origin in these individuals; it is however identical by state (IBS) because of mutations and recombination (Powell et al 2010). This approach can be used to elucidate demographic history such as bottlenecks and admixture (Gusev et al 2008).

To avoid false positives in genomic ancestry analysis using IBD, accessions that do not have clear racial designation and coordinates of the collection sites were removed. Unimputed GBSv2.7 SNPs and DArT's 1st pipeline SNPs were merged, and filtered for less than 20% of missing data before the analysis. The remaining missing genotypic values in the final SNP dataset were then replaced by the major allele for each position before converting the hapmap file to PLINK format.

Mutual information content is a non-linear correlation measure of the association between two variables by means of their conditional entropy (Hausser & Strimmer, 2008). In order to eliminate redundancy caused by linkage, mutual information (MI) content was calculated for every given pair of SNPs, as done by Lewis et al (2011) in 100,000 bp windows along each chromosome. One SNP from each group sharing MI > 0.5 was kept for IBD analysis.

The IBD calculated for all accessions was used to construct a relatedness network among landraces. First, genomic ancestry of landraces were filtered for IBD values between 0.02 and 0.5, as values larger than 0.5 are the expected values for clonal populations and values below 0.02 indicate that there is no significant relationship among individuals. IBD networks were then done with igraph package in R (Csardi & Nepust 2006). The IBD network was plotted in a map using (South 2011) and Geosphere (Hijmans et al 2015) packages, and colored according to the admixture group of the accessions.

A multilevel community detection algorithm was run to determine the number of communities within the relationship network using igraph package in R (Csardi & Nepust 2006).
The community detection procedure was taken as the following: to begin, each accession (node) in the network was assigned to a community of its own, a locally greedy iteration process was then used to reassign nodes to communities (i.e. each node was moved to the community where it achieved the highest contribution to modularity), and when nodes could no longer be reassigned, each module or community was re-considered a node for the next iteration. The iteration process restarted with the merged nodes, with which merged communities were, again, established. The algorithm stopped either when there was only a single node left or when the modularity could not be increased (Blondel et al, 2008). For each community, the most connected accessions and the highest number of connections were determined.

Allele Frequency Spectrum and Functional Enrichment

All SNPs, GBSv2.7 and DArT 1st pipeline, were classified into two large categories: genetic (SNPs located in coding regions) and intergenic (SNPs located in introns). The classification was done according to the B73 genotype annotation. The annotation file (GFF) used includes genes (Exon, intron, CDS, transcript, 3’UTR, and 5’UTR) and intergenic regions.

Unimputed SNP datasets from the GBSv2.7 and DArT's 1st pipeline, filtered based on 20% of missing data, were merged with redundant sites removed (detail procedure in previous section). Only SNPs within gene coding regions were kept for allele frequency comparisons. The counts of nucleotide bases for each SNP and its allele frequency were calculated for the entire population as well as for each mega-environment and admixture subgroups. In order to identify genomic regions that show significant differentiation of allele frequency between and among subgroups, a Fisher Exact test from the stats package (R Core Team 2013) was conducted using nucleotide counts for the SNP loci. SNPs were ranked according to the p-values and the top 500 SNPs were used for subsequent enrichment analysis (all p-values were under 1e-10).
Genome annotation (available in the Maize Genetics and Genomics Database http://www.maizegdb.org/) was used to assign gene ontological terms to each gene/SNP. Gene ontology (GO) enrichment was conducted using BinGO plugin (Maere et al. 2005) for Cytoscape software (Shannon et al. 2003). Enrichment was carried out using the whole genome annotation as reference/background dataset. A hypergeometric test followed by a Bonferroni (FWER) correction were conducted to assess over-representation. Only biological process GO terms were considered.

A two-dimensional clustering was computed to obtain the groups of GO terms and differentiation between mega-environments using the Heatmap.2 function from the gplots package in R (Warnes, 2015). The pattern of GO enrichment was used to cluster comparisons, constructing a dendrogram of comparisons with the same differentiation. GO terms that showed the same pattern of enrichment among comparisons were further clustered.

Results

Distribution of samples

Geographical distribution of accessions in CIMMYT’s “Breeder’s Core Collection” was mapped according to latitude and longitude coordinates of collection records together with other reference (collection sites, country of origin and elevation). Details are included in Table S1.

