UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

ESCUELA NACIONAL DE ESTUDIOS SUPERIORES, UNIDAD LEÓN

Evolutionary and Population Genomics of *Agave angustifolia*

TESIS

Que para obtener el título de

Licenciado en Ciencias Agrogenómicas

PRESENTA

Eddy Jesús Mendoza Galindo

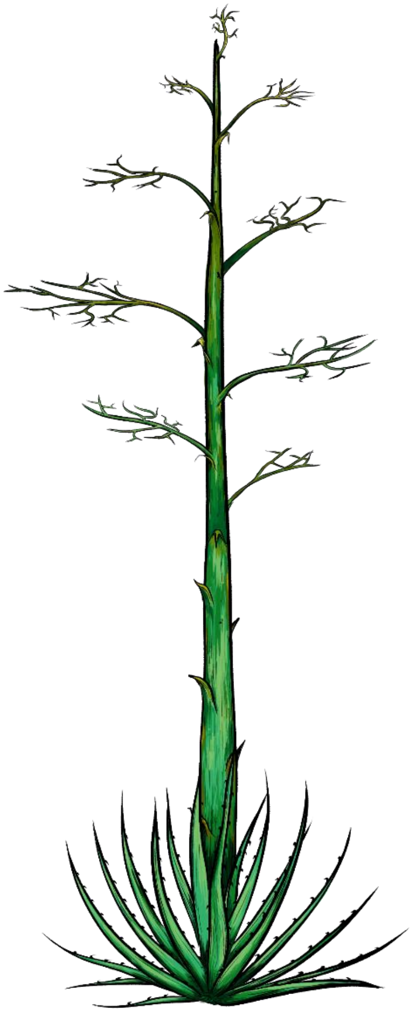
Tutora Externa: Dra. Tania Hernández Hernández

Desert Botanical Garden

Tutor Interno: Dr. Antonio Hernández López

ENES León, UNAM

León, Guanajuato. Septiembre 2022.



AGRADECIMIENTOS Y DEDICATORIA

Aunque se trate de un texto académico, llegar a escribir estas palabras significa que no llegué solo hasta aquí. Quiero que quede grabado en este texto mi reconocimiento a las personas que hicieron esto posible, que me lo hicieron posible.

A mi pá y má, por creer en mí, su incalculable apoyo, amor y enseña. Al chai y a Alan por darme más que un motivo para seguir. Sin ustedes yo no sería la persona que soy hoy. Este es un trabajo conjunto. *Estando juntos, ¿qué ha de pasar?*

A mi tío Iván por guiarme con sus pasos (y por hacer que me diera cuenta de que mi camino eran los Agaves) y a mi padrino Juan por su sonrisa eternamente fortalecedora. Los tres me han enseñado más que trabajar duro y nunca rendirse, aunque a veces el mundo parezca imposible de vencer, me enseñaron a perseguir los sueños. Mil gracias hasta donde estés padrino, dedico este pequeño logro a tu sueño.

Mis tíxs Irene, Emanuel, Mine, Aure. Su calor siempre me acompaña.

Mis amigxs, la familia que me conoce mejor que nadie y agradezco por encontrar entre este mar de gente. Estoy seguro de que yo ni estaría aquí, vivo, sin ustedes. Soy muy feliz de crecer a su lado. En orden de aparición; al Posser, Alison, Monik (krajo), Fer, Mily, Licha piciosa, Rodillas, Carlos el Old Taco, mi Rapport, Mafer, Pau gabachita, changuito, Andi, Payo, Tona, Cordero, Gus.

Gracias a mi tía Juanita por recibirme y acompañarme, también por el cafecito. A la mamá de Licha por adoptarnos.

Abuelitxs, gracias por darme buenas raíces.

Gracias a mis dos pueblitos: Platanar y Ahuateno. Aprendí más de Agaves aquí que en la computadora.

Al sonso CRISPR y al Chevy por ser mis fieles compañeros. Perdón por regresarles tan poco en esta cansadísima aventura.

ACKNOWLEDGES

It is hard to mention how thankful I am for learning from such incredible researchers before and during my degree.

Thanks to Jorge Lugo and Martha Mora for introducing me to biology and plant science.
To Piña’s lab to teach me how to work with Agave.

To all my professors at Agrigenomics, I couldn’t be happier with what I have learned from you these years. Thanks to Toño for introducing me to population genetics and evolution. Thanks to Kalpana for making me fall in love with plant genetics. I will be forever proud to wear my Agrigenomics t-shirt.

To Stewart Gillmor and Jaime Alaniz for introducing me to Arabidopsis and Development.

To Diego Ortega and Selene Fernández for letting me be partially part of their teams this year. Thanks for teaching me how it is to be part of a computational biology lab, how ideas should be coded and for all your support. I would not have been able to do many things presented in this thesis without your guidance. Thanks to Eve Coss for being a fantastic big lab sister.

To Arturo Marí and lab, now I know how it feels to be part of a TErrific science family. Thanks to Miguel and all the VBC people. Without you, it would be tough to understand what I have done here and what is coming next.

To the International Cactaceae Academic Network and the Desert Botanical Garden for their intern sponsorship, which made this work possible.

To Dianita and Marilyn for the teamwork and what they have taught me. Thanks to Víctor Flores and José Cetz for their recommendations and help during many of the bioinformatic analysis.

Finally, as a special mention, thanks to Tania Hernández. More than all the support, her great advice and mentoring, I appreciate enormously working with such an admirable person. I would not be happier to be your student. I have learned more than science and to organize my random, crazy, fussy ideas. I hope this is just the beginning of the Agave adventures in the lab.

INDEX

ABSTRACT

The genetic health of *Agave angustifolia* might be threatened by the overuse of asexual propagation and wild populations' overexploitation. Therefore, embracing modern genomic technologies is imperative to understand its evolution better and develop conservation and management strategies for these plants. We analyzed the genomic ecosystem of *A. tequilana*, founding broad and punctual patterns that may influence the genome's evolution. Secondly, we implemented a genotyping-by-sequencing approach to build a phylogenetic tree of the Agave genus and to describe the genetic diversity and structure of *A. angustifolia* populations from Northern and Central Mexico. Finally, we identified potential biogeographical associations to Agave genetics that led us to predict which populations may be at risk under climate change. Together these results highlight the importance of studying Agave plants interactively with sequencing data and open a new landscape of opportunities to understand its biology.

INTRODUCTION

Mexico is a biologically megadiverse country (Villaseñor et al., 1998). Among the vast plant diversity found within many ecosystems, the *Agave* genus is one of the most iconic and representative of the Mexican flora (García-Mendoza et al., 2019). The *Agave* genus, which means noble (from the Greek word *αγαυη*), was firstly described by Carl Linnaeus (Nobel, 1998). It is possible that it was originated in Central/South America about 30 Mya, corresponding to the Late Eocene, where a sudden global drop in CO2 has probably conduced succulent arid-adapted lineages origin like cacti (Ramawat, 2009). According to the International Taxonomic Information System (itis.gov/, consulted in August 2022), the *Agave* genus belongs to the Asparagaceae family, within the Asparagales order. Currently, there are around 200 described *Agave* species, many of which are endemic to Mexico (García- Mendoza et al., 2019). Maguey is the common name for Agave, and it has been closely linked to the culture and history of Mexico (Enrique Vela, 2014; Figure 1).

Figure 1. The cultural importance of Agave as represented in the Mixtecan Codex. Modified from Enrique Vela, (2014).

Agaves are CAM plants dominant in xerophyte scrubs, tropical deciduous forests, spiny forests and grasslands around 1,000 and 2,000 meters above sea level (Josué & Mendoza, 2007; Yin et al., 2019). Their common growth form comprises a short stem covered by thick leaves arranged in a rosette from the apex to the basement. Leaves are generally succulent and fibrous, with serrated margins and a prominent apical spine (Josué & Mendoza, 2007). Many species are monocarpic that produce flowers after around seven years (Escobar-Guzmán et al., 2008). The inflorescence is branched and disproportional to the plant size, considered one of the biggest in flowering plants (Eguiarte et al., 2021). This physiological effort compromises plant survival, which causes many Agave plants to die after they flower and thereby, they are considered Monocarpic (Nobel, 1977).

Many diverse spirits in Mexico are maguey-based (Nobel, 1998). The traditional spirit Mezcal received his name from the Nahuatl word *mexcalli* which means cooked maguey (from *metl*, maguey; and *ixcalli* cooking). Tequila is, in principle, a type of Mezcal, and it is believed that its origin remotes to an Asian input of the distillation process (Valenzuela Zapata et al., 2008; Walton, 1977). As reported by the Tequila Regulatory Council (CRT), the total production of Agave plants for tequila production reached 2 million tons in 2021 (crt.org.mx/, consulted in August 2022). According to legal statements in the Domination of Origin of Tequila (SEGOB, 1999). *A. tequilana* Weber var. Azul is the only specie that can be used for Tequila production in Mexico.

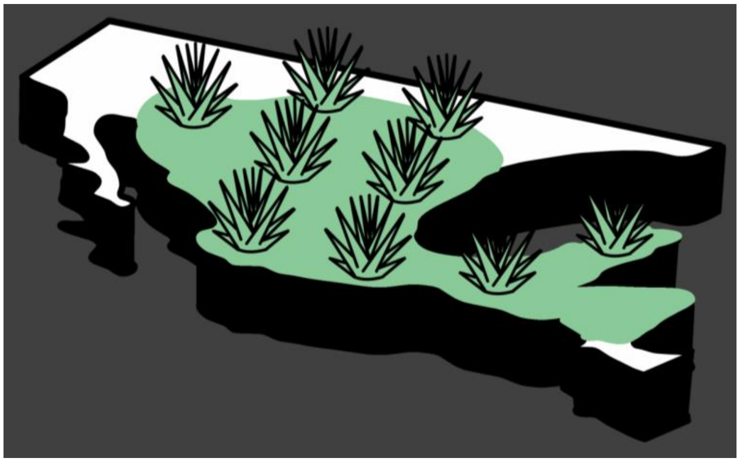


Figure 2. The natural distribution of *A. angustifolia* in Mexico. Modified from the IUCN Red List, ID:96899948/96899951 (iucnredlist.org/, Consulted in May 2022).

Although Tequila production is very well regulated, local and traditional Mezcal producers do not follow a management procedure and compete for resources against the big industries. The growing Tequila industry has enforced the pillage of plants from the wild and the exchange of varieties along the country, resulting in unorganized plant production in the center of Mexico (personal observation). Producers generally recognize landraces that are useful for Mezcal production that belong to species such as *A. angustifolia*, *A. rhodacantha* and probably *A. tequilana* (Zizumbo-Villarreal et al., 2013). Using AFLP markers, it was suggested that those species and landraces form a genetically mixed complex named the “Angustifolia complex” (Rivera-Lugo et al., 2018).

The over-exploitation of plants for Mezcal and Tequila production has raised concerns about their genetic status (Dalton, 2005). First, the plant production is ruled by asexual propagation mediated by rhizome suckers and bulbils (Abraham-Juárez et al., 2010; Arzate-Fernández & Mejía-Franco, 2011). Secondly, producers cut inflorescences at early stages to allow Agave accumulate sugars (Fructans; Gomez-Vargas et al., 2022) destined for fermentation and posterior distillation, then wait one year to harvest the stems. They call this process “capada”, which means sterilization. As a result, in most cultivars, sexual propagation is almost inexistent to preserve the homogeneous identity of the cultivars (personal observation). Even though seed production, if allowed, can reach many thousands of seeds (Nobel, 1977), we have reported extreme low seed viability and germination levels in the complex (Mendoza-Galindo Eddy & Mora-Herrera Martha E., 2021). The low seed viability can be due to environmental and physiological problems accumulated during human intervention. It has been reported that embryos may be defective or inexistent (Ramírez Tobías et al., 2016), probably a consequence of male and/or female gamete malformation (Gómez-Rodríguez et al., 2012; González-Gutiérrez et al., 2014), or selfing (Escobar-Guzmán et al., 2008).

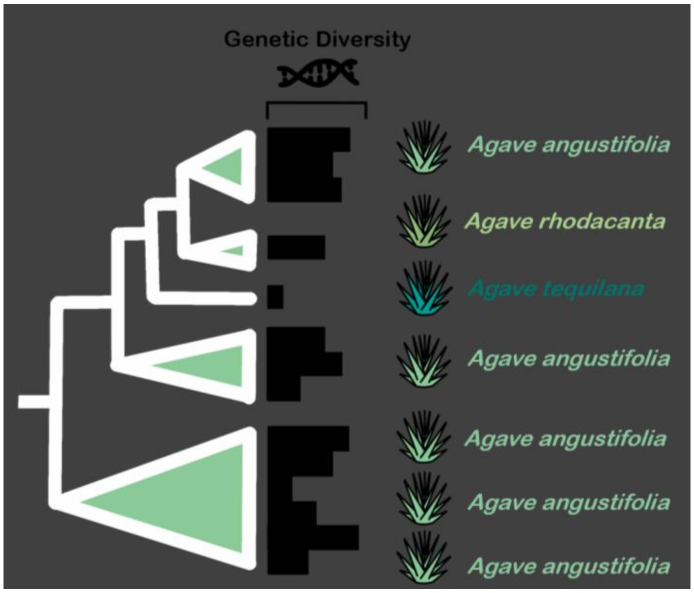


Figure 3. The Angustifolia genetic complex. Modified from Rivera-Lugo et al., (2018). Genetic diversity levels are illustrative.

Agave flowers first develop male structures, which is called protandry. Protandry arose as a mechanism to prevent selfing and inbreeding (Piven et al., 2001). Nevertheless, the pollinator-plant relationship is essential in establishing new, seed-originated Agave plants (Borbón-Palomares et al., 2018). Bats are the primary pollinator known for the complex, and their ecological relationship has been proposed as crucial for the genetic susceptibility of Agaves under climate change (Gómez-Ruiz & Lacher, 2019). It has been hypothesized that bats and Agaves may have co-evolved together, the reduction of flowers in the wild has decreased bat activity, and the lack of pollinators has contributed to poor sexual reproduction effectiveness (Eguiarte et al., 2021). For instance, integrative management of pollination and seed reproduction can avoid genetic erosion in the complex (Trejo-Salazar et al., 2016).

The ploidy levels in the complex make it challenging to study their genetics from a population perspective (Robert et al., 2008). Nevertheless, numerous efforts have been made in the complex and other Agave species using diverse genetic markers and sampling designs (Figure 4).

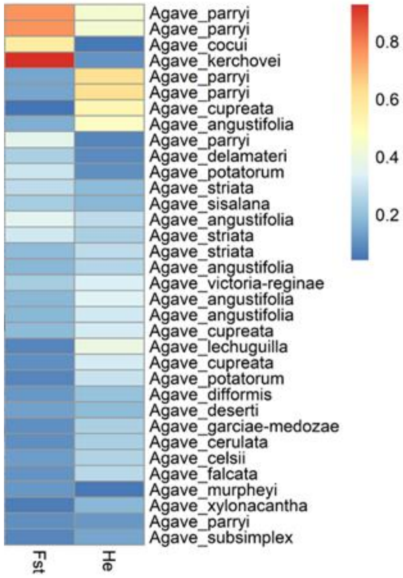


Figure 4. The genetics of Agave populations revealed by classical markers. Values were obtained from Eguiarte et al., (2013). Each row represents one population. Rows are ordered using a UPGMA algorithm. F_{ST} represents genetic divergence; He represents genetic diversity or expected Heterozygosity.

As reviewed by Eguiarte et al., (2013), the results from classical genetic markers indicate that *A. tequilana* cultivars have the lowest genetic diversity in the complex compared to the wild populations of *A. angustifolia* and the "wild-tolerated" *A. rhodacantha* (Figure 3). Nevertheless, the diversity of markers and experimental designs has made difficult to solve many discrepancies between the scientific reports. More recently, using SSR markers, it was possible to detect population structure in plantations from Jalisco, Mexico (Trejo et al., 2018), which was previously proposed using ISTR markers (Torres-Moran et al., 2013). Both studies support the Angustifolia complex hypothesis, suggesting the low-genetic-diversity-specie *A. tequilana* may be a subpopulation from *A. angustifolia* that was selected in the south of Jalisco (Torres-Moran et al., 2013; Trejo et al., 2018). This year it was published the first effort to study the genetics of *A. angustifolia* populations in Sonora, Mexico, using genomic tools. Authors found a narrow geographic population structure, and again, that cultivars are a selected subset of wild plants (Klimova et al., 2022). Nevertheless, there is lacking a study that involves the full natural distribution of the complex. Additionally, the lack of a reference genome of Agave imposes technical difficulties that have delayed understanding the complex's genetics at a finer level.

Even though no *Agave* genome has been published, there are significant efforts to understand the molecular signatures of the genus. Transcriptomic profiles are available for *A. lechuguilla* (Morreeuw et al., 2021), *A. sisalana* (Sarwar et al., 2019), *A. salmiana* (Cervantes-Pérez et al., 2018), *A. americana* (Yin et al., 2019), *A. deserti,* and *A.tequilana* (Gross et al., 2013); a chloroplast genome for *A. americana* (Yin et al., 2019), and a probe-target sequence for some species including nuclear and chloroplast DNA (Heyduk et al., 2016).

Genomes are essential to understand land plants' evolution and identify key genomic elements valuable to improve agronomic issues and conservational traits (Chen et al., 2018). Polyploidy, whole-genome rearrangements, transposable elements (TEs), and repetitive sequences are the main elements driving the evolution of plant genomes (Sahebi et al., 2018). Of these, TEs are a significant proportion of many crop genomes (Vitte et al., 2014), and it has been demonstrated their importance in shaping plant evolutionary fitness and physiology (Ariel & Manavella, 2021; Stitzer et al., 2021). They are classified into Class 1 TEs (cut-and-paste TEs or DNA TEs), and retrotransposons or Class 2 TES (copy-and-paste or RNA TEs) (Wells & Feschotte, 2020). Although the TE landscape has been examined using RNA-seq data (Gross et al., 2013), we still do not know how it looks from a genomic perspective and their importance in genome evolution.

In this way, analyzing the first Agave genome would allow us to fulfill must of the gaps still pending in understanding its evolution. Providing a more comprehensive viewpoint of its biology and genetics can also lead to developing technologies to exploit its potential better. Among many other ideas, the hypothesis that motivated this study was to prove that modern genomic technologies can allow the achievement of those goals.

OBJECTIVES

GENERAL:

Characterize the mechanisms underlying the genomic and population evolution of *Agave angustifolia*

SPECIFIC:

- Characterize the genomic landscape and the main mechanisms driving its evolution
- Build a phylogeny of the Agave genus
- Analyze population genetic variation and structure
- Identify putative biogeographical relationships associated with genetics

MATERIALS AND METHODS

GENOME ANALYSIS

The reference genome of *A. tequilana* and its annotation files were provided by Dr. José Cetz and Dr. Víctor Flores (LANGEBIO, Cinvestav Irapuato). The nomenclature of Genes and TEs was unified as follows:

Scaffolf+G(or TE)+MakerID

SYNTENY AND WHOLE-GENOME DUPLICATION ANALYSIS

Whole-genome duplication (WGD) events are common in plant evolution. A tool that has been useful in finding evidence of these events in genomes is synteny analysis, which aims to identify structurally conserved homologs (in the same genomic order; Tang et al., 2008). Thus, we performed a synteny analysis to look for evidence of WGD in the *Agave* genome.

For the intraspecies analysis, the mRNA sequences from the 16 largest scaffolds (>4 Mb) were aligned all-versus-all with BLAST (Camacho et al., 2009) (-evalue 1e-10 -num_alignments 5 -outfmt 6). Then, with default parameters, syntenic blocks were obtained with MCScanX (Wang et al., 2012). The collinearity output file was processed using custom bash, R, and python scripts to get the genome start/end coordinates for each syntenic block.

For the interspecies analysis, the protein-coding sequences and their genomic coordinates (FASTA and GFF files) for *Asparagus officinalis* were downloaded from the Phytozome platform (Goodstein et al., 2012; consulted in February, 2022), while the ones for garlic

(*Allium sativum*) are publicly available (Sun et al., 2020; https://doi.org/10.6084/m9.figshare.12570947.v1; consulted in February, 2022). Interspecies syntenic blocks shared with the 16 largest scaffolds were obtained using the python package JCVI (Tang et al., 2008) (jcvi.compara.catalog ortholog --no_strip_names). Then, we simplified the blocks following the pipeline (jcvi.compara.synteny screen --minspan=30 –simple). Both macro and microsynteny plots were generated from those selected shared blocks using JCVI.

To determine if there was a recent whole-genome duplication (WGD), we calculated synonymous substitutions per site (*Ks*) for each syntelog (synteny-conserved paralog). Using this approach one can distinguish peaks of syntelog pairs that share the same *Ks* value. Those peaks reflect evolutionary divergence among the two compared genomes, and for instance, the evidence of WGD. The above interspecies BLAST+MCScanX pipeline was repeated using all scaffolds. Shared syntenic blocks with asparagus and garlic were also re-calculated with JCVI using all scaffolds. The asparagus proteins file and its adjunct files (FASTA and GFF) were downloaded from ENSAMBL Plants (Cunningham et al., 2022). The syntenic gene pairs were aligned using ParaAT (-m mafft -f paml) (Z. Zhang et al., 2012). Then, we calculated *dS* (*Ks*) for each pair using the yn00 algorithm inside PAML (Yang, 2007).

GENOME ECOSYSTEM CHARACTERIZATION

We hypothesized that the genome can be understood as an ecosystem of different elements, dynamics, and relationships (Stitzer et al., 2021). To test this, we first characterized TE families. To construct a phylogeny for each TE class, we randomly selected 30,000 TE sequences (due a large amount of TEs in the *A. tequilana* genome). We then isolated TE protein domains in the REXdb plant database with TEsorter (R.-G. Zhang et al., 2022). Then, we isolated the Reverse Transcriptase domain for Class I TEs and the Transposase for Class II TEs. Alignments were built with MAFFT with the –auto flag (Katoh & Toh, 2010). Maximum Likehood trees were then built with IQTREE with 1000 bootstrap repetitions, enabling searching for the best model implementation (Minh et al., 2020). Trees were finally visualized in the iTOL web server (Letunic & Bork, 2021).

Secondly, we aimed to understand the gene-TE relationship. We calculated intergenic distances and isolated TEs located upstream genes. We measured their distance to their closest gene, length, and density of each TE upstream gene using BEDTOOLS (Quinlan & Hall, 2010). The mean expression of genes was calculated as described in the next section.

Then, we aimed to understand the physiological and adaptative contribution of each genomic element (Protein-coding genes, non-coding genes and TEs). To test this idea, we analyzed the transcriptomic profiles of *A. tequilana* main organs (leaf, stem and roots) based on the expression of each genomic element. Genes were classified as coding or non-coding with the online platform of CPC2 (Kang et al., 2017). Paired-end RNA-seq raw data were downloaded from the NCBI BioProject accession PRJNA193469 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA193469; consulted in March 2022). TRIMMOMATIC (Bolger et al., 2014) was used to remove adapters and low-quality reads (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 SLIDINGWINDOW:5:20 MINLEN:60). Afterwards, we indexed the *A. tequilana* genome and aligned the RNA-seq reads with STAR (--outSAMtype BAM SortedByCoordinate --outSAMattrIHstart 0 --alignMatesGapMax 120000 --outSAMstrandField intronMotif) (Dobin et al., 2013). For the gene expression quantification, we implemented the expression estimation mode of STRING TIE guided by the gene annotation (-e -f 0.3 -j 15 -c 2.5) (Pertea et al., 2015). For the TE expression quantification, we used KALLISTO (Bray et al., 2016) (previous genome indexing) against the consensus TE family FASTA file generated by REPEATMODELER (Flynn et al., 2020) using the following flags: --bias --fragment-length 200 --sd 50. Finally, differential expression analysis was carried out with the R package DESeq2 (Love et al., 2014).

GENOTYPING BY SEQUENCING

SAMPLING, DNA EXTRACTION AND SEQUENCING

A. angustifolia wild, cultivated, and botanical collection samples were collected from selected populations all over their natural range in the center-to-north pacific coast of Mexico. Dr. June Simpson donated 25 accessions from selected Agave species from CINVESTAV Irapuato. Approximately 300 mg of each sample was grounded using liquid nitrogen in a mortar. Samples of different managed varieties were provided by Dr. Danae Cabrera from Centro Universitario de Ciencias Biologicas y Agropecuarias (CUCBA) from Universidad de Guadalajara.

Approximately 300 mg of each sample was grounded using liquid nitrogen with a mortar and pistil. We extracted high-molecular-weight DNA using the DNeasy Plant QUIAGEN® Kit. We implemented a double-digest restriction protocol using BglII y DdeI. DNA was sequenced on a NovaSeq Illumina platform in the Genomic Servicies department at LANGEBIO, CINVESTAV Irapuato.

SNP CALLING AND POPULATION GENOMIC ANALYSIS

Raw reads quality was assessed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC (Ewels et al., 2016). Since further quality filtering was not necessary (https://github.com/somnya/agave_genomics/blob/main/popgen/multiqc_report_agave_gbs.html), raw reads were aligned to the reference genome using the default algorithm of the Burrows-Wheeler Aligner (Li & Durbin, 2009). Intermediate SAM files were converted to

BAM and sorted by coordinates using SAMTOOLS (Li et al., 2009). The variant calling procedure was conducted using the gstacks pipeline from STACKS (Rochette et al., 2019) with default parameters. The filtering and population genetic stats calculations were then conducted using the populations pipeline from STACKS with the following arguments: -p 2 -r 0.8 -R 0.8 --min-maf 0.1 --min-mac 2 -H -b 1000.

