UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

INSTITUTO DE ECOLOGÍA

BIOLOGÍA EVOLUTIVA

ESTRUCTURA POBLACIONAL DE LEVADURAS *SACCHAROMYCES* AISLADAS DE FERMENTACIÓN DE AGAVE

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN CIENCIAS BIOLÓGICAS

PRESENTA:

JOSÉ ANTONIO URBÁN ARAGÓN

TUTORA PRINCIPAL DE TESIS:

DRA. LUCÍA GUADALUPE MORALES REYES
INSTITUTO DE ECOLOGÍA, LIIGH, UNAM

COMITÉ TUTOR:

DRA. ALICIA GONZÁLEZ MANJARREZ
INSTITUTO DE FISIOLOGÍA CELULAR, UNAM

DRA. ALICIA MASTRETTA YANES
INSTITUTO DE ECOLOGÍA, UNAM, CONABIO

CD. MX. **ABRIL, 2021**

COORDINACIÓN DEL POSGRADO EN CIENCIAS BIOLÓGICAS
ENTIDAD INSTITUTO DE ECOLOGÍA
OFICIO CPCB/295/2021
ASUNTO: Oficio de Jurado

M. en C. Ivonne Ramírez Wence
Directora General de Administración Escolar, UNAM
P r e s e n t e

Me permito informar a usted que, en la reunión ordinaria del Subcomité de Biología experimental y Biomedicina del Posgrado en Ciencias Biológicas, celebrada el **día 15 de febrero de 2021,** se aprobó el siguiente jurado para el examen de grado de **MAESTRO EN CIENCIAS BIOLÓGICAS** en el campo de conocimiento de **Biología Evolutiva** del estudiante **URBÁN ARAGÓN JOSÉ ANTONIO** con número de cuenta **310534084,** con la tesis titulada: **"ESTRUCTURA POBLACIONAL DE LEVADURAS *SACCHAROMYCES* AISLADAS DE FERMENTACION DE AGAVE"** realizada bajo la dirección de la **DRA. LUCÍA GUADALUPE MORALES REYES,** quedando integrado de la siguiente manera:

Presidente:	DR. LUIS ENRIQUE EGUIARTE FRUNS
Vocal:	DR. JUAN SERVANDO NÚÑEZ FARFÁN
Vocal:	DRA. MARÍA DEL CARMEN ÁVILA ARCOS
Vocal:	DRA. MARÍA SOLEDAD FUNES ARGÜELLO
Secretario:	DRA. MARIA ALICIA GONZÁLEZ MANJARREZ

Sin otro particular, me es grato enviarle un cordial saludo.

A T E N T A M E N T E
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Cd. Universitaria, Cd. Mx., a 13 de abril de 2021

COORDINADOR DEL PROGRAMA

DR. ADOLFO GERARDO NAVARRO SIGÜENZA

POSGRADO EN CIENCIAS BIOLOGICAS
COORDINACIÓN

AGRADECIMIENTOS INSTITUCIONALES

Primeramente, me gustaría agradecer a Posgrado en Ciencias Biológicas de la Universidad Nacional Autónoma de México por su enorme apoyo y la oportunidad de hacer mi posgrado en este programa.

Este trabajo fue apoyado por los proyectos CONACYT números 284992 y 103000, los proyectos PAPIIT de DGAPA-UNAM números IA201019 y IN209021 y la beca CONACYT número 873507.

Por último, agradezco de forma muy especial a la Dra. Lucía Guadalupe Morales Reyes, quien fuera el Tutor Principal de este proyecto y a los miembros de mi Comité Tutor: Dra. Alicia González Manjarrez y Dra. Alicia Mastretta Yanes. Muchas gracias por todas sus observaciones y preguntas. Este fue un proceso que disfruté mucho a su lado y sus contribuciones ayudaron a enriquecer este proyecto.

AGRADECIMIENTOS A TÍTULO PERSONAL

Me gustaría agradecer a todas las personas que me acompañaron en la elaboración de este proyecto de investigación. Para empezar, estoy sumamente agradecido con la Dra. Lucía Guadalupe Morales Reyes, quien me aceptó en su grupo de investigación y me permitió formar parte de un proyecto increíble y apasionante. Tu confianza y apoyo depositados en mí, así como tus enseñanzas, me han ayudado muchísimo a crecer como investigador pero, sobre todo, como ser humano. De igual manera, a todos los miembros del Laboratorio de Evolución de Genoma de Levaduras (Iván, Jorge, Aarón, Arte, María, Natalia, Héctor, Luis, Margareta) por hacer de mi estancia en el laboratorio un momento muy feliz.

A nuestros colaboradores en este proyecto: el Dr. Alexander de Luna Fors del CINVESTAV-LANGEBIO y el Dr. Eugenio Mancera Ramos del CINVESTAV-Unidad Irapuato, así como a los miembros de sus respectivos grupos de investigación. En especial, quiero dar gracias al Dr. Luis Fernando García Ortega, cuyo apoyo fue extremadamente importante para poder llevar a buen puerto este proyecto.

A los miembros del Comité Tutor de esta tesis: la Dra. Alicia González Manjarrez y la Dra. Alicia Mastretta Yanes, por sus extraordinarias observaciones y revisiones durante la realización de este proyecto. Disfruté muchísimo los tutorales a su lado y aprendí mucho de cada una de ustedes.

A los miembros del Comité Jurado: la Dra. María del Carmen Ávila Arcos, Dr. Juan Servando Núñez Farfán, el Dr. Luis E. Eguiarte Fruns, la Dra. María Soledad Funes Argüello, la Dra. Alicia González Manjarrez y la Dra. Alicia Mastretta Yanes por el tiempo dedicado y las observaciones para el mejoramiento de este proyecto de tesis.

Al Dr. Diego Ortega Del Vecchyo y a la Dra. María Ávila Arcos por ayudarme a esclarecer dudas del proyecto a lo largo del camino. Su ayuda ha sido invaluable en el desarrollo de este proyecto.

A las técnicas Carina Díaz y Alejandra Castillo del LIIGH por su ayuda en *wet-lab* y a Luis Aguilar, Alejandro de Leon, Carlos Flores y Jair García del LAVIS-UNAM por su apoyo técnico en tecnologías de la información.

Por último, y no por ello menos importante, quiero agradecer infinitamente a mi familia, en especial, a mis padres, Patricia del Socorro Aragón Durand y José Antonio Urbán Carrillo. Gracias a su apoyo constante, he llegado hasta aquí, los amo profundamente. Al resto de mi familia que me han acompañado en todo momento quienes también han sido sumamente importantes para mí.

Con todo el amor, a mi novia Mariana Gamiochipi Arjona, quien ha emprendido este viaje conmigo. Durante este trayecto, tú me has escuchado y me das dado mucho ánimo a perseguir mis sueños. En los baches a lo largo del camino, tú has sabido encauzarme por el camino correcto y eso ha significado el mundo para mí. Me has enseñado a no conformarme y a siempre sacar lo mejor de mi en todas las facetas de mi vida. El esfuerzo y dedicación que pones a tus metas han sido mi más importante inspiración para seguir adelante. Te amo muchísimo, gracias por acompañarme en esta aventura y sé que aún nos faltan muchísimas más.

A mis padres por su apoyo, confianza y cariño durante todas las etapas de mi vida,

los amo con todo mi corazón.

A Mariana, por acompañarme en esta aventura y

ser un ejemplo a seguir, te amo muchísimo.