The Caribbean group was composed mainly of 16 Haiti Yellow accessions and one Haiti White accession collected in Haiti. In addition, there were also five Argentino accessions from Cuba, 10 Cubano Amarillo accessions (collected from Bolivia, Panama, and Ecuador), 16 Early Caribbean samples from Virgin British Islands, 25 Chandelle accessions from Dominican Republic and Venezuela, and 32 Tuson and 20 Coastal Tropical Flint accessions that came from all islands. Caribbean samples were classified as lowland race as the average altitude was at 114 m.a.s.l.
The northern South American group was composed of 153 accessions that were classified as
Costeño (19, 3.9%), Guaribero (11, 2.25%), Puya (7, 1.4%), Puya grande (10, 2.05%), Cariaco (9, 1.8%), Cateto norte (9, 1.8%), and Chococeno (8, 1.6%). There are also small numbers of Negrito (4, 0.8%) and Cuban Flint and Cubano Amarillo (3, 0.6%) accessions for this northern South American group. According to Vigoroux et al., 2009, all these accessions were classified as the Tropical Lowland cluster. As for our classification using bioclimatic variables, the majority were Dry Lowland and Wet Lowland with an average altitude of 438.7 m.a.s.l.

Middle and southern South American accessions were collected mainly from Uruguay, Argentina, and Brazil. These landrace accessions are mostly Dentado (124, 12.5%), Cateto Sulino (101, 10.2%), Dentado Paulista (97, 9.8%) and Dentado Riograndense (82, 8.3%). Cristalino Colorado, Sapé, Cainga, Cateto Paulista, Tupy, Avati and Cuarenton together composed less than 10% of the Mid-South American accessions. The averaged altitude of these accessions was at 434 meters above sea level.

*Mega-environment adaptation classification*

Mexican and Central American (including Panama, Costa Rica, Nicaragua, Honduras, El Salvador, and Guatemala) accessions were classified into eight different mega-environment groups (Highlands, Wet Upper Mid-Altitudes, Mid-Altitudes, Dry Mid-Altitudes, Wet Lower Mid-Altitudes, Dry Lowlands, Lowlands, and Wet Lowlands; the landrace distribution in Mexico according to their mega-environment classifications is shown in Figure 2). Mexican and Central American Highlands had an average altitude of 2,247 m.a.s.l. There were four Mexican and Central American Middle altitude groups with an average altitude of 1,556 m.a.s.l, Mexican and Central American Wet Lowlands with an average altitude of 292 m.a.s.l., whereas Mexican and Central American Dry Lowlands showed an average altitude of 310 m.a.s.l.
Dry Lowland accessions were distributed across the northern coast (Sonora, Sinaloa, Tamaulipas, and Nuevo Leon) as well as the Yucatan Peninsula in the southeastern part of Mexico. Maize landrace accessions collected from Dry Lowland are mostly Tabloncillo, Tabloncillo perla, Reventador, Onaveno, and Dulcillo de Noroeste. Tabloncillo and Tabloncillo Perla accessions were also found in Wet Lowland and Mid-altitude environments in the Mexican state of Jalisco and Michoacan (Figure 2, Table S1).

Wet Lowland accessions were composed of Salvadoreño, Olotillo, Tepecintle, Nal-tel, Clavillo, Oloton, Zapalote Grande, Zapalote Chico and Vandeño. These can be found along the central and southern coastal regions, such as Chiapas, Veracuz, Nayarit, San Luis Potosí, Guerrero, Oaxaca, Morelos, Jalisco and Hidalgo (Figure 2, Table S1).

Tuxpeño accessions were distributed in both Dry and Wet Lowland mega-environment and also spread into Middle Altitude regions due to its agricultural significance (López-Pereira & Morris, 1994).

In central Mexico (collection sites come from Mexican States Puebla, Mexico, Oaxaca, Tlaxpan, Veracruz, Hidalgo, Chiapas, Michoacan, Chihuahua, Jalisco, and Queretaro), most of the accession collections were grouped as Middle Altitude and Highland Adaptation based on mega-environment variables as well as the altitude of collection sites (Figure 2, Table S1). Accessions that were identified as Highland Adaptation include Conico, Conico Norte, Chalqueño, Elote Conico, Cacahuacintle and Arrocillo Amarillo. Conico, Chalqueño, Cacahuacintle and Arrocillo Amarillo were highland specific accessions. In addition to their highland origins, Conico Norte and Elote Conico were also found in Middle altitude environments.