Tajima's D calculation was implemented within genomic windows of 10 kilobases with VCFTOOLS (Danecek et al., 2011).

Towards construct the *Agave* genus phylogenetic tree, we first converted the variant calling file (VCF) into a PLINK's PED file and then to a FASTA format using a public script (https://github.com/gungorbudak/ped2fasta/blob/master/ped2fasta.pl). Then, we used IQTREE to build a maximum-likehood tree with the previous arguments

To characterize the genetic diversity and structure of the Angustifolia complex, we removed all species not belonging to it in the following procedures. The original dataset was pruned by linkage disequilibrium using PLINK (--indep-pairwise 50 10 0.1) (Chang et al., 2015). It is known that closer SNPs in a genomic region can be inherited together and linked (Hahn, 2018). Since we selected one SNP within a genome window of 50 SNPs, we named this SNP dataset "unlinked" after that.

Principal component analysis and genetic distances calculations were executed using the unlinked SNP dataset in PLINK. To add more evidence of population structure, we also performed an ADMIXTURE analysis using K values from 2 to 8, ten random seeds per K value and 200 bootstraps (Alexander et al., 2009).

BIOGEOGRAPHIC RELATIONSHIPS

ISOLATION BY DISTANCE

To see if the genetic structure can be explained by isolation by distance, we calculated the correlation between genetic distances and geographic distances. We used the PCA eigenvectors as input to calculate the "genetic distance" matrix. Geographic distances were measured using the geographic decimal coordinates from each sample. Both genetic matrices were constructed using the dist() function in R. Hierarchal clustering of genetic and geographical distances was performed in R using the function hclust(). A mantel test, which measures the correlation between two distance matrices, was performed with the R package ade4 (Dray & Dufour, 2007) with 1000 repetitions.

RESTRICTION-ASSOCIATED WIDE GENOME ASSOCIATION TO BIOCLIMATIC DATA

To statistically test for a correlation amongst genetic variation and climate, we performed an association analysis. Historical climate data (1970-2000) was downloaded at a resolution of 2.5 minutes from the WorldClim online database, one of the most comprehensive and detailed weather databases with high spatial resolution used for research and modeling (https://www.worldclim.org/data/bioclim.html). Coordinates from the sampling points were used to isolate the climate values using the R package raster (https://github.com/rspatial/raster). Bioclimate values for each sample were implemented as a phenotype to run an association analysis in PLINK (--linear "interaction"). SNPs with a P value lower than 0.001 were considered as “associated”.

To functionally characterize the genetic association to climate, we performed a Gene Ontology (GO) Enrichment analysis. A custom script and BEDTOOLS were used to identify genes cointaining associated SNPs. We identified Asparagus protein homologs (Harkess et al., 2017) with BLAST (-evalue 1e-30 -num_alignments 5) since they are functionally annotated. The resulting homolog IDs were used as input in the ShinyGO web server (Ge et al., 2020) to execute the GO enrichment analysis with default parameters.

Population structure can bias many association results (Hahn, 2018). To remove population structure background noise we used PCA eigenvectors as phenotypes for each sample to execute another GWAS in PLINK. PCA-associated SNPs (P value < 0.05) were removed from the list of SNPs associated with climate to create a candidate SNP dataset.

GENOMIC OFFSET

To predict the vulnerability of the Angustifolia complex under climate change, we implemented the workflow proposed by Aguirre-Liguori et al., (2021).

Forecast climate data was downloaded from WorldClim at a resolution of 2.5 minutes from the CMCC-ESM2 Global Climate Model that assembles a climate change in future scenarios (Lovato et al., 2022). For the mild scenario, the 245 Shared Socio-economic Pathway (SSP) between 2041-2060 was chosen. For the high-risk scenario, the 585 SSP between 2061-2080 was chosen.

The Specie’s Distribution Models were built in the R package dismo (https://github.com/rspatial/dismo). Models for current and future scenarios were constructed using the bioclimatic data and the bioclim() function. The difference of the suitability values between the current and future models was defined as δ.

In order to develop the machine learning input dataset for the workflow, we calculated minor allele frequencies of the Candidate SNP dataset using the sampling locations as subpopulations. A custom python and R script was then used to shape the data frame.

The reference SNP dataset was constructed using SNPs unrelated to the climate or PCA (P value > 0.8). The candidate SNP list was also extracted to avoid intersections.

The GradientForest R package (Ellis et al., 2012) was used afterward in the machine learning approach. Models for both candidate and reference SNPs were built with the gradientForest() function using the bioclimatic values as predictor variables and the allele frequencies as response variables (ntree = 500, maxLevel = log2(0.368*nrow(candidate)/2), trace=T, corr.threshold=0.50). In this way, we build the models using sampling locations as populations.

Allele turnover was tested for both SNP datasets using the cumimp() function and the most representative bioclimate variable as a predictor.

To calculate the Genomic Offset (the vulnerability to climate change), we used the gradient forest model from the candidate SNPs, and the two forecasted bioclimatic values for each location to predict allele frequencies in such scenarios (predict()). Euclidian distances between the present and forecasted allele frequencies were calculated for each subpopulation.

CODE AVAILABILITY

The bioinformatic workflow and all custom scripts mentioned before are publicly available at https://github.com/somnya/agave_genomics.

RESULTS AND DISCUSSION

I - THE GENOMIC ECOSYSTEM OF *A. TEQUILANA*

TWO MAJOR WHOLE GENOME DUPLICATION EVENTS PROCEED THE ORIGIN OF THE MODERN AGAVE GENOME

The genome assembly of *Agave tequilana* is composed of 11,948 Scaffolds and 37,653 predicted genes (*in prep*). To visualize the overall architecture and composition of the assembly, we calculated the accumulative size representation of three proposed scaffold size thresholds: >2.5 bp (50 Scaffolds), >3 million bp (30 Scaffolds), and >4 million bp (16 Scaffolds) (Supplementary Figure 1). There was no significant difference between the percentage of the assembly representation of the threshold size cuttings proposed. Also, we noted scaffold size gap from 3.4 to 4.2 million basepairs. Thereby, we selected the 16 largest scaffolds for further visualization.

We observed a chromosome-like distribution of genes and TEs in some scaffolds. That is, there is a low or null presence of TEs and genes in what appears to be the centromeric regions, in the middle of the scaffold. Scaffolds S00 to S04 show a high peak in TE distribution (mainly composed of Class I TEs) while low or null presence of genes in centromeric-like regions (Figure 5A). Still, some scaffolds like S05, S07 and S13 have an extremely low density of genes. In general, Class I TEs are the most overrepresented in the genome.

The distribution of SNPs is correlated to genes. Tajima's D values are generally higher than 0, with few exceptions (Figure 5B). This means genes harbor high levels of common and expected diversity that may be simply associated with genetic drift. The proportion of SNPs within TE regions and the type of diversity (D value) they represent was not calculated and is still unknown. Also, identifying genes subjected to rare variation (i.e., those with D values lower than 0) would be interesting to understand if such genes are crucial players in some of the *Agave* physiological characteristics

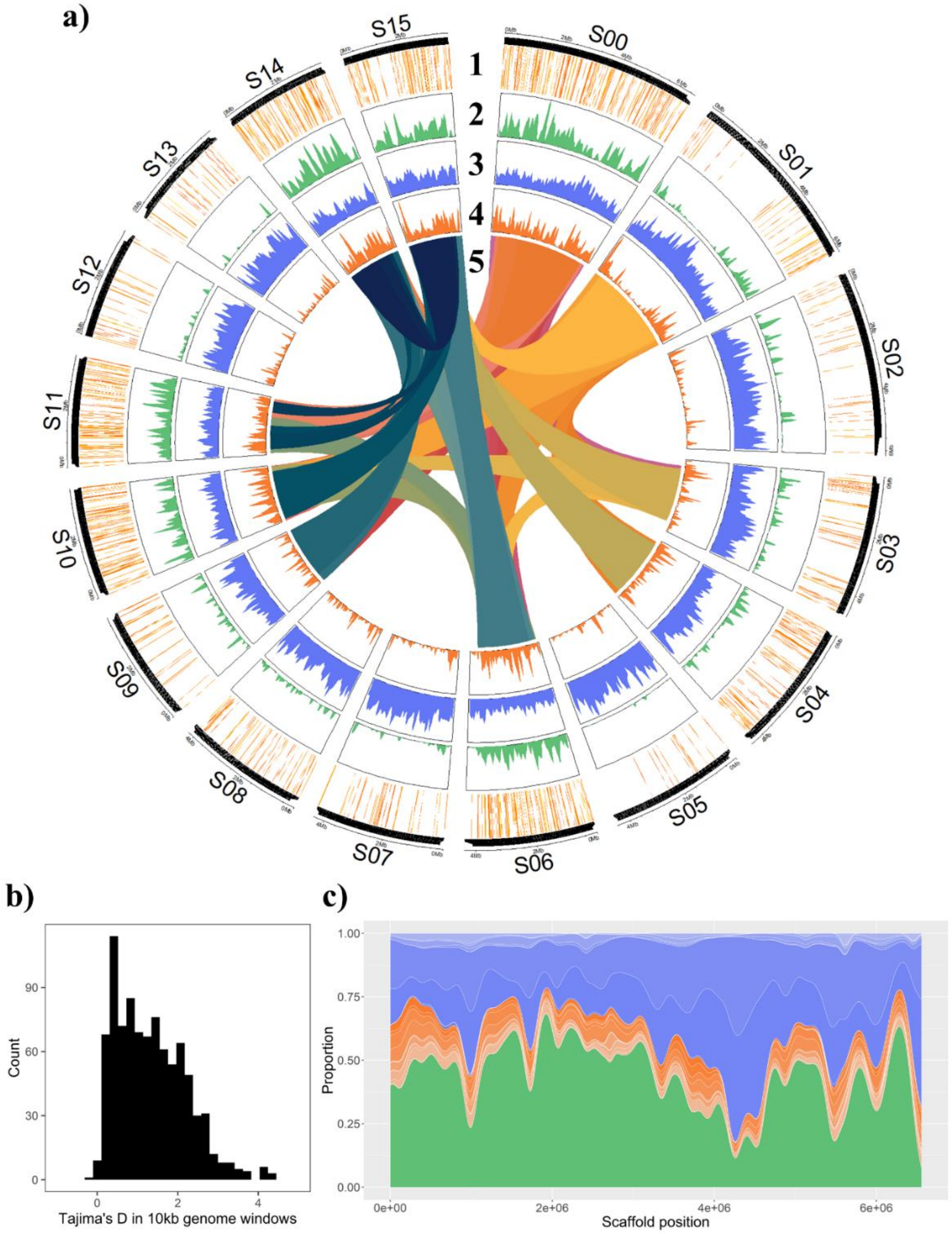


Figure 5. The genomic ecosystem of *A. tequilana*. a) The 16 longest scaffolds of the assembly; 1: heatmap of Tajima's D values in 10 kb windows (yellow-red); 2: exon density (green); 3: Class I TE density (blue); 4: Class II TE density (orange); 5: syntenic blocks (rainbow). b) Summary of genome-wide Tajima's D analysis. c) A close-up of the scaffold S00. Color codes are the same as in a). Different opacity levels correspond to different TE families.

Syntenic blocks were identified using paralogs, genes that may have diverged but still share similarities (Tang et al., 2008). It is observed that the presence of syntenic blocks also correlates with the distribution of genes. In this way, scaffolds with low gene representation do not share syntenic regions with other scaffolds. On the other hand, many scaffolds share most of their composition with at least other three scaffolds. For example, scaffolds S00, S01, S14 and S15 are all represented by scaffolds S03, S04, S06, S09, S10 and S11. Since we are visualizing the largest scaffolds, it may be possible that many of them are part of the same chromosome, the analysis is biased to miss annotation or assembling issues (as a result of the high repetitive characteristic of the genome) or those are duplicated chromosome fragments coming from a recent whole genome duplication (WGD). We sought evidence comparing the genes within those syntenic blocks to investigate the last hypothesis.

To gain insight into the possible WGD's origin, we also identified the syntenic blocks shared with Asparagus (*Asparagus officinalis*) and Garlic (*Allium sativa*), members of the same family and order, respectively. For each pair of syntelogs (Reyes-Chin-Wo et al., 2017); paralogs that share genomic ordering, we calculated the number of synonymous substitutions (*Ks*).

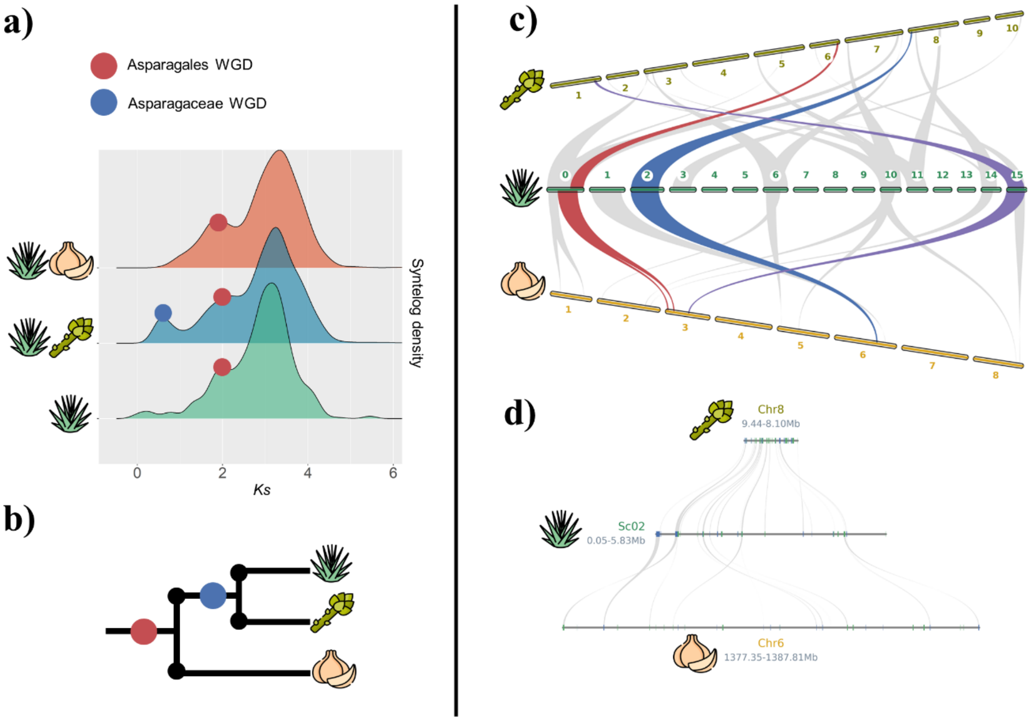


Figure 6. WGD events preceding the origin of the modern genome. a) *Ks* analysis of syntelog pairs. b) Proposed model of the WGD events in the Asparagales order, c) macrosynteny and d) microsynteny representations.

When comparing Agave vs. Agave syntelogs (intraspecies *Ks* analysis), we observe that most syntelogs pairs have a *Ks* value of around three, and a small fraction have a *Ks* value of approximately two (Figure 6A). Interestingly, few syntelogs pairs have values close to 0 and 1. Based on this evidence, we cannot assume the syntenic blocks observed before originated from a whole genome duplication preceding the origin of the modern *A. tequilana* genome. Nevertheless, the fact that some of the pairs have not diverged enough (referring to the small peaks near *Ks* values of 0), implies that including closer species would be needed to discard the possibility of a newer genome duplication completely. Contrary to our expectations, we could not see a peak of the density of syntelogs pairs close to *Ks* values of 0, which would have suggested a recent duplication before the origin of the modern genome.

We then compared Agave vs. Asparagus (interspecies analysis). The Asparagus plant was the first one in the family to be sequenced (Harkess et al., 2017). The genome assembly and annotation are now the best characterized within the taxa. When comparing the syntelog divergence, we noted the same density peaks around *Ks* = 2 and *Ks* = 3 were also shared. Besides this, we found another prominent density peak around *Ks* = 0.5. Given that the time divergence of Agave and the asparagus remains unknown, those Ks values correlate to the idea that a WGD occurred before this taxa's origin. Therefore, we suggest there was a WGD in the Asparagaceae family (Figure 6A).

Finally, we compared Agave vs. Garlic as the last interspecies analysis. Although Garlic is not the most recent or the only sequenced plant within the Asparagales order, it has the only fully public genome assembly and annotations (X. Sun et al., 2020). When we compared Garlic vs. Agave syntelogs pairs, we also found the same pattern of a prominent density peak around *Ks* = 3 and a small one around *Ks* = 2. Given that Garlic would be the most divergent plant in this analysis (and within the Asparagales order), we propose the peak we observe in all three comparisons around *Ks* = 2 is a remanent of an ancient WGD in the Asparagales order. This idea was previously discussed (X. Sun et al., 2020) and together with our results, we named this WGD the Asparagales WGD (Figure 6A).

Including closer, intermediate and far species in this analysis would allow us to refine the relative timing where these genome duplications happened. Although some of those species have been sequenced already (*Yucca;* Heyduk et al., 2021), the genome annotations and assembly are not accessible. Nevertheless, we suggest two WGD events preceding the origin of the modern *A. tequilana* genome: the Asparagales and Asparagaceae WGDs (Figure 6B). To gain a deeper understanding of those events, a gene family duplication and expansion analysis (de Bie et al., 2006) would let us identify genomic and physiological insights into the role of WGD in the origin of this species. Some possible hypotheses we can think about to focus on while doing such analysis would be: Have the genes involved in Fructan/CAM metabolism been subjected to an expansion? Which gene families have been reduced in the origin of the Agave genome? And so on.

TE FAMILIES HAVE NICHE-LIKE CHARACTERISTICS

At this part, we know genes can tell us a partial but complicated story about the origin of the Agave genome. More extensive genomic rearrangements are common in plants and better tolerated than in animals (Zhao & Schranz, 2019). Still, plant genomes are highly repetitive because of the activity of TEs (Marí-Ordóñez et al., 2013). We then aimed to look up punctual evolutionary patterns in the Agave genome. For this purpose, we implemented the concept of the genomic ecosystem for its outstanding representation of the mobilome and genome and their relationships as a system (Stitzer et al., 2021).

TEs comprise around 80% of the *A. tequilana* genome. When we looked at the family-specific distribution of TEs along the largest scaffold (Figure 5C), it was appreciated that TEs distribute depending on their identity. To further characterize the TE landscape in the Agave genome, we randomly selected 30 000 TE sequences and extracted TE-related protein domains. Of the 1,517 TE identified proteins, 1433 corresponded to Class I TEs or the so-called Retrotransposons and 84 corresponded to Class II TEs or the so-called DNA TEs. Maximum-likehood (ML) phylogenetic trees of the TE proteins revealed Superfamily classification explains most of the evolutionary relationships between the Agave TEs. We choose the Reverse Transcriptase domain and the Transposase domain to build the trees for Class I and II TEs, respectively. Class I TEs are mainly composed of the superfamilies Gypsy and Copia, some LINE elements and Pararetrovirus-like TEs that seem related to Gypsy TEs (Figure 7A). Class II TEs are divided into three main superfamilies: hAT, MuDr and CACTA (Figure 7B). The history of each TE family may constrain the expansion of TEs in the Agave genome. TEs are complex genomic structures composed of many protein domains (functional or not) (Wells & Feschotte, 2020). Using only one representative domain to infer evolutionary relationships can lead to miss placing of some TEs into the wrong category. In order to build a more comprehensive phylogenetic tree of the Agave TEs, it would be essential to incorporate the whole TE structure, story and composition into the analysis. Current analytical methods like machine learning could be used to achieve this aim. Anyway, a polished TE annotation would also be necessary.

It is well established that TE activity act as a selection pressure and mutation source (Kidwell & Lisch, 1997). TEs can jump inside or near genes and knock out or modify how they are regulated (Iwasaki et al., 2019). Following the concept of the genomic ecosystem where the relationships between all the organismic communities (in this case, genomic elements) shape the system economy, which we interpret as evolving capability (Stitzer et al., 2021), we tested the idea that TEs possess niche-like characteristics. A niche-like “phenotype” would be interpreted as evidence of differential dynamics and characteristics among the TE families. For that, we characterized TEs upstream of all the annotated genes. We identify 69,421 TEs within a 10-kilobase window upstream of the 37,653 genes (Figure 7C). The Class I Gypsy and Copia were again the most common superfamilies upstream genes, while the Class II ot1 and Ginger were the ones with less representation. The mean trend for the localization of most of the TEs was around five kilobases upstream genes. DNA and SINE/LINE elements are generally closer to genes than other retroelements. Notably, some families have more than one distribution peak (i.e., DNA/Ginger, DNA/Novosib, LTR/Ngaro, LTR/ERV1, LINE/R1). This may be due to family-specific lengths and age, where young TEs are generally longer, and older TEs are highly degraded and shorter (Marí-Ordoñez et al., 2013). Longer TEs will then be more distant to genes than older TE sequences that are highly degraded and fixed near promoter regions. This observation coincided with the distribution of TE lengths, where families that are evenly distributed upstream genes also have normal-distributed sizes. Due to retroelements' repetitive and copy-paste behavior, many Copia/Gypsy TEs have uncommon long lengths. ANOVA tests for counts, distance and length, showed significant differences over the TE families ($p < 0.001$).

It was recently proved TE proximity negative correlates to gene expression (Edger et al., 2019) since TEs and genes are both methylated in a row when close enough (Stitzer et al., 2021), hence we also inquired about gene expression using a prey-predator analogy. In general, mean gene expression values of genes downstream TE families were statistically different ($p < 0.05$). Some of the outliers in the analysis were related to genes close to Gypsy/Copia elements. Since gene expression depends on many other mechanisms than TE closeness, we cannot further discuss why this happened. To gain more information on the relationship between TE closeness and gene expression, we need to identify tissue/time-specific patterns.

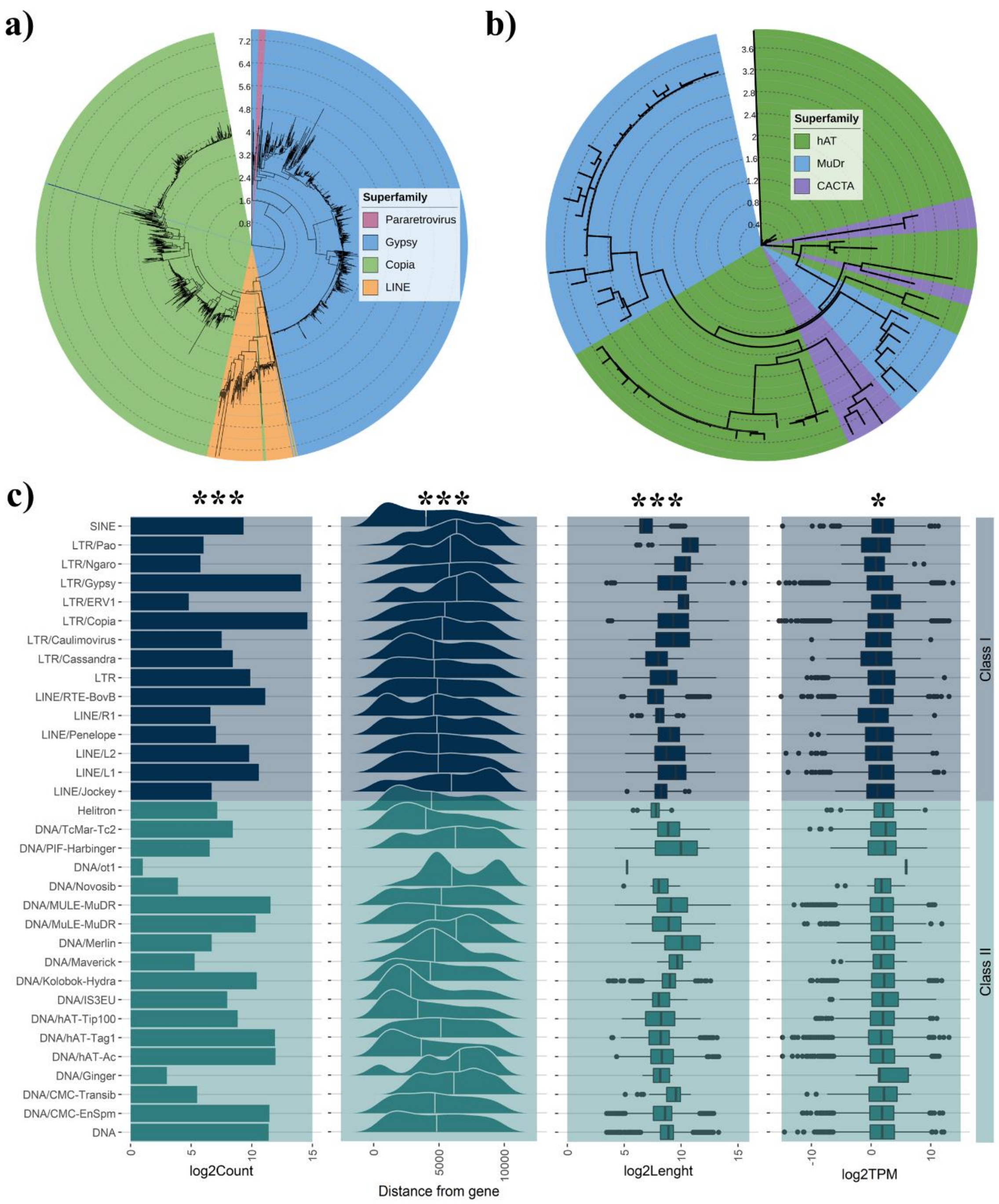


Figure 7. The TE landscape of the *A. tequilana* genome. a) Class I TE ML phylogeny based on the Retrotranscriptase protein domain. JTT+F+R5 was selected as the best nucleotide substitution model. b) Class II TE ML phylogeny based on the Transposase protein domain. VT+F+G4 was selected as the best nucleotide substitution model. c) Family-based characterization of TEs upstream genes. Asterisks show statistical ANOVA differences (*** = p value < 0.000; * = p value < 0.05).