A mis amigos, que también fueron piedra angular en este recorrido…

ÍNDICE

RESUMEN

Saccharomyces cerevisiae es uno de los organismos modelos eucariontes por excelencia. Este hongo unicelular ha sido considerado como el “caballo de batalla” de la biotecnología, ya que se utiliza en diferentes e importates procesos industriales y biotecnológicos, como son la fermentación de bebidas alcohólicas y alimentos, en la fabricación de bioetanol y en la producción de medicamentos. Debido a su maleable genoma, su capacidad para hibridar con otras especies del género y a que ha pasado por diferentes eventos de domesticación, el estudio de la evolución de *S. cerevisiae* es relevante. El análisis de su diversidad genética muestra que puede dividirse en dos grandes grupos: cepas de origen natural (suelo, frutas, cortezas de árboles, animales, entre otros) y cepas de origen antropogénico y/o industrial (fermentaciones, producción de alimentos, bioetanol, entre otros). En el caso de las cepas antropogénicas, algunos autores se han dado a la tarea de estudiar ciertos grupos de *S. cerevisiae* dedicados específicamente a las fermentaciones de algún alimento o bebida, donde el clado asociacio al vino es monofilético y poco diverso, mientras que el clado de la cerveza es polifilético y muy diverso. La fermentación de jugo de agave cocido es importante para México, porque es la base de la producción de destilados de agave. En el caso de la fermentación de agave para producción de mezcal, previamente sólo un estudio había evaluado la diversidad genética de las cepas de fermentación de agave en un contexto general. Otros estudios se habían enfocado en estudiar características fisiológicas y morfológicas de algunos aislados, así como caracterizaciones moleculares puntuales. Por otro lado, Peter y colegas (2018) secuenciaron y analizaron 1,011 aislados de *S. cerevisiae* de diferentes partes del mundo. Este estudio incluyó nueve cepas de fermentación de agave donde siete de nueve provenían del estado de Tamaulipas. Si bien este estudio fue uno de los primeros en analizar aislados de fermentación de agave en un contexto global, el número de cepas mexicanas de fermentación de agave fue bajo, no se cubrió a todas las regiones productoras de mezcal y, en consecuencia, probablemente se subestimó la diversidad. En esta tesis, se muestrearon, aislaron y secuenciaron 137 cepas de fermentación de agave de diferentes regiones de México. De las 137 cepas, 118 cepas fueron secuenciados con tecnología BGI de lecturas cortas, 14 con Illumina NextSeq y cinco con Illumina MiSeq. Posteriormente, se añadieron 170 secuencias de *S. cerevisiae* de Peter y colaboradores (2018). Análisis filogenéticos y de componentes principales, evaluaciones de su estructura poblacional y una prueba de Mantel identificaron que la mayoría de las cepas de fermentación de agave forman un grupo monofilético, genéticamente diferenciado de clados de *S. cerevisiae* de otras partes del mundo. Además, se observó que las siete cepas previamente analizadas por Peter y colegas (2018) representan una pequeña porción de la diversidad presente en el grupo. Se encontró una correlación entre la agrupación de cepas y la localización geográfica de las cepas y se identificó una alta presencia de introgresiones provenientes de *Saccharomyces paradoxus* (*S. paradoxus*). Finalmente, la mayoría de los aislados tienen una estructura poblacional bien definida con mosaicismo en algunos *clusters* de cepas.

ABSTRACT

Saccharomyces cerevisiae is one of the best model eukaryotic organisms. This unicellular fungus is the “workhorse” of biotechnology since this species participates in the fermentation of alcoholic beverages and foods, the production of bioethanol, the manufacture of medicines, among others. Because of its malleable genome, its proclivity to hybridize with other species of the genus, and its passage through several domestication trajectories, the research on *S. cerevisiae* evolution is highly relevant. The analysis of *S. cerevisiae* genetic diversity shows that the species isolates can be divided into two major groups: strains of natural origin and strains of anthropogenic/industrial origin (fermentations, food production, bioethanol, among others). In the case of anthropogenic strains, some authors have studied certain *S. cerevisiae* groups that are involved in the fermentation of a particular food or beverage. For example, we know that the wine clade of strains is monophyletic and has low genetic diversity, whereas *S. cerevisiae* beer strains are polyphyletic and highly diverse. The fermentation of agave juice is very important for Mexico because this industrial activity is the foundation of diverse agave distilled spirits. In the case of agave´s fermentation for mezcal production, only one study had assessed the genomic diversity of the agave fermentation strain in a broad context. Most of the research had focused on the prevalence of the *S. cerevisiae* species in this process, some morphological and physiological characteristics of the isolates, and some molecular hallmarks of the strains involved in the fermentation of agave. On the other side, Peter and coworkers (2018) sequenced and analyzed 1,011 *S. cerevisiae* genomes from various regions of the world. The 1,011 genomes paper included nine Mexican agave fermentation strains and seven of the nine isolates came from the state of Tamaulipas. The 1,011 genomes paper included nine Mexican agave fermentation strains and 7 of the 9 isolates came from the state of Tamaulipas. Even though this article was one of the first studies to compare Mexican agave fermentation strains to isolates from different regions of the world, we think that the number of Mexican agave strains evaluated was low, the authors did not cover all the regions involved in the production of mezcal and therefore, the diversity of the agave fermentation isolates was probably underestimated. For this dissertation, we sampled, isolated, and sequenced 137 agave fermentation isolates from different regions of Mexico. Of the 137 strains, 118 strains were sequenced with BGI technology, 14 with Illumina NextSeq, and five with Illumina MiSeq. Then, we added 170 genome sequences of *S. cerevisiae* strains from the Peter and coworkers (2018) article. Different analyses such as phylogenetic inferences, principal component analysis, population structure analysis, and a Mantel test identified that most of the strains form a monophyletic group, which is genetically differentiated from other *S. cerevisiae* clades. Additionally, we also found that the 7 strains, previously described by Peter and coworkers (2018), only represent a tiny fraction of the agave fermentation group´s diversity. We also found a correlation between the clustering of the strains and their geographic location and a high presence of introgressions from *Saccharomyces paradoxus* (*S. paradoxus*). Finally, we found that the population structure of most of these strains is well defined with some isolates showing signs of mosaicism and admixture.

INTRODUCCIÓN

Saccharomyces cerevisiae es un buen modelo para estudiar organismos eucariontes

Desde la primera mitad del siglo XX, la levadura *S. cerevisiae* ha sido un organismo modelo ampliamente usado para estudiar aspectos importantes de la biología molecular y celular de las células eucariontes (Marsit S. *et al*., 2017). La trascendencia de este organismo se vio reflejada al ser el primer organismo eucarionte en tener su genoma secuenciado (Goffeau A. *et al*., 1996; Marsit S. *et al*., 2017). Su genoma tiene un tamaño de ~12 Mb distribuidos en 16 cromosomas (Goffeau A. *et al*., 1996). De igual manera, *S. cerevisiae* cuenta con una mitocondria cuyo tamaño oscila entre 73 y 96 kb (de Chiara M. *et al*., 2020). La especie más cercana a *S. cerevisiae* es *S. paradoxus* con una divergencia nucleotídica de 10-15% entre ambas especies (Liti G. *et al*., 2006). Adicionalmente, este hongo unicelular ha sido denominado como el “caballo de batalla” de la biotecnología, ya que participa en procesos industriales y biotecnológicos sumamente importantes, como la fermentación de bebidas alcohólicas y alimentos en la la fabricación de bioetanol y medicamentos, entre otros procesos (Liu L. *et al*., 2013; Marsit S. *et al*., 2017).

De igual manera, con el advenimiento de las tecnologías más avanzadas de secuenciación, *S. cerevisiae* y otras levaduras del mismo género se han convertido en buenos modelos para estudiar las dinámicas genómicas evolutivas involucradas

en la hibridación, inestabilidad genómica, los cambios estructurales a nivel genoma, el aislamiento reproductivo entre especies, poliploidización y aparición de aneuplodías (desviaciones de un múltiplo del número de cromosomas normales en una célula) (Gabaldón T., 2020; Marsit S. *et al*., 2017).

