Plastome genomics in South American maize landraces: chloroplast lineages parallel the geographic structuring of nuclear gene pools

Mariana Gabriela López [1,2,3Ψ], Mónica Fass [1,2,Ψ], Juan Gabriel Rivas [2], José Carbonell-Caballero [4], Pablo Vera [2], Andrea Puebla [2], Raquel Defacio [5], Joaquín Dopazo [6], Norma Paniego [1,2], Horacio Esteban Hopp [2,7], Verónica Viviana Lia [1,2,7,*]

[1]Consejo Nacional de Investigaciones Científicas y Técnicas–CONICET. Argentina; [2]Instituto de Biotecnología, Centro de Investigaciones en Ciencias Veterinarias y Agronómicas (CICVyA), Instituto Nacional de Tecnología Agropecuaria (INTA); Instituto de Agrobiotecnología y Biología Molecular (IABIMO), INTA-CONICET, Argentina; [3] Instituto de Biomedicina de Valencia (IBV-CSIC). c/ Jaume Roig 11, 46010, Valencia, Spain; [4]Gene Regulation, Stem Cells and Cancer Program. Centre for Genomic Regulation (CRG). Dr. Aiguader 88, 08003 Barcelona, Spain; [5]Estación Experimental Agropecuaria INTA Pergamino, Pergamino Buenos Aires, Argentina; [6]Clinical Bioinformatics Area, Fundación Progreso y Salud. CDCA, Hospital Virgen del Rocío. c/Manuel Siurot s/n, 41013, Sevilla, Spain; [7]Facultad de Ciencias Exactas y Naturales Universidad de Buenos Aires, Intendente Güiraldes 2160, (1428), Ciudad Autónoma de Buenos Aires, Argentina.

**For correspondence. E-mail lia.veronica@inta.gob.ar*

Ψ These authors contributed equally to this work.

- **Background and Aims** The number of plastome sequences has increased exponentially during the last decade. However, there is still little knowledge of the levels and distribution of intraspecific variation. The aims of this study were to estimate plastome diversity within *Zea mays* and analyze the distribution of haplotypes in connection with the landrace groups previously delimited for South American maize based on nuclear markers.
- **Methods** We obtained the complete plastomes of 30 South American maize landraces and three teosintes by means of Next Generation Sequencing and used them in combination with data from public repositories. After quality filtering, the curated data was employed to search for SNPs, INDELs and cpSSRs. Exact permutational contingency tests were performed to assess associations between plastome and nuclear variation. Network and Bayesian phylogenetic analyses were used to infer evolutionary relationships among haplotypes.
- **Key Results** Our analyses identified a total of 124 polymorphic plastome loci, with the intergenic regions *psbE-rps18*, *petN-rpoB*, *trnL_UAG-ndhF* and *rpoC2-atpI* exhibiting the highest marker densities. Although restricted in number, these markers allowed the discrimination of 27 haplotypes in a total of 51 *Zea mays* individuals. Andean and lowland South America landraces differed significantly in haplotype distribution. However, overall differentiation patterns were not informative with respect to subspecies diversification, as evidenced by the scattered distribution of maize and teosinte plastomes in both the network and Bayesian phylogenetic reconstructions.

- **Conclusions** Knowledge of intraspecific plastome variation provides the framework for a more comprehensive understanding of evolutionary processes at low taxonomic levels and may become increasingly important for future plant barcoding efforts. Whole-plastome sequencing provided useful variability to contribute to maize phylogeographic studies. The structuring of haplotype diversity in the maize landraces examined here clearly reflects the distinction between the Andean and South American lowland gene pools previously inferred based on nuclear markers.

Key words: *Zea mays*, whole plastome sequencing, maize landraces, intraspecific variation, cpSSR, maize dispersal.

INTRODUCTION

Latin American maize comprises ca. 260 landraces grown across a wide variety of agroecosystems, from the lowlands of southern South America to the highlands of Mexico and the Andes (Salhuana and Pollak, 2005). This locally adapted germplasm constitutes an invaluable source of genetic diversity to cope with the increasing production demands and changing environmental conditions of modern agriculture.

After its initial domestication in southwestern Mexico approximately 9000 years before present (BP) (Matsuoka *et al.*, 2002), maize (*Zea mays* L. *ssp. mays*) rapidly spread along the continent. The archaeological record suggests that maize dispersal into South America was already ongoing by 6500 years BP (Grobman *et al.,* 2012) and that cultivation of the crop had become widespread after about 5000 years BP (Haas *et al.*, 2013). More recently, archaeogenomic evidence revealed that the domestication process was not fully complete by 5300 years BP (Ramos-Madrigal *et al.*, 2016; Vallebueno-Estrada *et al.*, 2016), thus implying that maize germplasm first entering South America was only partially domesticated. Subsequently, landrace differentiation within South America progressed without direct input from maize conspecific wild relatives, the teosintes *Z. mays* ssp *parviglumis* Iltis & Doebley, *Z. mays* ssp. *mexicana* (Schrader) Iltis and Z. *mays* subsp. *huehuetenangensis* (Iltis & Doebley) Doebley, whose distribution is restricted to Mexico and Guatemala (Doebley, 1990).

Two main genetic groups with distinctive geographical distribution have been traditionally recognized for extant South American landraces: the Andean and the Tropical Lowland groups (Vigouroux *et al.*, 2008). Additionally, Bracco *et al.* (2016) reported the occurrence of a third, locally adapted, lowland gene pool (i.e. Northeastern Argentina Flours) associated with Guaraní indigenous communities of middle South America, which was genetically and spatially differentiated from both the Andean and Tropical Lowland gene pools. In agreement with these findings, genomic, linguistic, archaeological, and paleoecological data suggest that the southwestern Amazon was a secondary improvement center for partially domesticated maize, from which deeply structured lineages evolved, originating an Andean and a Lowland group (Kistler *et al.*, 2018). Following establishment of these gene pools, a Pan-American lineage, which had received the contribution of teosintes for a longer period, also dispersed from the domestication center spanning from northern Mexico to lowland South America (Kistler *et al.*, 2018).

Chloroplast markers have been the tool of choice for most plant phylogeographic studies. But, the low levels of intraspecific variation generally observed when using a handful of plastid markers have often resulted insufficient to provide resolution below the species level. With the advent of next generation sequencing, complete plastomes from across the land plant tree of life are being sequenced rapidly (Tonti-Filippini *et al.*, 2017), giving rise to new opportunities to exploit their advantages for population genetics studies.

The wealth of genomic data currently available for maize has led to neglect of the potential utility of plastomes in providing a less complex alternative to reconstruct the crop's trajectories along the continent. Early research on chloroplast intraspecific variation within the genus *Zea* using RFLP techniques showed little diversity and a poor prospect for population genetics studies (Doebley *et al.*, 1987). However, posterior analysis of chloroplast simple sequence repeats (cpSSR) revealed high haplotype diversity and allowed resolution of sectional and racial affiliations (Provan *et al.*, 1999; Bird, 2012). The first complete maize plastome was 140.387 base-pairs long and was assembled from sequencing clones of two hybrids (Maier *et al.*, 1995). Since then, only eight additional chloroplast genomes have been reported. Seven of them correspond to breeding germplasm, including the hybrid cultivar Zhengdan958 (Liu *et al.*, 2019), the fertile inbred lines A188, B37, and B73, and the cytoplasmic male-sterile (*cms*) lines B37 *cms-C*, B37 *cms-S*, and B37 *cms-T* (Bosacchi *et al.*, 2015). The remaining plastome is a partial sequence obtained from a 5300-year-old maize sample collected at the San Marcos Cave, Valley of Tehuacan, Mexico (sample SM10, Pérez-Zamorano *et al.*, 2017). Yet, no reports of chloroplast genomes are currently available for extant landraces, nor are there estimates of intraspecific plastome variation at the sequence level. Similarly, no complete plastomes have been published to date for *Z. mays* ssp. *parviglumis* or *Z. mays* ssp. *mexicana*.