Typical Middle Altitude accessions were composed of Ancho, Pepitilla, Bolita and Celaya landraces. These accessions were also found in all Mexican mega-environments. The wide spread of these accessions perhaps was due to the broad adaptability of these mid-Altitude accessions.
Central American accessions were also classified into mega-environments; most of them were assigned to Wet Middle Altitude and Wet Lowland environments due to their collection location near the Pacific and Caribbean coasts, with a few exceptions from the mountain regions. For instance, collection records of Nal-tel, Tepecintle and Oloton accessions can be found from Southern Mexico to Costa Rica with a wide range from Wet Lowlands to Highlands. Salvadoreño and Clavillo are Wet Lowlands races; Salvadoreño can be seen from Guatemala to Panama, while Chococeño was collected from Costa Rica to South America.

**Population structure**

PCA showed that the first two components discriminated our accessions, but there were no differences including more principle components (data not shown). These first two components gave us information about the environmental effect over the sample differentiation and adaptation. The first component built a gradient between highland samples and lowland samples, but there was not discrimination among races. The second component separated a group of South American Lowland accessions from Mexican Lowland accessions.

After filtering, in total 51,451 SNPs merged from DArT 1st pipeline and the unimputed GBSv2.7 dataset were used to assess the admixture of maize landrace accessions in Latin America. A total of 4,294 accessions were used for ADMIXTURE analysis. Cross-validation analysis did not show a conclusive minimum value of K (Figure S2); however, based on previous reports and correspondence with reported groups (Vigoroux et al 2008), K=3 and K=6 were selected to proceed.

The mixed ancestry of Latin American maize landrace was apparent. When only three ancestral clusters (K=3) were allowed, maize landraces were clustered with mega-environment classification in large- Highland (majority in red), Lowland (majority in blue), and Caribbean and
South American accessions (composed of green and blue components) (Figure 3A). Though a large proportion of common components could be identified for different mega-environment adaptation groups, none of the landrace accessions showed a complete fixation of genomic component. The admixture plot gradually transits from having a larger proportion of red component in the Mexican Highlands to a mix of blue and red components for the Middle Altitudes, and finally a larger proportion of blue in the Mexican Lowland accessions. In detail, for the Mexican Highland material, 60% of the landrace accessions showed > 75% of red component (top panel in Figure 3A), as opposed to the Mexican Lowland classification where 70% of the Wet Lowland accessions and more than 50% of Dry Lowland accessions were made of mostly blue component (more than 75% of blue; see Figure 3A). Possessing very little Highland red component, Caribbean accessions were also found to have a large proportion of the blue component, suggesting either shared ancestry or parallel genomic adaptation with Mexican Lowland maize (Figure 3A).

ADMIXTURE illustrated a unique green component mixed with Mexican Lowland blue component for South American accessions (Figure 3A). This group of maize landraces was mostly Cateto, Dentado, and Cuarenton, with a number of indigenous ones, such as Cainga, Avati, and Tupy. Indigenous landraces are supposed to show a decreased introgression from US germplasm to Brazilian germplasm (Bonavia 2008). At K=3, different Lowland adaptations (Dry versus Wet Lowland) could not be well distinguished.

When fitting ADMIXTURE with a higher degree of population stratification, the admixture of maize landraces adapted to the Mid-Altitude and Lowlands became visible (Figure 3B, C; Table S2). The admixture plot showed the same transition as K=3 (Figure 3A), from a larger proportion of red component in Highlands to Blue component in Lowlands. However, with a higher degree of admixture, a larger proportion of purple component in Caribbean samples can be seen, as well as
two clearly differentiated South American groups: the northern South American group with majority of orange component and the Southern group with majority of green component.

The yellow component that was not observed using K=3, appears to be common in all mega-environment adaptations when using K=6, likely reflecting the component of the shared genomic ancestry. Illustrated in Figure 3C, the landraces showing the highest amount of yellow were classified into Dry Lowlands adaptation; and their collection locations were mostly found in the northern part of Mexico. The landrace accessions in this Dry Lowland group include Reventador (typical pop corn), Tabloncillo, Tabloncillo perla, Cristalino de Chihuaha, Dulcillo de Noroeste, Ancho and Onaveno landraces (eight row flint maize). The distribution of these races ranged from the Balsas River to northern Mexico, suggesting the mixture of the lowland maize component with the ancient gene pool. Like Reventador, these eight-row maize landraces historically came from the cross between Teosinte and lowland accessions (from Chapalote or Nall tel complex).