EACH GENOMIC ELEMENT HAS A DIFFERENT ROLE IN SHAPING THE TRANSCRIPTIONAL PROFILE OF TISSUES

Knowing that TE and Genes (and their relationship) can tell different stories about the evolution of the genome from a broad and specific perspective, we wondered how important each element would be for the species' physiological functioning and adaptative potential. For this purpose, we re-assembled the *A. tequilana* transcriptome (Gross et al., 2013). The transcriptome comprises adult plants' leaf, stem and root samples (Figure 8A). In this manner, we hypothesized that we should expect a clear clustering of the samples based on the tissue identity when using multivariate analysis. A poor contribution of the genomic element's physiological functioning and adaptative potential would be reflected in the absence of tissue clustering and differential expression. For that, we divided genes into protein-coding and non-coding, resulting in 32,155 coding genes and 5,408 non-coding genes.

The multivariate and differential analysis revealed Coding genes contribute the most to shaping the transcriptional profile of the Agave organs (Figure 8B). Principal Component Analysis (PCA) showed that samples cluster depending on the organ identity. Euclidean distances between samples suggest transcriptional profiles of leaves and roots are the most different (differences around values of 150). In contrast, stem samples are intermediate between those two (differences around values of 120). Interestingly, we found some samples from the same organ identity may be slightly different (difference values of around 50). Differences within the same organ may be due to environmental factors that alter the transcriptional response (Palande et al., 2022). All these observations were already expected. Development and ontogeny relationships between the plant organs are stablished early in embryogenesis, where root apical meristems and shoot apical meristems role the way the plant will grow after germination (Armenta-Medina & Gillmor, 2019. As a side observation, this information about tissue-specific gene expression could further characterize pathways of interest like CAM and Fructan metabolism, biotic and abiotic stress responses, etc.

Non-coding genes also have importance in shaping the transcription profiles (Figure 8C). We observe an apparent clustering in the PCA. Despite this, Euclidean differences and differential expression are reduced. The highest Euclidean difference is around a value of 50, which is the amount of difference we observed previously within samples from the same organ. In the same way, differential expression of non-coding genes has less statistical support even though fold changes remain like coding genes. A more detailed classification of these non-coding genes would be essential to characterize their biological importance further. Many plant non-coding genes have been studied due to their importance in specific physiological events such as age and flowering (Buendía-Monreal & Gillmor, 2017), germination (Fedak et al., 2016) and vernalization (Csorba et al., 2014). Their specificness may imply the entire population of non-coding genes is not as important as we observed with coding genes. Besides, this first examination of Agave non-coding genes could allow further identification of previously described genes. For example, developing biotechnological approaches to study and modify the flowering process in Agave is a goal that has been pursued by academic and industry institutions for a long time.

TE family-based transcriptional profiles are the worst defined in the comparison (Figure 8D). Regarding some root samples that cluster together apart from the rest, there is no well-defined clustering based on the plant organs. Most of the samples show a slight clustering respecting their organ. Nevertheless, they are mixed in a supergroup. As observed before, Euclidean distances do not show considerable distances between the organs. Differential expression, although kept, has no sufficient statistical support. These observations are consistent with previous knowledge of plant TE activity. TE activity is generally not tissue/time-dependent but stress-related (L. Sun et al., 2020). In this case, similar clustering of some samples could be due to similar environmental conditions rather than their cell identity. This hypothesis is supported by the differential expression of some TE families in different tissues. Those differentially expressed TEs do not have statistical support given their specificity to some samples. It should be mentioned that using family-level quantification may lead to copy number bias in the results. Due to the highly repetitive behavior of TEs, we cannot assume the expression levels are a result of proper TE activity or a result of the high copy number of the family. Including stress-induced transcriptomes and a refined quantification and pipeline could enhance our understanding of Agave TE activity.

Our results suggest each genomic element, coding genes, non-coding genes, and TEs, contribute differently to defining the transcriptional profile of the *Agave* organs. From an evolutionary perspective, these differences can be considered a differential contribution to the adaptative potential of the *Agave* genome. Thus, if a mutation is introduced in a TE region, selection will not have the same influence as on a gene. Genetic drift would then be the main force driving the variation of TE-related polymorphisms along the genome. As expected, genetic variation in gene regions would have more substantial effect on the fitness of the specie under environmental stresses. To further comprehend how genetic variation can influence the evolution of the *Agave* plant, we then characterized its genetics at a population-level.

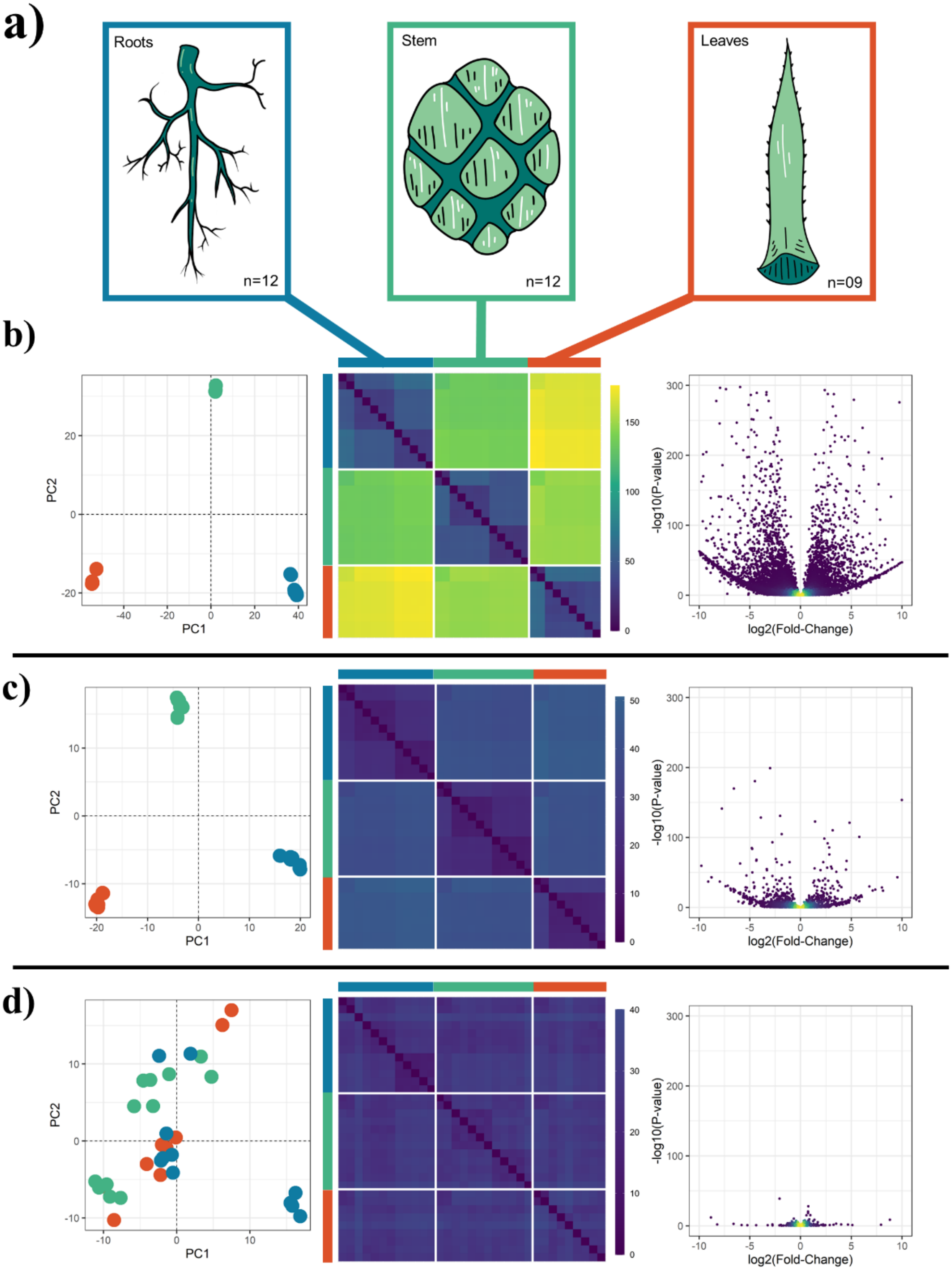


Figure 8. Transcriptional profiles of Agave organs based on each genomic element. a) Tissue samples used in this experiment. b) Protein-coding genes. c) Non-coding genes. d) TEs.

II - POPULATION GENOMICS OF *A. ANGUSTIFOLIA*

THE FIRST AGAVE PHYLOGENY USING WHOLE GENOME DATA

At the time of the writing of this thesis, there has not been told an evolutionary story of the Agave genus based on whole genomic data. Understanding the phylogenetic relationships of *A. tequilana* to their relatives is essential to discuss its origin and the domestication process behind it. For this aim, we collected 175 plants belonging to the so-called “Angustofolia complex” (*A. tequilana*, *A. angustifolia*, *A. rhodacantha* and its relative *A. furcroydes*; Piven et al., 2001) and 40 botanical garden plants belonging to 25 different species from the Agave genus (Table 1). In total, we visited eight states from the north-western coast and center of Mexico (Table 2; Figure 9).

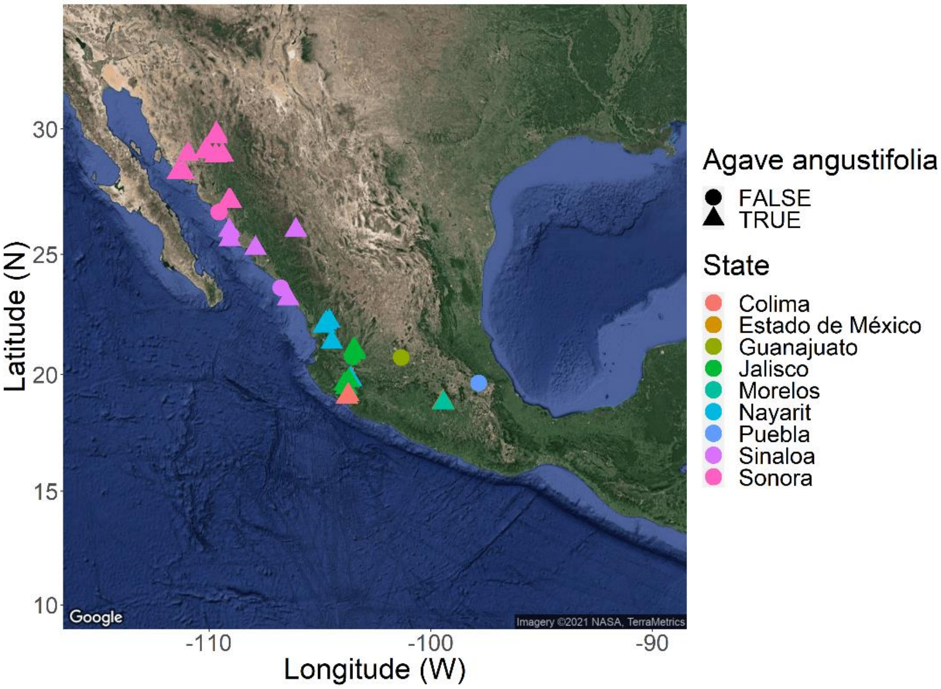


Figure 9. Map of the sampling design.

We implemented a genotyping by sequencing (GBS) reference-based approach to identify high-quality genome-wide polymorphic genomic markers in our sampled plants (see methods). GBS ensures a complexity reduction by randomly sampling non-repetitive genomic regions from giant genomes as those from *Agave* (Elshire et al., 2011; Robert et al., 2008). We obtained a mean per-site depth of 6X (Supplementary Figure 2). The SNP calling pipeline resulted in 72,770 representative SNPs with coverage higher than 80% in all the *Agave* species. The identified SNPs are evenly distributed along the reference genome (Figure 5A).

We built a maximum-likelihood tree with the support of 1000 bootstrap repetitions (Figure 10). Based on the information from this tree, we find evidence to support the existence of the Angustifolia complex (Rivera-Lugo et al., 2018). Although some samples from the same species cluster together, we found no evidence to delimitate those species genetically. In this case, we cannot discuss if they belong to the same species. A taxonomical review would be necessary to solve this incognita.

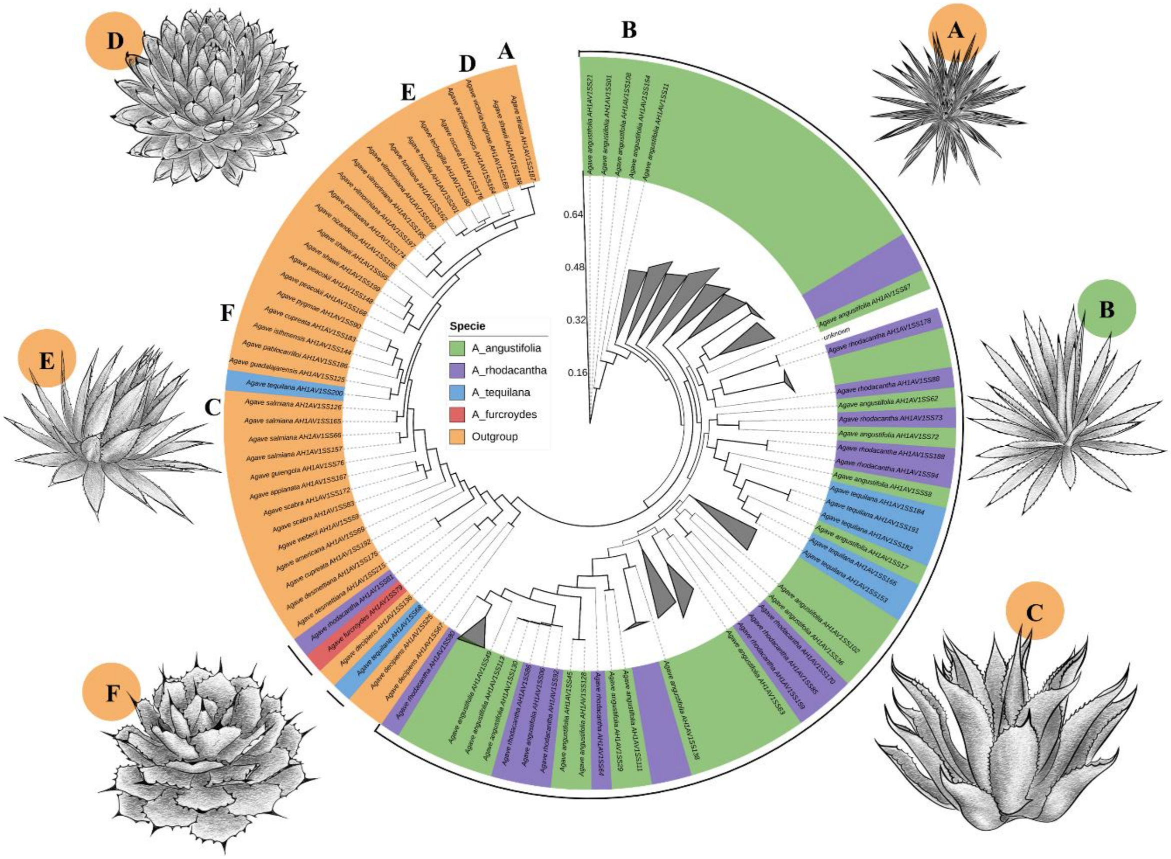

D
A
B
E
C
F
B
D
A
E
F
C
Agave striata AH1AV1SS187
Agave shawii AH1AV1SS198
Agave victoria-reginae AH1AV1SS169
Agave arcedianoensis AH1AV1SS164
Agave oscura AH1AV1SS176
Agave lechuguilla AH1AV1SS180
Agave horrida AH1AV1SS201
Agave funkiana AH1AV1SS162
Agave vilmoriniana AH1AV1SS160
Agave vilmoriniana AH1AV1SS195
Agave parrasana AH1AV1SS197
Agave nizandensis AH1AV1SS174
Agave shawii AH1AV1SS185
Agave shawii AH1AV1SS95
Agave peacokii AH1AV1SS199
Agave peacokii AH1AV1SS148
Agave pygmae AH1AV1SS168
Agave cupreata AH1AV1SS90
Agave isthmensis AH1AV1SS183
Agave pablocarrilloi AH1AV1SS144
Agave guadalajarensis AH1AV1SS186
Agave tequilana AH1AV1SS125
Agave tequilana AH1AV1SS200
Agave salmiana AH1AV1SS126
Agave salmiana AH1AV1SS165
Agave salmiana AH1AV1SS65
Agave salmiana AH1AV1SS157
Agave guiengola AH1AV1SS76
Agave applanata AH1AV1SS167
Agave scabra AH1AV1SS172
Agave scabra AH1AV1SS83
Agave weberii AH1AV1SS59
Agave americana AH1AV1SS69
Agave cupreata AH1AV1SS192
Agave desmettiana AH1AV1SS175
Agave desmettiana AH1AV1SS215
Agave rhodacantha AH1AV1SS81
Agave furcroydes AH1AV1SS79
Agave decipiens AH1AV1SS136
Agave tequilana AH1AV1SS68
Agave decipiens AH1AV1SS25
Agave decipiens AH1AV1SS67
Agave rhodacantha AH1AV1SS80
Agave angustifolia AH1AV1SS49
Agave angustifolia AH1AV1SS112
Agave angustifolia AH1AV1SS130
Agave rhodacantha AH1AV1SS86
Agave angustifolia AH1AV1SS06
Agave rhodacantha AH1AV1SS92
Agave angustifolia AH1AV1SS45
Agave angustifolia AH1AV1SS128
Agave rhodacantha AH1AV1SS64
Agave angustifolia AH1AV1SS29
Agave angustifolia AH1AV1SS111
Agave angustifolia AH1AV1SS138
Agave angustifolia AH1AV1SS63
Agave rhodacantha AH1AV1SS159
Agave rhodacantha AH1AV1SS85
Agave rhodacantha AH1AV1SS170
Agave angustifolia AH1AV1SS36
Agave angustifolia AH1AV1SS102
Agave tequilana AH1AV1SS153
Agave tequilana AH1AV1SS166
Agave angustifolia AH1AV1SS17
Agave tequilana AH1AV1SS182
Agave tequilana AH1AV1SS191
Agave tequilana AH1AV1SS184
Agave angustifolia AH1AV1SS58
Agave rhodacantha AH1AV1SS94
Agave rhodacantha AH1AV1SS188
Agave angustifolia AH1AV1SS72
Agave rhodacantha AH1AV1SS73
Agave angustifolia AH1AV1SS62
Agave rhodacantha AH1AV1SS88
Agave rhodacantha AH1AV1SS178
unknown
Agave angustifolia AH1AV1SS87
Agave angustifolia AH1AV1SS21
Agave angustifolia AH1AV1SS01
Agave angustifolia AH1AV1SS108
Agave angustifolia AH1AV1SS154
Agave angustifolia AH1AV1SS11
0.64
0.48
0.32
0.16
Specie
🟩 A_angustifolia
🟪 A_rhodacantha
🟦 A_tequilana
🟥 A_furcroydes
🟧 Outgroup

Figure 10. The Agave genus phylogeny. a-d are illustrations of some selected species studied here. Monophyletic branches with three or more individuals were compacted. Line thickness represents bootstrap support.

The most ancestral characters in the tree belong to individuals from Jalisco. Most early genotypes from the tree all form part of the *A. angustifolia* species. Then, two branches converge, one comprising a mixed group of samples from the complex and the other covering all the outgroup species. It must be noted that two *A. tequilana* samples are localized within the outgroup. This is also the case for *A. furcroydes*, previously identified as the wild closest species (Piven et al., 2001). The origin of the tequila plant is still uncovered, even though previous ideas suggest its domestication from Jalisco landraces (Trejo et al., 2018). Some tequila plant samples are placed outside the complex, which can mean introgression from other species. This behavior has been previously reported when using different genetic markers (Jiménez-Barron et al., 2020), nevertheless, the presence of introgression in the tequila plant needs to be tested carefully. It may be also possible that the bioinformatic pipeline and sequencing protocol are sources of bias for this to happen, giving that we used a tequila plant as reference during the SNP calling.

Specie	Classification	Number of Individuals
Agave angustifolia	Angustifolia complex	145
Agave furcroydes	Angustifolia complex	1
Agave rhodacantha	Angustifolia complex	21
Agave tequilana	Angustifolia complex	7
Agave americana	Outgroup	1
Agave applanata	Outgroup	1
Agave arcedianoensis	Outgroup	1
Agave cupreata	Outgroup	2
Agave decipiens	Outgroup	5
Agave desmettiana	Outgroup	2
Agave funkiana	Outgroup	1
Agave guadalajarensis	Outgroup	1
Agave guiengola	Outgroup	1
Agave horrida	Outgroup	1
Agave isthmensis	Outgroup	1
Agave lechugilla	Outgroup	1
Agave nizandesis	Outgroup	1
Agave oscura	Outgroup	1
Agave pablocarrilloi	Outgroup	1
Agave parrasana	Outgroup	1
Agave peacokii	Outgroup	2
Agave pygmae	Outgroup	1
Agave salmiana	Outgroup	4
Agave scabra	Outgroup	2
Agave shawii	Outgroup	3
Agave striata	Outgroup	1
Agave victoria-reginae	Outgroup	1
Agave vilmoriniana	Outgroup	3
Agave weberii	Outgroup	1
Unknown	Unknown	1

Table 1. Summary of Agave species and samples sequenced in this study.

Different growth forms and anatomies like those from *A. striata* and *A. victoria-reginae* are placed in the last extreme of the phylogeny (Figure 10A, D), as previously reported with other markers. Adding evidence of those species to be early divergent species in the genus. In the same way, species with similar characteristics and growth forms, such as *A. horrida* and *A. lechugilla*; *A. cupreata* and *A. isthmensis;* within others, are placed together as expected. A careful taxonomic revision of our phylogenetic tree is needed to understand better the Agave genus's possible history and the Angustifolia complex's origin. It should be noted that our approach reduces the outgroup species-specific polymorphism pool. Around 80% of our samples belong to the complex (and is our reference genome). When looking for representative SNPs, we only keep those polymorphic in all our experiments. Despite those observations, we were able to build the first Agave phylogeny using genome-wide data, and it is possible to improve it.

State	Number of Individuals
Estado de México	7
Colima	1
Guanajuato (CINVESTAV Irapuato)	38
Jalisco	82
Nayarit	11
Puebla	1
Sinaloa	22
Sonora	46
Unknown	6

Table 2. Summary of the geographical sampling design used in this study.

THE GENETIC VARIATION AND STRUCTURE OF THE ANGUSTIFOLIA COMPLEX IN THE NORTH-WEST AND CENTER OF MEXICO

As our focus in this thesis is on the Angustifolia complex, we removed samples that did not belong to it. To eliminate genetic redundance, we removed all but one SNP in linkage disequilibrium per genome window (see methods). After filtering, we kept 29,617 SNPs from the 175 samples of the complex.

To explore the genetic structure of the populations, we conducted a PCA. Of the 20 components, most of the variance (around 25%) was explained by the first two components (Figure 11). Based on the eigenvectors from the first two components, we observed a significant clustering of the samples according to their location sampling. Thus, we suggest that the complex’s genetics is structured.

First, we noted all samples from the west-northern states form a compact and well-separated cluster (Sonora, Sinaloa and Nayarit). Samples from the central states (Estado de Mexico and Guanajuato), cluster together with a single sample from Colima and some from Jalisco. The central cluster also comprises plants from the species *A. rhodacantha*, *A. tequilana* and *A. furcroydes*, supporting the complex hypothesis and what we found in the phylogeny. It should be noted that samples from different species are arranged compactedly and near Estado de

Mexico and Colima samples. In contrast, Jalisco samples show an expanded non-compacted distribution. These observations correlate to what was described using classic markers in Jalisco (Trejo et al., 2018), and using GBS in Sonora (Klimova et al., 2022).