In this study, we sequenced the complete plastomes of 30 South American maize landraces, one *Z. mays* ssp. *mexicana* and two *Z. mays* ssp. *parviglumis* individuals to estimate plastome intraspecific diversity and analyze haplotype distribution in connection with the landrace groups previously delimited for South American maize based on nuclear markers (Bracco et al., 2016). In addition, we used NGS data from the landraces and teosintes included in maize HapMap 2 (Beissinger, 2016) to extract chloroplast reads and enlarge our dataset for comparative purposes.

MATERIALS AND METHODS

A total of 34 individuals were sequenced in this study using a long-PCR approach: 30 South American maize landraces (31 individuals), two *Z. mays ssp. parviglumis* and one *Z. mays ssp. mexicana*. Maize individuals previously analyzed by Bracco et al. (2016) and Rivas (2015) were chosen to represent the five genetic groups identified on the basis of nuclear SSR data. These are: Highland Mexico and US (HM-US), Tropical Lowland (TL), Andean, NEA (North Eastern Argentina) Flours and NEA Popcorns. Four admixed individuals were also included for comparison (Fig. 1, Supplementary data Table S1).

In addition, bam files from 42 landraces and teosintes available from HapMap 2 (Beissinger 2016) were used to extract chloroplast reads, perform variant calling and consensus assembly. Some of these accessions were also examined in Bracco *et al.* (2016) and were thus assigned here to the corresponding SSR genetic group (Fig. 1, Supplementary data Table S1). Intraspecific analysis also included published sequences from Bosacchi *et al.* (2015) and Perez-Zamorano *et al.* (2017). The original sequence by Maier *et al.* (1995) was not considered for analysis since it presents many sequencing errors, as previously pointed out by Bosacchi *et al.* (2015). The details of the different accessions are provided in Supplementary data Table S1. Map data was downloaded from Natural Earth (Free vector and raster map data, naturalearthdata.com).

Plastome sequencing

A long-PCR strategy was employed to sequence complete chloroplast genomes. Total DNA was extracted from seedlings as previously described (Lia *et al.* 2007). Fourteen primer pairs were designed with Primer3Plus (Untergasser *et al.* 2012) using as reference the B73 chloroplast genome sequence from Bosacchi *et al.* (2015) (KF241981.1). The resulting amplicons covered the 140.4 KB of maize plastome (Supplementary data Table S2).

Amplifications were performed in a 25 ul reaction mix, including 300 uM of each dNTP, 0.4 uM of each primer, 1X PCR Reaction Buffer (New England Biolabs), 2.5 units of LongAmp Taq DNA Polymerase (New England Biolabs), 1 ul of DNA template (ca. 100 ng) and sterile double-distilled water. PCR reactions were carried out in a Mastercycler EP S cycler (Eppendorf), with the following conditions: 1 min denaturation at 94ºC, and 30 cycles of: 20 sec at 94ºC, 50 sec at 55-60ºC, 1 min per Kb at 65ºC. PCR products were separated on 0.8% agarose gels, stained with ethidium bromide 2% and visualized with a trans-UV system (Gel Doc XR+ BIO-RAD).

Each PCR amplicon was purified with the Exonuclease I-FastAP Thermosensitive AP enzymes (ThermoFisher) and quantified with QUBIT fluorometer using Qubit dsDNA BR assay kit (ThermoFisher). The 14 amplicons from each individual were then normalized to the same concentration and mixed, yielding a single sample for Illumina sequencing.

Library preparation and sequencing

Library preparation was performed with Nextera XT DNA Sample Preparation Kit (ThermoFisher), according to the manufacturer’s instructions. Briefly, individual samples were diluted at 0.5 ng/ul, fragmented into 500 base-pair segments with a tagmentation enzymatic reaction, and subjected to adapter and index incorporation. After two steps of library clean up and normalization, paired-end sequencing (250 bp) was conducted on a MiSeq Illumina instrument (Illumina) at UGB Service (Unidad de Genómica y Bioinformática – INTA, Argentina).

Quality control, mapping, variant calling and annotation

Read quality assessment was conducted with FastQC v0.11.5 (Andrews, 2010). Low scoring sequences and short reads were removed using Trimmomatic 0.36 (Bolger *et al.*, 2014).

Output paired and unpaired files were mapped with the BWA-MEM algorithm of Burrows-Wheeler Aligner software (ver 07.13, http://bio-bwa.sourceforge.net/bwa.shtml#13), using KF241981.1 as reference genome.

A coverage analysis with a 10 kb sliding window was conducted to check complete plastome assemblies for accuracy (https://github.com/mrmckain/Fast-Plast).

Variant calling of 76 samples (34 bam files from this study, plus 42 from Hapmap 2) was performed using GATK (McKenna *et al.*, 2010). The calling parameters were adjusted to obtain both single nucleotide polymorphism (SNP) and indel markers. Following GATK best practices (https://software.broadinstitute.org/gatk/best-practices/), files were realigned with the IndelRealigner tool, then paired and unpaired files were merged to increase deep coverage with the aim to improve variant calling. Alignments were inspected with Integrative Genomic Viewer ver. 2.3.91 (Robinson *et al.*, 2011) and evaluated with Qualimap tool (García-Alcalde *et al.* 2012). Coverage was evaluated using, DepthOfCoverage, FindCoveredIntervals and DiagnoseTargets tools from GATK .

Variants were called with HaplotypeCaller GVCF mode. The obtained raw variants were hard filtered using the parameters recommended by GATK (https://software.broadinstitute.org/gatk/documentation/article.php?id=2806), all the reads with quality (MQ) lower than 40 were filtered out.

Consensus sequences were obtained using VCFtools (https://vcftools.github.io/perl_module.html). Fasta files were also manually curated with BioEdit (Hall, 1999), inspecting Indels to identify cpSSR.

Plastomes were annotated with the AnnoBDT software included in Verdant (http://verdant.iplantcollaborative.org/plastidDB/index.php). Annotation files were evaluated and curated with tlb2asn program from ncbi (https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/).

The annotation and classification of genetic variants based on the effect of annotated genes were carried out using SnpEff v4.0 (Cingolani *et al.*, 2012). Genetic variants were classified into frameshift, missense, synonymous, intronic, intergenic and structural variants in intergenic regions and were represented using CIRCOS plot (Krzywinski *et al.*, 2009). A list of the polymorphisms identified by the above procedures is provided in Supplementary data Table S3. These analyses were performed in a local server at IMPaM, Sistema Nacional de Computación de Alto Desempeño (SNCAD) ID 924, Ministerio de Ciencia, Tecnología e Innovación Productiva (MINCyT), Argentina.

Assembly of the data matrix for diversity and network analysis

After initial variant calling, a more in-depth manual curation was performed to exclude positions with more than one variant in frequencies higher than 10% within individuals. In addition, individuals with >10% missing data and/or ambiguous calling were excluded from further analyses. This resulted in only 11 out of 42 HapMap 2 samples being part of the final data matrix, which ultimately consisted of the consensus sequences of the 34 samples sequenced in this study, 11 samples from HapMap 2, the B73, B37 *cms*-C, B37 *cms*-T, and B37 *cms*-S plastome sequences from Bosacchi *et al.* (2015), the SM10 plastome sequence from Perez-Zamorano *et al.* (2017), and the *Z. mays huehuetenangensis* sequence from Orton *et al.* (2017).

Multiple sequence alignment (MSA) was conducted using MAFFT v6.857b (Katoh and Toh, 2008) under the FFT-NS-2 default strategy.

Haplotype networks and genetic diversity

Haplotype networks were constructed using the median-jonining network procedure (Bandelt *et al.* 1999) implemented in popART (Leigh and Bryant, 2015), with epsilon=0. To this end, the manually curated SNP, cpSSR and indels were extracted from the MSA, with cpSSR and indels being codified as multi-state characters. The singletons present in the consensus sequences obtained from public repositories were not taken into account because no quality scores were available. Given that popART does not consider positions with more than 5% missing data, and to avoid further reduction of the number of characters, we imputed the data matrix using the Bayesian approach of the PHASE v1 software (Stephens *et al.*, 2001). The algorithm was run five times with different seeds (burn-in:500, iterations=500), assuming a non-stepwise mutation model for cpSSR (-d1), and a fix recombination rate equal to 0 (-MR4 -R0). Consistency across runs was checked as suggested in the program documentation. The resulting data matrix is provided in Supplementary data Table S4.