Las cepas naturales e industriales de *S. cerevisiae* presentan diferencias sustantivas

A lo largo de la historia evolutiva de *S. cerevisiae*, se han domesticado aislados de esta especie, a tal grado que se ha propuesto clasificarlas como cepas de origen silvestre o natural y cepas de origen antropogénico o industrial (Marsit S. *et al*., 2017; Peter J. *et al*., 2018). Las cepas de origen industrial o antropogénico son aquellas que se encuentran bajo un proceso de domesticación y, en consecuencia, se han adaptado bien a un medio en particular (Borneman A.R. *et al*., 2011). La domesticación se puede definir como la crianza y selección artificial de especies silvestres para obtener variantes mejor adaptadas a ambientes antropogénicos y con características útiles para los humanos, aunque generalmente con una menor adaptación a los ambientes naturales (Gallone B. *et al*., 2018). Se han descrito tres vías de domesticación de organismos: “comensal”, “de la presa” y “dirigida” (Steensels, J. *et al*., 2018). Estas vías han sido utilizadas, principalmente, para describir y evaluar los procesos de domesticación de animales y plantas. En la vía “comensal”, los animales salvajes se sienten atraídos por los desechos humanos y comienzan a habituarse al nicho humano. El ser humano no juega un papel

importantes en las etapas tempranas de esta vía y, después, ya se iniciaban esquemas de apareamiento estrictos (Steensels, J. *et al*., 2018). Por otro lado, la vía “de la presa” consiste en mantener en cautiverio a los animales que se cazaban. Por útlimo, la vía “dirigida” es la más reciente y consiste en saltarse las etapas de habituación y manejo para empezar inmediatamente con esquemas de apareamiento estricto y, así seleccionar los caracteres deseados en los organismos (Steensels, J. *et al*., 2018). En un contexto de domesticación de microorganismos, la vía “comensal” es la que mejor encajó con este proceso por miles de años hasta que se empezaron a seleccionar cultivos de ciertos organismos mejor adaptados para ciertos procesos (vía “dirigida”) (Steensels, J. *et al*., 2018). Por último, es importante mencionar que algunos microorganismos como las *S. cerevisiae* involucradas en la fermentación de oliva han estado involucradas en un proceso de cuasi-domesticación, la cual consiste en desarrollar adaptaciones a un medio antropogénico sin que haya consecuencias positvias o negativas para los seres humanos (Pontes A. *et al*., 2019).

Las múltiples trayectorias de domesticación de las diferentes subpoblaciones de *S. cerevisiae* han resultado en un proceso de diversificación dentro de la misma especie, que ha traído como consecuencia la separación y divergencia de las cepas industriales de las naturales (Marsit S. *et al*., 2017). Así se han encontrado nuevos ORFs no descritos previamente en el genoma de referencia de *S. cerevisiae* (Borneman A.R. *et al*., 2011). Conforme se han secuenciado más cepas de diferentes orígenes y lugares, la comunidad científica se ha dado cuenta que hay diferencias substanciales-tanto a nivel genómico como fenotípico entre los aislados

naturales y antropogénicos de *S. cerevisiae* (Borneman A.R. *et al*., 2011; Marsit S. *et* al., 2017; Peter J. *et al*., 2018). Gracias al aumento en el número de secuencias genómicas disponibles, se ha empezado a construir el pangemona (colección de todos los ORFs descritos en todos los individuos secuenciados de la especie) de *S. cerevisiae* donde se han incluido ORFs base y ORFs variables (Gang L. *et al*., 2018; Peter J. *et al*., 2018). Los ORFs base son aquellos que se han descrito en todas las cepas secuenciadas de *S. cerevisiae* mientras que los ORFs variables presentan diferencias considerables entre las diversas poblaciones de esta especie. Por ejemplo, se ha descrito que muchos ORFs variables son resultado de eventos de intercambio genético entre diferentes especies, principalmente, hibridaciones y transferencia horizontal de genes (Peter J. *et al*., 2018).

A nivel genómico, se ha reportado que las cepas industriales son más proclives a generar aneuploidías (Gilchrist C. & Stelkens R., 2019; Marsit S. *et al*., 2017), a mostrar evidencia reciente de entrecruzamiento (*admixture*) entre cepas (Tilakaratna V. & Bensasson D., 2017), a mostrar ploidías mayores a 2 (Peter J. *et al*., 2018), a hibridar más con otras especies cercanas, y a mostrar presencia de introgresiones provenientes de otras especies (D´Angiolo M. *et al*., 2018; Peter J. *et al*., 2018). Estas características están íntimamente relacionadas con la plasticidad genómica y la rápida adaptación que estas cepas de *S. cerevisiae* deben experimentar en ambientes con cambios constantes (Giannakou, K. *et al*., 2020). Por lo tanto, se ha encontrado evidencia que estas propiedades han jugado, en la mayoría de las veces, un rol adaptativo importante en los aislados industriales de *S. cerevisiae* (Gilchrist C. & Stelkens R., 2019; Marsit S. *et al*., 2017; Peter J. *et al*.,

2018). En contraste, las cepas de *S. cerevisiae* de origen silvestre o natural, si bien también se entrecruzan entre sí y con otras especies (Dujon, B.A. & Louis E.J., 2017), no muestran evidencia de entrecruzamiento reciente con otros aislados de *S. cerevisiae* (Tilakaratna V. & Bensasson D., 2017), normalmente son diploides (Peter J. *et al*., 2018) y la presencia de introgresiones es más baja en las cepas silvestres que en las cepas de origen antropogénico. Por ejemplo, Peter y colaboradores (2018) reportaron para *S. cerevisiae* que la gran mayoría de las introgresiones provenientes de *S. paradoxus* se concentran en 4 grupos/clados de *S. cerevisiae*: ***Alpechin***, ***Brazilian Bioethanol*** (Bioetanol de Brasil), ***French Guiana*** (Guyana Francesa) y ***Mexican Agave*** (Agave Mexicano). *S. paradoxus* es la especie del género *Saccharomyces* más cercana a *S. cerevisiae*. Ambas especies divergieron de un ancestro común hace ~5-10 millones de años; sus genomas son colineares y guardan una identidad de ~90% (Dori-Bachash M. *et al*., 2011). A diferencia de *S. cerevisiae*, *S. paradoxus* ha sido aislado mayoritariamente de ambientes silvestres, tiene una estructura poblacional más definida e íntimamente relacionada con su geografía y son fenotípicamente menos diversas que la *S. cerevisiae* (Alsammar H. & Delneri, D., 2020). *S. paradoxus* se agrupa en tres linajes principales: Norteamérica, Oriente Lejano y Europa. Una diferencia clave con *S. cerevisiae* es la alta divergencia nucleotídica que puede existir entre individuos de diferentes linajes (1.5-4.6%) (Alsammar H. & Delneri, D., 2020). A pesar de la alta divergencia entre *S. cerevisiae* y *S. paradoxus*, las barreras precigóticas entre ambas especies son débiles y permite que hibriden entre sí (Marsit S. *et al*., 2017).

La hibridación tiene implicaciones evolutivas importantes en las cepas *Saccharomyces*

El ciclo de vida canónico de *S. cerevisiae* se compone de fases entrelazadas de expansión clonal asexual, reproducción sexual y quinisciencia donde las células pasan por fases haploides y diploides (haplo-dibionte) (Fischer G. *et al*., 2020). Células haploide de *S. cerevisiae* con tipos de apareamiento (*mating-types*) distintos (a y α) se aparean para formar células diploide que proliferan mitóticamente cuando los nutrientes son abudantes (Fischer G. *et al*., 2020). Por otro lado, las células diploides pueden hacer meiosis (ciclo sexual) en condiciones de escasez de nutrientes y producir tétradas compuestas de cuatro esporas haploides en un ascus (Fischer G. *et al*., 2020). Sin embargo, el ciclo de vida de *S. cerevisiae* puede ser mucho más complejo que ello e involucrar apareamiento entre especies diferentes del género *Saccharomyces* para dar lugar a hibridaciones interespecie (Fischer G. *et al*., 2020; Marsit S. *et al*., 2017).

La hibridación interespecie es un proceso mediante el cual dos especies son capaces de aparearse y producir progenie, la mayoría de las veces infértil (Marsit S. *et al*., 2017). La presencia de hibridaciones en levaduras tiene implicaciones evolutivas importantes. Por ejemplo, evidencia filogenética apunta a que el clado de duplicación de genoma completo (WGD por sus siglas en inglés) en *S*accharomyceteae surgió a partir un evento de hibridación entre dos especies ancestrales-probablemente diploides que dieron lugar a un alotetraploide que, mediante recombinación y pérdida masiva de genes, resultó en un linaje donde se había duplicado el número de cromosomas- y no de un evento de duplicación de

genoma completo, como originalmente se pensaba (Marcet-Houben M. & Gabaldón T., 2015; Wolfe K.H., 2015). En términos de adaptación, diversificación y especiación, la hibridación interespecie juega un papel importante en el proceso evolutivo de *Saccharomyces*, y por ello, el estudio de sus híbridos es relevante (Gabaldón T., 2020; Marsit S. *et al*., 2017).