Secondly, given that Jalisco samples show a major contribution to the genetic structure of the complex, we wondered if this structure correlates to their landrace assignation. We annotated all Jalisco and *A. tequilana* samples established on what local producers from the sampled locations called them. For tequila plant varieties, names were provided by Katia Gil (CINVESTAV Irapuato, personal communication). We observed a trend of the Jalisco samples to organize according to their landrace assignation. *A. rhodacantha* landraces like "Cenizo", "Chico Aguiar" and "Pencudo" are placed near the central/mixed group, proving that the phenotypical similarities between *A. rhodacantha* and tequila plants relate to their genetic similarities. On the other hand, we observed a directional orientation of "Amarillo" and "Lineño" samples towards one of the three directions of the PCA. In the same way, "Ixtero" and "Barranqueño" samples comprise the remaining path in which Jalisco samples are structured. Based on these observations and what we will discuss further in this text about the genetic variation of Jalisco samples, we conducted a deeper analysis and discussion of these plants. The resulting work was published in a scientific article attached at the end of this thesis.

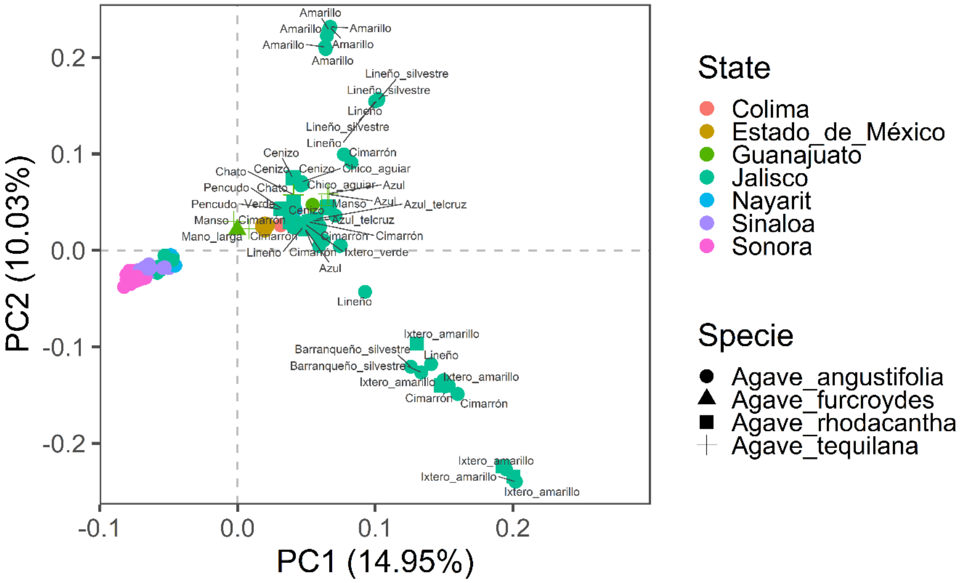


Figure 11. The genetic structure of the Angustifolia complex suggested by a principal component analysis. Landrace names provided by local producers in Jalisco are indicated in some samples.

Admixture analyses are incredibly helpful in understanding the genetic structure suggested by the PCA. This Bayesian approach aims to identify shared genomic proportions across the population, starting with the premise of a *K* number of ancestral populations (Alexander et al., 2009). More than a test of ancestry, this analysis is helpful to find patterns of stratification in the population. Saying this, we performed an explorative ADMIXTURE analysis in the complex. To find the best fit for *K*¸, we choose *K*=5 for its lowest cross-validation error rate (from 2 to 8) and *K*=2 as its reference. Firstly, we did not note any general pattern according to the state pertinence as the PCA suggested. Instead, we observed punctual patterns (Figure 13C). The influence of the green putatively ancestral population explains the same ancestry footprint in *K*=5 and 2. Samples with solid input from the blue population in *K*=2 will have a mixed pattern of ancestry in *K*=5. The blue population influences almost all mixed samples, except for some of them that are fully represented by this population in various states. The orange population is more influential in a small group of individuals from Jalisco. The yellow population strongly influences a few individuals from Jalisco, Sinaloa and Sonora, and it fully represents one individual from Estado de Mexico. Finally, the purple population has more influence in some samples from Sonora, although it is also present in most of the mixed samples. The patterns observed here do not fully capture the expected population structure. It would be necessary to add robustness to this analysis by selecting representative SNPs and enhancing the statistical support (more replicates and randomly selected seeds).

To characterize the complex's genetic variation, we returned to the original SNP dataset. Because we cannot define subpopulations based on our genetic structure analysis, we calculated nucleotide diversity (π) and heterozygosity per individual (He and Ho; Figure 13B). The values for genetic diversity are low compared to other angiosperms (Warschefsky & von Wettberg, 2019), but higher than values reported for the complex using other genetic markers (Eguiarte et al., 2013). In general, individuals with high heterozygosity also have high nucleotide diversity values. Nevertheless, the method used here does not consider the missing data. We wondered if missing data plays a role in the genetic diversity values presented here. For that, we looked for a correlation between these two values. We found a low, non-significant but positive correlation (r^2=0.08, p = 0.47; Figure 12). To discuss such individual values deeper, we plan to implement the algorithm from Pixy (Korunes & Samuk, 2021).

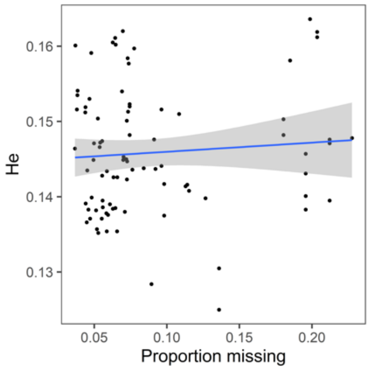


Figure 12. SNP coverage bias.

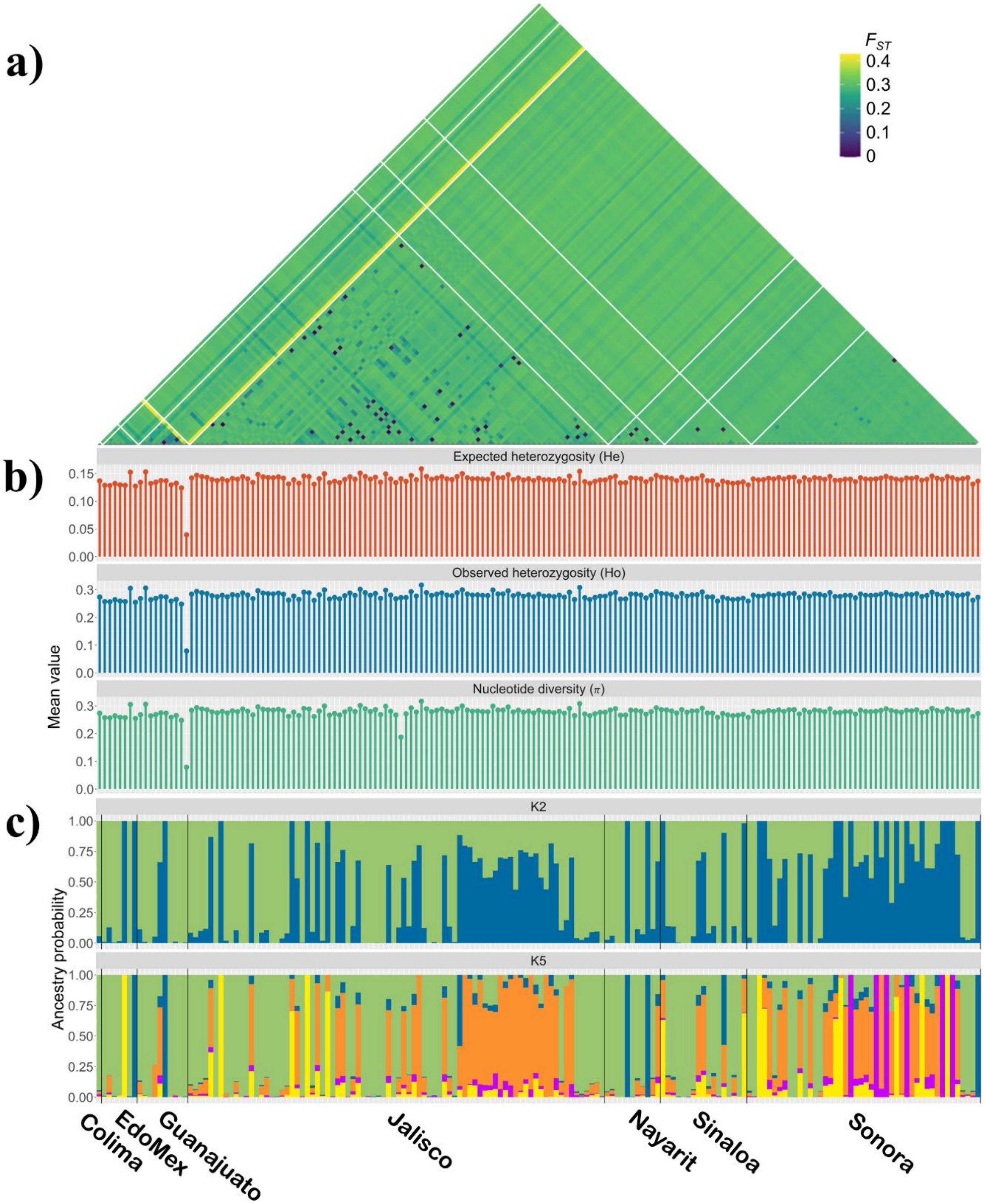


Figure 13. Genetic diversity and differentiation across the Angustifolia complex. a) Pairwise F_{ST} calculation. b) Genetic variation. c)ADMIXTURE analysis. Samples are vertically oriented.

Next, we wondered how divergent the samples from each other within the complex are an if this differentiation correlates to the suggested genetic structure. To do so, we calculated pairwise F_{ST} values (see methods; Figure 13A). Briefly, northern samples, such as those from Sinaloa and Sonora, are well differentiated from the rest of the samples and themselves (values around 0.3). Such values show a moderate but significant differentiation compared to other Agave populations (Eguiarte et al., 2013). In contrast, central samples show low differentiation against the rest of the samples and themselves. It is well appreciated that many pairwise values are almost 0. However, we cannot argue that differentiated samples are clones. We imply central Agaves are subjected to genetic erosion due to their management. Clonal propagation in those states can be the leading cause of this behavior (Arzate-Fernández & Mejía-Franco, 2011).

The most divergent sample was the tequila-producing variety “Manso” (in the phylogeny, its genetic diversity and differentiation). This variety does not have spines. Interestingly, such phenotype is desirable in large-scale Agave plantations. From personal experience, working with Agaves in the field is arduous labor. Its existence suggests that this plant may have been through rigorous human selection. Getting easy-going working characteristics (which is the meaning of being “Manso”) has been the aim of domestication all over human history (Purugganan, 2022). We would need to genotype more Manso samples to suggest its low levels of genetic diversity are caused by the human selection pressure.

Contrary to what we expected, we could not fully relate the genetic structure and the patterns of genetic variation and differentiation in the Angustofolia complex. The origin of the tequila plant should be discussed and revised, giving the implications of the specie definition and its exploitation on the legal rules of spirit production in Mexico (SEGOB, 1999). Recalling the general pattern we observed for the Tajima’s D values, the complex may have been through a strong bottleneck (Figure 5B). More than an economic issue, the implications of expanding our knowledge in Agave genetics would have an essential role in conserving the complex.

III - BIOGEOGRAPHICAL RELATIONSHIPS TO AGAVE GENETICS

THERE IS A POSITIVE CORRELATION BETWEEN GEOGRAPHICAL DISTANCES AND GENETIC DISTANCES

Plants are sessile, and because of that, their genetics are highly influenced by environmental pressure and their dispersal behavior (Bustamante et al., 2016). In the previous chapter, we found interesting genetic patterns that may explain the evolution of the Angustifolia complex. Even though the competition of the sampling of the natural distribution of *A. angustifolia* is still pending to go further in the discussion, we wondered if the environment plays a role in the evolutionary and genetic patterns we have found.

The geographical isolation of subpopulations commonly leads the population to structure (Welsh & Mohamed, 2011). We performed a hierarchal clustering of geographic and genetic distances (the last one was calculated using the eigenvectors from the first two components from the PCA). We observed a similar pattern in both clustering analyses. There are two main clusters: the central and the north-western, with an evident intersection (data not shown). The F_{ST} differences also suggested this pattern. We performed a Mantel Monte-Carlo test with 1000 repetitions to test mathematically if those differences correlate. We found a positive and statistically significant correlation between genetic and geographic distances (r = 0.1917, p-value = 0.0009; Figure 14). The shape we observed is a common characteristic of big populations and correlates to what has been seen in Mezcal-fermenting yeasts (Urbán-Aragón, 2021). In this case, we suggest that the complex may be subjected to isolation by distance, where geographical barriers may play a significant role in stratifying the population structure (Jaynes et al., 2022).

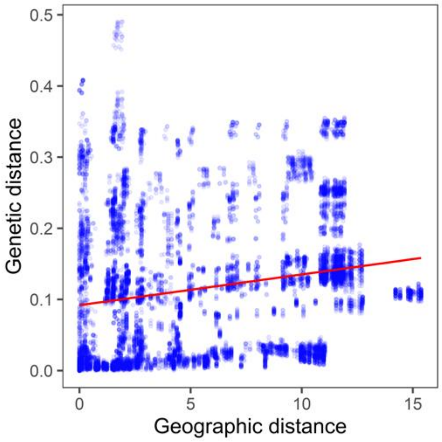


Figure 14. Isolation by distance.

AGAVE GENETICS ARE ASSOCIATED WITH CLIMATE

Knowing that the complex genetics may be exposed to geographical isolation, we wondered if such biogeographical barriers may influence the local adaptation of the population to their environmental conditions. Agave plants are highly resilient to harsh conditions (Garcia-Moya et al., 2011). One of their most striking characteristics, the CAM metabolism, provides Agave plants with anatomic and physiological advantages that make them feasible models to study adaptation (Yin et al., 2019). To elucidate the relationship between Agave genetics and their adaptations to climate, we performed a Genome-Wide Association Analysis (GWAS). It is essential to mention that due to the approach here implemented (ddRAD seq), we should consider a not complete representation of the genome polymorphism in our analysis. We will then name this analysis a Restricted-site GWAS or REGWAS. As input for REGWAS, we obtained historical bioclimate data from the WorldClim server. We calculated per-SNP association for each of the 19 variables and 165 samples with available location data (Figure 15A).

There was a differential association across the 19 bioclimatic variables (Figure 15B). After applying a threshold of p<=0.001, we found 102,930 highly supported associations to bioclimate data (redundancy is highly expected). Of the 19 variables, the temperature-related variables bio_3 (Isothermality), bio_4 (Temperature Seasonality) and bio_7 (Temperature Annual Range) have the majority of SNPs associated with them (more than 7500), followed by the precipitation-related variables bio_12 (Annual Precipitation) and bio_16 (Precipitation of Wettest Quarter). Those highly associated SNPs are located in 820 non-redundant genes. We then annotated those genes functionally by identifying their Asparagus homologs. Once we identified the Gene Ontology (GO) terms linked with those genes, we performed a GO Enrichment analysis to understand the most represented biological processes involved in the putative polygenic association with climate. Within many enriched processes, those related to polysaccharide metabolism and fiber biogenesis were the most representative (Figure 15C).

In essence, our REGWAS results seem promising for studying the genetic adaptation of Agave plants to climate. We performed a preliminary pilot REGWAS analysis using the conditional "tequilero" characterization as phenotype (where 1 was a tequila plant and 0 was not a tequila plant). Compared to our Climate REGWAS, the results from the pilot experiment were redundant and extremely noisy (data not shown). In contrast, the merged association we listed in the Climate REGWAS presents many characteristics that suggest true associations. First, many associated SNPs from the most representative variables (bio_3, bio_4 and bio_7) share the exact genomic location and similar statistic levels of significance. Secondly, all bioclimatic variables that showed low genetic association (bio_1, bio_9, etc.) were constant throughout the genome. Hence, we observed no outliers in such variables. Third, as expected in a standard recombination scenario, we observed association peaks. The last means a highly associated SNP shares the same significance level with their genomic neighborhood. In this way, closely located SNPs share similar p values because they are also associated with themselves. Linkage disequilibrium assures those polymorphic genetic markers are generally inherited together through the population (Xu et al., 2019). This observation reinforces the genetic response to climate and correlates with many studies that show the same pattern (Sasaki et al., 2020).

The association to temperature dynamics is probably related to plant metabolism, cell wall biogenesis, and maintenance. This makes sense when we think Agave plants must survive hot environments where thick and fibrous leaves are essential to preserving water and resources. Transcriptomic analysis shows how CAM metabolism in Agave depends on many other molecular mechanisms (Yin et al., 2019). When mining the time-dependent expression level of the 6G-FFT homolog (methods not shown), a key enzyme in the Fructan metabolism (Gomez-Vargas et al., 2022), we observed it is mainly expressed during the night and thereby may be coupled to CAM metabolism (Supplementary figure 3). It would not be surprising that such important pathways as CAM and Fructan metabolism are connected and linked in the climate adaptation of Agave. They are, in fact, a partial result of natural selection. Additional and integrative analysis would be needed to clarify the relationship between Agave genetics and the climate selection pressure. At this point, much of the associated polymorphism we observed may be due to the suggested population structure. The bias of population structure represented in this analysis can also serve us to understand the role of local adaptation since it is now well known that population genetics may be stratified depending on their level of structure.

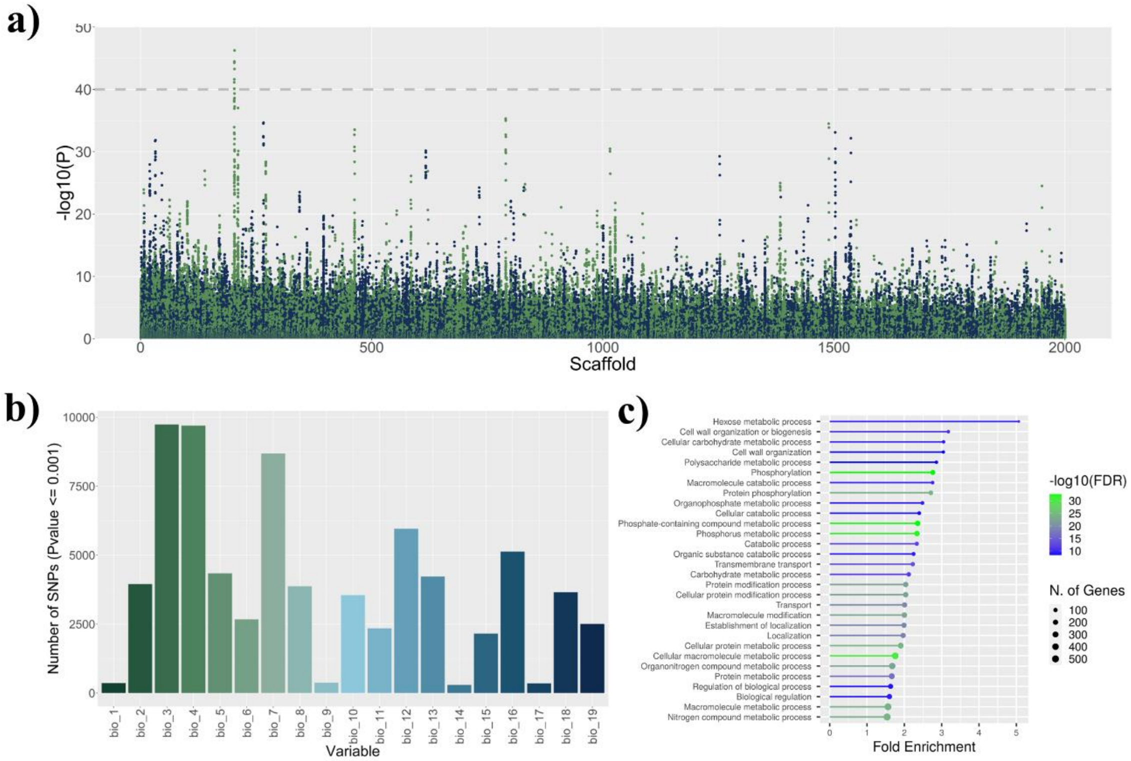


Figure 15. Climate REGWAS. a) Manhattan plot for the first 2000 scaffolds. b) Number of highly associated SNPs to each bioclimate variable analyzed. c) Gene Ontology enrichment analysis from the genes associated with climate.

THE GENOMIC RESPONSE TO CLIMATE CHANGE

So far, we have analyzed the Agave genomics from the present to the past and from a broad to a minuscule perspective. To have an additional perspective of the putative evolutionary path of *A. angustifolia* in the future, we tested the predictive power of the following model:

$$\gamma = a + \beta x_i \ + \epsilon$$

Where:

γ is the predicted frequency of each allele,

a is the known allele frequency,

β is the modeled slope,

x_i are the bioclimatic variable values, and

ϵ is the residual error.

To test the model, we implemented a Machine Learning algorithm commonly used in ecology studies called Gradient Forest (Ellis et al., 2012). Following the model, we used the allele frequencies (per sampling location) as response variables and the values from the 19 bioclimate variables as predictor variables. Knowing that population structure may influence our Climate REGWAS, we performed a new REGWAS using the PCA eigenvectors as a phenotype to identify those SNPs associated with the population structure. To build a refined list of Candidate SNPs, we removed the structure-associated SNPs from our original climate-associate SNPs dataset. As a reference, we selected SNPs not associated with climate (p>0.8) and not associated with the population structure. This procedure resulted in 1,138 SNPs in the Candidate set and 5724 SNPs in the Reference set.

First, we asked the Gradient Forest model to measure the contribution of each variable to the model. Surprisingly, variables related to precipitation (bio_12, bio_16 and bio_13) were now the most important in our model (Figure 16A). Since we have already removed SNPs that may be associated with population structure, we believe the first observation of the temperature-related variables being the most associated with genetics was a result of local adaptation. Thus, some populations from the complex may strongly associate with the temperature dynamics because of the specific characteristics of their habitat. Still, bios 3, 4 and 7 are at the top of the contribution, so we can still suggest Agave plants prefer environments with smooth temperature changes. As a proxy of what we can expect from our model and to compare the Candidate and Reference SNPs datasets, we modeled the allele turnover under the increase of the most important bioclimatic variable (bio_12; Annual precipitation). The Candidate dataset showed high sensitivity to the rise of the annual precipitation (Figure 16B). In contrast, the Reference dataset, considering that is four times bigger, is generally static under the increase in the annual precipitation (Figure 16C). Nevertheless, we can appreciate a similar average behavior of both datasets. A mathematical/statistical analysis would be necessary to calculate the difference in sensibilities. Anyway, we proceeded to use the Candidate SNP dataset based on these observations.

We obtained two forecasted Global Climate Models (GCM) from the CMCC-ESM2 reference (Lovato et al., 2022). First, we built a Specie's Distribution Model (SDM) based on the collecting points and the historical climate data (Figure 17A). We used the resulting SDM to model future species distribution maps under the mild and high-risk climate change scenarios. We then calculated δ, representing the difference between the current and forecasted SDMs.

We predicted allele frequencies in the mild and high-risk climate change scenarios using our Gradient Forest model. Next, we calculated the Euclidean distance between the current and forecasted allele frequencies. For our aim, we consider the dimension of this difference as the Genomic Offset (Aguirre-Liguori et al., 2021), which symbolizes the vulnerability of the populations under climate change. The premise is that the current genetic X environmental (GXE) relationship is the optimal fitness point of the populations simply because plants exist in such conditions. In this way, a big difference in the allele frequencies implies the population should undercome massive changes to reach the same GXE relationship as it is now.

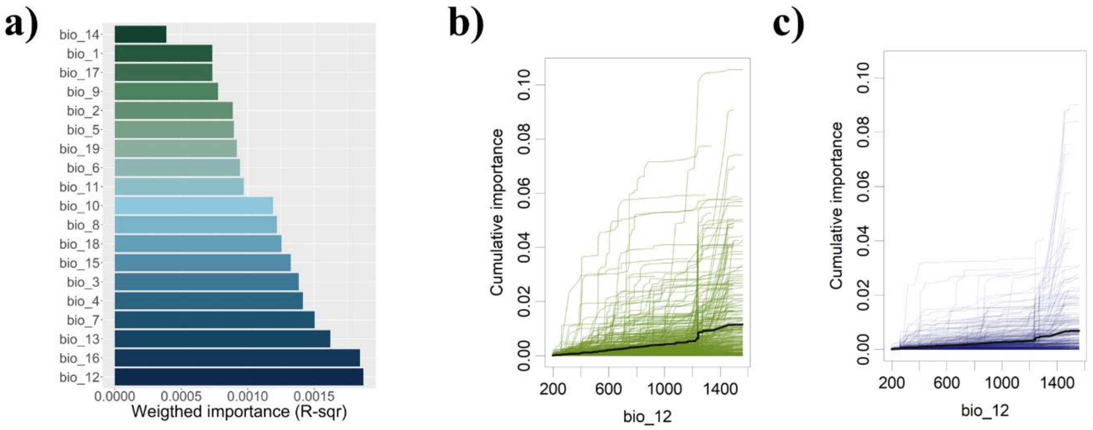


Figure 16. Gradient Forest model performance to predict Agave genetics under climate change. a) Per-variable contribution to the model. b) Allele turnover of the Candidate SNP dataset under bio_12 increasing. c) Allele turnover of the Reference SNP dataset under bio_12 increasing.

The mild scenario (SSP 245, from 2041 to 2060) suggests a big portion of the current SDM will be lost due to climate change in the northern states (Figure 17B). The biggest lost part is between Chihuahua and Durango, where water availability and vegetation are not as good as near the coastal areas. The most negative δ region is in Sinaloa, which implies climate change may affect this region badly even though it is close to the ocean. Positive δ regions are also observed from the north to the center along the mountain range. Such areas may be gained due to the conditioning of high-altitude zones after climate change.