Genetic diversity indices (number of haplotypes, number of segregating sites, haplotypic diversity, mean number of pairwise differences) were estimated with DNAsp v5.10.01 (Librado and Rozas 2009). Haplotype frequency distributions among the groups defined based on nuclear markers were compared by means of Fisher's exact tests as implemented in the R stats package (R Core Team, 2013).

Bayesian phylogenetic analysis

Phylogenetic analyses were conducted with BEAST v2.6.0 (Bouckaert et al., 2019) using two publicly available plastomes of *Zea luxurians* (GenBank KR873424.1, Verdant TK096) as outgroups. The SM10 sample was not considered for this analysis due to the large number of undetermined positions at invariant regions.

Site models were averaged using bModelTest (Bouckaert and Drummond, 2017). Analyses were conducted under strict, uncorrelated lognormal relaxed (ucld), and random local clock models. Model comparisons were performed with the SS/PS method as implemented in BEAST model-selection package 1.5.3. The ucld model had the highest support, but the standard deviation (ucld.stdv) was much higher than 1, indicating that a random clock model was a better fit to our data (Drummond and Bouckaert, 2015). Although SS/PS analyses failed to provide stable marginal likelihood estimates for the random local clock model, evaluation of the posterior probabilities of the number of rate changes clearly rejected a strict clock model, with 2 and 3 rate changes having the highest posterior probabilities. Summarizing, two lines of evidence support the use of the random clock model: that the uncorrelated lognormal has a high coefficient of rate variation and that the random local clock supports multiple rate changes (Drummond and Suchard, 2010, Duchêne, pers. comm.).

Preliminary analysis using an Extended Bayesian Skyline Plot tree prior failed to reject a constant population size (sum(indicators.alltrees) 95% HPD: 0-3), and so reconstructions were made under a constant-size coalescent tree prior, with the prior for the constant population size hyperparameter set to a lognormal distribution with mean 1 and standard deviation 1.25.

Analyses were run at least twice for a minimum of 50 million generations, with sampling every 5000 generations. All operator settings were left as default. Chain convergence to stationarity was assessed in Tracer v.1.7.1 (Rambaut *et al.*, 2018). The first 10% of each run was discarded as burn-in after checking for effective sample size (ESS) > 200. TreeAnnotator v2.5.1 was used to compute a maximum-clade-credibility tree, which was visualized in FigTree 1.3.1 (Rambaut, 2009).

The multiple sequence alignment used for this analysis is provided in Supplementary data Fasta File S1. One of the IRs was manually removed to avoid duplicated sequences.

Data availability

Raw sequence data is available at the NCBI Sequence Read Archive (BioProject ID: PRJNA604083, BioSample accessions SAMN13951573-SAMN13951606). Processed data matrices are available as supplementary files.

RESULTS

Complete plastomes from 30 maize landraces (31 individuals), one Z. *mays* ssp. *mexicana* and two Z. *mays* ssp. *parviglumis* specimens were obtained through NGS sequencing, with an average depth of 1466.8 X and an average of 1,002,617.6 mapped reads per sample. The fraction of the plastome with sequencing depth >30 X (10 Kb windows) was 100% for all individuals, except for *Z. mays* ssp. *parviglumis* M7, in which 6.7% of the windows did not reach 30 X (Supplementary data Table S5).

The length of the plastomes ranged from 140,452 to 140,464, with the B73 sequence used as reference being 140,447 bp long. The short single copy (SSC) regions varied from 12,527 to 12,541 bp and the long single copy regions (LSC) from 82,380 to 82,455 bp. No variation was observed in the length of the inverted repeats (Supplementary data Table S5).

Variant calling, including 42 maize and teosinte samples from HapMap 2 (Supplementary data Table S1), yielded a total of 124 polymorphic loci (80 SNPs, 34 cpSSR, 9 indels and 1 inversion) that surpassed quality filters and manual curation (Fig. 2, Supplementary data Table S3). The 34 cpSSR were mononucleotide repeats, 25 coincide with those reported by Provan *et al.* (1999) and nine represent novel findings.

All polymorphic loci were located within the SSC or the LSC regions, with four regions concentrating the highest density of markers (*psbE-rps18*; *petN-rpoB*; *trnL_UAG-ndhF*; and *rpoC2-atpI*) (Supplementary data Table S3). The 34 cpSSR lay within non-coding regions (79% intergenic, 21% intronic), whereas all indels were intergenic. Conversely, the inversion produced a missense mutation in the *ccsA* gene. Of the 80 SNPs, 56 were found in non-coding regions (80% intergenic, 20% intronic), 11 resulted in synonymous changes, and 13 resulted in missense mutations (Supplementary Information Table S3). The number of transversions was higher than the number of transitions (Ts/Tv = 0.6).

Haplotype diversity and relationships

Prior to haplotype analysis, we applied additional filtering steps in order to reduce missing data and eliminate unreliable variation (see methods). The final data matrix was then composed by 99 characters (80 SNPs, 11 cpSSR, 7 indels, 1 inversion) and 51 taxa, including B73, B37 *cms*-C, B37 *cms*-T, and B37 *cms*-S (Bosacchi *et al.* 2015), the SM10 archaeological sample (Pérez-Zamorano *et al.*, 2017), and *Z. mays* ssp. *huehuetenangensis* (Orton *et al.*,2017) (Supplementary data Table S4).

In total, 27 plastome haplotypes were identified, with most of them (23/27) being unique to a single individual (Supplementary data Table S1, Fig. 3). Diversity indices based on different sets of markers are presented in Table 1. Despite their lower number, indels exhibited the highest proportion of parsimony informative loci (87%), followed by SNP (69%) and cpSSR (63%). However, SNP variation accounted for 24 of the 27 haplotypes identified here. Restricting the data matrix to SNP and cpSSR resulted in the loss of a single haplotype, H13, which differed from H12 by an 83-bp indel. When only SNPs were considered, H14 collapsed into H1, and H8 into H3.

The median-joining network shows haplotype relationships based on the 99-character set, as well as their occurrence in the different landrace groups defined by Bracco *et al.* (2016) (Fig. 3). As a general pattern, haplotypes found in maize and teosintes were intermixed along the network. H1, the most frequent haplotype, was prevalent within the NEA Flours, and was also present in individuals from five other groups, including the inbred line B73. In addition, it was the closest to the haplotype found in the archaeological sample SM10 (H23) and the only one shared with ssp. *parviglumis*. Interestingly, H1 was not found in any of the Andean landraces. In contrast, H2 was dominant among Andean samples and absent from the lowland groups (i.e. NEA Flours group and TL).

Haplotypes from the TL and HM-US groups were widespread across the network, whereas haplotypes from NEA popcorns did not show any clear associations with those from the other lowland groups. The two haplotypes present in *Z. mays* ssp. *mexicana* clustered together (H24, H25), and the haplotype found in *Z. mays* ssp. *huehuetenangensis* (H26) occupied a relatively central position in the network.

The *cms* lines exhibited remarkably distinct haplotypes (H20, H21, H22) separated from the rest by a high proportion of mutational changes. This cluster also encompassed the haplotypes found in landraces Dentado Blanco (H13) and Hickory King (H12), none of which could be assigned to the genetic groups previously reported by Bracco *et al.* (2016).

Network reconstruction with the 91- and 80-character data sets (i.e., SNP+cpSSR, and SNP) yielded similar patterns with the consequent decrease in the number of haplotypes. There was also a reduction in the number of internal loops and the position of the SM10 haplotype resulted slightly different, although it maintained its affiliation with H1 (Supplementary data Fig. S1).

To gain further insights into the diversification of *Z. mays* plastomes, we obtained a Bayesian phylogenetic tree using two *Zea luxurians* plastomes as outgroups. The aligned data matrix, excluding one IR and positions with gaps, was 118089 bp long (Supplementary data Fasta File S1). Overall, the resulting tree topology (Supplementary data Fig. S2) agreed well with the relationships obtained in the network analysis (Fig. 3). *Zea mays* plastomes formed a monophyletic group, with teosinte haplotypes distributed across the clusters.