Un evento recurrente en las cepas industriales de *Saccharomyces* es la hibridación interespecie (Gabaldón T., 2020; Marsit S. *et al*., 2017). Se ha propuesto que la hibridación es una estrategia habitual de las levaduras en entornos antropogénicos para lidiar con ambientes altamente estresantes y cambiantes (Gabaldón T., 2020; Marsit S. *et al*., 2017; Peter J. *et al*., 2018). Estos eventos pueden suceder debido a las débiles barreras precigóticas, las cuales permiten que dos genomas con bagajes evolutivos coexistan en un organismo. Si bien hay cierto aislamiento reproductivo y las barreras pos-cigóticas (incompatibildades genéticas como las de Bateson-Dobzhansky-Müller y ausencia de recombinación homóloga) entre diferentes especies son fuertes (Marsit S. *et al*., 2017), la hibridación interespecie es capaz de generar progenie viable y fértil de vez en cuando (D'Angiolo M. *et al*., 2020, Marsit S. *et al*., 2017). Por lo tanto, sería conveniente interpretar como una sola población a todas las especies del complejo *Saccharomyces.*

La inestabilidad genómica en los organimos híbridos derivada de la convivencia de dos genomas divergentes (Marsit S. *et al*., 2017), resulta inicialmente en una duplicación de genoma completo e incremento en la heterocigosidad, y posteriormente en la aparición de introgresiones genómicas, pérdidas de ADN

ribosomal de una de las especies parentales, en la pérdida de heterocigosidad, en la pérdida de mitocondria de una de las especies parentales y en la pérdida de genes y enventualmente en la aparición de aneuploidías y de rearreglos cromosómicos (**Figura 1A**) (Charron G. *et al*., 2019; D'Angiolo M. *et al*., 2020; Dujon B.A. & Louis E.J., 2017; Gabaldón T., 2020; Gilchrist C. & Stelkens R., 2019; Marsit S. *et al*., 2020). El caso de las introgresiones es interesante porque, para algunas cepas industriales como las del grupo de **Alpechin** (fermentación de aceitunas) (**Figura 1B**), han funcionado como estrategia adaptativa para ese ambiente en particular. Por ejemplo, D´Angiolo y colaboradores (2020) describieron que las introgresiones provenientes de *S. paradoxus* en el clado de **Alpechin** estaban enriquecidas con genes de interacción con el ambiente, además de compartir algunos genes introgresados con otros clados de *S. cerevisiae* como **Mexican Agave, French Guiana** y **Brazilian Bioethanol** (D´Angiolo M. *et al*., 2020)**.**

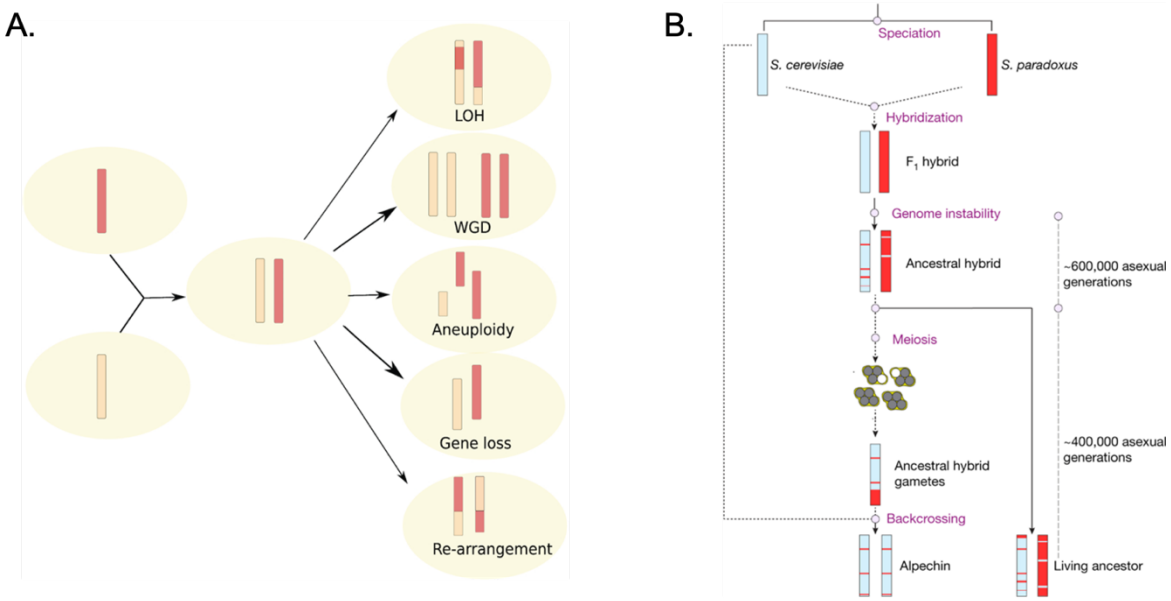


Figura 1. **Implicaciones de la hibridación interespecie en levaduras del género *Saccharomyces.*** El **panel A** (imagen tomada de Gabaldón T., 2020) muestra las implicaciones y consecuencias de la hibridación interespecie: pérdida de heterocigosidad (LOH), duplicación de genoma completo (WGD), aneuploidía (*Aneuploidy*), pérdida de genes (*gene loss*) y rearreglos cromosómicos (*re-arrangement*). El **panel B** (imagen tomada y modificada de D´Angiolo M. *et al*., 2020) muestra una ruta propuesta por D´Angiolo y colegas (D´Angiolo M. *et al*., 2020) mediante la cual se generaron introgresiones de *S. paradoxus* en cepas de *S. cerevisiae* involucradas en la fermentación de aceitunas (**Alpechin**). Se cree que la mayoría de las introgresiones en levadura se forman a partir de retrocruzas de híbridos ancestrales (*backcrossing)* con alguna de las dos cepas parentales (Dujon B.A. & Louis E.J., 2017).

S. cerevisiae ha sido domesticada varias veces a lo largo de la historia

Los inicios de la domesticación de *S. cerevisiae* están relacionados con los primeros eventos de fermentación (Marsit S. *et al*., 2017). La evidencia arqueológica más antigua de la fermentación de una bebida se remonta a China, hace aproximadamente 9,000 años (Marsit S. *et al*., 2017; McGovern P.E. *et al*., 2004). De igual manera, se tiene registro de las primeras fermentaciones de cerveza, vino y pan en Sumeria (7,000 A.C.), Irán (5,400-5,000 A.C.) y el antiguo Egipto (3,000 A.C.), respectivamente (Marsit S. *et al*., 2017). A lo largo del tiempo, las cepas involucradas en las fermentaciones de estos tres productos se involucraron en diferentes trayectorias de domesticación que dieron lugar a diferentes historias evolutivas.

Las cepas de fermentación de vino forman un grupo monofilético, y las observaciones de diversos estudios apuntan a que las cepas de vino sufrieron un cuello de botella y/o altos niveles de endogamia antes de experimentar una expansión poblacional relacionada con el aumento de la fermentación de vino en diversas regiones del mundo (Gonçalves M. *et al*., 2016; Marsit S. & Delquin S., 2015; Marsit S. *et al*.,2017; Schacherer J. *et al*., 2009).

Por otro lado, las cepas de *S. cerevisiae* involucradas en la fermentación de cerveza y pan tienen un origen polifilético (Marsit S. *et al*., 2017). En el caso de las

cepas de cerveza, los diferentes orígenes de los distintos tipos de cerveza han influido en las historias evolutivas de cada uno de los subclados de cerveza (Gallone B. *et al*., 2016; Marsit S. *et al*., 2017). Por último, las cepas de pan son polifiléticas y autotetraploides (Bigey F. *et al*., 2020; Marsit S. *et al*., 2017; Randez-Gil F. *et al*., 2013). Algunas cepas de levadrura de pan están relacionadas con cepas de vino; otras son cercanas a las cepas de sake y las de vida libre asociadas a cortezas de robles, y otras son cepas únicas que no tienen parentesco con ningún otro clado. Esta evidencia sugiere que las cepas de pan han surgido varias veces en diversas partes del mundo (Randez-Gil F. *et al*., 2013; Marsit S. *et al*., 2017).

Las cepas domesticadas de *S. cerevisiae* involucradas en la fermentación de varios productos han seguido estas trayectorias distintas, debido en gran medida, a las diferentes condiciones de fermentación a las que han sido expuestas a lo largo del tiempo y a los diferentes caminos que esta especie siguió con la dispersión de los distintos grupos humanos a lo largo de la historia (Marsit S. *et al*., 2017).