The high-risk scenario (SSP 585, from 2061 to 2080) shows a similar pattern as the mild does (Figure 17C). The most notable differences are more positive δ regions in the northern and central coast regions. Also, a negative δ region becomes prominent and bigger in the north of Jalisco. Positive δ regions are generally punctual and widely distributed, while negative δ regions have more extensive spans and are located in specific regions.

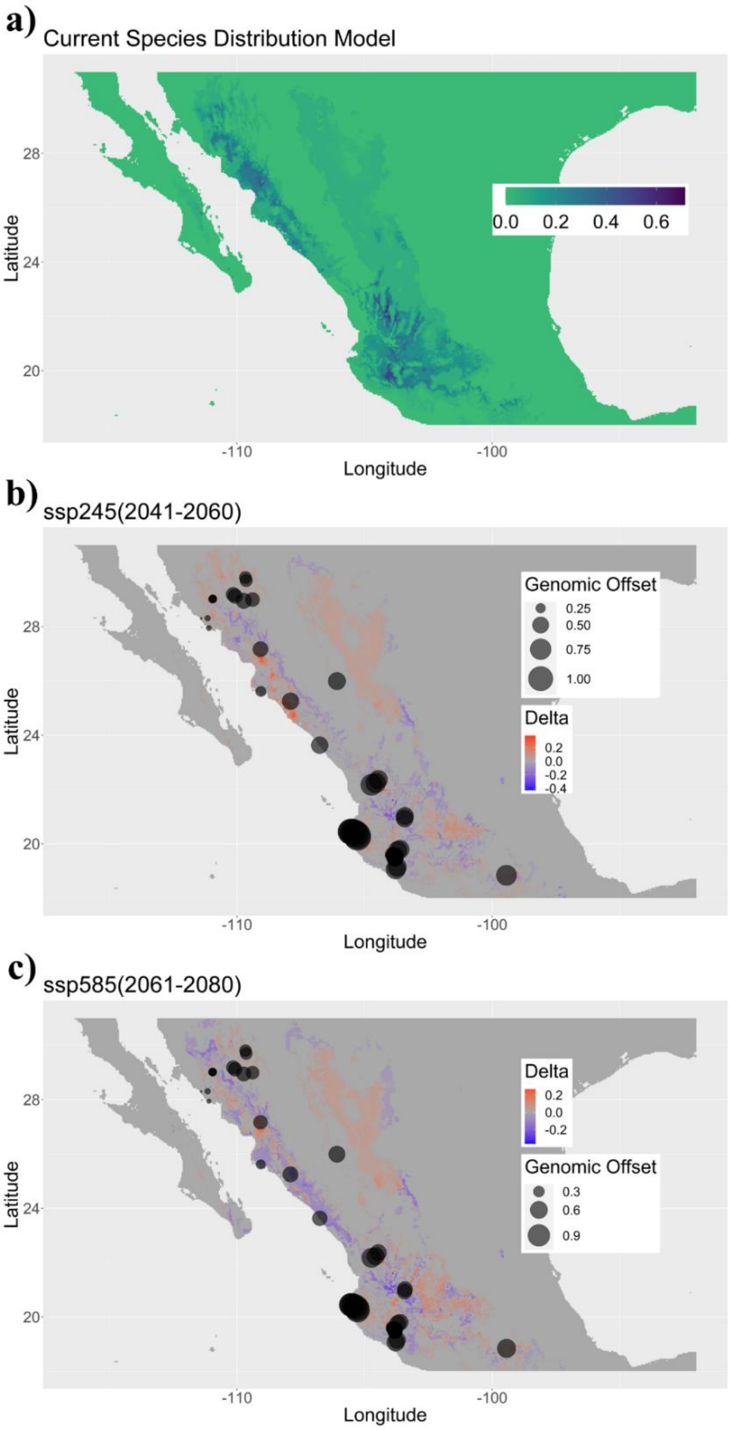


Figure 17. Genomic offset of Agave under climate change. a) Current predicted SDM. b) Genomic Offset and δ under the mild climate change scenario. c) Genomic Offset and δ under the high-risk climate change scenario.

The Genomic Offset values reveal a striking pattern where central populations may be at risk compared to northern populations (inclusively, there are northern populations with Genomic Offset values close to 0). Central populations also have low genetic differentiation and diversity. They show high stratification, admixture and signs of human selection, telling us we should take care of what is happening in that region. Agave plantations in central Mexico are becoming intensive (Mendoza-Galindo Eddy & Mora-Herrera Martha E., 2021). Without a proper management and conservation plan, the impact of genetic erosion can lead to several phytosanitary, economic and exosystemic concerns that need to be issued now.

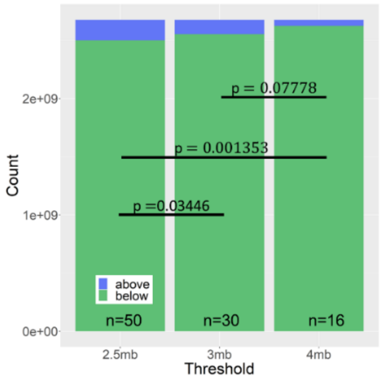
As a final remark, positive δ regions may also influence the Genomic Offset prediction. Regions that will be potentially more sustainable for Agaves after climate change imply a big difference in allele frequencies. Following the aim of understanding which populations may be threatened, it would be necessary to define if they belong to a negative or positive δ region. Our approach also captures allele frequencies without considering mutation, migration, recombination, genetic drift, and other evolutionary forces that strongly impact populations (Hahn, 2018). The results from this analysis should not be discussed individually, as some suggested (Aguirre-Liguori et al., 2021). Nevertheless, we obtained a convincing perspective of what we can expect from the evolution of Agave in the future.

In conclusion, we observed many patterns that suggest *A. angustifolia* evolution is complex and may be subject to genomic and population mechanisms that relate to each other. We found evidence that suggests big-scale rearrangements and TE dynamics may proceed to the origin of the modern genome. From a population-level perspective, we found that *A. angustifolia*, *A. tequilana* and *A. rhodacanatha* form a complex that cannot be genetically limited into different species. Finally, central populations from the complex showed the lowest genetic health and may be at risk. Together these results highlight the importance of introducing genomic studies in non-model and vital plants such as Agave, opening a new landscape for conservation and management strategies of this species.

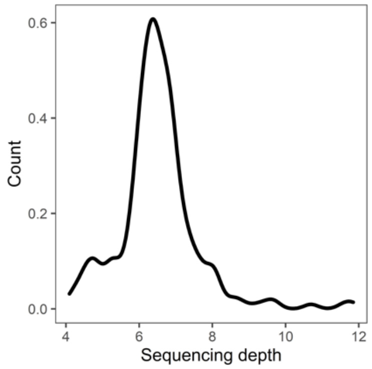
FURTHER DIRECTIONS

As discussed in the first chapter, TEs may contribute substantially to the evolution of Agave. A recently formulated idea proposes using SNP data to infer duplication events of genes, which can also be used for TEs (Jaegle et al., 2022). We could take advantage of this method to 1- Measure the impact of false SNPs coming from TEs and 2- Measure indirect TE content in the population. From this information, we could understand how Agave TEs contribute to the population-level diversity and structure and if those false discoveries may bias our GBS approach.

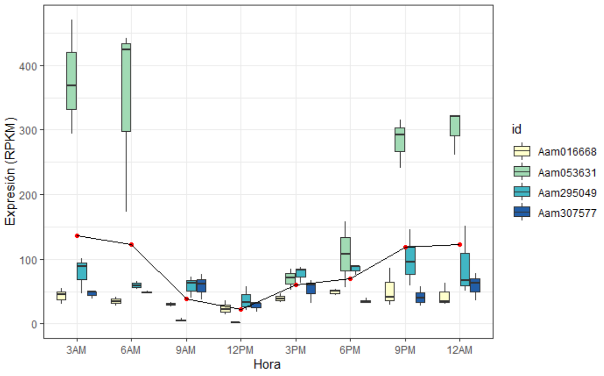
SUPPLEMENTARY FIGURES



Supplementary Figure 1. Genome representativity of the proposed Scaffold thresholds.



Supplementary Figure 2. Sequencing depth overview of our GBS approach.



Supplementary Figure 3. The expression level of *A. americana* 6G-FFT homologs through the day. Data was obtained from Yin et al., (2019).

REFERENCES

Abraham-Juárez, M. J., Martínez-Hernández, A., Leyva-González, M. A., Herrera-Estrella, L., & Simpson, J. (2010). Class i KNOX genes are associated with organogenesis during bulbil formation in Agave tequilana. *Journal of Experimental Botany*, *61*(14), 4055–4067. https://doi.org/10.1093/jxb/erq215

Aguirre-Liguori, J. A., Ramírez-Barahona, S., & Gaut, B. S. (2021). The evolutionary genomics of species' responses to climate change. In *Nature Ecology and Evolution* (Vol. 5, Issue 10, pp. 1350–1360). Nature Research. https://doi.org/10.1038/s41559-021-01526-9

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664.

Ariel, F. D., & Manavella, P. A. (2021). When junk DNA turns functional: transposon-derived non-coding RNAs in plants. *Journal of Experimental Botany*, *72*(11), 4132–4143. https://doi.org/10.1093/jxb/erab073

Arzate-Fernández, A.-M., & Mejía-Franco, R. (2011). CAPACIDAD EMBRIOGÉNICA DE CALLOS INDUCIDOS EN EJES EMBRIONARIOS CIGÓTICOS DE Agave angustifolia Haw. *Revista Fitotecnia Mexicana*, *34*(2), 101–106.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Borbón-Palomares, D. B., Laborin-Sivirian, F., Tinoco-Ojanguren, C., Peñalba, M. C., Reyes-Ortega, I., & Molina-Freaner, F. (2018). Reproductive ecology of Agave colorata: the importance of nectar-feeding bats and the germination consequences of self-pollination. *Plant Ecology*, *219*(8), 927–939. https://doi.org/10.1007/s11258-018-0847-x

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Buendía-Monreal, M., & Gillmor, C. S. (2017). Convergent repression of miR156 by sugar and the CDK8 module of Arabidopsis Mediator. *Developmental Biology*, *423*(1), 19–23. https://doi.org/https://doi.org/10.1016/j.ydbio.2017.01.007

Bustamante, E., Búrquez, A., Scheinvar, E., & Eguiarte, L. E. (2016). Population genetic structure of a widespread bat-pollinated columnar cactus. *PLoS ONE*, *11*(3). https://doi.org/10.1371/journal.pone.0152329

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Cervantes-Pérez, S. A., Espinal-Centeno, A., Oropeza-Aburto, A., Caballero-Pérez, J., Falcon, F., Aragón-Raygoza, A., Sánchez-Segura, L., Herrera-Estrella, L., Cruz-Hernández, A., & Cruz-Ramírez, A. (2018). Transcriptional profiling of the CAM plant Agave salmiana reveals conservation of a genetic program for regeneration. *Developmental Biology*, *442*(1), 28–39. https://doi.org/10.1016/j.ydbio.2018.04.018

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), s13742-015-0047–0048. https://doi.org/10.1186/s13742-015-0047-8

Chen, F., Dong, W., Zhang, J., Guo, X., Chen, J., Wang, Z., Lin, Z., Tang, H., & Zhang, L. (2018). The sequenced angiosperm genomes and genome databases. *Frontiers in Plant Science*, *9*(April), 1–14. https://doi.org/10.3389/fpls.2018.00418

Csorba, T., Questa, J. I., Sun, Q., & Dean, C. (2014). Antisense COOLAIR mediates the coordinated switching of chromatin states at FLC during vernalization. *Proceedings of the National Academy of Sciences*, *111*(45), 16160–16165. https://doi.org/10.1073/pnas.1419030111

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., … Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Research*, *50*(D1), D988–D995. https://doi.org/10.1093/nar/gkab1049

Dalton, R. (2005). Saving the agave. *Nature*, *438*(7071), 1070–1071. https://doi.org/10.1038/4381070a

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & Group, 1000 Genomes Project Analysis. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

de Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, *22*(10), 1269–1271. https://doi.org/10.1093/bioinformatics/btl097

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dray, S., & Dufour, A.-B. (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, *22*(4), 1–20. https://doi.org/10.18637/jss.v022.i04

Edger, P. P., Poorten, T. J., VanBuren, R., Hardigan, M. A., Colle, M., McKain, M. R., Smith, R. D., Teresi, S. J., Nelson, A. D. L., Wai, C. M., Alger, E. I., Bird, K. A., Yocca, A. E., Pumplin, N., Ou, S., Ben-Zvi, G., Brodt, A., Baruch, K., Swale, T., … Knapp, S. J. (2019). Origin and evolution of the octoploid strawberry genome. *Nature Genetics*, *51*(3), 541–547. https://doi.org/10.1038/s41588-019-0356-4

Eguiarte, L. E., Aguirre-Planter, E., Aguirre, X., Colín, R., González, A., Rocha, M., Scheinvar, E., Trejo, L., & Souza, V. (2013). From Isozymes to Genomics: Population Genetics and Conservation of Agave in México. *Botanical Review*, *79*(4), 483–506. https://doi.org/10.1007/s12229-013-9123-x

Eguiarte, L. E., Jiménez Barrón, O. A., Aguirre-Planter, E., Scheinvar, E., Gámez, N., Gasca-Pineda, J., Castellanos-Morales, G., Moreno-Letelier, A., & Souza, V. (2021). Evolutionary ecology of Agave: distribution patterns, phylogeny, and coevolution (an homage to Howard S. Gentry). *American Journal of Botany*, *108*(2), 216–235. https://doi.org/10.1002/ajb2.1609

Ellis, N., Smith, S. J., & Pitcher, C. R. (2012). Gradient forests: calculating importance gradients on physical predictors. *Ecology*, *93*(1), 156–168. https://doi.org/https://doi.org/10.1890/11-0252.1

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE*, *6*(5), e19379-. https://doi.org/10.1371/journal.pone.0019379

Enrique Vela. (2014). El Maguey. *Arqueología Mexicana, Número 57*.

Escobar-Guzmán, R. E., Hernández, F. Z., Vega, K. G., & Simpson, J. (2008). Seed production and gametophyte formation in Agave tequilana and Agave americana. *Botany*, *86*(11), 1343–1353. https://doi.org/10.1139/B08-099

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Fedak, H., Palusinska, M., Krzyczmonik, K., Brzezniak, L., Yatusevich, R., Pietras, Z., Kaczanowski, S., & Swiezewski, S. (2016). Control of seed dormancy in Arabidopsis by a cis-acting noncoding antisense transcript. *Proceedings of the National Academy of Sciences*, *113*(48), E7846–E7855. https://doi.org/10.1073/pnas.1608827113

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

García-Mendoza, A. J., Martínez, I. S. F., & Gutiérrez, D. S. (2019). Four new species of Agave (Asparagaceae, Agavoideae) from southern Mexico. *Acta Botanica Mexicana*, *126*, 1–18. https://doi.org/10.21829/abm126.2019.1461

Garcia-Moya, E., Romero-Manzanares, A., & Nobel, P. S. (2011). Highlights for Agave Productivity. In *GCB Bioenergy* (Vol. 3, Issue 1, pp. 4–14). Blackwell Publishing Ltd. https://doi.org/10.1111/j.1757-1707.2010.01078.x

Ge, S. X., Jung, D., & Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*, *36*(8), 2628–2629. https://doi.org/10.1093/bioinformatics/btz931

Gómez-Rodríguez, V. M., Rodríguez-Garay, B., & Barba-Gonzalez, R. (2012). Meiotic restitution mechanisms involved in the formation of 2n pollen in Agave tequilana Weber and Agave angustifolia Haw. *SpringerPlus*, *1*(1). https://doi.org/10.1186/2193-1801-1-17

Gómez-Ruiz, E. P., & Lacher, T. E. (2019). Climate change, range shifts, and the disruption of a pollinator-plant complex. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-50059-6

Gomez-Vargas, A. D., Hernández-Martínez, K. M., López-Rosas, M. E., Alejo Jacuinde, G., & Simpson, J. (2022). Evidence for Light and Tissue Specific Regulation of Genes Involved in Fructan Metabolism in Agave tequilana. *Plants*, *11*(16), 2153. https://doi.org/10.3390/plants11162153

González-Gutiérrez, A. G., Gutiérrez-Mora, A., & Rodríguez-Garay, B. (2014). Embryo sac formation and early embryo development in Agave tequilana (Asparagaceae). *Journal of the Korean Physical Society*, *3*(1), 1–11. https://doi.org/10.1186/2193-1801-3-575

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(D1), D1178–D1186. https://doi.org/10.1093/nar/gkr944

Gross, S. M., Martin, J. A., Simpson, J., Abraham-Juarez, M. J., Wang, Z., & Visel, A. (2013). De novo transcriptome assembly of drought tolerant CAM plants, Agave deserti and Agave tequilana. *BMC Genomics*, *14*(1), 1–14. https://doi.org/10.1186/1471-2164-14-563

Hahn, M. W. (2018). *Molecular population genetics*. Oxford University Press.

Harkess, A., Zhou, J., Xu, C., Bowers, J. E., van der Hulst, R., Ayyampalayam, S., Mercati, F., Riccardi, P., McKain, M. R., Kakrana, A., Tang, H., Ray, J., Groenendijk, J., Arikit, S., Mathioni, S. M., Nakano, M., Shan, H., Telgmann-Rauber, A., Kanno, A., … Chen, G. (2017). The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nature Communications*, *8*(1), 1279. https://doi.org/10.1038/s41467-017-01064-8

Heyduk, K., McAssey, E. v., Grimwood, J., Shu, S., Schmutz, J., McKain, M. R., & Leebens-Mack, J. (2021). Hybridization History and Repetitive Element Content in the Genome of a Homoploid Hybrid, Yucca gloriosa (Asparagaceae). *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.573767

Heyduk, K., McKain, M. R., Lalani, F., & Leebens-Mack, J. (2016). Evolution of a CAM anatomy predates the origins of Crassulacean acid metabolism in the Agavoideae (Asparagaceae). *Molecular Phylogenetics and Evolution*, *105*, 102–113. https://doi.org/10.1016/j.ympev.2016.08.018

Iwasaki, M., Hyvärinen, L., Piskurewicz, U., & Lopez-Molina, L. (2019). Non-canonical RNA-directed DNA methylation participates in maternal and environmental control of seed dormancy. *ELife*, *8*, e37434. https://doi.org/10.7554/eLife.37434

Jaegle, B., Pisupati, R., Soto-Jiménez, L. M., Burns, R., Rabanal, F. A., & Nordborg, M. (2022). Extensive gene duplication in Arabidopsis revealed by pseudo-heterozygosity. *BioRxiv*, 2021.11.15.468652. https://doi.org/10.1101/2021.11.15.468652

Jaynes, K. E., Myers, E. A., Gvoždík, V., Blackburn, D. C., Portik, D. M., Greenbaum, E., Jongsma, G. F. M., Rödel, M.-O., Badjedjea, G., Bamba-Kaya, A., Baptista, N. L., Akuboy, J. B., Ernst, R., Kouete, M. T., Kusamba, C., Masudi, F. M., McLaughlin, P. J., Nneji, L. M., Onadeko, A. B., … Bell, R. C. (2022). Giant Tree Frog diversification in West and Central Africa: Isolation by physical barriers, climate, and reproductive traits. *Molecular Ecology*, *31*(15), 3979–3998. https://doi.org/https://doi.org/10.1111/mec.16169

Jiménez-Barron, O., García-Sandoval, R., Magallón, S., García-Mendoza, A., Nieto-Sotelo, J., Aguirre-Planter, E., & Eguiarte, L. E. (2020). Phylogeny, Diversification Rate, and Divergence Time of Agave sensu lato (Asparagaceae), a Group of Recent Origin in the Process of Diversification. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.536135

Josué, A., & Mendoza, G. (2007). *Los agaves de México*. https://www.researchgate.net/publication/26549731

Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., & Gao, G. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research*, *45*(W1), W12–W16. https://doi.org/10.1093/nar/gkx428

Katoh, K., & Toh, H. (2010). Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics*, *26*(15), 1899–1900. https://doi.org/10.1093/bioinformatics/btq224

Kidwell, M. G., & Lisch, D. (1997). Transposable elements as sources of variation in animals and plants. *Proceedings of the National Academy of Sciences*, *94*(15), 7704–7711. https://doi.org/10.1073/pnas.94.15.7704

Klimova, A., Mondragón, K. Y. R., Freaner, F. M., Aguirre-Planter, E., & Eguiarte, L. E. (2022). Genomic Analyses of Wild and Cultivated Bacanora Agave (Agave angustifolia var. pacifica) Reveal Inbreeding, Few Signs of Cultivation History and Shallow Population Structure. *Plants*, *11*(11). https://doi.org/10.3390/plants11111426

Korunes, K. L., & Samuk, K. (2021). pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, *21*(4), 1359–1368. https://doi.org/https://doi.org/10.1111/1755-0998.13326

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. https://doi.org/10.1093/nar/gkab301

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, 1000 Genome Project Data Processing. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lovato, T., Peano, D., Butenschön, M., Materia, S., Iovino, D., Scoccimarro, E., Fogli, P. G., Cherchi, A., Bellucci, A., Gualdi, S., Masina, S., & Navarra, A. (2022). CMIP6 Simulations With the CMCC Earth System Model (CMCC-ESM2). *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002814. https://doi.org/https://doi.org/10.1029/2021MS002814

Love, M., Anders, S., & Huber, W. (2014). Differential analysis of count data–the DESeq2 package. *Genome Biol*, *15*(550), 10–1186.

Marí-Ordóñez, A., Marchais, A., Etcheverry, M., Martin, A., Colot, V., & Voinnet, O. (2013). Reconstructing de novo silencing of an active plant retrotransposon. *Nature Genetics*, *45*(9), 1029–1039. https://doi.org/10.1038/ng.2703

Mendoza-Galindo Eddy, & Mora-Herrera Martha E. (2021). Germinación de semillas de Agave angustifolia en diferente madurez fisiológica de la inflorescencia. *PCTI*, *197*.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Morreeuw, Z. P., Escobedo-Fregoso, C., Ríos-González, L. J., Castillo-Quiroz, D., & Reyes, A. G. (2021). Transcriptome-based metabolic profiling of flavonoids in Agave lechuguilla waste biomass. *Plant Science, 305*(January). https://doi.org/10.1016/j.plantsci.2020.110748

Nobel, P. S. (1977). Water relations of flowering of Agave deserti. *Botanical Gazette, 138*(1), 1–6.

Nobel, P. S. (1998). *Los incomparables agaves y cactos.* Trillas.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

Piven, N. M., Barredo-Pool, F. A., Borges-Argáez, I. C., Herrera-Alamillo, M. A., Mayo-Mosqueda, A., Herrera-Herrera, J. L., & Robert, M. L. (2001). Reproductive biology of henequén (Agave fourcroydes) and its wild ancestor Agave Angustifolia (Agavaceae). i. Gametophyte development . *American Journal of Botany*, *88*(11), 1966–1976. https://doi.org/10.2307/3558424

Purugganan, M. D. (2022). What is domestication? *Trends in Ecology & Evolution*, *37*(8), 663–671. https://doi.org/https://doi.org/10.1016/j.tree.2022.04.006

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramawat, K. G. (2009). *Desert plants: biology and biotechnology*. Springer Science & Business Media.

Ramírez Tobías, H. M., Niño Vázquez, R., Aguirre Rivera, J. R., Flores, J., De-Nova Vázquez, J. A., & Jarquin Gálvez, R. (2016). Seed viability and effect of temperature on germination of Agave angustifolia subsp. tequilana and A. mapisaga; two useful Agave species. *Genetic Resources and Crop Evolution, 63*(5), 881–888. https://doi.org/10.1007/s10722-015-0291-x

Reyes-Chin-Wo, S., Wang, Z., Yang, X., Kozik, A., Arikit, S., Song, C., Xia, L., Froenicke, L., Lavelle, D. O., Truco, M.-J., Xia, R., Zhu, S., Xu, C., Xu, H., Xu, X., Cox, K., Korf, I., Meyers, B. C., & Michelmore, R. W. (2017). Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nature Communications, 8*(1), 14953. https://doi.org/10.1038/ncomms14953

Rivera-Lugo, M., García-Mendoza, A., Simpson, J., Solano, E., & Gil-Vega, K. (2018). Taxonomic implications of the morphological and genetic variation of cultivated and domesticated populations of the Agave angustifolia complex (Agavoideae, Asparagaceae) in Oaxaca, Mexico. *Plant Systematics and Evolution*, *304*(8), 969–979. https://doi.org/10.1007/s00606-018-1525-0

Robert, M. L., Yoong Lim, K., Hanson, L., Sanchez-teyer, F., Bennett, M. D., Leitch, A. R., & Leitch, I. J. (2008). *Wild and agronomically important Agave species (Asparagaceae) show proportional increases in chromosome number, genome size, and genetic markers with increasing ploidy*. https://academic.oup.com/botlinnean/article/158/2/215/2418219

Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21), 4737–4754. https://doi.org/https://doi.org/10.1111/mec.15253

Sahebi, M., Hanafi, M. M., van Wijnen, A. J., Rice, D., Rafii, M. Y., Azizi, P., Osman, M., Taheri, S., Bakar, M. F. A., Isa, M. N. M., & Noor, Y. M. (2018). Contribution of transposable elements in the plant's genome. *Gene*, *665*(December 2017), 155–166. https://doi.org/10.1016/j.gene.2018.04.050

Sarwar, M. B., Ahmad, Z., Rashid, B., Hassan, S., Gregersen, P. L., Leyva, M. D. la O., Nagy, I., Asp, T., & Husnain, T. (2019). De novo assembly of Agave sisalana transcriptome in response to drought stress provides insight into the tolerance mechanisms. *Scientific Reports*, *9*(1), 1–14. https://doi.org/10.1038/s41598-018-35891-6

Sasaki, E., Kawakatsu, T., Ecker, J. R., & Nordborg, M. (2020). Common alleles of CMT2 and NRPE1 are major determinants of CHH methylation variation in Arabidopsis thaliana. *PLOS Genetics*, *15*(12), e1008492-. https://doi.org/10.1371/journal.pgen.1008492

SEGOB. (1999). Declaración de Protección a la Denominación de Origen Tequila. In *Diario Oficial de la Federación*.