Haplotype counts for the different categories used in this study are provided in Table 2. Comparison of frequency distributions showed statistically significant differences between NEA Flours and Andean landraces (Fisher's exact test, p<0.001). Pairwise comparisons among the remaining groups were not conducted due to the small sample sizes.

DISCUSSION

Chloroplast genomes have long been a fundamental source of phylogenetic characters in plant systematic and evolutionary studies. Indeed, the number of plastome sequences has increased exponentially during the last decade (Tonti-Filippini *et al.* 2017). However, there is still little knowledge of the levels and distribution of intraspecific variation. Here, we present the first report of whole plastome diversity in maize landraces and teosintes, which integrates data from different taxonomic levels and provides novel information for population genetics analyses.

The number of polymorphic loci identified in this study (i.e. 80 SNP, 34 cpSSR, 9 indels and 1 inversion) is on the lower end of the range reported for other plant species (e.g., Tang *et al.* 2004; Young *et al.* 2011; Sancho *et al.* 2018; Nock *et al.* 2019). Using a sampling intensity similar to the one applied here, two and three times more loci were observed for *Brachypodium distachyon* (Sancho *et al.* 2018) and *Macadamia integrifolia* (Nock *et al.* 2019), respectively. This disparity was even larger when comparing the mean number of pairwise differences among haplotypes, with *B. distachyon* and *M. integrifolia* showing two to six times higher estimates than those obtained here for *Z. mays* (18.875). Interestingly, these differences did not translate into a comparable increase in the number of haplotypes, suggesting that the absolute number of polymorphic loci may not be a direct evidence of the real information content. Although under the infinite-alleles model, with neutral mutation and random genetic drift, the number of haplotypes is expected to increase almost linearly with θ (Hartl and Clark, 2007), demographic factors, such as population size changes or population structure, may help explain the different dynamics among species.

Twenty-four of the 27 haplotypes found in our survey could be defined by considering SNP variants only. In addition, cpSSR are potentially more informative than SNP, but they showed high levels of intra-individual variation. These results suggest that SNP are a more robust option for defining haplotype variants, at least with a technique as sensitive as NGS sequencing.

Taking into account the relatively low number of polymorphic sites, the sequencing strategy used in this work may seem too costly in terms of sequencing effort per data point, even when excluding the invariable inverted repeats. However, in comparison to whole-genome sequencing approaches, reconstructing plastomes using the long PCR method has the strength of facilitating variant calling by avoiding the confounding effects of organellar gene transfer to the nucleus, a phenomenon that could explain the large number of ambiguities found in the Hapmap 2 samples. In this context, the variants identified here can serve as a reliable catalog to map sequence data from the growing number of low-coverage maize resequencing projects. Plastome haplotypes could then be reconstructed as an inexpensive by-product of genome-wide analyses, complementing nuclear markers for evolutionary inferences.

According to our results, the highest density of markers is concentrated in the following regions: *psbE-rps18*; *petN-rpoB*;*trnL_UAG-ndhF*; and *rpoC2-atpI*. The first three regions coincide with the most consistently variable regions across monocots, while *rpoC2-atpI* has not been mentioned so far (Shaw *et al.* 2014). It still remains to be determined whether these patterns are exclusive to *Zea mays* or are present at higher taxonomic levels. In any case, this idiosyncratic behavior is in line with recent proposals that encourage the use of whole plastome sequences, instead of one or a few plastid markers, for the next generation of plant DNA barcodes (Hollingsworth *et al.*

2016). In addition, given that plastid markers are the tools of choice for most plant systematics studies, knowledge of intraspecific variation may allow for a more comprehensive understanding of evolutionary processes at low taxonomic levels, particularly when species delimitation is unclear.

Most of the haplotypes (85%) found in landraces and teosintes were unique to a single individual, suggesting high levels of underlying genetic diversity. The distribution of mutational steps is also indicative of a variable gene pool, with *cms* inbred lines showing the most pronounced differentiation. The distinctiveness and early divergence of the *cms*-S organellar genomes have already been highlighted by some authors, who interpreted that the affiliation between the *cms*-S cytotype and *Z. mays ssp. mexicana* is a consequence of the introgression of this teosinte into maize (Doebley 1990; Allen 2005). By contrast, the *cms*-T cytotype has been observed in Latin American maize landraces (Weissinger *et al.* 1983), but has not yet been reported in any teosinte accession. Although our analyses do not show the proposed relationship between *cms*-S and ssp. *mexicana*, probably due to a very limited sampling, they do reveal the connection of *cms*-S with two maize landraces, Dentado Blanco and Hickory King (Fig. 3). If *cms* plastids and mitochondria were always co-transmitted maternally -as suggested by Bosacchi *et al.* (2015)- these findings may point to yet unexploited sources of cytoplasmic male sterility in maize.

Regarding the distribution of plastome haplotypes within the genetic groups previously delimited for South American maize landraces, the most relevant features are: the prevalence of H1 in the NEA Flours, the absence of this haplotype in the Andean races and the almost exclusive dominance of H2 among the latter. Such pattern is consistent with the nuclear genetic differentiation reported for these groups (Bracco *et al.* 2016), as it also reflects their geographic and environmental separation. In this regard, the NEA Flours represent maize germplasm associated with Guarani people, whose area of influence encompassed the lowland areas of middle South America, including northeastern Argentina, Paraguay, southern Bolivia and southwestern Brazil. In turn, the Andean germplasm is distributed over the western part of the continent and cultivated at altitudes often above 2000 meters above sea level.

Although both the NEA Flours and Tropical Lowland groups are associated with the lowland regions of South America, they exhibit substantial genetic differentiation (Bracco *et al.* 2016). Adding to these differences and in contrast to the homogeneity of Andean and NEA Flour landraces, the haplotypes identified in HM-US and Tropical Lowland showed lower frequencies and were sparsely distributed across the network. The same is true for teosinte specimens, all of which showed different haplotypes. Though limited in number, the haplotypes found in the NEA Popcorns do not conform to the general pattern observed for the NEA Flours, i.e., a prevalence of H1. Paterniani and Goodman (1977) have emphasized that popcorn maize was only cultivated in the region by the Guaraní people, and not by the other indigenous groups of the same area, implying that popcorn cultivation is much more recent than that of NEA Flours, and may thus be the result of a more recent introduction.

In line with the proposals by Kistler *et al.* (2018), the diversity of haplotypes within the Tropical Lowland group may be seen as concordant with the expectations for the Pan American lineage. The latter exhibited an excess of shared ancestry with ssp. *parviglumis* (Kistler *et al.* 2018), which is consistent with Tropical Lowland maize showing the lowest value of the genetic drift parameter relative to the hypothetical ancestor of all landraces (Bracco *et al.* 2016). The dominance of a single high-frequency haplotype in both NEA Flours and Andean landraces suggests that the two groups experienced a rather severe bottleneck, as might be expected for the lineages derived from the Amazonian secondary improvement center proposed by Kistler et al. (2018).

Previous studies based on cpDNA have shown the clustering of maize landraces into two or three distinctive groups, albeit with different affiliations relative to teosintes (Doebley 1990; Provan *et al.* 1999; Bird 2012). The phylogenetic relationships obtained here for *Z. mays* plastomes show that subspecies have not yet attained reciprocal monophyly and that all haplotypes may be envisaged as part of a single gene pool. Although differentiation may already be apparent in the nuclear genome, its signals would still be undetectable for the slowly evolving chloroplast genomes. In our data set, H1 was the only haplotype shared by both maize and teosintes. It was found in several maize landraces and in the *Z. mays* ssp. *parviglumis* M7 individual from Toliman, Jalisco. Interestingly, it has recently been proposed that the Jalisco-southern Pacific Coast region of Mexico is a more likely area for maize domestication than the Balsas river valley (Moreno-Letelier *et al.* 2020). Moreover, in the same study, ssp. parviglumis populations from Toliman and Guachinango were the only two to show signs of gene flow with maize landraces. The comparative antiquity of H1 is also supported by its widespread distribution in different genetic groups (i.e., NEA Flours, NEA popcorns, Tropical Lowland, teosintes, and inbred lines), along with a relatively large number of derived haplotypes. The presence of H1 in B73 probably stems from the contribution of Yellow Dent, a landrace involved in the origin of all three maize heterotic groups (van Heerwaarden *et al.* 2012). Moreover, the close relationship between H1 and the haplotype from the 5300-year-old cob from San Marcos cave (SM10) suggests that H1, and its derivatives, were captured early in the domestication process. Considering that the genomic constitution of SM10 is intermediate to those of Balsas teosintes and maize landraces (Vallebueno-Estrada *et al.* 2016), the presence of H1 in South America also agrees well with the notion that a partial domesticate was the first to enter into the continent.