La diversidad de *S. cerevisiae* de diferentes regiones del mundo ha sido estudiada desde una perspectiva de genómica de poblaciones

El uso extendido de *S. cerevisiae* en diferentes procesos industriales y biotecnológicos ha propiciado su diversificación en distintos subclados (Marsit S. *et al*., 2017; Peter J. & Schacherer J., 2016). Cada uno de los subclados industriales ha tenido trayectorias de domesticación diferentes, influenciadas por sus características particulares (Marsit S. *et al*., 2017). Para conocer mejor su evolución, se ha concluido que se necesita un enfoque poblacional para estudiar los

patrones evolutivos de sus distintos clados y así elucidar las fuerzas evolutivas que han moldeado a sus genomas, incluyendo a otras especies cercanas, ya se ha observado que el flujo génico entre distintas especies puede ser considerable (Marsit S. *et al*., 2017; Peter J. *et al*., 2018).

La diversidad genética de *S. cerevisiae* en procesos de fermentación de agave ha sido pobremente estudiada

En el caso de cepas de *S. cerevisiae* involucradas en fermentación de agave, existen pocos estudios donde se haya evaluado su diversidad genética. Flores-Berrios y colaboradores (2005) compararon a las *S. cerevisiae* involucradas en la fermentación de vino y de mezcal y determinaron que ambos grupos de cepas se diferencian genéticamente entre sí. Kirchmayr y colegas (2017) evaluaron la prevalencia de especies de levaduras en la fermentación de agave de dos palenques (fabricas tradicionales de mezcal) oaxaqueños durante dos años y observaron que *S. cerevisiae* era la especie predominante. En un contexto global, Peter y colaboradores (2018) incluyeron siete cepas clasificadas como *S. cerevisiae* de fermentación de agave de Tamaulipas, más dos sin clasificar en su análisis de 1,011 genomas de la especie. Sin embargo, no se muestreó otras regiones del país donde también se fermenta agave para la producción de mezcal. En consecuencia, probablemente, no abarcaron todo el abánico de diversidad genómica presente en este grupo de aislados de *S. cerevisiae*.

La fermentación de agave para la fabricación de mezcal se lleva a cabo después de un evento de cocción prolongada de las piñas de agave, y sin inoculación de cepas industriales (Kirchmayr M.R. *et al*., 2017). Por lo tanto, las

levaduras involucradas en este proceso provienen del entorno natural y su proceso de domesticación o, mejor dicho, de cuasi-domesticación, está influenciado por características del lugar donde ocurre la fermentación como la geografía, el clima, la altitud, el tipo de suelo, la vegetación, el agua y la fauna del lugar y, sobre todo, los insectos. El caso de los insectos como las abejas, algunas especies de avispas y *Drosophilas* es sumamente interesante porque han sido propuestos como los principales vectores que transportan a *S. cerevisiae* de un lugar a otro (Di Paola M. *et al*., 2020; Meriggi N. *et al*.,2020). En el caso de la fermentación de vino, se ha demostrado que la piel de uva o la uva intacta no tiene un alto número de células de *S. cerevisiae* (Di Paola M. *et al*., 2020). Por lo tanto, se cree que los insectos (piel y tracto intestinal), en algunos casos, son los medios que transportan a las levaduras de su "vida libre" hacia los tanques de fermentación, ya que el alto contenido de azúcares y compuestos volátiles presentes en los tanques atrae a los insectos (Di Paola M. *et al*., 2020; Meriggi N. *et al*.,2020). Además de vectores de transporte, los insectos han sido propuestos como reservorios de hibridación interespecie entre diferentes especies *Saccharomyces* (Di Paola M. *et al*., 2020). Los aislados del género *Saccharomyces* de vida libre no suele aparearse entre sí porque no suelen esporular y germinar (Meriggi N. *et al*.,2020). Estas condiciones son necesarias para que el apareamiento pueda ocurrir entre especies. Sin embargo, el tracto intestinal de los insectos es un ambiente ideal para la esporulación y germinación, lo cual promueve el apareamiento de las levaduras y, en consecuencia, la hibridación interespecie (Meriggi N. *et al*.,2020).

La fermentación de agave para producción de mezcal se lleva a cabo en varias y diversas regiones de México que abarcan los estados de México, Oaxaca,

Puebla, Guerrero, Michoacán, Sonora, Durango, Jalisco, San Luis Potosí, Guanajuato, Tamaulipas y Zacatecas. Al evaluar la diversidad genómica y la estructura poblacional de las levaduras del género *Saccharomyces* involucradas en la fermentación de agave, se podrá conocer más sobre la historia evolutiva de estos aislados y qué aspectos han coadyuvado en su diferenciación con respecto a otras poblaciones de levaduras del resto del mundo.

OBJETIVOS

Objetivo general

- Analizar genomas de levaduras aisladas de fermentaciones de agave de varios lugares de México para evaluar y analizar su estructura poblacional e historia evolutiva.

Objetivos particulares

- Secuenciar genomas de aislados identificados como *Saccharomyces* de fermentaciones de agave de diferentes lugares de México.
- Clasificar las muestras en especies *(S. cerevisiae, S. paradoxus,* híbrido u otras).
- Limpiar los datos de secuenciación, mapear a un archivo de genomas concatenados de especies *Saccharomyces.*
- Crear una matriz de SNPs, llamando y filtrando variantes de baja calidad.
- Identificar características genómicas a gran escala de nuestras cepas de mezcal (ploidía, cigosidad, niveles de heterocigosidad, entre otros).
- Realizar estudios de inferencia filogenética y estructura poblacional (ADMIXTURE y PCA).

ANTECEDENTES

La fermentación de agave para la fabricación de mezcal es un proceso que se diferencia al de otros destilados de agave como el tequila porque, hasta la fecha, se lleva a cabo de forma artesanal en palenques (fábricas de mezcal) y sin involucrar pasos automatizados ni inoculación de levaduras industriales (**Figura 2**) (Kirchmayr M.R. *et al*., 2017). Al no inocular, la fermentación del mosto y jugo de agave ocurre gracias a los distintos microorganismos localizados en los tanques y/o en el ambiente donde se lleva a cabo este proceso (Kirchmayr M.R. *et al*., 2017).

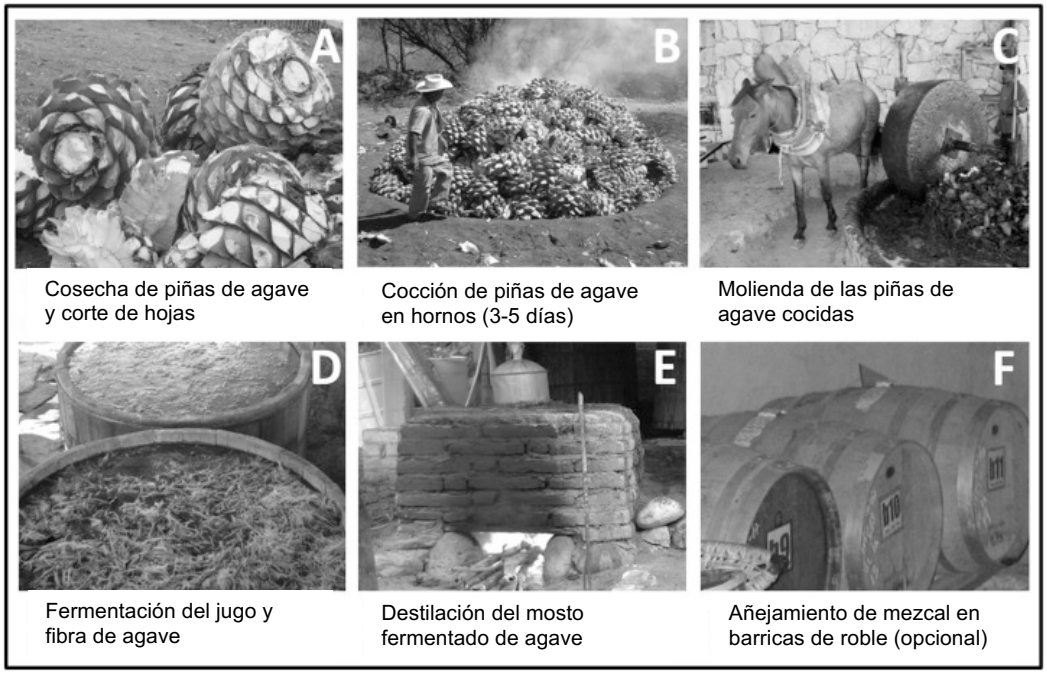


Figura 2. **Diagrama de pasos de fermentación de agave para mezcal.** La imagen muestra los diferentes pasos de la fermentación de agave para la fermentación de agave. Primero, se da la recolección de las piñas maduras de agave (A), se cuecen las piñas por 3-5 días (B), se muele las piñas dando lugar al mosto y jugo de agave (C), se fermentan el mosto y jugo de agave cocido (D), se destila el mosto de agave fermentado (E) y se añeja en barricas de roble (F). La fermentación "espontánea" de agave es de sumo interés porque no se inocula el tanque con ninguna cepa de levadura industrial. Algunos productores "inoculan" con pulque o pedazos de corteza de árboles. Imagen tomada de Kirchmayr M.R. *et al*., 2017.