Even when K=6, the red component mostly remained for the Chalqueño, Conico, Conico Norte, and Cachhuacintle, the conical shaped, Highland maize (top panel in Figure 3B). Most of these accessions were collected from high altitude (on average 2,145 m.a.s.l.) regions in central States of Mexico, indicating specific genomic components related to Highland mega-environment adaptation and the relative little amount of gene flow between highland accessions and other landraces. In accessions like Ancho and Pepitilla that have spread from Highland into Wet Upper Mid-Altitude, ADMIXTURE recapitulated the introgression of yellow and blue components. While Wet Lowland accessions were mostly Tuxpeño-like and Highland accessions were Conico-like, in our analysis Mid-Altitude landraces were the most genetically diverse maize accessions.

Wet Lowland accessions were mostly found in low altitude regions along the Pacific and Caribbean coasts, as well as some central regions in Mexico (Figures 2, 3C). This group of maize
accessions included samples from Nal-tel, Salvadoreño, Olotillo, Oloton, Clavillo, Dzit-Bacal, Tepecintle, Zapalote Grande, and Zapalote Chico. In Vigoroux et al. (2008) they were also grouped as Wet Mexican Lowland and Central American Wet Lowlands. Among these, Nal-tel is a typical representative of Wet lowland landraces. Regardless of its wide distribution spanning Mexico, Costa Rica, Panama, and El Salvador, Nal-tel accessions found in Lowland environments with an average altitude of 406 m.a.s.l. (primarily in the blue component) showed little genomic composition from other mega-environments. In contrast, Nat-tel accessions collected from Mid-Altitude environments like Celaya and Bolita showed a slightly larger proportion of Highland component (red component).

The blue component was also found as an important fraction of Northern South American samples, such as, Costeño, Cariaco, and Guaribero landraces, which also showed a lower yellow component, and higher purple and orange components. Suggesting this region as a point of diversification or mixture of landraces.

Caribbean accessions formed another clear group of landraces (Haiti, Argentino, Tuson, Cubano, Chandelle, Coastal Tropical Flint, Sant Croix, and Early Caribbean) distributed from northern Venezuela to Cuba. These accessions were composed of more than 70% of the purple component, where Haiti landraces seem to be the typical Caribbean accessions (>90% purple component). There was a gradient of blue component from Venezuela to Haiti, suggesting the spread of the ancestral maize of those races from northern South America to the Caribbean complex. The same admixture of blue and purple components was observed in Suriname and French Guyana accessions (Puya, Puya grande, Cateto Norte, and Negrito landraces); those accessions also showed less than 10% of the orange component. A unique purple component from “Cubano” accessions (Cubano Amarillo, Cubano Colorado, and Cubano Blanco) appeared in a
number of accessions in Bolivia and Paraguay, which could be supportive of the history of
germplasm introduction by agricultural extension before World War II (Vigouroux et al. 2008).

Middle South American accessions (mostly from Brazil) constituted the orange admixture
cluster. Though this study did not include U.S. accessions, evidence suggests that his large orange
component might have been a result of the post-Columbian translocation of southeastern US
germlasm to Brazil and the secondary introgression contributed from the purple component of
“Cubano” accessions. These accessions are dent races like the Dentado Paulista group and the
Dentado Riograndense group. Dentado Rojo, Dentado Blanco and Dentado Riograndense
accessions however were introgressed with a larger proportion of green component from North
Argentinian accessions (Cateto group, mainly).

Southern South American Cateto Sulino, Cristalino Colorado, Cateto Sulino Grosso, and
Cuarenton landraces possessed a unique green component.

**Identical-by-Descent analysis**

After filtering, we identified 1,852 IBD relationships among 656 accessions from 131
landraces, in which IBD values were used to represent the degree of connection in any given two
nodes (accessions) (Figure 4A). Taking all filtered IBD relationships into account, the most
connected nodes were Zapalote Grande from Chiapas (Mexico), a Coastal Tropical Flint accession
from Antigua and Barbuda, a Dente Riograndense Rugoso accession from Brazil and a Tuxpeño
accession from Puebla (Mexico). These four landraces together covered 49.13% of the network, i.e
49% of the edges in the network can connect to these four accessions (Table 1). These accessions
further showed the highest values of centrality as the majority of the shortest paths that connect
only two nodes pass through one of these accessions. The fact that most of the accessions had an
IBD relationship with these four central accessions suggests their pivotal roles in the diversification of maize landraces.