In conclusion, the structuring of plastome variation in the maize landraces examined here clearly reflects the distinction between the Andean and lowland NEA Flours gene pools previously inferred based on nuclear markers. This scenario is in accordance with the differentiation of Andean and lowland lineages from a secondary South American domestication center, as postulated by Kistler *et al.* (2018). However, our results are also compatible with the existence of two independent diffusion waves accounting for the diversity of South American landraces, as previously suggested by Freitas *et al.* (2003). In the light of our findings, extending plastome analyses to a higher number of landraces with wider geographical distribution may provide significant insights to elucidate the pattern of maize diffusion into South America.

ACKNOWLEDGEMENTS

We thank Dr. Graciela Gonzalez and Ing. Ana María Miante Alzogaray for kindly providing the teosinte seeds and maize landrace images used in this study. We are grateful to Dr. Sebastian Duchêne for his valuable suggestions regarding BEAST analyses. We are also thankful to Dr. Silvia Pietrokovsky for revising the English of the manuscript. MGL, PV and JGR performed DNA extractions and library preparations. MGL, MF and VVL analyzed and interpreted the data. MGL, MF and VVL drafted and edited the manuscript. VVL and HE conceived and designed the study. RD selected the accessions to be included in the study and coordinated sample collection. NP and AP provided technical support for NGS Sequencing. JD and JCC contributed to the work by the interpretation and discussion of the data. All authors reviewed and approved the final manuscript.

FUNDING

This work was supported by the Agencia Nacional de Promoción Científica y Técnica (PICT 2012 0325, PICT 2016 1101), the Consejo Nacional de Investigaciones Científicas y Tecnológicas (PIP 11220120100416CO 2013-2015), the Instituto Nacional de Tecnología Agropecuaria (PNBIO 1131044), and the DEANN Project.

LITERATURE CITED

Allen JO. 2005. Effect of teosinte cytoplasmic genomes on maize phenotype. *Genetics* 169: 863–880.

Andrews S. 2010. *FASTQC. A quality control tool for high throughput sequence data*. https://github.com/s-andrews/FastQC. 12 Feb. 2021.

Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* 16: 37–48.

[dataset] Beissinger T. 2016. Maize and Teosinte BAM files. CyVerse Data Commons. DOI 10.7946/P2QP4N.

Bird RM. 2012. The maze of *Zea*: I. Chloroplast SSRs and evolution. *Maydica* 57: 194–205.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.

Bosacchi M, Gurdon C, Maliga P. 2015. Plastid genotyping reveals the uniformity of cytoplasmic male sterile-T maize cytoplasms. *Plant Physiology* 169: 2129–2137.

Bouckaert RR, Drummond AJ. 2017. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evolutionary Biology* 17: 42. https://doi.org/10.1186/s12862-017-0890-6

Bouckaert R, Vaughan TG, Barido-Sottani J, *et al.* 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* 15 (4): e1006650. https://doi.org/10.1371/journal.pcbi.1006650

Bracco M, Cascales J, Hernandez JC, Poggio L, Gottlieb AM, Lia VV. 2016. Dissecting maize diversity in lowland South America: genetic structure and geographic distribution models. *BMC Plant Biology* 16: 186. https://doi.org/10.1186/s12870-016-0874-5

Cingolani P, Platts A, Wang LL, *et al.* 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 6: 80–92.

Doebley J. 1990. Molecular evidence and the evolution of maize. *Economic Botany* 44: 6–27.

Doebley J, Renfroe W, Blanton A. 1987. Restriction site variation in the *Zea* chloroplast genome. *Genetics* 117: 139–13947.

Drummond AJ, Bouckaert RR. 2015. *Bayesian evolutionary analysis with BEAST*, 1st edn. Cambridge: Cambridge University Press.

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biology* 8: 114. https://doi.org/10.1186/1741-7007-8-114

Freitas FO, Bendel G, Allaby RG, Brown TA. 2003. DNA from primitive maize landraces and archaeological remains: implications for the domestication of maize and its expansion into South America. *Journal of Archaeological Science* 30: 901–908.

García-Alcalde F, Okonechnikov K, Carbonell J, *et al.* 2012. Qualimap: Evaluating next-generation sequencing alignment data. *Bioinformatics* 28: 2678–2679.

Grobman A, Bonavia D, Dillehay TD, Piperno DR, Iriarte J, Holst I. 2012. Preceramic maize from Paredones and Huaca Prieta, Peru. *Proceedings of the National Academy of Sciences of the United States of America* 109: 1755–9.

Haas J, Creamer W, Huamán Mesía L, Goldstein D, Reinhard K, Rodríguez CV. 2013. Evidence for maize (*Zea mays*) in the Late Archaic (3000-1800 B.C.) in the Norte Chico region of Peru. *Proceedings of the National Academy of Sciences of the United States of America* 110: 4945–9.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41: 95–98.

Hartl DL, Clark AG. 2007. *Principles of population genetics*, 4th edn. Sunderland: Sinauer Associates, Inc.

van Heerwaarden J, Hufford MB, Ross-Ibarra J, Heerwaarden J Van. 2012. Historical genomics of North American maize. *Proceedings of the National Academy of Sciences of the United States of America* 109: 12420–5.

Hollingsworth PM, Li DZ, Van Der Bank M, Twyford AD. 2016. Telling plant species apart with DNA: From barcodes to genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371 371(1702):20150338. doi: 10.1098/rstb.2015.0338

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* 9: 286–298.

Kistler L, Yoshi Maezumi S, De Souza JG, *et al.* 2018. Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. *Science* 362: 1309–1313.

Krzywinski M, Schein J, Birol I, *et al.* 2009. Circos: An information aesthetic for comparative genomics. *Genome Research* 19: 1639–1645.

Leigh JW, Bryant D. 2015. Popart: full-feature software for haplotype network construction. *Methods* *in Ecology and Evolution* 6: 1110–1116.

Lia V V., Bracco M, Gottlieb AM, Poggio L, Confalonieri VA. 2007. Complex mutational patterns and size homoplasy at maize microsatellite loci. *Theoretical and Applied Genetics* 115: 981–991.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451–1452.

Liu D, Cui Y, Li S, *et al.* 2019. A new chloroplast DNA extraction protocol significantly improves the chloroplast genome sequence qQuality of foxtail millet (*Setaria italica* (L.) P. Beauv.). *Scientific Reports* 9: 16227. https://doi.org/10.1038/s41598-019-52786-2

Maier RM, Neckermann K, Igloi GL, Koössel H. 1995. Complete sequence of the maize chloroplast genome: Gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *Journal of Molecular Biology* 251: 614–628.

Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez G J, Buckler E, Doebley J. 2002. A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences of the United States of America* 99: 6080–6084.

McKenna A, Hanna M, Banks E, *et al.* 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20: 1297–1303.

Moreno-Letelier A, Aguirre-Liguori JA, Piñero D, Vázquez-Lobo A, Eguiarte LE. 2020. The relevance of gene flow with wild relatives in understanding the domestication process. *Royal Society Open Science* **7**: 191545. https://doi.org/10.1098/rsos.191545

Nock CJ, Hardner CM, Montenegro JD, *et al.* 2019. Wild origins of macadamia domestication identified through intraspecific chloroplast genome sequencing. *Frontiers in Plant Science* 10: 334. https://doi.org/10.3389/fpls.2019.00334

Orton LM, Burke S V., Wysocki WP, Duvall MR. 2017. Plastid phylogenomic study of species within the genus *Zea*: rates and patterns of three classes of microstructural changes. *Current Genetics* 63: 311–323.