Se han realizado pocos estudios para evaluar la diversidad de levaduras en los tanques de fermentación de agave. Kirchmayr y colaboradores (2017) estimaron la prevalencia de diferentes especies de levadura mediante la examinación de

propiedades morfofisiológicas y la secuenciación de la subunidad 26 del ADN ribosomal en dos palenques en Oaxaca durante dos años. En ambas fábricas, se identificó que *S. cerevisiae* era la especie más prevalente durante los dos años. Por otro lado, Flores-Berrio y colaboradores (2005) y colaboradores determinaron, a través de análisis de caracterización molecular por polimorfismos en la longitud de fragmentos amplificados (AFLP por sus siglas en inglés), que las *S. cerevisiae* involucrada en la fermentación de agave son muy diferentes a las *S. cerevisiae* que fermentan vino.

En un contexto global, Peter y colegas llevaron a cabo un estudio ambicioso donde secuenciaron 1,011 genomas de aislados de *S. cerevisiae* de lugares y orígenes distintos (Peter J. *et al*., 2018). Este estudio incluyó cepas pertenecientes a grupos naturales (suelo, robles, ríos, etc) y antropogénicos (fermentaciones y procesos industriales). Entre sus observaciones, Peter y colaboradores (2018) identificaron que la estructura filogenética de los 1,011 aislados agrupa a las cepas naturales y antropogénicas en dos particiones diferentes (**Figura 3**). De igual manera, el linaje más divergente (rama más larga en **Figura 3**) proviene de Taiwán, lo que refuerza la hipótesis de un sólo evento de origen de *S. cerevisiae* en Asia del Este (Duan S.F. *et al*., 2018; Peter J. *et al*., 2018). Adicionalmente, hicieron análisis estructurales del genoma donde observaron que la mayoría de las cepas son diploides; sin embargo, las cepas industriales son más proclives a aumentar su ploidía y generar aneuploidías (Peter J. *et al*., 2018). Esto sucede porque son estrategias útiles para la adaptación a ambientes estresantes y cambiantes en intervalos de tiempo relativamente cortos (Peter J. *et al*., 2018). Por último, se identificaron distintas trayectorias de domesticación con historias evolutivas

diferentes. Las cepas de vino son monofiléticas y tienen una diversidad genética promedio relativamente baja (π= 1 x 10^{-3}), mientras que las de cerveza son polifiléticas y tiene un nivel de diversidad promedio más alta (π= 2.8 x 10^{-3}) (Peter J. *et al*., 2018; Peter J. & Schacherer J., 2016). Finalmente, se reportaron cuatro grupos con un alto contenido de introgresiones provenientes de *S. paradoxus*: Alpechin, Mexican Agave, Brazilian Bioethanol y French Guiana Human (Peter J. *et al*., 2018)**.**

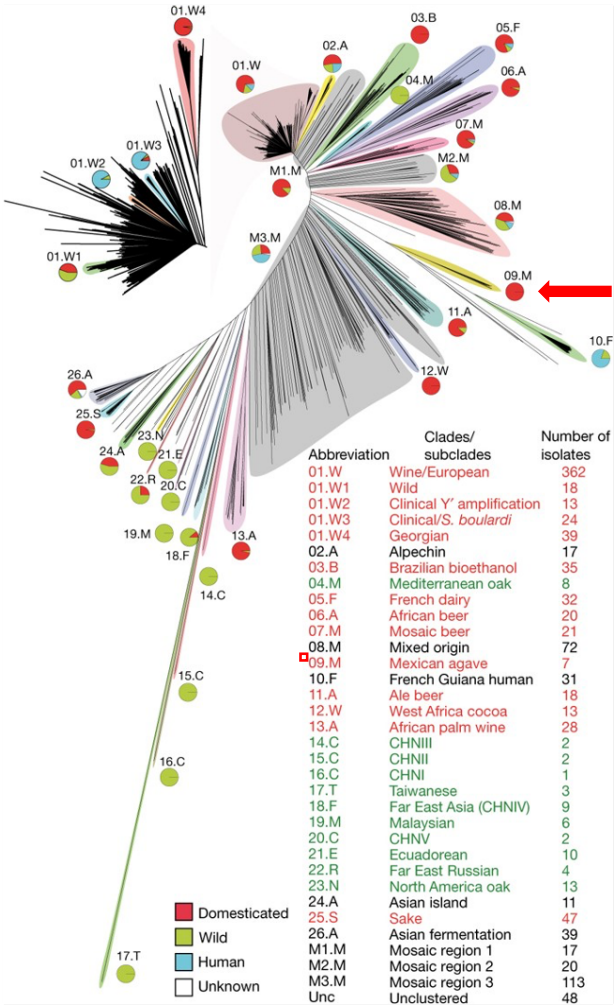


Figura 3. **Árbol *Neighbor-Joining* de los 1,011 aislados de *S. cerevisiae* evaluados.** La imagen muestra el árbol filogenético de las 1,011 cepas de *S. cerevisiae* evaluadas por Peter y colaboradores (2018). En la parte superior, se agrupan las cepas domesticadas (círculos rojos) y en la inferior, las cepas naturales (círculos verdes). Al lado izquierdo del árbol, se observa un *zoom* del grupo monofilético de vino (1.W). La rama más larga (17.T) representa al linaje Taiwanés (*Taiwanese*) de *S. cerevisiae*. Esta rama probablemente representa al linaje más antiguo de *S. cerevisiae*, lo que refuerza la hipótesis de un origen de *S. cerevisiae* en el Lejano Oriente. Este árbol incluye siete cepas de fermentación de agave de Tamaulipas, las cuales fueron agrupadas en el grupo Agave Mexicano (flecha y rectángulo rojos) (*Mexican Agave*). Imagen tomada y modificada de Peter J. *et al*., 2018.

El artículo de los 1,011 genomas es el único que ha reportado y analizado el genoma completo de cepas de fermentación de agave, comparándolo con aislados de diferentes partes del mundo. Peter y colaboradores (2018) incluyeron siete cepas de fermentación de agave (flecha roja en **Figura 3**) de Tamaulipas y las clasificaron como ***Mexican Agave.*** Además, incluyeron otras dos cepas de agave de México (una de Jalisco y otra de lugar desconocido) que fueron clasificadas como ***Unclustered***. Es importante precisar que varios estados de la República Mexicana fermentan agave para la producción de mezcal (Estado de México, Oaxaca, Puebla, Guerrero, Michoacán, Sonora, Durango, Jalisco, San Luis Potosí, Guanajuato, Tamaulipas y Zacatecas). Tambien hay que considerar que la variedad de especies de agave utilizadas para la fermentación es muy grande y cambia dependiendo de la región y el estado (López-Romero J.C. *et al*., 2018), y que México es un país megadiverso donde coexisten diferentes tipos de climas, ecosistemas y condiciones geográficas (Sarukhan, J., 2008) que podrían influir sobre la diversidad genética de las levaduras involucradas en fermentación de mezcal.

Tomando en cuenta estas ideas, creemos que la diversidad de cepas de *S. cerevisiae* de México reportada por Peter y colegas (2018) está subrepresentada, por lo que nos dimos a la tarea de hacer un muestreo de cepas de fermentación de agave de diferentes partes de la República Mexicana y secuenciar sus genomas. Este estudio es uno de los primeros en hacer una evaluación grande y representativa de la estructura filogenética y poblacional de *S. cerevisiae* involucradas en la fermentación de agave. Estudiar estos aislados nos permitirá

conocer los orígenes de este grupo de *S. cerevisiae*, su relación con otros clados y su trayectoria de domesticación.