To assess if these relations had any historical significance, we performed a multi-level community analysis. This analysis allowed us to identify particular modules or clusters along the networks. Community analysis yielded seven communities in the network; and when we compared the central node (hub) of the identified communities, each of the four highly connected nodes represented the central node of an individual community (Table 1). Moreover, these communities seem to have a particular composition of races according to ADMIXTURE groups (Figure 4B-G, Table S3). For example, community number 2, which clustered around the Coastal Tropical Flint accessions from Antigua and Barbuda, consisted mostly of Caribbean accessions (Figure 4C). Furthermore, this community shared 364 edges with community 3, and 364 with community 1, suggesting the migration event of maize landraces through the islands, instead of the continental landmass.

Communities 1 and 3 depicted the clustering of the Mexican Lowlands and South American U.S. accessions with Zapalote Grande and Tuxpeño, showing the importance of these two races to the breeding of modern maize (Figures 4B, 4D). Also, despite the lack of US accessions, southern South American accessions, derived from northern U.S. (green panel form ADMIXTURE), showed a high proportion of shared IBD with Mexican accessions, supporting that U.S. accessions were derived from lowland ancestral Mexican landraces. Furthermore, the IBD connections of South American accessions in these two communities are indicative of a secondary translocation from the U.S. to South America.

Community 6 largely consisted of Mexican lowland, Mexican highlands, Brazilian and Caribbean accessions. This shared IBD in community 6 might be supportive of the translocation of
maize from Cuba to mid South America, as well as its later introgression to southern Brazilian germplasm (Rio grande do Sul region).

Lastly, only four landraces were found in the community 5. However, these were connected with high levels of IBD, due to their background and to the proximity of their collection sites. Admixture panels suggested misclassification of those races and therefore, this community was discarded from analysis.

**SNP Annotation**

Overall, GBSv2.7 and DArT SNPs had similar proportion of SNPs located in genetic regions; this proportion of genetic DNA was also consistent among 10 chromosomes- 15% of the SNPs located within the genes (Figure S3) with higher concentration in sub-telomeric regions than in pericentromeric regions (Figure S4). This corresponded to the percentage of genes found in the genome (Schnable et al 2009).

**Allele Frequency Spectrum and Functional Enrichment**

The allele frequency (AF) spectrum was build based on 52,000 SNPs from the DArT 1st pipeline and GBSv2.7 unimputed datasets. The SNP variation was higher towards both ends of the maize chromosomes (Figure 5A). Eighteen different comparisons were made between mega-environment adaptations. Resulted from the Fisher’s exact test, AF spectrum showed highly differentiated allele frequencies between Highland accessions and all other accessions (Figure 5).

In general, most of the highly differentiated SNPs can be found in sub-telemetric regions (Figure 5). Centromeric regions are in theory lower in diversity, thus the low differentiation in allele frequencies. The highest differences in allele frequencies can also be found in the regions that are responsible for flowering time variation (Buckler et al 2009), for example the inversion in the
Comparisons involving highland accessions showed the typical pattern related with flowering time (Figure 5S). Comparisons involving highland accessions showed differentiation in the centromeric region of Chromosomes 1, 3, 4, 5, 8, and 9. However, we observed a number SNPs showing high differentiations in allele frequency when comparing Highlands with Lowlands and Highlands with Brazilian accessions that were not found when comparing Highlands with Caribbean accessions. South American US derived samples showed differences to other races in the middle of chromosome 3 and 10. Comparisons between Caribbean accessions and other populations commonly found regions at the middle of the chromosome 1, 2, 5, 6, and 7 with highly differentiated allele frequencies.

Top 500 SNPs based on the p-values of Fisher exact test were selected for GO term enrichment analysis. Only terms related with Biological Process were taken into consideration. Our cluster analysis with GO enrichment showed that pairwise comparisons involving Highland accessions vs Lowland and Brazilian accessions were grouped together, and comparisons involving Caribbean samples vs others were grouped together in another cluster. Comparisons that included lowland accessions were not grouped, suggesting interesting adaptations in Highland material and Caribbean material (Figure 6A). That clustering is the same as shown by allele frequencies differentiation.