Paterniani E, Goodman MM. 1977. *Races of Maize in Brazil and Adjacent Areas*. Centro Internacional de Mejoramiento de Maiz y Trigo.

Pérez-Zamorano B, Vallebueno-Estrada M, González JM, *et al.* 2017. Organellar Genomes from a ~5,000-Year-old archaeological maize sample are closely related to NB genotype. *Genome Biology and Evolution* 9: 904–915.

Provan J, Lawrence P, Young G, *et al.* 1999. Analysis of the genus *Zea* (Poaceae) using polymorphic chloroplast simple sequence repeats. *Plant Systematics and Evolution* 218: 245–256.

R Core Team. 2013. R: A language and environment for statistical computing. **R** Foundation for Statistical Computing, Vienna, Austria. URL http://www.**R**-project.org/.

Rambaut A. 2009. *FigTree, a graphical viewer of phylogenetic trees*. Institute of Evolutionary Biology University of Edinburgh.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7 (E Susko, Ed.). *Systematic Biology* 67: 901–904.

Ramos-Madrigal J, Smith BD, Moreno-Mayar JV, *et al.* 2016. Genome sequence of a 5,310-year-old maize cob provides insights into the early stages of maize domestication. *Current Biology* 26: 3195–3201.

Rivas JG. 2015. *Caracterización de la diversidad genética de razas nativas de maíz (Zea mays ssp mays) del Noroeste Argentino mediante descriptores morfométricos y marcadores moleculares*. PhD Thesis, University of Buenos Aires, Argentina.

Robinson JT, Thorvaldsdóttir H, Winckler W, *et al.* 2011. Integrative genomics viewer. *Nature Biotechnology* 29: 24–26.

Salhuana W, Pollak LM. 2005. Latin American Maize Project (LAMP) and Germplasm Enhancement of Maize (GEM) project: Generating useful breeding germplasm. *Maydica* **51**: 339–355.

Sancho R, Cantalapiedra CP, López-Alvarez D, *et al.* 2018. Comparative plastome genomics and phylogenomics of *Brachypodium*: flowering time signatures, introgression and recombination in recently diverged ecotypes. *New Phytologist* 218: 1631–1644.

Shaw J, Shafer HL, Rayne Leonard O, Kovach MJ, Schorr M, Morris AB. 2014. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. *American Journal of Botany* 101: 1987–2004.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* 68: 978–989.

Tang J, Xia H, Cao M, *et al.* 2004. A comparison of rice chloroplast genomes. *Plant Physiology* 135: 412–420.

Tonti-Filippini J, Nevill PG, Dixon K, Small I. 2017. What can we do with 1000 plastid genomes? *Plant Journal* 90: 808–818.

Untergasser A, Cutcutache I, Koressaar T, *et al.* 2012. Primer3-new capabilities and interfaces. *Nucleic Acids Research* 40: 1–12.

Vallebueno-Estrada M, Rodríguez-Arévalo I, Rougon-Cardoso A, *et al.* 2016. The earliest maize from San Marcos Tehuacan is a partial domesticate with genomic evidence of inbreeding. *Proceedings of the National Academy of Sciences of the United States of America* 113: 14151–14156.

Vigouroux Y, Glaubitz JC, Matsuoka Y, Goodman MM, Sánchez G. J, Doebley J. 2008. Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. *American Journal of Botany* 95: 1240–1253.

Weissinger AK, Timothy DH, Levings CS, Goodman MM. 1983. Patterns of mitochonidral DNA variation in indigenous maize races of Latin America *Genetics* 104.

Young HA, Lanzatella CL, Sarath G, Tobias CM. 2011. Chloroplast genome variation in upland and lowland switchgrass. *PLoS ONE* 6: e23980. https://doi.org/10.1371/journal.pone.0023980

Figure Legends

Figure 1. Geographic origin of the maize landraces and teosintes used for network and Bayesian phylogenetic analyses. Landrace samples are color-coded according to the nuclear genetic groups delimited by Bracco *et al.* (2016). Yellow: Highland Mexico and US; orange: Andean; green: Tropical Lowland; blue: NEA Flours; light blue: NEA popcorns; black: admixed. Teosintes are highlighted in light orange and the archaeological maize cob SM10 is presented in pink. Samples from HapMap2 are marked with an asterisk. Haplotype variants are given between parentheses.

Figure 2. Distribution of intraspecific variation in the *Zea mays* plastome. Inner circles correspond to the individuals sequenced in this study (pink), and the samples from HapMap 2 (maize landraces: dark green; teosintes: light green). Radial dots indicate polymorphic sites color-coded according to their functional annotation on the reference genome **KF241981.1** (black: missense mutation; green: synonymous mutation; yellow: intronic variant; blue: intergenic variant; orange: structural variant in intergenic region). Genes are depicted in black and grey. Intergenic regions are shown in green. The absence of dots indicates missing data. The complete list of markers is provided in Supplementary data Table S3.

Figure 3. A. Median-joining network of complete *Zea mays* plastomes. Circle size is proportional to haplotype frequencies. Colors denote category (inbred line, teosinte, ancient sample) or previous group assignment of landraces based on nuclear markers according to Bracco *et al.* (2016) (Andean, NEA Flours, NEA Popcorns, Tropical Lowland, Highland Mexico and US). Dashes represent one mutational step and black circles indicate missing vectors. Edge lengths are not to scale. B. Morphological diversity in south American landraces carrying H1 and H2 haplotypes.

TABLES

Table 1. Diversity indices for 51 *Zea mays* plastomes

	All markers (99)	SNP + cpSSR (91)	SNP (80)	cpSSR + indel (19)
Number of Haplotypes	27	26	24	15
Parsimony informative sites	69	62	55	14
Haplotype diversity	0.905	0.904	0.891	0.793
Average number of pairwise differences	18.875	16.858	14.405	4.470

Table 2. Haplotype distribution in the *Zea mays* categories used in this study.

	H1	**H2**	H3	H4	H5	**H6**	H7	**H8**	**H9**	**H10**	**H11**	H12	**H13**	H14	H15	**H16**	**H17**	H18	H19	H20	H21	H22	H23	H24	H25	H26	H27
Landraces[a]																											
Andean (N=10)	0	**8**	0	0	0	**0**	0	**0**	**1**	**0**	**0**	0	**0**	0	0	**1**	**0**	0	0	0	0	0	0	0	0	0	0
NEA Flours (N=9)	7	**0**	1	0	0	**0**	0	**0**	**0**	**0**	**0**	0	**0**	0	0	**0**	**1**	0	0	0	0	0	0	0	0	0	0
Tropical Lowland (N=4)	1	**0**	0	0	0	**0**	1	**1**	**0**	**0**	**0**	0	**0**	0	1	**0**	**0**	0	0	0	0	0	0	0	0	0	0
HM-US (N=4)	0	**0**	1	1	0	**0**	0	**0**	**0**	**0**	**1**	0	**0**	0	0	**0**	**0**	1	0	0	0	0	0	0	0	0	0
NEA Popcorns (N=3)	1	**1**	0	0	0	**1**	0	**0**	**0**	**0**	**0**	0	**0**	0	0	**0**	**0**	0	0	0	0	0	0	0	0	0	0
Admixed (N= 11)	2	**0**	2	1	1	**0**	0	**0**	**0**	**1**	**0**	1	**1**	1	0	**0**	**0**	0	1	0	0	0	0	0	0	0	0
Teosintes																											
ssp. mexicana (N=2)	0	**0**	0	0	0	**0**	0	**0**	**0**	**0**	**0**	0	**0**	0	0	**0**	**0**	0	0	0	0	0	0	1	1	0	0
ssp. parviglumis (N=2)	1	**0**	0	0	0	**0**	0	**0**	**0**	**0**	**0**	0	**0**	0	0	**0**	**0**	0	0	0	0	0	0	0	0	0	1
ssp.	0	**0**	0	0	0	**0**	0	**0**	**0**	**0**	**0**	0	**0**	0	0	**0**	**0**	0	0	0	0	0	0	0	0	1	0

huehuetenangensis (N=1)																													
SM10 (ancient cob)																									▨				
Inbred Lines																													
B73	▨																												
B37C																						▨							
B37S																							▨						
B37T																								▨					

Haplotypes exclusive to the samples sequenced in this study are highlighted in bold. Shaded squares indicate the haplotype of individual specimens. [a] Groups were delimited based on the SSR analysis of Braco et al. (2016).