Stitzer, M. C., Anderson, S. N., Springer, N. M., & Ross-Ibarra, J. (2021). The genomic ecosystem of transposable elements in maize. *PLOS Genetics*, *17*(10), e1009768.

Sun, L., Jing, Y., Liu, X., Li, Q., Xue, Z., Cheng, Z., Wang, D., He, H., & Qian, W. (2020). Heat stress-induced transposon activation correlates with 3D chromatin organization rearrangement in Arabidopsis. *Nature Communications*, *11*(1), 1886. https://doi.org/10.1038/s41467-020-15809-5

Sun, X., Zhu, S., Li, N., Cheng, Y., Zhao, J., Qiao, X., Lu, L., Liu, S., Wang, Y., Liu, C., Li, B., Guo, W., Gao, S., Yang, Z., Li, F., Zeng, Z., Tang, Q., Pan, Y., Guan, M., …

Liu, T. (2020). A Chromosome-Level Genome Assembly of Garlic (Allium sativum) Provides Insights into Genome Evolution and Allicin Biosynthesis. *Molecular Plant*, *13*(9), 1328–1339. https://doi.org/https://doi.org/10.1016/j.molp.2020.07.019

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., & Paterson, A. H. (2008). Synteny and Collinearity in Plant Genomes. *Science*, *320*(5875), 486–488. https://doi.org/10.1126/science.1153917

Torres-Moran, M. I., Velasco-Ramirez, A. P., Hurtado-de la Pena, S. A., Rodriguez-Garcia, A., & Mena-Munguia, S. (2013). Variability and genetic structure in a commercial field of tequila plants, Agave Tequilana Weber (Agavaceae). *American Journal of Agricultural and Biological Science*, *8*(1), 44–53. https://doi.org/10.3844/ajabssp.2013.44.53

Trejo, L., Limones, V., Peña, G., Scheinvar, E., Vargas-Ponce, O., Zizumbo-Villarreal, D., & Colunga-GarcíaMarín, P. (2018). Genetic variation and relationships among agaves related to the production of Tequila and Mezcal in Jalisco. *Industrial Crops and Products*, *125*, 140–149. https://doi.org/10.1016/j.indcrop.2018.08.072

Trejo-Salazar, R. E., Eguiarte, L. E., Suro-Piñera, D., & Medellin, R. A. (2016). Save Our Bats, Save Our Tequila: Industry and Science Join Forces to Help Bats and Agaves. *Natural Areas Journal*, *36*(4), 523–530. https://doi.org/10.3375/043.036.0417

Urbán-Aragón, J. A. (2021). *Estructura poblacional de levaduras Saccharomyces aisladas de fermentación de Agave*. UNAM.

Valenzuela Zapata, A. G., Regalado Pinedo, A., & Mizoguchi, M. (2008). Influencia asiática en la producción de mezcal en la costa de Jalisco. El caso de la raicilla. *México y La Cuenca Del Pacífico*, *11*(33), 91–116. http://www.redalyc.org/articulo.oa?id=433747603006

Villaseñor, J. luis, Ibarra, G., & Ocaña, D. (1998). Strategies for the Conservation of Asteraceae in Mexico. *Conservation Biology*, *12*(5), 1066–1075. https://doi.org/https://doi.org/10.1046/j.1523-1739.1998.97171.x

Vitte, C., Fustier, M. A., Alix, K., & Tenaillon, M. I. (2014). The bright side of transposons in crop evolution. *Briefings in Functional Genomics and Proteomics*, *13*(4), 276–295. https://doi.org/10.1093/bfgp/elu002

Walton, M. K. (1977). The evolution and localization of mezcal and tequila in Mexico. *Revista Geográfica*, 113–132.

Warschefsky, E. J., & von Wettberg, E. J. B. (2019). Population genomic analysis of mango (Mangifera indica) suggests a complex history of domestication. *New Phytologist*, *222*(4), 2023–2037. https://doi.org/https://doi.org/10.1111/nph.15731

Wells, J. N., & Feschotte, C. (2020). A Field Guide to Eukaryotic Transposable Elements. *Annual Review of Genetics*, *54*, 539–561. https://doi.org/10.1146/annurev-genet-040620-022145

Welsh, A. B., & Mohamed, K. I. (2011). Genetic Diversity of Striga hermonthica Populations in Ethiopia: Evaluating the Role of Geography and Host Specificity in Shaping Population Structure. *International Journal of Plant Sciences*, *172*(6), 773–782. https://doi.org/10.1086/660104

Xu, S., Stapley, J., Gablenz, S., Boyer, J., Appenroth, K. J., Sree, K. S., Gershenzon, J., Widmer, A., & Huber, M. (2019). Low genetic variation is associated with low mutation rate in the giant duckweed. *Nature Communications*, *10*(1), 1243. https://doi.org/10.1038/s41467-019-09235-5

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Yin, H., Guo, H. B., Weston, D. J., Borland, A. M., Ranjan, P., Abraham, P. E., Jawdy, S. S., Wachira, J., Tuskan, G. A., Tschaplinski, T. J., Wullschleger, S. D., Guo, H., Hettich, R. L., Gross, S. M., Wang, Z., Visel, A., & Yang, X. (2019). Correction: Diel rewiring and positive selection of ancient plant proteins enabled evolution of CAM photosynthesis in Agave (BMC Genomics (2018) 19 (588) DOI: 10.1186/s12864-018-4964-7). *BMC Genomics*, *20*(1), 1–16. https://doi.org/10.1186/s12864-019-5663-8

Zhang, R.-G., Li, G.-Y., Wang, X.-L., Dainat, J., Wang, Z.-X., Ou, S., & Ma, Y. (2022). TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. *Horticulture Research*, *9*, uhac017. https://doi.org/10.1093/hr/uhac017

Zhang, Z., Xiao, J., Wu, J., Zhang, H., Liu, G., Wang, X., & Dai, L. (2012). ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. *Biochemical and Biophysical Research Communications*, *419*(4), 779–781. https://doi.org/https://doi.org/10.1016/j.bbrc.2012.02.101

Zhao, T., & Schranz, M. E. (2019). Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proceedings of the National Academy of Sciences*, *116*(6), 2165–2174. https://doi.org/10.1073/pnas.1801757116

Zizumbo-Villarreal, D., Vargas-Ponce, O., Rosales-Adame, J. J., & Colunga-GarcíaMarín, P. (2013). Sustainability of the traditional management of Agave genetic resources in the elaboration of mezcal and tequila spirits in western Mexico. *Genetic Resources and Crop Evolution*, *60*(1), 33–47. https://doi.org/10.1007/s10722-012-9812-z

PAPER

plants

MDPI

Article

Genomic and Morphological Differentiation of Spirit Producing *Agave angustifolia* Traditional Landraces Cultivated in Jalisco, Mexico

Dánae Cabrera-Toledo [1,†], Eddy Mendoza-Galindo [2,†], Nerea Larranaga [1], Alfredo Herrera-Estrella [3], Marilyn Vásquez-Cruz [3] and Tania Hernández-Hernández [3,4,*]

[1] Laboratorio Nacional de Identificación y Caracterización Vegetal (LaniVeg), Departamento de Botánica y Zoología, Universidad de Guadalajara, Zapopan 45200, Mexico
[2] Agrogenomic Sciences, ENES Leon UNAM, Guanajuato 37689, Mexico
[3] LANGEBIO-UGA, Guanajuato 36821, Mexico
[4] Research and Collections, Desert Botanical Garden, Phoenix, AZ 85008, USA
* Correspondence: thernandez@dbg.org
† These authors contributed equally to this study.

Abstract: Traditional agave spirits such as mezcal or tequila are produced all over Mexico using different species of *Agave*. Amongst them, *A. angustifolia* is the most popular given its agricultural extension. *A. angustifolia* is a wild species extensively distributed from North to Central America, and previous studies show that it is highly related to the tequila agave *A. tequilana*. In different regions of Mexico, *A. angustifolia* is cultivated under different types and levels of management, and although traditional producers identify several landraces, for the non-trained eye there are no perceivable differences. After interviews with producers from different localities in Jalisco, Mexico, we sampled *A. angustifolia* plants classified as different landraces, measured several morphological traits, and characterized their genetic differentiation and diversity at the genome-wide level. We included additional samples identified as *A. tequilana* and *A. rhodacantha* to evaluate their relationship with *A. angustifolia*. In contrast with previous studies, our pool of ca 20K high quality unlinked SNPs provided more information and helped us to distinguish different genetic groups that are congruent with the ethnobotanical landraces. We found no evidence to genetically delimitate *A. tequilana*, *A. rhodacantha* and *A. angustifolia*. Our large genome level dataset allows a better understanding of the genetic identity of important *A. angustifolia* traditional and autochthonous landraces.

Keywords: *A. angustifolia*; *A. tequilana*; *A. rhodacantha*; tequila; mezcal; mescal; *Agave*

check for updates

Citation: Cabrera-Toledo, D.; Mendoza-Galindo, E.; Larranaga, N.; Herrera-Estrella, A.; Vásquez-Cruz, M.; Hernández-Hernández, T. Genomic and Morphological Differentiation of Spirit Producing *Agave angustifolia* Traditional Landraces Cultivated in Jalisco, Mexico. *Plants* **2022**, *11*, 2274. https://doi.org/10.3390/plants11172274

Academic Editors: Xiangdong Liu and Muhammud Qasim Shahid

Received: 9 July 2022
Accepted: 19 August 2022
Published: 31 August 2022

1. Introduction

The use of *Agave* for spirits production is recent in human history. Although it has been suggested that *Agave* plants have been used as a source of food or fiber production for at least 9000 years [1,2], the distillation of spirits using *Agave* did not begin until the 16th century and the cultivation and production of spirits increased with the popularity of tequila as recently as the early 19th century [3–8]. This indicates the extremely recent initiation of the management and domestication processes of *Agave* with spirit production purposes, especially for some landraces such as the tequila plant. This, together with the long life-cycle particular of agaves, explains the incipient domestication status of cultivars and landraces.

From all spirits produced in Mexico using *Agave*, tequila is the most emblematic. Made from the "blue agave" plant *Agave tequilana* var. "Azul", tequila has been produced traditionally since the 18th century in the state of Jalisco (Figure 1). After the Denomination of Origin of Tequila (DOT) obtained in 1977 [9], the number of plants cultivated for tequila production has increased enormously [10]. The blue agave is now massively cultivated and used in a highly industrialized system owned by transnational corporations [11]. Unfortunately, this industrialization also brought expensive social and environmental consequences, including the erosion of the diversity of traditional landraces [11–15]. Since the DOT established that only blue agave can be used to make tequila, producers shifted their efforts to grow only blue agave landrace, severely mining the diversity of landraces that used to be part of traditional tequila agroecosystems 400 years ago [8,11,16]. In addition, to preserve the integrity of the blue agave, producers have historically recurred almost exclusively to asexual propagation, a strategy that has also been adopted for the cultivation of other landraces, resulting in an extreme loss of genetic diversity [17,18]. These current conditions urge the development and promotion of traditional and new landraces locally adapted to regional climatic regimes and conditions [11,14,15,19,20].

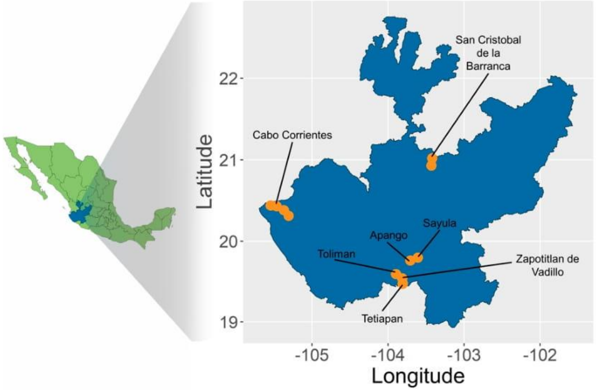


Figure 1. Area of study (Jalisco, Mexico) and localities where samples were collected.

Mexico is the center of the diversity and domestication of *Agave* [21], where 53 species are used for spirit production, most widely known as "mezcal" [22,23]. However, *A. angustifolia* is the most extensively distributed and cultivated for this purpose. *A. tequilana, A. angustifolia* and *A. rhodacantha* belong to a morphological and genetic complex of species known as *A. angustifolia* complex [24,25]. Previous studies using traditional molecular markers have proposed wild populations of *A. angustifolia* in the state of Jalisco as possible wild closest relatives of the blue agave [17,18,26]. Ethnobotanical evidence shows that Southern Jalisco is the nucleus of the greatest diversity of spirit producing landraces within the *A. angustifolia* complex, and it is likely that this area was where traditional farmers initiated the selection of *Agave* germplasm for spirit production [8]. Traditional mezcal-producing farmsteads in this region are known as "mezcaleras" and currently sustain around 20 landraces [18,24]. Another important region in the state of Jalisco where mezcal is traditionally produced is the North Coast, but, until now, ethnobotanical documentation in this region has not been performed and landraces have not been studied (Figure 1). In the North Coast, producers recognize about 10 different landraces from the same *A. angustifolia* complex (Huerta-Galván, unpublished data). As it happened with tequila, mezcal has also gained worldwide recognition and explosive demand. Over the period between 2011 and 2019, mezcal production in Mexico increased over 700% [27], with the US being the highest buyer of this spirit around the world [28]. Under this increasing pressure, it is important to document, understand, and preserve alternative autochthonous landraces and traditional agroecosystems, avoiding the adoption of procedures such as massive, industrialized monoculture by clonal reproduction as is the case with the blue agave.

Although some species and landraces included in the *A. angustifolia* complex have been studied using traditional genetic markers such as AFLPs and microsatellites [18,24–26], none of the Jalisco traditional landraces have been evaluated using a genomic approach with next generation sequencing techniques. The implementation of a low-cost, high-throughput genotyping-by-sequencing (GBS) approach allows the identification of thousands of markers per sample [29], and it is a method particularly useful when working with species with large genomes such as *Agave* [30]. In this study, we implement GBS to evaluate if there is a congruence amongst the ethnobotanical description, morphology, and genetic association of wild, semi-wild and established landraces within the *A. angustifolia* complex in Jalisco. We visited 12 mezcaleras both in the North Coast and in Southern Jalisco and interviewed with local producers. We sampled several individual plants that were referred to by them as different landraces or semi-wild landraces ("Cimarrón", "Sierreño" or "Barranqueño"). We also sampled *A. angustifolia* plants in the wild, as well as additional *A. tequilana* and *A. rhodacantha* accessions, to evaluate their genetic identity and characterize their genetic diversity in relation to geography and ethnobotanical characterization. Finally, we conducted several morphological measurements on a subset of our cultivated sampled plants that were classified by producers as belonging to the widest recognized landraces. This study contributes with the largest available dataset at the genomic level for local Jalisco traditional landraces. Our pool of SNPs allowed us to better understand the genetics underneath the cultivation, propagation and domestication processes of species, varieties, and landraces within the *A. angustifolia* complex in the region.

2. Results

2.1. Sampling and SNP Calling

In total, we collected 87 samples in the state of Jalisco, Mexico between 2020 and 2021 (Figures 1 and 2 and Table 1). With our sequencing approach, we generated about 5 million reads per sample with a mean per-site depth of 6.59X. Our SNP calling procedure using a reference genome of *A. tequilana* var. "Azul" allowed us to generate a matrix of 67,175 high-quality SNPs, with a mean individual SNP coverage of 0.91%. Linked SNPs within a window size of 50, a step of 10 and square-r threshold of 0.1 were pruned, leaving a final matrix of 19,983 SNPs after filtering, suggesting about 70% of the original SNP dataset was linked. According to this dataset, the *A. angustifolia* Jalisco complex studied with our samples showed low genetic variation, with a mean expected heterozygosity value (H_e) of 0.1311 ± 0.0110 and mean nucleotide diversity (π) of 0.2622 ± 0.0220.

2.2. Genetic Differentiation and Population Structure

The pairwise F_{ST} mean value (0.3747 ± 0.0740) showed moderate-to-high genetic divergence across individuals. To further explore the population structure, we carried out an Identity by State (IBS) analysis to compare pairwise genetic distances. We did not find a full congruence of the clusters generated using genetic distances with the landraces. To verify, we conducted an Identity by Descent (IBD) test. Without exception, all pairwise IBD values suggested each sample is genetically related to each other. Therefore, we omitted genetic distances in the following procedures.

We carried out a Principal Component Analysis (PCA) using our filtered SNP dataset. The first three components of the PCA explained 13.9321%, 11.1120% and 10.4036% of the variance, respectively. The Euclidean distances of the eigenvectors from the three components were used to build a UPGMA genetic dendrogram (Figure 3). Four major groups were clearly defined, with observed heterozygosity values, while expected heterozygosity values are lower in the groups including a single landrace (Table 2). The first and most divergent one corresponds to what we call Cluster 1, including all the "Amarillo" landrace samples coming from four different mezcaleras in Cabo Corrientes (North Coast; localities and samples information in Table 1). The next group corresponds to what we designate as Cluster 2, with samples from Southern Jalisco. Cluster 2 includes all the plants sampled in Tolimán (Southern Jalisco) that were classified by local producers as "Lineño" landrace. According to our results, these samples group together in a close relationship with two genetically similar individuals classified one as "Cimarrón" (coming from the same farmstead in Tolimán) and another wild plant coming from a nearby area (AP2 from Apango, Jalisco, 30 km northeast from Tolimán). The following group is what we call Cluster 3 and includes two genetically separated subgroups: 3a and 3b. Subcluster 3a includes several samples that are related by producers to similar types: "Ixtero Amarillo" from Southern Jalisco, with samples from multiple mezcaleras in different locations (IALA1, IALA10 from Zapotitlán de Vadillo, IAMG from Tolimán, TET8 and TET11 from Tetapán,) and "Ixtero Amarillo Barranqueño" landrace (TOL8) from Tolimán. Subcluster 3a is highly related to 3b, which includes a mixture of different landraces, as well as a wild plant from Southern Jalisco: "Ixtero Amarillo" (IAMG1 from Tolimán, TET10 from Tetiapán, IATE2 from Zapotitlán), "Lineño" (TET7 from Tetiapán, TOL9 from Tolimán), "Cimarron" (TET3 and TET1 from Tetiapán), as well as a wild plant from Apango (AP5).

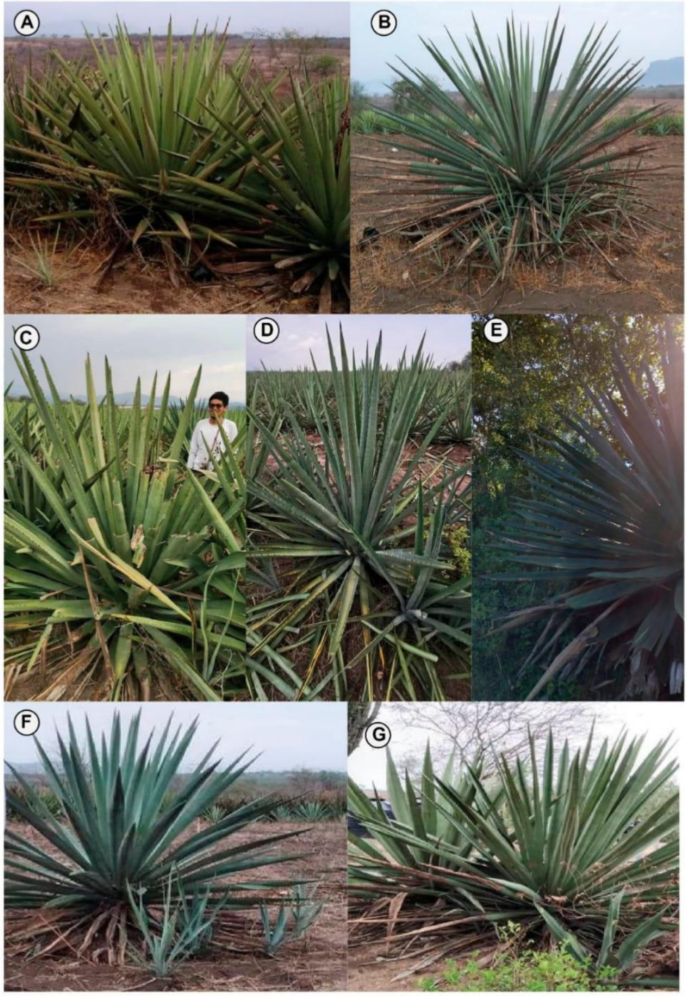


Figure 2. *A. angustifolia* landraces of Jalisco. (**A**) Ixtero Amarillo, (**B**) Ixtero Verde, (**C**) Lineño, (**D**) Barranqueño, (**E**) Cenizo, (**F**) Azul Telcruz, (**G**) Cimarrón.

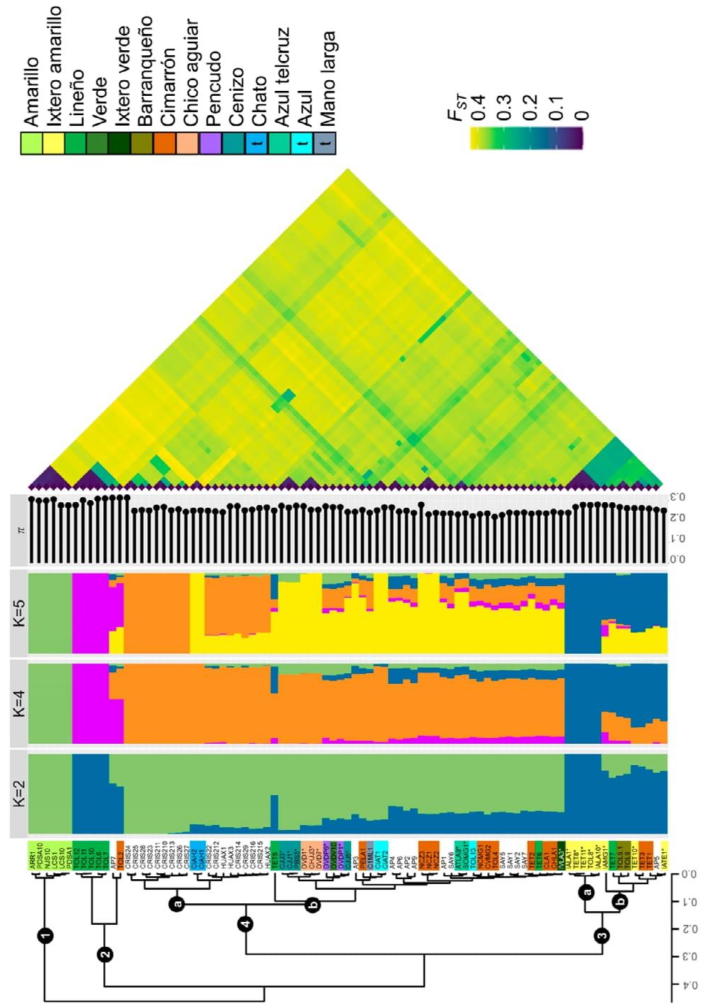


Figure 3. UPGMA, ADMIXTURE, π, and *Fst* analyses results. Landraces are color coded. t = *A. tequilana*, * = *A. rhodacantha* or *A. aff. rhodacantha*.

Table 1. Samples and information in this study. Municipality or Locality name where collected, species name, landrace according to producer, code used in this study, voucher number if available, and number of accessions (*N*).