A few GO terms were commonly found responsible for most of the population differentiation, such as development, flowering cycle, response to pathogens, response to UV-light and water deprivation. In addition to these common terms, seed dormancy and epigenetic development play an important role in the genomic differentiation between Wet Lowland and Dry Lowland maize (Figure 6C).
Terms related with light signaling, flowering cycle and development, response to blue light, and pathogens were enriched in the comparison between Highlands and Lowlands. The comparisons between lowlands and Brazilian samples also clustered together but their differences were mainly related to pigments, the circadian cycle, response to bacteria, response to UV-C and flowering cycle. The GO terms underlying the allele frequency differentiation between Lowland and U.S. derived accessions were on the genes involved in DNA repair, root hair development, lipid storage, response to UV, flowering cycle and development. Allele frequencies changes between Brazilian and US accessions were associated with jasmonic acid pathway (Figure 6C).

All comparison, with an exception in the comparison between the Caribbean and South American U.S. derived accessions, showed the importance of genes related determination of floral meristem to allele frequency differentiation. Similarly, genes regarding the regulation of inflorescence meristem growth seems responsible for allele frequency differentiation for all comparison, but the comparison between Caribbean and Brazilian accessions (Figure 6C).

Discussion

According to the general estimate of the physiological behavior of wild maize under cultivation, researches concluded that this cereal must be of tropical or subtropical origin, since it tolerates almost any amount of heat but not of cold; that it must have come from an open region, since it requires direct sunlight and not shade; neither drought resistance nor supporting wet soils, it is, however, adapted to a change of seasons, with a rapid growth until maturity in order to finish its cycle in less than a year (Grobman 2004). Narrow ecological requirements of wild maize had to evolve in order to adapt maize to different locations with opposite environments, such as, highland and lowland elevations, wet or dry seasons, light incidence according to latitude and longitude. Population structure results showed the differentiated genomic composition related to adaptation of
landraces according to their environmental conditions, and we could find the signature/admixture of each group. Six components of admixture were found and corresponded to highland vs lowland adaptation, Caribbean complex differentiation, and South American translocations from U.S. materials. Accessions from Peru and United States were not included in the panel due to their absence in the Breeder’s Core Collection.

Our population structure results supported the idea of two different domestication origins of maize (Goodman and Bird 1977), as they separated clearly the races of Brazil and Caribbean complex, from The Mexican and Central American samples. These results were also supported by SSR data (Vigouroux et al 2008). A translocation of maize from South American in pre-historic time has been identified (Wellhousen et al 1951), followed by Teosinte introgression (Grobman 2004).

Primitive maize landraces, Chapalote (Mexican western) and Nal-tel (Yucatan), are closely related to “eight row” landraces and are similar to the southwestern U.S. samples and some Peruvian races in Los Gavilanes cave (Sauer 1993; Grobman 2008). This suggests the pre-historic translocation of wild maize and there is clear evidence that plants were moved from South America to Mesoamerica between the fourth and fifth millenium BC (e.g. *Manihot esculenta*, domesticated as cassava), but the reverse movement has not been proven (Pope et al 2001). This signature of genome admixture still remains today as a resource of maize genetic diversity preserved from pre-historic time, and was clear as the yellow component in admixture plots. However, Peruvian accessions are crucial to solve this controversy.

Northern and Southern South American groups (Green and orange in this study) were previously documented by Vigouroux et al. (2008) using 96 microsatellite markers. Phylogeny from individual plants suggests that maize spread into northern South American from Mexico, races of lowland middle South American were derived from Andes, as well as post-Columbian translocation of Southeastern U.S. germplasm to Brazil and the subsequent introgression result in the clade of
middle South America and Southeast U.S (Figure 4A in Vigouroux et al. 2008). U.S. derived group might be due to the translocation after the European colonization (Timothy et al., 1961).

Caribbean accessions had a majority of purple component and a little blue component. Placing Northern South America (Panama, Colombia, and Venezuela) as a maize spanning center, evidence supported the spread of maize from Venezuela to Trinidad and Tobago, and from those islands to the entire Caribbean complex. Cuban races were the most pure landraces, as the last colonized island, showing more than 90% of purple component in our admixture plots. A few Tuxpeño accessions collected in Cuba and Haiti are product of later translocations from Yucatan, Mexico (Smith & Beltran 2004).