Figure 1

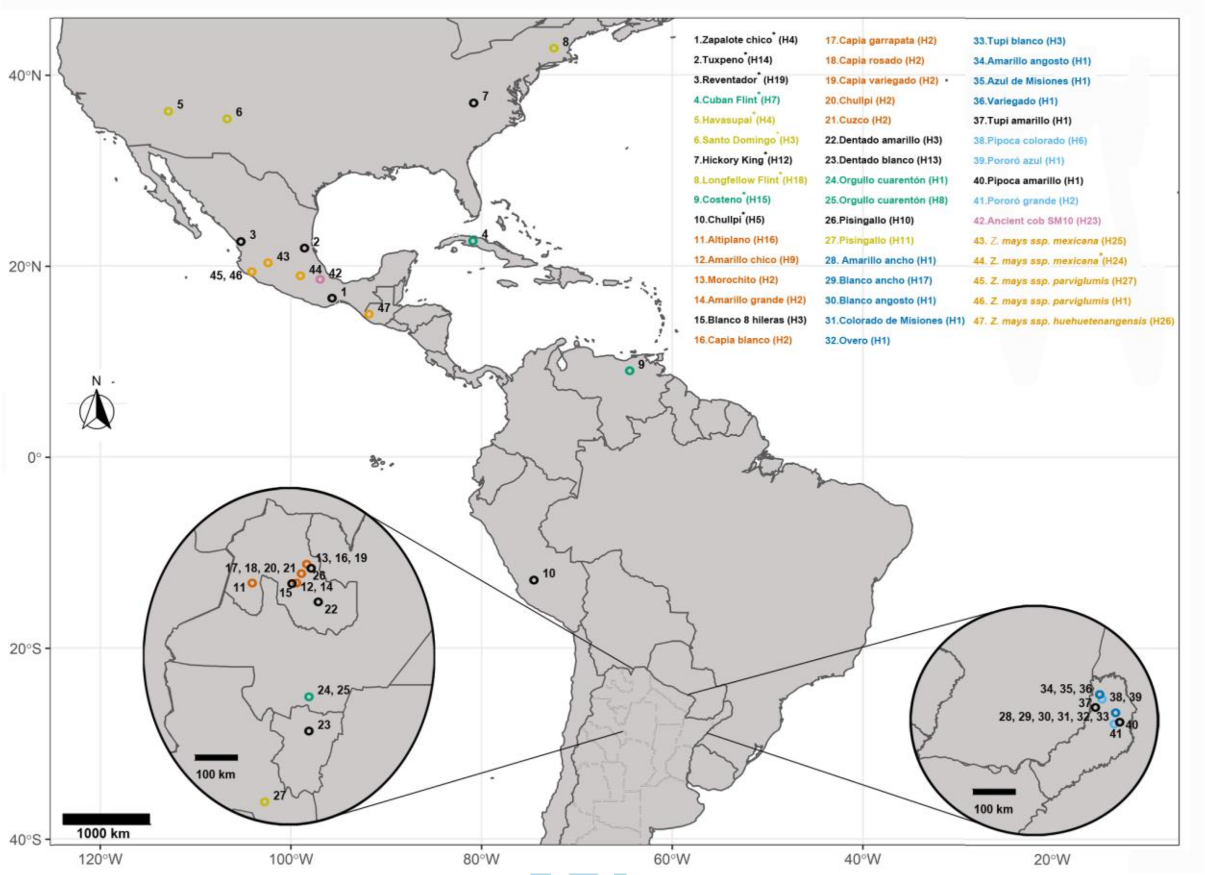

1.Zapalote chico* (H4)
2.Tuxpeno*(H14)
3.Reventador* (H19)
4.Cuban Flint*(H7)
5.Havasupai*(H4)
6.Santo Domingo*(H3)
7.Hickory King*(H12)
8.Longfellow Flint*(H18)
9.Costeno*(H15)
10.Chullpi*(H5)
11.Altiplano (H16)
12.Amarillo chico (H9)
13.Morochito (H2)
14.Amarillo grande (H2)
15.Blanco 8 hileras (H3)
16.Capia blanco (H2)
17.Capia garrapata (H2)
18.Capia rosado (H2)
19.Capia variegado (H2)
20.Chullpi (H2)
21.Cuzco (H2)
22.Dentado amarillo (H3)
23.Dentado blanco (H13)
24.Orgullo cuarentón (H1)
25.Orgullo cuarentón (H8)
26.Pisingallo (H10)
27.Pisingallo (H11)
28. Amarillo ancho (H1)
29.Blanco ancho (H17)
30.Blanco angosto (H1)
31.Colorado de Misiones (H1)
32.Overo (H1)
33.Tupi blanco (H3)
34.Amarillo angosto (H1)
35.Azul de Misiones (H1)
36.Variegado (H1)
37.Tupi amarillo (H1)
38.Pipoca colorado (H6)
39.Pororó azul (H1)
40.Pipoca amarillo (H1)
41.Pororó grande (H2)
42.Ancient cob SM10 (H23)
43. Z. mays ssp. mexicana (H25)
44. Z. mays ssp. mexicana*(H24)
45. Z. mays ssp. parviglumis (H27)
46. Z. mays ssp. parviglumis (H1)
47. Z. mays ssp. huehuetenangensis (H26)
40°N
20°N
0°
20°S
40°S
120°W
100°W
80°W
60°W
40°W
20°W
N
1000 km
100 km
100 km
17, 18, 20, 21
13, 16, 19
26
11
15
12, 14
22
24, 25
23
27
34, 35, 36
38, 39
37
28, 29, 30, 31, 32, 33
40
41
45, 46
43
44
42
47