Municipality or Locality	Species	Landrace	Code	Voucher	*N*
Cabo Corrientes	*A. angustifolia*	Amarillo	LCS1, LCS10	MPDM-218	2
Cabo Corrientes	*A. angustifolia*	Amarillo	PDSA1, PSDA10	MPDM-243	2
Cabo Corrientes	*A. angustifolia*	Amarillo	NJS10	-	1
Cabo Corrientes	*A. angustifolia*	Amarillo	ARR1	MPDM-215	1

Table 1. *Cont.*

Municipality or Locality	Species	Landrace	Code	Voucher	*N*
Tolimán	*A. angustifolia*	Ixtero Amarillo Barranqueño	TOL8	-	1
Tequila	*A. tequilana*	Azul	CIAT1-2	-	2
Tolimán	*A. tequilana*	Azul	TOL13	-	1
Tolimán	*A. rhodacantha*	Azul Telcruz	SOMG1	-	1
Zapotitlán de Vadillo	*A. rhodacantha*	Azul Telcruz	ATLA9	MPDM-273	1
Cabo Corrientes	*A. rhodacantha*	Cenizo	CJJ1, CJJ3, CJJ5	MPDM-219	3
Cabo Corrientes	*A. rhodacantha*	Cenizo	PBS5	MDPM-223	1
Tequila	*A. tequilana*	Chato	CIAH1-2	-	2
Cabo Corrientes	*A. rhodacantha*	Chico Aguiar	CHJJ3	PCR-9688	1
Cabo Corrientes	*A. rhodacantha*	Chico Aguiar	DVD1, DVD7	MPDM-235	2
Zapotitlán de Vadillo (Tetiapán)	*A. angustifolia*	Cimarrón	TET1-3	-	3
Tolimán	*A. angustifolia*	Cimarrón	TOL1-4	-	4
Tolimán	*A. angustifolia*	Cimarrón Verde	CVMG	-	1
Zapotitlán de Vadillo (Tetiapán)	*A. rhotacantha*	Ixtero Amarillo	TET8-11	-	4
Tolimán	*A. rhodacantha*	Ixtero Amarillo	IAMG1	-	1
Zapotitlán de Vadillo	*A. rhodacantha*	Ixtero Amarillo	IALA1, IALA10	MPDM-271	2
Zapotitlán de Vadillo	*A. rhodacantha*	Ixtero Amarillo	IATE1	-	1
Zapotitlán de Vadillo	*A. aff. rhodacantha*	Ixtero Verde	IVLA3	MPDM-269	1
Zapotitlán de Vadillo (Tetiapán)	*A. angustifolia*	Lineño	TET4-7	-	4
Tolimán	*A. angustifolia*	Lineño	TOL5-6	DCT-23, DCT-25, DCT-26	3
Tolimán	*A. angustifolia*	Lineño Ixtero	TOL7	-	1
Tolimán	*A. angustifolia*	Lineño silvestre	TOL9-12	-	4
Tolimán	*A. angustifolia*	Negro Cimarrón	NCMG1	DCT-24	1
Zapotitlán de Vadillo	*A. rhodacantha*	Negro Cimarrón	NCZ1-3	MPDM-276	3
Cabo Corrientes	*A. rhodacantha*	Pencudo	DVDP6, DVDP1	MPDM-238	2
Cabo Corrientes	*A. aff. rhodacantha*	Verde	DVDV10	MPDM-237	1
San Gabriel (Apango)	*A. angustifolia*	wild	AP1-9	DCT-21, DCT-22	9
Milpillas (entronque Huaxtla)	*A. angustifolia*	wild	HUAX1-3	LMCC-150	3
San Cristobal de la Barranca	*A. angustifolia*	wild	CRIS2_1-16	LMCC-151	16
Sayula	*A. angustifolia*	wild	SAY1-7	DCT-20	7

Table 2. Genetic diversity indexes for the Clusters obtained with the PCA based dendrogram. Expected heterocigosity (He), observed Heterocigosity (Ho) and, inbreeding coefficient (Fis) and nucleotide diversity (pi).

Cluster	n	Ho	He	Fis	pi
1	6	0.2884	0.1546	−0.8011	102.7727
2	7	0.3017	0.2155	−0.2893	125.0220
3	14	0.3144	0.2667	−0.1198	151.5476
4	60	0.2850	0.3337	0.1543	182.6654
3a	5	0.3204	0.1708	−0.8420	113.8000
3b	9	0.3109	0.2946	−0.0520	165.0458
4a	20	0.2711	0.3157	0.1326	174.0885
4b	40	0.2925	0.3300	0.1222	180.4139

The fourth cluster is the largest, comprising a cohesive group with highly related samples, mostly coming from plants growing in the wild. This group comprises a more genetically diverse set of samples from different regions. It is split in two major subgroups: 4a and 4b with excess of homozygotes and the highest values of Pi (Table 2). Cluster 4a includes all the wild plants sampled in San Cristobal de la Barranca (CRIS and HUAX coded samples), a region in the Northeast of Jalisco, geographically close to Tequila, the locality of origin of the blue agave. This cluster of wild plants also includes samples CIAH1 and CIAH2, that come from two plants classified as *A. tequilana* var. "Chato". Cluster 4b is the most complex of all clusters. It includes several different landraces as well as plants from the wild. The most divergent sample within 4b is TET5, classified as "Lineño" by producers in Tetiapán (Zapotitlán, Southern Jalisco). Two subgroups in addition to TET5 can be found within 4b. The first includes all samples from different landraces in the

North Coast, all collected in Cabo Corrientes: "Cenizo" (CJJ3, CJJ1), "Amarillo" (PDSA), and "Chico Aguiar" (samples coded as DVD, CHJJ, CJJ). The other subgroup includes all samples of plants from Southern Jalisco: samples collected in the wild from Apango (coded as AP) and Sayula (coded as SAY), several plants classified as "Cimarron" collected in Zapotitlán de Vadillo (PELA1, CLA1, CHLA1, TET2), Tolimán (TOL4), Tetiapán (TET2); "Negro Cimarrón" from Zapotitlán de Vadillo (NCZ) and Tolimán (NCMG1), "Cimarrón Verde" from Tolimán (CVMG2), "Azul Telcruz" from Zapotitlán de Vadillo (ATLA9) and Tolimán (SOMG1), "Lineño" from Tetiapán (TET6), "Ixtero verde" from Zapotitlán de Vadillo (IVLA3); as well as *A. tequilana* var. "Mano Larga" (C1ML1) and *A. tequilana* var. "Azul" (CIAT1-2).

The arrangement of our samples in the genetic clusters described above was also congruent with the results of the analyses of the genetic structure with ADMIXTURE. For these analyses, we selected *K* values from two to eight and ran ten repetitions with a random seed each. We observed that *K* = 5 had the lowest cross-validation error rate from the analysis. Then, we looked for the most representative repetitions with CLUMPAK. We selected the fifth repetition from *K* = 2 (7/10), the first repetition from *K* = 4 (8/10) and the eighth repetition from *K* = 5 (10/10) (see Figure 3). When *K* = 2, we observed that Clusters 1–4 also conform to different genetic entities, with the largest Cluster 4 as well as Cluster 3b, which include wild samples, having the highest diversity. Cluster 1, 2 and 3a show the most cohesiveness and less degrees of diversity; "Lineño" and "Ixtero Amarillo" landraces are mainly formed by one population, while "Amarillo" landraces form the counterpart. The supergroup shows a mixed ancestry. *K* = 4 was useful to separate the "Ixtero Amarillo", "Amarillo" and "Lineño" landraces into three different ancestry groups. The fourth group was mainly present in the mixed supergroup and in the sister group of "Ixtero Amarillo". When including a fifth ancestral group (*K* = 5), we noted *A. tequilana* and *A. rhodacantha* landraces such as "Chato", "Azul" and "Chico Aguiar" are fully represented.

2.3. Morphological Differentiation

We gathered morphological data for 20 genotypes in our dataset (i.e., 23% of the total number of samples), focusing on several variables that have been useful to study traditional *Agave* landraces [24]. In total, we obtained measurements of 17 traits, which include Munsell leaf colors that were converted to xyY coordinates. After landrace identification by farmstead owners, all the 20 individuals were measured, most of them from Cabo Corrientes (North Coast) and four individuals from Zapotitlán de Vadillo (Southern Jalisco).

The measurements were used to build a heatmap (Figure 4a), which shows all retained morphological variables (only maximum leaf width was removed) in columns and samples in rows, ordered by its Euclidean distance and the UPGMA hierarchical clustering method. Although, some patters within variables related to different landraces could be seen, individuals included in each of them present morphological variation. As an example, the highest values of the three-color coordinates as well as the simple variables related to the terminal thorn are shown by the two individuals of "Ixtero Amarillo", but the variables "terminal thorn length/terminal thorn base width" as well as "teeth length" present a variation among them. For this last variable, "Amarillo" individuals also show high variability.

Using our morphological measurements, we built a UPGMA dendrogram based on Euclidean distances. This dendrogram showed groups that are highly congruent with landrace determination as well as geography (Figure 4a). Three major groups with high bootstrap support were distinguished according to morphology. The first one and clearly separated from the pool of other individuals is the two "Ixtero Amarillo" plants from Zapotitlán de Vadillo. The second group is large and includes several landraces from Cabo Corrientes ("Amarillo", "Verde", and "Pencudo"), as well as two individuals measured in Zapotitlán de Vadillo identified as "Azul Telcruz" and "Ixtero Verde", respectively. The last group is also large and includes plants from Cabo Corrientes identified as "Chico Aguiar", "Pencudo", and "Cenizo".

The dendrogram built with morphological data was compared both visually and numerically (cophenetic and Baker correlation matrix indexes) with a genetic tree built with a reduced SNPs dataset that only includes the samples for which we had morphological measurements. This last dendrogram of genetic distances shows highly cohesive and supported grouping patterns that follow landrace classification. However, there is no full congruence amongst morphological and genetic dendrograms, as the congruence is more evident only in some groups (Figure 4b). Genetic and morphological distance matrices showed a Mantel correlation of 0.54 (*p*-value = 0.001) and the derived dendrograms a cophenetic and Baker matrix correlation of 0.54 and 0.7, respectively (with near 0 values meaning that the two trees are not statistically similar), showing high equivalence among them.

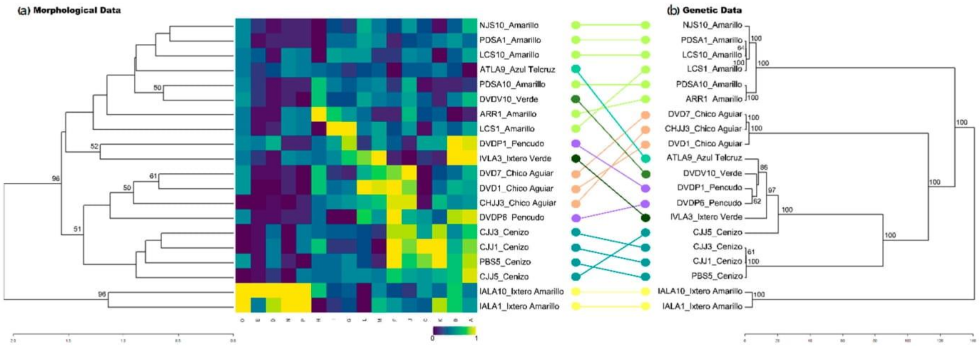


Figure 4. Dendrograms obtained with the morphological (**a**) and genetic (**b**) 23% subset data. (**a**) Morphological dendrogram and heatmap with normalized values to a range 0–1. Variables are: (A) plant length, (B) leaf length, (C) leaf width at middle, (D) terminal thorn length, (E) terminal thorn base width, (F) number of lateral teeth, (G) distance between teeth, (H) teeth length, (I) Distance between teeth/leaf length, (J) number of lateral teeth/leaf length, (K) terminal thorn length/terminal thorn base width, (L) leaf length/terminal thorn length, (M) leaf length/leaf width at middle and Munsell coordinates (N) xyY.x, (O) xyY.y and (P) xyY.Y.

3. Discussion

As it happened with the history of tequila production in the region of west-central Mexico, the specific mezcaleras visited for this study began to be managed at least during the second half of the 18th century (Joya Hildegardo, *pers. com*). Even though producers identify several landraces, only few are commonly known and less ambiguously identified by most of them in each region, which might indicate the generational time elapsed since the start of management and the strength of their genetic identity. That is the case of landraces “Amarillo”, “Verde”, “Cenizo” and “Chico Aguiar” in the North Coast, as well as “Ixtero Amarillo”, “Ixtero Verde” and “Lineño” in the South. In contrast, other plants grown in these agroforest systems are classified using terms such as "Cimarrón”, “Sierreño” or “Barranqueño”. These terms refer to plants that farmers recently found in the wild and identified as good candidates to bring to their cultivars. This is the mechanism by which landraces probably start and then producers keep reproducing them asexually for generations. For example, the terms “Negro Cimarrón” or “Cimarrón Verde” indicate plants that have recently been brought from the wild or are under incipient management. Plants under these different stages of management suggested by their ethnobotanical denomination might show different levels of genetic diversity and differentiation [31]. Our results show more differentiation at the genomic level for the oldest and more extensively used landraces, and they also show the mixed ancestry for the most recent ones. In addition, we found that there is a morphological cohesiveness amongst the widely recognized landraces in comparison with the “Cimarrón”, “Sierreño” or “Barranqueño” plants, which might also be related to a particularly genetic blueprint. These last plants might show a relationship with individuals in the same geographic region and a mixed genetic component with wild plants.

3.1. Genetic Differentiation within the A. angustifolia Complex in Jalisco

This is the first study to report the genome wide characterization of traditional spirit-producing landraces included in the *A. angustifolia* complex in Jalisco, Mexico, together with morphological data for several samples, and the integration of these data with ethnobotanical information related to the management and domestication processes for spirit producing landraces suggested by their common names. After deep ethnobotanical explorations, Colunga-García Marín and Zizumbo-Villareal showed that southern Jalisco might be the nucleus of the greatest number and diversity of landraces used for the production of *Agave* spirits, particularly within the *A. angustifolia* complex [8]. However, *A. angustifolia* is used in other areas of the state to produce distillates, such as in the North Coast, where we are reporting new molecular and morphological data for local varieties.

The *A. angustifolia* complex comprises four species and five varieties, including three species used for mezcal production [25]. In this study we included, in addition to *A. angustifolia* landraces, two other species growing in Jalisco: *A. tequilana* and *A. rhodacantha*, although due to taxonomic difficulties in distinguishing the latter, we cautiously consider some of our specimens as *A. aff. rhodacantha*. Our approach, using GBS sequencing and SNPs calling, gave us access to a significantly larger number of loci to analyze (19,983 unlinked loci) in comparison with previous studies using traditional markers such as AFLPs or microsatellites [25,32]. Our results confirm the close relationship of *A tequilana*, *A. angustifolia* and *A. rhodacantha*, but do not show a clear differentiation of these three species in separate supported clusters. In contrast, our samples of *A. tequilana* and *A. rhodacantha* or *A. aff. rhodacantha* are mixed and intercalated amongst a diverse arrangement of *A. angustifolia* wild and landraces samples, suggesting the extensive gene flow and poor genetic limits existing amongst entities within the complex. However, a more conclusive assessment of the identity of these taxonomic units and landraces requires the inclusion of more specimens from different localities in the study, as well as morphological data for both vegetative as well as reproductive structures.

3.2. A. tequilana Genetic Background and Domestication

Previous information on the genetics within the *A. angustifolia* complex coming from AFLP studies showed the genetic differentiation but very close relationship of *A. angustifolia* and *A. tequilana*, even suggesting that *A. tequilana* is a type of *A. angustifolia* [33]. Additionally, using AFLPs, studies of cultivars in Oaxaca, Mexico, Rivera-Lugo and collaborators showed *A. tequilana* as a cohesive group closely related to *A. angustifolia* var. "Espadin" [25]. These results are congruent with the taxonomical and ethnobotanical literature, which suggest the origin of the cultivar from wild populations of *A. angustifolia* growing on the semi-arid slopes in Cocula and Tecolotlán in Central Jalisco [2,8].

Given its importance in tequila production, *A. tequilana* and the suggested landraces have been studied in more depth than *A. angustifolia*. The study of the genetics of *A. tequilana* with traditional genome-wide genetic techniques such as AFLPs and RAPDs were inconclusive, showing either low or significant diversity depending on the technique used [17,32]. These molecular markers have a high mutation rate and can change from one generation to the next [33], or even in between bulbils produced asexually from the same mother plant [34]. By using microsatellites, Trejo et al. [26] studied several cultivars and populations in the *A. angustifolia* complex in Jalisco, finding that *A. tequilana* var "Azul" samples were more closely related to the *A. angustifolia* populations from southern Jalisco. In addition, they found that the *A. tequilana* var "Sigüin" and *A. tequilana* var "Chato" as well as the landrace "Ixtero Amarillo" are closely related.

For our study, we included in our analyses two accessions corresponding to the *A. tequilana* var. "Chato" (CIAH-1, CIAH-2), and two accessions of *A. tequilana* var. "Azul"

(CIAT-1, CIAT-2, TOL13), as well as one accession of *A. tequilana* var. "Mano Larga" (C1ML-1). Interestingly, and similarly to Trejo et al. [26], our results suggest an independent domestication of these three varieties from different wild *A. angustifolia* backgrounds. We also found more genetic variation in these *A. tequilana* landraces than previously thought, which deserves more in-depth studies. Our results indicate that the *A. tequilana* var. "Chato" in the cluster 4a (Figure 3) emerged from populations of *A. angustifolia* in San Cristobal de la Barranca and Milpillas (CRIS and HUAX coded samples), a region located to the north of Guadalajara city, geographically close to Tequila, the locality of origin of the blue agave. On the other hand, our samples of *A. tequilana* var. "Azul" and *A. tequilana* var. "Mano Larga" both group in a totally different cluster, 4b, which is the largest in our study and includes several landraces and wild plants collected in the south of Jalisco, suggesting that these two other *A. tequilana* landraces might have their origins in that region.

3.3. Traditional Landraces and Their Morphological Identity

Agave spirits are produced all over Mexico, applying both industrialized or semi-industrialized as well as traditional and artisanal methods using different species of *Agave*. In most localities, particularly where spirits are distilled traditionally, producers still recur to wild plants and follow locally developed non-industrialized methods. In other regions where the production of spirits has deeper heritage and history, families of farm holders and producers inherit a tradition of cultivating and propagating plants that were initially brought from the wild generations ago. It has been suggested that landraces such as "Ixtero Verde", "Ixtero Amarillo", "Cenizo" or "Lineño" have been cultivated in the region from at least 100–150 years [24].

Some inherent biological characteristics of *Agave* spirit producing crops provide particularities to their domestication processes and establishment of landraces. The long life- cycles of *Agave* plants (around 6 to 10 years or more), and the fact that reproductive structures that allow sexual crossing are cut before maturity during their cultivation, make breeding practices practically inexistent. Plants recognized as landraces might come from one or a few genetic lines of plants found originally in the wild, and then propagated over and over asexually through clonal rhizome plantlets or bulbils. Thus, we expect the genetic variation within landraces to be significantly reduced. We could also expect that morphological characteristics are highly consistent; however, the recognition and naming of landraces by producers involves a social process of acceptance of the group of morphological traits, particularly if it occurs amongst different farm holders over larger geographic areas. For example, although most landraces are cultivated by single producers, others such as "Lineño" and "Ixtero Amarillo" are more widely recognized and distributed, suggesting a longer history of propagation and cultivation [27].

In the case of the *A. angustifolia* complex, the morphological characteristics of several landraces have been studied extensively in Jalisco by Vargas-Ponce and collaborators. These authors report the presence of at least 24 traditional landraces cultivated in the state, most of which are grown in Southern Jalisco. In contrast, Central Jalisco is currently dominated by monovarietal plantations of *A. tequilana* var "Azul" for tequila production, and traditional landraces have become very scarce [4,8]. Our study extends to the analyses of landraces in another region in Jalisco, the North Coast, where the spirit produced is often named "Raicilla". To our knowledge, this is the first time that these landraces are studied. For our morphological analyses, we used the same variables as Vargas-Ponce et al. [24]. With this set of morphological variables, Vargas-Ponce et al. [24] found that the landraces in southern Jalisco are already morphologically differentiated. Barrientos-Rivera et al. [35] study *A. angustifolia* landraces in Guerrero, Mexico using a similar set of variables, and they were also able to separate the individuals belonging to populations of the "Sacatoro" landrace from the population of the "Espadín" landrace included in their study. With this set of morphological variables, we were also able to confirm that the studied landraces already show morphological differentiation and cohesiveness at this organismic level, showing the adequacy and sufficiency of the morphological variables selected to study the differentiation of landraces in the *A. angustifolia* complex in Jalisco, as well as in other regions of Mexico.

Although with a limited sampling, our analyses of morphological data clearly differentiated landraces that form strongly supported clusters: "Cenizo", and "Chico Aguiar" from the North Coast and "Ixtero Amarillo" from the South (Figure 4b). The plants classified as "Amarillo" landraces show the highest diversity, and they form a large cluster such as "Pencudo" from the North Coast, but also "Azul Telcruz" and "Ixtero Verde" from Southern Jalisco. Unfortunately, for these last landraces we have very limited sampling, and the inclusion of more plants coming from different farms and regions is necessary to better assess both their morphological and genetic establishment.

3.4. Genetic Differentiation of Landraces

Our sampling design allowed for the collection and comparison of wild plants and plants growing under different levels of domestication and management (tolerated, encouraged and cultivated) at the genomic level. We generated a set of approximately 20K high-quality SNPs by RAD sequencing. The restriction site associated (RAD) DNA sequencing is one of the most common methods to genotype by sequence representative loci from all the genomes in plant populations [36]. With our dataset, we were able to characterize the genetic differentiation and variation in samples of plants within the *A. angustifolia* complex assigned to several autochthonous landraces under different levels of acceptance and width of use in Jalisco. We also included a significant number of wild plants from the same geographical region, which allowed us to have a better picture of the domestication process in the wild genetic context. With the genomic information obtained in this study, we were able to fine discriminate among producer's varieties, showing that several landraces have a genomic basis. The use, preservation and promotion of these traditional *Agave* landraces has been proposed as an alternative to reduce the impact of the extensive monovarietal cultivation of *A. tequilana* var. "Azul" [24].

Archaeological evidence shows that *Agave* species and landraces have been used in the Jalisco and surrounding regions for at least 2500 years, and that they have been very important in both daily and ceremonial life [15,37]. For example, the word "Ixtero" gives the name to one of the most representative and ancient landraces in southern Jalisco: "Ixtero Amarillo" [37]. The word "Ixtero" is derived from the Náhuatl language (Mexican original native prehispanic language) "ichtli" or "ixtl", name of the fiber produced with *Agave* spp., and refers to the first use of this specific landrace [37]. According to Colunga-GarcíaMarín and Zizumbo-Villarreal [8], with the introduction of the distillation process, the agaves, first selected for food or their fibers, were then submitted to a second selection pressure to provide the best germplasm for the production of distilled spirits [8]. In this context, "Ixtero Amarillo" has been cultivated mostly by shoots for probably more than 200 generations of agave plants (considering 12 years to mature on average, Miguel Partida, *pers. com*).

Based on genomic data, our results show a clear differentiation in the landraces that are best recognized by farmers: two in Southern Jalisco ("Ixtero Amarillo" and "Lineño") and one in the North Coast ("Amarillo"), that conform separate supported clusters (Cluster 1, 2 and 3), even when the plants were sampled from different parcels or farms. These groups, particularly "Amarillo", show small Ho values (Table 2), indicating the low genetic diversity expected particularly for a group that has been reproduced vegetatively for generations. However, the "Ixtero Amarillo" and "Lineño" groups (Clusters 2 and 3) show higher genetic diversity, and in our dendrogram, they are genetically close to plants recognized under landrace names such as "Barranqueño" or "Cimarrón". These terms are relatively generic and used by farmers to refer to plants recently brought from the wild. Thus, both the ethnobotanical information contained on the designation of landraces and the genetics are congruent with a still incipient process of domestication and homogenization of these landraces. Unfortunately, we only have morphological information for two "Ixtero Amarillo" plants and none for the "Lineño" to allow for the confirmation of this intermediate level of domestication at the morphological level.

The "Amarillo" landrace, the most clearly differentiated landrace in our study both genetically and morphologically, sustains 95% of the Raicilla spirit production on the North Coast (Pedro Jiménez, *pers. com*). In contrast with the scenario found in Southern Jalisco, there are no historical or anthropological studies on the management of *Agave* in the Cabo Corrientes municipality. However, producers share similar traditions with the ones in Southern Jalisco, such as the Filipino distillation method, which incorporates stills made with hollow trunks. In both regions, producers refer to their spirit factories as "taverns" and the fermented beverage that they distillate as "tuba". Thus, the clear genomic differentiation of "Amarillo" and common cultural traits related to the elaboration of spirits between the North Coast and Southern Jalisco suggest a similar human selection process in both regions and are congruent with the hypothesis of Bruman [6,7], suggesting a quick spread of Filipino distillation techniques from Southern Jalisco to the North.

The "Lineño" landrace is better represented in Tolimán, the neighboring municipality of Zapotitlán de Vadillo, but it is also frequent in other municipalities not included in this study (e.g., Tonaya, [8]). This landrace is recognized as one of the highest yielders, with a shorter production cycle and great capacity to produce shoots. It is, together with "Ixtero Amarillo", the most mentioned and recognized by all interviewed farmers in both municipalities, and it is currently the most common in Tolimán, where it is grown in monocultures with a high density of plants, sometimes together with *A. tequilana*, and also reproduced mostly by bulbs.

Finally, samples that came from landraces used with less frequency by producers are all grouped within the large Cluster 4 (Figure 3, left), which are closely related to most of our wild samples. This lack of differentiation between cultivated and wild populations was already noticed in the sampling of *A. angustifolia* var. *pacifica* in Sonora, the north of Mexico [31]. This is congruent with affirmations from the oldest farmers interviewed, who narrate that they used to have an active and constant practice of germplasm selection from the surrounding hills. However, this is not the case for recent generations of farmers, who seek more established landraces.