Community and admixture analyses showed that there is a major population from where most of the landraces are derived. Tuxpeño, which is largely predominant on Mexico, could have migrated to Brazil Uruguay and Argentina and derived into Dentado landraces (Dentado, Dentado blanco, Dentado blanco Paulista, Dentado semi-riograndense, Dente riograndense rugoso), which are largely predominant in the orange population, the communities 1 and 3 (Figure 1, 4C and Table S2). Zapalote chico and Zapalote grande are representative races of community 6, this community is however composed by a large fraction of the Blue population and even fractions of the other populations. This could indicate that the 6th community has a big role in the history of maize landraces, and that the Zapalote landraces later derived into Tuxpeño, which could have come form, a mixture of Teocintle and Zapalote chico. Thus the genetic admixture is telling a similar history to the one presented by historians and anthropologists (Bonavia 2008).

Looking into functionality of genomic differentiation, flowering cycle differences had been documented in maize domestication and improvement processes (Hufford et al 2012). Flowering cycle is related with the circadian cycle and both are dependent on the light intensity. Changes between tropical and subtropical regions were evidenced in comparisons including Caribbean
samples and Southern South American samples. Jasmonic acid and Salicylic acid pathways are related to photomorphogenesis and the response to pathogens such as bacteria, fungi, and nematodes, and were identified in adaptations to Lowland environments probably due to the higher prevalence of pathogens in Lowland regions (mainly Wet Lowland regions) (The CIMMYT maize program 2004).

Genes identified as adapted to different environments, such as Highland environments, are clue for breeding programs improvement. Landraces history and evolution, and its inclusion in breeding programs, will help us to challenge the climate changes and guarantee global food security.

Nowadays, genotypic information has a huge impact in the classification of landraces. Racial designation in legacy data is an intensive manual curation effort that is unfortunately prone to errors. Examples can be seen in the misclassification of Nat-tel for one accession that shows a large proportion of red component from the Highland adaptation and indeed was collected from highland region (2,063 m.a.s.l in Oaxaca), and two misclassified Tepecintle samples showing much larger orange component than is supposed to be in southern Brazilian material. Those races were collected in Northern Brazil. Vandeño and Pepitilla panels included few samples from Dominican Republic, which showed high levels of the purple component and low levels of the blue one, suggesting more relatedness with Caribbean material. Classification of landraces could be confirmed by genomic data, instead of lists of passports for each sample.

Acknowledgements

This work was supported by the Bill and Melinda Gates Foundation and the International Centre of Maize and Wheat Improvement (CIMMYT).


Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81


Table 1. Community analysis for IBD network. Seven communities were found. Number of nodes, node with the highest degree (Location and Race) and its number of connections are shown.

<table>
<thead>
<tr>
<th>Community</th>
<th>Number of nodes</th>
<th>Number of Connections</th>
<th>Location</th>
<th>Race</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>109</td>
<td>121</td>
<td>Mexico (Puebla)</td>
<td>Tuxpeño</td>
</tr>
<tr>
<td>2</td>
<td>107</td>
<td>173</td>
<td>Antigua &amp; Barbuda</td>
<td>Coastal Tropical</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Flint</td>
</tr>
<tr>
<td>3</td>
<td>117</td>
<td>157</td>
<td>Brazil</td>
<td>Dente</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Riograndense</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rugoso</td>
</tr>
<tr>
<td>4</td>
<td>67</td>
<td>72</td>
<td>Mexico (Oaxaca)</td>
<td>Zapalote Chico</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4</td>
<td>Mexico (Morelos)</td>
<td>Tuxpeño, Pepitilla,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ancho</td>
</tr>
<tr>
<td>6</td>
<td>224</td>
<td>459</td>
<td>Mexico (Chiapas)</td>
<td>Zapalote Grande</td>
</tr>
<tr>
<td>7</td>
<td>28</td>
<td>13</td>
<td>Tobago</td>
<td>Tuson</td>
</tr>
</tbody>
</table>
Figure 1. Distribution of Central and South American maize samples according to their location in countries. A total of 1699 samples are shown. Majority of samples belong to Brazil, Argentina, Uruguay, Venezuela, and Guatemala.
Figure 2. Geographical distribution of the comprehensive set of Mexican maize (*Zea mays* subsp. *mays*) landrace accessions analyzed in this study. Samples are colored according to the environment. Environments were defined based on 19 climatic variables and altitude. Highland samples (red dots) grow over 1700 m.a.s.l., and lowland samples (dots in blue) grow below 800 m.a.s.l.
samples are grouped