Figure 2

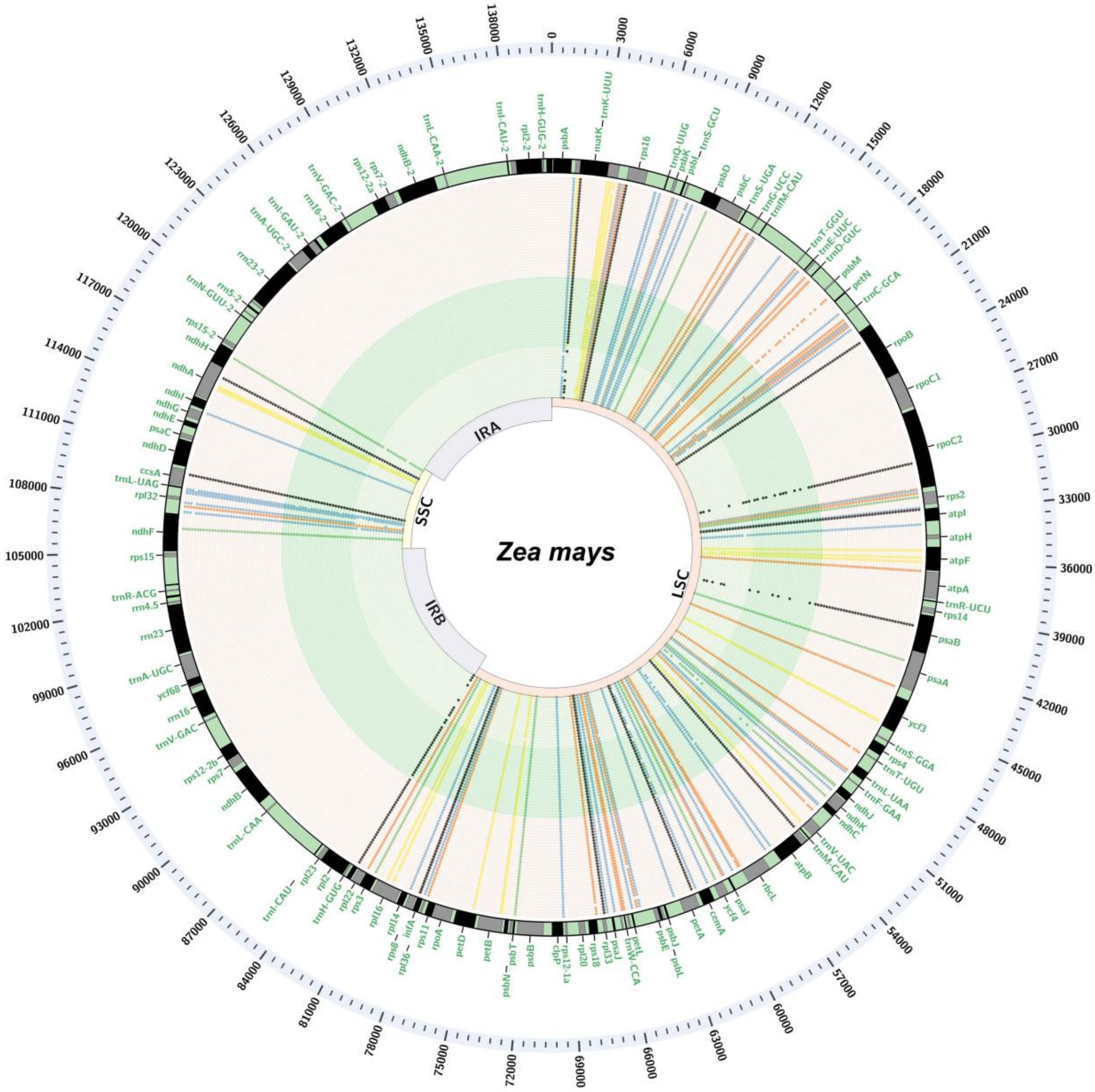
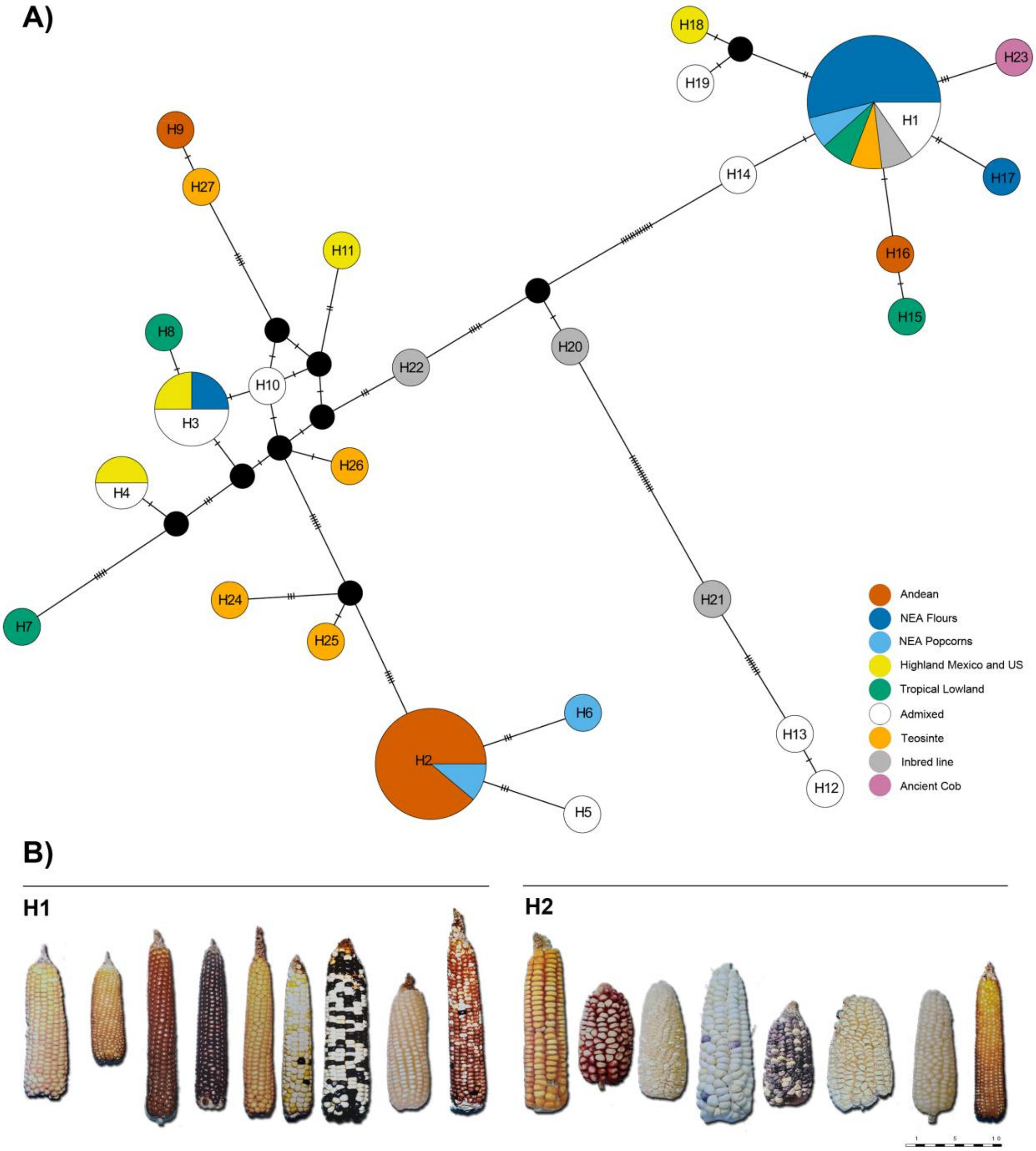

Zea mays
IRA
IRB
SSC
LSC
0
3000
6000
9000
12000
15000
18000
21000
24000
27000
30000
33000
36000
39000
42000
45000
48000
51000
54000
57000
60000
63000
66000
69000
72000
75000
78000
81000
84000
87000
90000
93000
96000
99000
102000
105000
108000
111000
114000
117000
120000
123000
126000
129000
132000
135000
138000
psbA
matK
trnK-UUU
rps16
trnQ-UUG
psbK
psbI
trnS-GCU
psbD
psbC
trnS-UGA
trnG-UCC
trnfM-CAU
trnT-GGU
trnE-UUC
trnD-GUC
psbM
petN
trnC-GCA
rpoB
rpoC1
rpoC2
rps2
atpI
atpH
atpF
atpA
trnR-UCU
rps14
psaB
psaA
ycf3
trnS-GGA
rps4
trnT-UGU
trnL-UAA
trnF-GAA
ndhJ
ndhK
ndhC
trnV-UAC
trnM-CAU
atpB
rbcL
psaI
ycf4
cemA
petA
psbJ
psbL
psbE
petL
trnW-CCA
psaJ
rpl33
rps18
rpl20
rps12-1a
clpP
psbB
psbT
psbN
petB
petD
rpoA
rps11
rpl36
infA
rps8
rpl14
rpl16
rps3
rpl22
rpl2
trnH-GUG
rpl23
trnI-CAU
trnL-CAA
ndhB
rps7
rps12-2b
trnV-GAC
rrn16
ycf68
trnA-UGC
rrn23
rrn4.5
trnR-ACG
rps15
ndhF
rpl32
trnL-UAG
ccsA
ndhD
psaC
ndhE
ndhG
ndhI
ndhA
ndhH
rps15-2
trnN-GUU-2
rrn5-2
rrn23-2
trnA-UGC-2
trnI-GAU-2
rrn16-2
trnV-GAC-2
rps12-2a
rps7-2
ndhB-2
trnL-CAA-2
trnI-CAU-2
rpl2-2
trnH-GUG-2

Figure 3


A)
H1
H2
H3
H4
H5
H6
H7
H8
H9
H10
H11
H12
H13
H14
H15
H16
H17
H18
H19
H20
H21
H22
H23
H24
H25
H26
H27
Andean
NEA Flours
NEA Popcorns
Highland Mexico and US
Tropical Lowland
Admixed
Teosinte
Inbred line
Ancient Cob
B)
H1
H2
1
5
10