Selection criteria for spirit production vary between producers. Some valuable traits are linked to the production of larger, heavier heads, high sugar content, resistance to pests, diseases and foragers, reproductive precociousness, flavor, or meeting commercial criteria [24]. The generation of SNPs sites allowed us to evaluate the genetic variability and differentiation in these landraces in comparison with wild plants globally along the genome and not only in a few loci. The future availability of high-quality reference genomes for major landraces in the *A. angustifolia* complex, together with a more extensive study of landraces with an increased sampling, will allow the detection of loci of interest related to the desired valuable traits, as well as a correlation to regional abiotic conditions.

4. Conclusions

In this study, we report for the first time a genome-wide characterization of Jalisco spirit producing *A. angustifolia* autochthonous landraces using a GBS approach. This geographic area is highly relevant because it holds the greatest diversity of landraces in the *A. angustifolia* complex; it is also the putative center of origin of agave distillation in general, and the domestication of the blue agave tequila plant in particular. Our approach, using GBS sequencing and SNPs calling, gave us access to a significantly larger number of loci to analyze in comparison with previous studies using traditional markers such as AFLPs or microsatellites. Our data show that the more extensively used and widely recognized traditional landraces such as "Ixtero", "Lineño" or "Amarillo" have an important degree of differentiation, both at the genetic and morphological level. However, the differentiation is not absolute and there is a very close relationship with wild genotypes. This might be the result of the extremely recent management and domestication of landraces, together with the important use of clonal propagation. We document the genetic basis of the domestication process in the area, where farm holders recur to wild plants to select and propagate new local varieties that reduce their genetic diversity after several generations of clonal propagation. This method promotes the acquisition of locally adapted germplasm and the diversification of available landraces.

The information we provide enables a better characterization of traditional landraces in the state of Jalisco, Mexico. We hope that it helps in technologizing their use and management and provides tools to foment and protect traditional landraces and farming systems to preserve the agrobiodiversity they hold.

5. Materials and Methods

5.1. Sample Collection

In total, we collected 87 samples in the state of Jalisco, Mexico during 2020 and 2021 (Figure 1). We visited twelve mezcaleras in three municipalities: Cabo Corrientes (four localities) on the North Coast, and Zapotitlán de Vadillo (two localities) and Tolimán (one locality) in Southern Jalisco. In addition, we collected several wild populations around both North and South areas of study, as well as central Jalisco near Tequila, a locality of putative origin of the *A. tequilana* var. "Azul". When sampling cultivated plants, producers were interviewed to document the landraces they use and how they recognize them. With their help, we selected several mature plants per landrace and took morphometric measurements. Our pool of samples includes 14 landraces from both *A. angustifolia* and *A. rhodacantha* (Table 1, Figure 2). Additionally, we included five *A. tequilana* samples of plants growing on the CINVESTAV Irapuato botanical garden whose origin are private blue agave destined plantations in Tequila, Jalisco (Simpson J. *pers. com*).

In total, twelve producers were visited and interviewed in a semi-structured way to document the landraces they use and how they recognize them. With their help, 10 mature plants per accession; that is, near to the reproductive state; were selected. Though most of the plants are vegetatively propagated, producers usually obtain them from different mother plants on the same farm. Efforts were made to have each variety represented by more than one accession in order to avoid clones, except in the cases where varieties are maintained by a single producer. To assess the wild genetic background of the landraces studied, we sampled *A. angustifolia* plants growing in the wild. In this case, plants sampled were located far enough from farmsteads, far from residential or agricultural properties, and evidently growing under isolation and not under cultivation. When wild plants were isolated, we collected only tissue samples for DNA extraction. When wild plants were found in colonies or populations, ten individuals separated by more than 10m were sampled and an herbarium voucher specimen was prepared (Table 1).

5.2. Morphological Data and Analyses

We gathered morphological information for 20 samples within our genetic dataset (i.e., 23%). The morphological variables measured have previously shown to be useful in the study and characterization of traditional agave landraces [24]: plant length (cm), leaf length (cm), maximum leaf width (cm), leaf width at the middle of the leaf (cm), terminal thorn length (cm), terminal thorn base width (cm), number of lateral teeth, distance between teeth (cm) and teeth length (cm). Munsell leaf colors were also recorded and converted to xyY coordinates. In addition, the other 5 variables were calculated as combinations of the previous ones: distance between teeth/leaf length, number of lateral teeth/leaf length, terminal thorn length/terminal thorn base width, leaf length/terminal thorn length and leaf length/leaf width at middle. Variables highly correlated (>0.9 Pearson coefficient) were removed. All the measurements were taken from plants growing in mezcalera agroforest producing systems from Cabo Corrientes and Zapotitlán de Vadillo, thanks to the owner's collaboration. To explore the congruence amongst morphological features and landrace information provided by producers, a normalized heatmap in the range (0–1) was constructed in R with BBmisc [38], palette [39] and stats (R core) packages, and the resulting UPGMA Euclidean distance dendrogram was compared both visually and numerically (cophenetic and Baker correlation matrix indices) with a genetic tree built just for the plants, to which we had morphological data using R package dendextend [40]. The genetic tree was also built with UPGMA clustering using the Euclidean distances from the first three principal components of the PCA obtained with the unlinked SNP. To evaluate support, both morphological and genetic trees were bootstrapped 100 times using the R package ape [41]. A mantel test was performed among the morphological and genetic pairwise distance matrices (vegan R package, [42]). All calculations were performed in R using the following packages: vcfR [43], VariantAnnotation [44], BBmisc [38], adegenet [45], ape [41], paletteer [39], dendextend [40] and vegan [42].

5.3. DNA Sequencing and Bioinformatic Analyses

Tissue samples of plants collected in the field were either immediately frozen on liquid N or maintained at 4 °C for 24 to 36 h and then transferred to a −70 °C storage. Approximately 300 mg of each sample was grounded using liquid nitrogen and high-molecular-weight DNA was extracted (DNeasy Plant 96 QUIAGEN® Kit). Genomic DNA extracted was checked for degradation using a 1% agarose gel electrophoresis, and the quality was determined using NanoDrop 8000 (Thermo Scientific (Waltham, MA, USA)). Samples were standardized and library preparation was performed at the Genomics Services Laboratory in Langebio, Mexico. Samples were double digested using the restriction enzymes BglII y DdeI and sequenced using the NovaSeq Illumina platform for 5 million single end 100bp reads per sample (1 × 100).

Per-base and per-read quality score statistics were calculated for each fastq file through FastQC v0.11.1 [46] and MultiQC v1.0 [47]. After demultiplexing, mean quality scores for all libraries were $Q \geq 50$ and no adapters were present. The unpublished *Agave tequilana* reference genome was provided by LANGEBIO CINVESTAV Irapuato (Irapuato, Mexico) (*in prep*) and was used as a reference to align sequenced reads using the default algorithm of the Burrows–Wheeler Aligner v0.7.15 (bwa mem [48]), and the BAM file was sorted and indexed using SAMTOOLS v1.9 [49]. These BAM files were used for the SNPs calling process and assessed for sequencing depth using SAMTOOLS (samtools depth). Loci were identified using the *gstacks* pipeline with default parameters in STACKS v2.3e [50].

5.4. Population Genetics and Statistical Analyses

Population genetics statistics were computed using the populations program in STACKS v2.3e [50] with the following arguments: 80 percent of individuals within and across populations were required to process a locus, a minimum number of 2 populations a locus must be present to process a locus, a minimum minor allele frequency of 0.1 and a minimum allele count of 2 was required to process a nucleotide site at a locus (-p 2 -r 0.8 -R 0.8–min-maf 0.1–min-mac 2 -H). Finally, we removed all SNPs that did not pass these filters. Population genetic parameter estimations (H_e, F_{ST} and $\boldsymbol{\pi}$) were calculated using 1000 bootstrap repetitions, *p*-value F_{ST} correction and kernel smoothing (–fst_correction '*p*-value'–bootstrap 1000–fstats–k).

To select representative unlinked SNPs, the original dataset was pruned by linkage disequilibrium (LD) using PLINK v1.9 [51] with standard parameters (–indep-pairwise 50 10 0.1). PLINK was also used with this dataset to calculate Nei's genetic distances (–ibs), to perform identity-by-descent analysis (–genome) and to calculate the principal component analysis and genetic distances. We selected the eigenvectors from the first three components to build a hierarchical clustering (hclust (); method = "average") within R. Finally, we carried out ADMIXTURE analyses using K values ranging from 2 to 8, ten random seeds and 200 bootstraps per K value (ADMIXTURE v1.3.0 [52]) and selected the best K values, as well as identify the most representative repetitions with the lowest error rates using the CLUMPAK online pipeline (http://clumpak.tau.ac.il/; accessed on 14 May 2022 [53]). Genetic diversity indexes for the different groups were calculated as follow for two partitions of the data set; individuals of the groups 1, 2, 3, 4 and groups 1, 2, 3a, 3b, 4a, 4b obtained after analyzing the PCA based dendrogram. Expected heterocigosity (*He*), observed Heterocigosity (*Ho*) and inbreeding coeffient (*Fis*) were calculated with the

R packages hierfstat [54] and adegenet [45,55]. Nei Pi nucleotide diversity was calculated with the R package PopGenome [56].

Author Contributions: Conceptualization, T.H.-H. and D.C.-T.; methodology, all authors; field work and laboratory work, M.V.-C., T.H.-H. and D.C.-T.; formal analysis, E.M.-G. and N.L.; investigation, T.H.-H. and D.C.-T.; resources, A.H.-E., T.H.-H. and D.C.-T.; data curation, T.H.-H., E.M.-G. and D.C.-T.; writing—original draft preparation T.H.-H., D.C.-T. and E.M.-G.; writing—review and editing, all authors; supervision, T.H.-H., A.H.-E. and D.C.-T.; project administration, A.H.-E., T.H.-H. and D.C.-T.; funding acquisition, A.H.-E., T.H.-H. and D.C.-T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Mexican Ministry of Agriculture and Rural Development (SADER, formerly SAGARPA) and the Mexican National Council of Science and Technology (Conacyt) grant number SAGARPA-CONACYT 292546. The Comisión Nacional Para el Conocimiento y Uso de la Biodiversidad provided funds for travel expenses for interviews through the Project GEF 9380-Agrobiodiversidad Mexicana. T.H.-H. and E.M.-G. were funded by Desert Botanical Garden (DBG). N.L. and M.V.-C. were supported by a post-doctoral grant financed by CONACYT (Mexico).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The manuscript describing the *A. tequilana* draft genome is still in preparation. However, access and permission to access the data before publication can be arranged by directly contacting Dr. Alfredo Herrera-Estrella (alfredo.herrera@cinvestav.mx) or Dr. Selene Fernández-Valverde (selene.valverde@cinvestav.mx).

Acknowledgments: We wish to thank the technical staff of the Genomic Services Unit of the UGA-Langebio for generating the GBS data. Field work, Pablo Carrillo-Reyes, Ofelia Vargas-Ponce, María Magdalena Padilla del Muro, Isaac Bizarro-Martín, Estefanía Acosta-Pérez, Luis Manuel Cardona-Cruz, and Lucero Morales-Cervantes. Voucher preparation and taxonomic identification of some samples, Pablo Carrillo-Reyes and Abisaí García-Mendoza. DNA extraction, Pablo Luis Hernández-Cervantes. Suggestions on the Admixture analysis, Diego Ortega-del-Vecchyo. To all the maguey and mezcal producers and families in the municipalities visited.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Callen, E.O. Food Habits of Some Pre-Columbian Mexican Indians. *Econ. Bot.* **1965**, *19*, 335–343. [CrossRef]
2. Gentry, H.S. *Agaves of Continental North America*; University of Arizona Press: Tucson, AZ, USA, 1982.
3. Walton, M.K. The evolution and localization of mezcal and tequila in Mexico. *Geográfica* **1977**, *85*, 113–132.
4. Valenzuela-Zapata, A.G. *El Agave Tequilero, Su Cultivo e Industria*, 2nd ed.; Monsato: Guadalajara, México, 1997.
5. Luna-Zamora, R. *La Historia del Tequila, de Sus Regiones y Sus Hombres*, 2nd ed.; CONACULTA: México City, México, 1999.
6. Bruman, H.J. Aboriginal Drink Areas of New Spain. Ph.D. Thesis, University of California, Berkley, CA, USA, 1940.
7. Bruman, H.J. *Alcohol in Ancient Mexico*; University of Utah Press: Salt Lake, UT, USA, 2000.
8. Colunga-GarcíaMarin, P.; Zizumbo-Villarreal, D. El tequila y otros mezcales del centro-occidente de México: Domesticación, diversidad y conservación de germoplasma. In *En lo Ancestral Hay Futuro: Del Tequila, los Mezcales y Otros Agaves*; Centro de Investigación Científica de Yucatán: Mérida, México, 2007; pp. 113–131.
9. SEGOB. *Declaración de Protección a la Denominación de Origen Tequila*; Diario Oficial de la Federación: Ciudad de México, México, 1999.
10. CRT. Consejo Regulador del Tequila. *Inf. Estadístico* **1995**, *1999*, 2022. Available online: https://www.crt.org.mx/EstadisticasCRTweb/ (accessed on 16 June 2022).
11. Tetreault, D.; McCulligh, C.; Lucio, C. Distilling Agro-Extractivism: Agave and Tequila Production in Mexico. *J. Agrar. Chang.* **2021**, *21*, 219–241. [CrossRef]
12. Hostettler, S. *Land Use Changes and Transnational Migration*; EPFL: Lausanne, Switzerland, 2007.
13. Bowen, S.; Zapata, A.V. Geographical Indications, Terroir, and Socioeconomic and Ecological Sustainability: The Case of Tequila. *J. Rural Stud.* **2009**, *25*, 108–119. [CrossRef]
14. Grzybowska, N.S.; Gerritsen, P. *Construyendo Poderes Locales: Microdestilerias y Agave Azul en el sur de Jalisco*; Editorial Universitaria; Centro Universitario de la Costa Sur–Universidad de Guadalajara: Autlán de Navarro, México, 2013.

15. Zizumbo-Villarreal, D.; Vargas-Ponce, O.; Rosales-Adame, J.J.; Colunga-GarcíaMarín, P. Sustainability of the Traditional Management of Agave Genetic Resources in the Elaboration of Mezcal and Tequila Spirits in Western Mexico. *Genet. Resour. Crop Evol.* **2013**, *60*, 33–47. [CrossRef]
16. Gerritsen, P.R.W.; Martínez-Rivera, L.M. *Agave Azul, Sociedad y Medio Ambiente: Una Perspectiva de La Costa Sur de Jalisco*; Universidad de Guadalajara, Centro Universitario de la Costa Sur: Autlán de Navarro, México, 2010; ISBN 607-450-213-7.
17. Gil-Vega, K.; González Chavira, M.; Martínez de la Vega, O.; Simpson, J.; Vandemark, G. Analysis of Genetic Diversity in Agave Tequilana Var. Azul Using RAPD Markers. *Euphytica* **2001**, *119*, 335–341. [CrossRef]
18. Vargas-Ponce, O.; Zizumbo-Villarreal, D.; Martínez-Castillo, J.; Coello-Coello, J.; Colunga-GarcíaMarín, P. Diversity and Structure of Landraces of Agave Grown for Spirits under Traditional Agriculture: A Comparison with Wild Populations of *A. Angustifolia* (Agavaceae) and Commercial Plantations of *A. tequilana*. *Am. J. Bot.* **2009**, *96*, 448–457. [CrossRef]
19. Ibarrola-Rivas, M.J. Sustainability Analyses of Agave Production in Mexico. Master's Thesis, University of Groningen, Groningen, The Netherlands, 2010; p. 53.
20. Torres-García, I.; Rendón-Sandoval, F.J.; Blancas, J.; Moreno-Calles, A.I. The Genus *Agave* in Agroforestry Systems of Mexico. *Bot. Sci.* **2019**, *97*, 263–290. [CrossRef]
21. Eguiarte, L.E.; Jiménez Barrón, O.A.; Aguirre-Planter, E.; Scheinvar, E.; Gamez, N.; Gasca-Pineda, J.; Castellanos-Morales, G.; Moreno-Letelier, A.; Souza, V. Evolutionary Ecology of *Agave*: Distribution Patterns, Phylogeny, and Coevolution (an Homage to Howard S. Gentry). *Am. J. Bot.* **2021**, *108*, 216–235. [CrossRef]
22. Colunga-GarcíaMarín, P.; Zizumbo-Villarreal, D.; Martínez-Torres, J. Tradiciones en el aprovechamiento de los agaves mexicanos: Una aportacion a la proteccion legal y conservación de su diversidad biológica y cultural. In *En lo Ancestral Hay Futuro: Del Tequila, los Mezcales y Otros Agaves*; Centro de Investigaciones Científicas de Yucatán: Mérida, México, 2007; p. 248.
23. Torres, I.; Casas, A.; Vega, E.; Martínez-Ramos, M.; Delgado-Lemus, A. Population Dynamics and Sustainable Management of Mescal Agaves in Central Mexico: *Agave potatorum* in the Tehuacán-Cuicatlán Valley. *Econ. Bot.* **2015**, *69*, 26–41.
24. Vargas-Ponce, O.; Zizumbo-Villarreal, D.; Colunga-García Marin, P. In Situ Diversity and Maintenance of Traditional Agave Landraces Used in Spirits Production in West-Central Mexico. *Econ. Bot.* **2007**, *61*, 362–375. [CrossRef]
25. Rivera-Lugo, M.; García-Mendoza, A.; Simpson, J.; Solano, E.; Gil-Vega, K. Taxonomic Implications of the Morphological and Genetic Variation of Cultivated and Domesticated Populations of the *Agave angustifolia* Complex (Agavoideae, Asparagaceae) in Oaxaca, Mexico. *Plant Syst. Evol.* **2018**, *304*, 969–979. [CrossRef]
26. Trejo, L.; Limones, V.; Peña, G.; Scheinvar, E.; Vargas-Ponce, O.; Zizumbo-Villarreal, D.; Colunga-GarcíaMarín, P. Genetic Variation and Relationships among Agaves Related to the Production of Tequila and Mezcal in Jalisco. *Ind. Crops Prod.* **2018**, *125*, 140–149. [CrossRef]
27. El Mezcal Va Por los Mercados Internacionales. Líderes Mexicanos, October 20, 2021. Available online: https://lideresmexicanos.com/tendencias/el-mezcal-va-por-los-mercados-internacionales/ (accessed on 23 June 2022).
28. Gobierno del Estado de Oaxaca. Estados Unidos, principal comprador de mezcal en el mundo. In *Coordinación General de Comunicación Social y Vocería del Estado*; Gobierno del Estado de Oaxaca, Oaxaca, México. LA, California, 27 de Junio de 2019. Available online: https://www.oaxaca.gob.mx/comunicacion/estados-unidos-principal-comprador-de-mezcal-en-el-mundo/ (accessed on 23 June 2022).
29. Elshire, R.J.; Glaubitz, J.C.; Sun, Q.; Poland, J.A.; Kawamoto, K.; Buckler, E.S.; Mitchell, S.E. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* **2011**, *6*, e19379. [CrossRef]
30. Palomino, G.; Dolezel, J.; Mendez, I.; Rubluo, A. Nuclear Genome Size Analysis of *Agave tequilana* Weber. *Caryologia* **2003**, *56*, 37–46. [CrossRef]
31. Klimova, A.; Ruiz Mondragón, K.Y.; Molina Freaner, F.; Aguirre-Planter, E.; Eguiarte, L.E. Genomic Analyses of Wild and Cultivated Bacanora *Agave* (*Agave angustifolia* var. pacifica) Reveal Inbreeding, Few Signs of Cultivation History and Shallow Population Structure. *Plants* **2022**, *11*, 1426. [CrossRef]
32. Gil-Vega, K.; Díaz-Quezada, C.E.; Nava-Cedillo, A.; García-Mendoza, A.; Simpson, J. Análisis AFLP del género *Agave* refleja la clasificación taxonómica basada en caracteres morfológicos y otros métodos moleculares. In *En lo Ancestral Hay Futuro: Del Tequila, los Mezcales y Otros Agaves*; Centro de Investigaciones Científicas de Yucatán: Mérida, México, 2007; pp. 23–39.
33. Zietkiewicz, E.; Rafalski, A.; Labuda, D. Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. *Genomics* **1994**, *20*, 176–183. [CrossRef]
34. Abraham-Juárez, M.J.; Ramírez-Malagón, R.; del C. Gil-Vega, K.; Simpson, J. AFLP analysis of genetic variability in three reproductive forms of *Agave tequilana*. *Rev. Fitotec. Mex.* **2009**, *32*, 171–175. [CrossRef]
35. Barrientos Rivera, G.; Esparza Ibarra, E.L.; Segura Pacheco, H.R.; Talavera Mendoza, Ó.; Sampedro Rosas, M.L.; Hernández Castro, E. Morphological characterization of *Agave angustifolia* and its conservation in Guerrero, Mexico. *Rev. Mex. Cienc. Agrícolas* **2019**, *10*, 655–668. [CrossRef]
36. Deschamps, S.; Llaca, V.; May, G.D. Genotyping-by-Sequencing in Plants. *Biology* **2012**, *1*, 460–483. [CrossRef] [PubMed]
37. Carrillo-Galván, G. Domesticación de Agaves Productores de Fibra en el Centro-Occidente de México: Una Aproximación Etnobotánica y Morfológica. Master's Thesis, Centro de Investigaciones Científicas de Yucatán (CICY), Mérida, México, 2011.
38. Bischl, B.; Lang, M.; Bossek, J.; Horn, D.; Richter, J.; Surmann, D. *BBmisc: Miscellaneous Helper Functions for B. Bischl*; Version 1.12; R Package; R Foundation: Vienna, Austria, 2017.

39. Hvitfeldt, E. Paletteer: Comprehensive Collection of Color Palettes. Available online: https://cran.r-project.org/web/packages/paletteer/index.html (accessed on 10 May 2022).
40. Galili, T. Dendextend: An R Package for Visualizing, Adjusting and Comparing Trees of Hierarchical Clustering. *Bioinformatics* **2015**, *31*, 3718–3720. [CrossRef] [PubMed]
41. Paradis, E.; Schliep, K. Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R. *Bioinformatics* **2019**, *35*, 526–528. [CrossRef]
42. Oksanen, J.; Blanchet, F.G.; Kindt, R.; Legendre, P.; Minchin, P.R.; O'hara, R.; Simpson, G.L.; Solymos, P.; Stevens, M.H.H.; Wagner, H. *Package 'Vegan'*; Version 2.5.7; Community Ecology Package; R Foundation: Vienna, Austria, 2013; Volume 2, pp. 1–295.
43. Knaus, B.J.; Grünwald, N.J. Vcfr: A Package to Manipulate and Visualize Variant Call Format Data in R. *Mol. Ecol. Resour.* **2017**, *17*, 44–53. [CrossRef] [PubMed]
44. Obenchain, V.; Lawrence, M.; Carey, V.; Gogarten, S.; Shannon, P.; Morgan, M. VariantAnnotation: A Bioconductor Package for Exploration and Annotation of Genetic Variants. *Bioinformatics* **2014**, *30*, 2076–2078. [CrossRef]
45. Jombart, T.; Ahmed, I. Adegenet 1.3-1: New Tools for the Analysis of Genome-Wide SNP Data. *Bioinformatics* **2011**, *27*, 3070–3071. [CrossRef]
46. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*; ScienceOpen, Inc.: Burlington, MA, USA, 2010.
47. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report. *Bioinformatics* **2016**, *32*, 3047–3048.
48. Li, H.; Durbin, R. Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef]
49. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M. Twelve Years of SAMtools and BCFtools. *Gigascience* **2021**, *10*, giab008. [CrossRef]
50. Rochette, N.C.; Rivera-Colón, A.G.; Catchen, J.M. Stacks 2: Analytical Methods for Paired-end Sequencing Improve RADseq-based Population Genomics. *Mol. Ecol.* **2019**, *28*, 4737–4754.
51. Chang, C.C.; Chow, C.C.; Tellier, L.C.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets. *Gigascience* **2015**, *4*, s13742-015. [CrossRef] [PubMed]
52. Alexander, D.H.; Novembre, J.; Lange, K. Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome Res.* **2009**, *19*, 1655–1664. [CrossRef] [PubMed]
53. Kopelman, N.M.; Mayzel, J.; Jakobsson, M.; Rosenberg, N.A.; Mayrose, I. Clumpak: A Program for Identifying Clustering Modes and Packaging Population Structure Inferences across K. *Mol. Ecol. Resour.* **2015**, *15*, 1179–1191. [CrossRef] [PubMed]
54. Goudet, J.; Jombart, T. Package "hierfstat", Version 1.5-11; Estimation and Tests of Hierarchical F-Statistics. Available online: https://cran.r-project.org/web/packages/hierfstat/index.html (accessed on 20 April 2022).
55. Jombart, T. Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* **2008**, *24*, 1403–1405. [CrossRef] [PubMed]
56. Pfeifer, B.; Wittelsbürer, U.; Ramos-Onsins, S.E.; Lercher, M.J. PopGenome: An Efficient Swiss Army Knife for Population Genomic Analysis in R. *Mol. Biol. Evol.* **2014**, *31*, 1929–1936. [CrossRef]