Caribbean

Admixture of samples for Environments (A and B) and Races (C) using K=3 (A), and K=6 (B and C). Distribution of samples is shown in the Americas map, and each sample is colored according to the major K in the admixture analysis. K=6 is enough to discriminate environments. Races belonging to the same environment show similar admixture composition. Mexican highlands and lowlands are distributed according to the altitude, and had a high red and a blue component respectively. Brazilian group showed the highest orange component and Southern South American samples, classified as Northern US translocation showed the highest green component. Caribbean samples are grouped by the purple component.
Figure 4. Network of relations between samples based on Identical-by-Descent (IBD) value. A. IBD network filtered by 0.2-0.5, darkest lines represent IBD >0.25. B-G. Distribution of samples according to the community they belonged (1, 2, 3, 4, 6, 7 respectively). The color scheme matches that of the admixture analysis, Figure 3C.
Figure 5. Allele frequencies (AF) spectrum among comparisons. A. Distribution of SNPs used for AF comparisons, a higher amount of SNPs was observed in the telomeric regions than in pericentromeric regions. The distribution matches the distribution of the entire dataset of SNPs (Fig. S3). The window size was fixed at 200,000 bp. B. Distribution of the SNPs significance for each comparison. Top 500 SNPs are plotted, and significance is measured as the p-value result from Fisher exact test. Comparison order matches the clustering in figure 6.
Figure 6. Biological process GO terms enriched among the comparisons. A. Comparisons clustering based on p-value. B. GO-terms clustering. C. Heatmap showing Bonferroni corrected p-values from the enrichment. Blue regions showed the most significant terms for each comparison. 24 groups of terms were found. Majority of terms are related to flowering cycle (embryogenesis, clock regulation), response to drought (homeostasis), storage (starch, fatty acids and sugar), and response to biotic stress (fungi, bacteria).
Table S1. Samples information. ID, geographical information, mega environment, and race (txt file available in https://www.dropbox.com/s/gsldswe5jk3718eu/TableS1?dl=0)

Table S2. Population structure components according to admixture colors

<table>
<thead>
<tr>
<th>Color</th>
<th>Population/Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yellow</td>
<td>Ancient 8 row maize</td>
</tr>
<tr>
<td>Blue</td>
<td>Lowland</td>
</tr>
<tr>
<td>Red</td>
<td>Highland</td>
</tr>
<tr>
<td>Purple</td>
<td>Caribbean</td>
</tr>
<tr>
<td>Orange</td>
<td>Brazilian (from Southeastern US)</td>
</tr>
<tr>
<td>Green</td>
<td>South American (from Northern US)</td>
</tr>
</tbody>
</table>

Table S3. Number of samples in each community according to admixture grouping.

<table>
<thead>
<tr>
<th>Population/Community</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mexican lowlands</td>
<td>74</td>
<td>26</td>
<td>34</td>
<td>47</td>
<td>4</td>
<td>140</td>
<td>4</td>
</tr>
<tr>
<td>US</td>
<td>3</td>
<td>2</td>
<td>14</td>
<td>5</td>
<td>0</td>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>Mexican highlands</td>
<td>8</td>
<td>6</td>
<td>11</td>
<td>4</td>
<td>0</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>Brazilian</td>
<td>17</td>
<td>6</td>
<td>35</td>
<td>4</td>
<td>0</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>Caribbean</td>
<td>2</td>
<td>62</td>
<td>17</td>
<td>4</td>
<td>0</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>Others</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure S1. Theoretic network of mexican maize landraces relatedness according to history and phenotypic data (Wellhausen et al 1951). Reds are ancient races, greens are prehistoric races, purples are pre-colombian races, blues are modern incipients. Teosinte is in orange. The arrows show the direction of the genetic flow South American races are not included in this map.
Figure S2. Cross-validation error for the number of clusters created using admixture. There was not found a minimum point, however, K=3 and K=6 were selected according to reported grouping using SSRs.
Figure S3. GBS SNP annotation according to B73 information. B73 genotype genes correspond to less than 15% of the genome. The same proportion is observed for SNPs distribution. 85% of the genome is composed by repetitions, including transposons. Gene annotation included UTR and coding regions.
Figure S4. Distribution of SNPs along the maize genome. The SNPs are distributed along the 10 maize chromosomes, and more highly concentrated in sub-telomeric than pericentromeric regions. Also there are some SNPs with matches in the mitochondria, chloroplast and the chromosome 11 (contigs without position in the 10 maize chromosomes).
**Figure S5.** Allele frequency differences between Mexican highlands and Mexican wet lowlands. The typical flowering pattern is showed. Inversion in the chromosome 4 is the region with more changes in the allele frequencies.