METODOLOGÍA

Recolección, enriquecimiento y aislamiento de cepas *Saccharomyces* de fermentación de agave

Se recolectaron muestras de tanques de fermentación de agave en palenques localizados en 12 estados y distribuidos en siete regiones (73 palenques) (**Tabla 1-Anexo**) en las siguientes regiones del país (Aguirre X., *et al*., 2006):

- Noroeste Lejano: Sonora
- Noroeste: Durango
- Oeste: Jalisco
- Valles Centrales: San Luis Potosí, Guanajuato y Zacatecas
- Noreste Tamaulipas
- Cuenca del Río Balsas: Estado de México, Michoacán y Guerrero
- Centro-Sur: Puebla y Oaxaca

Posteriormente, se enriquecieron las muestras con ayuda del protocolo de Liti y colegas (2017) para obtener preferencialmente cepas del género *Saccharomyces* (trabajo realizado por Porfirio Gallegos).

Se realizó un análisis utilizando un protocolo de MALDI-TOF (*Matrix Assisted Laser Desorption Ionization Time-of-Flight*) (De la Torre-González F.J. *et al*., 2018) con el objetivo de identificar cepas de *S. cerevisiae* (trabajo realizado por Porfirio Gallegos). Es importante recalcar que este método no es capaz de diferenciar entre *S. cerevisiae* y *S. paradoxus*, por lo que se deben utilizar otros métodos para poder distinguir entre especies.

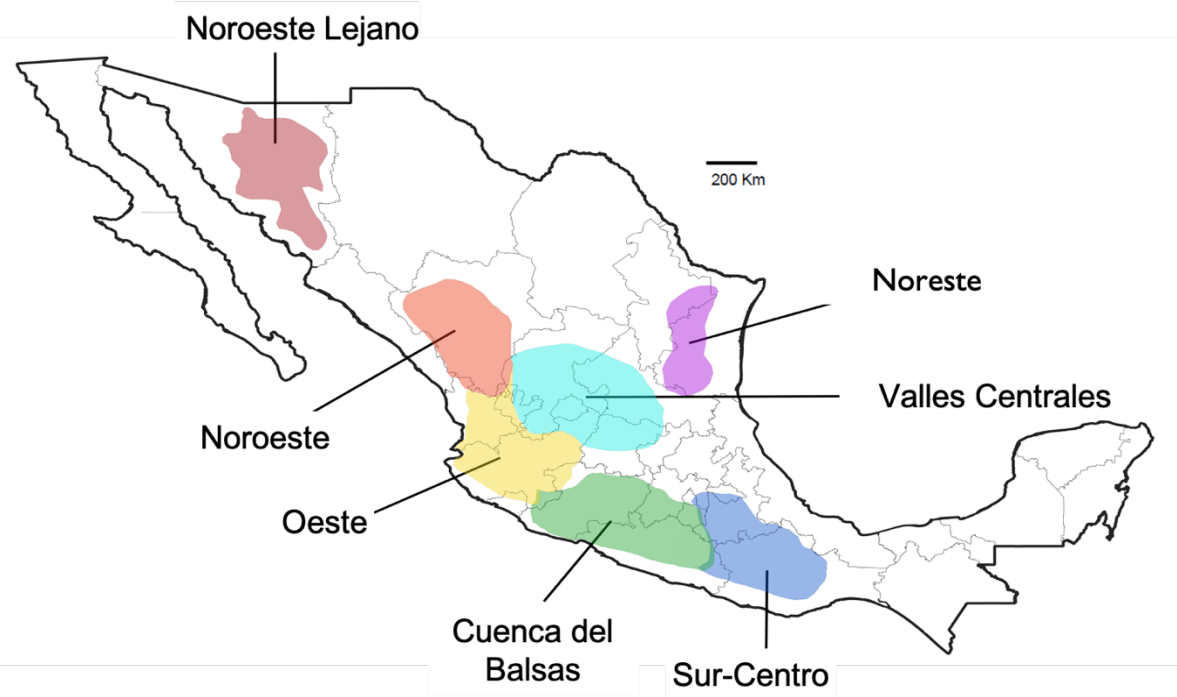


Figura 4. **Mapa de regiones de cosecha y fermentación de agave en México.** El mapa muestra las regiones de México donde se cultiva, cosecha y fermenta agave en México. La región Noroeste Lejano está compuesta por el estado de Sonora; la región Noroeste está compuesta por el estado de Durango; la región Oeste está compuesta por el estado de Jalisco; la región Valles Centrales está compuesta por los estados de San Luis Potosí, Guanajuato y Zacatecas; la región de Cuenca del Balsas está compuesta por el Estado de México, Michoacán y Guerrero; la región Sur-Centro está compuesta por los estados de Oaxaca y Puebla.

Extracción de ADN y secuenciación de genomas completos de las cepas *Saccharomyces.*

Se extrajo ADN de los aislados (**Tabla 1- Anexo**) utilizando el kit MasterPure™ Yeast DNA Purification con ayuda de las técnicas Alejandra Castillo y Carina Díaz del Laboratorio Internacional de Investigación sobre el Genoma Humano-UNAM. Los genomas de las cepas fueron secuenciados en dos tandas con las siguientes tecnologías: Illumina TruSeq PCR-Free Library Prep (plataformas MiSeq y NextSeq) y secuenciación de lecturas cortas DNBSeq™ (*DNA nanoballs and PCR-free Rolling Circle Replication*) del *Beijing Genomics Institute* (BGI). La tecnología de secuenciación del BGI es parecida a las plataformas de Illumina™ pero con un costo

30% menor (https://www.bgi.com/us/dnbseq-ngs-technology/). Adicionalmente, no existen sesgos inducidos por PCR, ya que esta plataforma no requiere de un paso de PCR. En su lugar, se lleva a cabo una *Rolling-Circle Replication* (RCR) donde se hacen copias del templado original de ADN, en lugar de hacer copias de la copia (https://www.bgi.com/us/dnbseq-ngs-technology/). Por último, es importante mencionar que se tuvo un problema con los tamaños de los insertos al momento de hacer las librerías y, en consecuencia, se tuvo un inserto muy chico; esto fue un error de BGI y se les comunicó en su momento. En total, se secuenciaron 19 cepas con las plataformas de Illumina (**Tabla 1-Anexo**) y 118 cepas con la tecnología de BGI (**Tabla 1-Anexo**). Adicionalmente, se extrajo DNA de unas cepas de *Kluyveromyces marxianus* (10% de las cepas totales) para mandar a secuenciar sus genomas a BGI y corroborar que no se había cometido algún error que resultara en contaminación cruzada.

Disponibilidad de código y datos

Se creó un repositorio en la cuenta de Gitlab del servidor del Laboratorio Internacional de Investigación sobre el Genoma Humano (LIIGH). En este repositorio, se pueden consultar los *scripts*, un resumen del *pipeline* (**Figura 5**) y datos necesarios para llevar a cabo los análisis descritos en la metodología de esta tesis. El enlace para acceder al repositorio es:

- http://app.liigh.unam.mx/lmorales/Master_Thesis_JAUA/tree/master

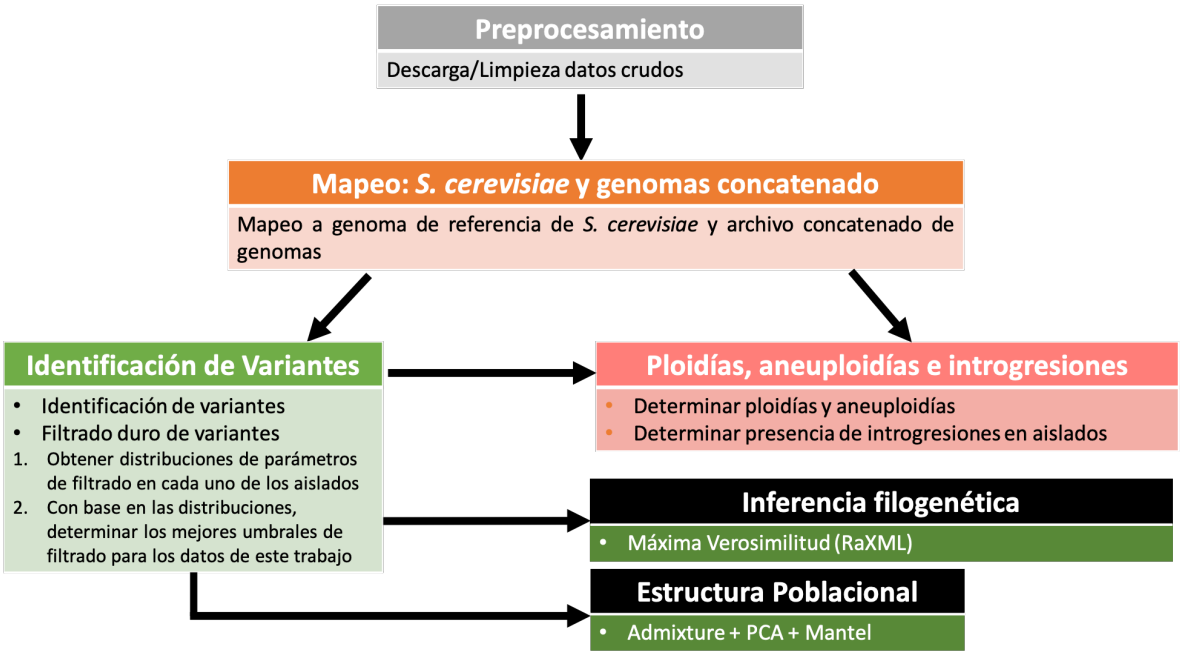


Figura 5. **Resumen del *pipeline* seguido para la realización del proyecto.** Este diagrama muestra, de forma resumida, los diferentes pasos que se siguieron para llevar a cabo la realización del proyecto.

Descarga, procesamiento, filtrado y mapeo de las lecturas de secuenciación de las cepas de fermentación de agave

Las lecturas de secuenciación crudas BGI de muestras de fermentación de agave **(Tabla 1-Anexo)** fueron descargadas de *Amazon Web Services* mientras que las lecturas de Illumina crudas se obtuvieron del servicio de secuenciación del Instituto Nacional de Medicina Genómica (INMEGEN). Por otro lado, los archivos crudos de las lecturas de 170 cepas de *S. cerevisiae* del trabajo de Peter y colegas (2018) (representantes de cada uno de los clados descritos en este artículo) fueron descargados del *Sequence Read Archive* (SRA). Posteriormente, se utilizaron *scripts* de bash "ID.sge" (Gitlab

bin/Cleaning_Sequencing_Reads_and_Mapping/Cleaning) que ejecuta el programa fastp v0.20.0 (Chen S. *et al*., 2018) para llevar a cabo el filtrado y la exlcusión de las lecturas de baja calidad de cada uno de los aislados. Los estadísticos de bases filtradas se encuentran descritos a mayor detalle en las **Tablas 1 y 2** del anexo. Los histogramas de frecuencia del número de bases secuenciadas posterior al filtrado fueron realizados con los *scripts* de R Markdown "Histograms_Table1.R e Histograms_Table2.R" (R Core Team, 2020) (Gitlab bin/BasesFiltered_and_Coverage/Mapping) y la librería ggplot2 (Wickham H., 2016). La longitud promedio de las lecturas de Illumina fue de 150 a 299 pb, mientras que la longitud promedio de las lecturas de BGI posterior al filtrado fue de 149 pb para todos los aislados de agave.

Las lecturas limpias fueron mapeadas con bwa-mem v0.7.8 (Li H., 2013) a un archivo que contiene los genomas de referencia de las siguientes especies del complejo *Saccharomyces*: *S. cerevisiae* S288C, *S. paradoxus* YPS138, *S. mikatae* IFO1815, *S. jurei* M1, *S. arborícola* H6, *S. uvarum* CBS7001, *S. eubayanus* FM1318 y *S. kudriavzevii*. Se utilizaron los *scripts* ID_vs_CONC.sge (Gitlab bin/Cleaning_Sequencing_Reads_and_Mapping/Cleaning) para mapear las lecturas. Las referencias de estos genomas pueden consultarse en el README del directorio data/ del sitio de Gitlab (Gitlab data/References) de este proyecto. Adicionalmente, se incluyeron los genomas de las especies *Kluyveromyces marxianus* DMKU31042 y *Pichia kudriavzevii* CBS573, con la finalidad de observar si algunas lecturas mapeaban con alta calidad a algunas de estos genomas, lo cual indicaría que posiblemente existieron eventos de contaminación cruzada. Se

marcaron las lecturas duplicadas con Picard v2.6.0 (http://broadinstitute.github.io/picard/) para obtener los archivos .bam que posteriormente serían utilizados para la identificación de variantes y la genotipificación de los aislados de fermentación de agave.

Análisis de cobertura de las muestras de agave y de Peter y colaboradores (2018)

Se estimó la cobertura en cada genoma del archivo concatenado para todas las muestras. Después, se estimó la mediana de la cobertura del genoma con mayor cobertura y ésta se utilizó para normalizar la cobertura. Posteriormente, se graficaron las coberturas normalizadas por cromosoma de cada unas de las cepas de agave y de Peter y colaboradores con la finalidad de observar cambios en la ploidía y aneuploidías. De igual manera, se cuantificaron los porcentajes de lecturas mapeadas con calidad de mapeo MAPQ > 20 a los diferentes genomas de referencia del archivo concatenado con el objetivo de observar si algunas de las cepas eran híbridos interespecie, principalmente entre *S. cerevisiae y S. paradoxus* o sí existían introgresiones de *S. paradoxus* en las *S. cerevisiae* de fermentación de agave.

Se graficaron los porcentajes lecturas mapeadas a *S. paradoxus* en gráficos de violín con ayuda de un *script* de R "Pct_Reads_to_SAPA_MexicanAgave_Alpechin_FrenchGuiana_BrazilianBioethanol.Rmd" (R Core Team, 2020) (Gitlab bin/Introgressions_from_SAPA) y la librería ggplot2 (Wickham H., 2016).

Identificación de variantes, filtrado y genotipificación de las cepas de fermentación de agave y de Peter y colaboradores (2018)

Se llevó a cabo la identificación de variantes de todas las cepas (**Tabla 1-2 de Anexo)** con el *script* de shell "VCalling_CONC_ID.sge" (Gitlab bin/Variant_Calling_Genotyping/VariantCalling_and_Genotyping) que ejecutan la herramienta HaplotypeCaller de GATKv4.1.1.0 (van der Auwera G.A. *et al*., 2013; Poplin R. *et al*., 2017). Luego, se extrajeron los sitios localizados en el genoma nuclear (cromosomas 1-16) de *S. cerevisiae* de cada una de las cepas de fermentación de agave (**Tabla 1-Anexo**) y de Peter y colegas (**Tabla 2-Anexo**) con los *scripts* "Obtain_matrix_SACE_from_Mexico_Mezcal.sge" y "Obtain_matrix_SACE_from_allSACE.sge" (carpeta Gitlab bin/Variant_Calling_Genotyping/SNPs_Matrices) en dos archivos g.vcf multi-muestras:

- Archivo multi-muestra g.vcf con 146 aislados de fermentación de agave: 137 muestras de fermentación de agave + 9 cepas mexicanas de agave reportadas por Peter y colaboradores (Peter J. *et al*., 2018).
- Archivo multi-muestra g.vcf con 307 cepas: 137 cepas de fermentación de agave + 170 cepas de distintos lugares y orígenes de Peter y colaboradores (2018).

Después de crear ambos archivos multi-muestra g.vcf, se realizó la genotipificación de éstos con los *scripts* "Obtain_ matrix_ SACE_ from_ Mexico_ Mezcal.sge" y "Obtain_matrix_SACE_from_allSACE.sge" (Gitlab

bin/Variant_Calling_Genotyping/SNPs_Matrices) que ejecuta la herramienta GenotypeGVCFs de GATK (van der Auwera G.A. *et al*., 2013; Poplin R. *et al*., 2017). Se seleccionaron exclusivamente los SNPs usando SelectVariants de GATK en los *scripts* "Obtain_matrix_SACE_from_Mexico_Mezcal.sge y Obtain_matrix_SACE_from_allSACE.sge" (Gitlab bin/Variant_Calling_Genotyping/SNPs_Matrices) (van der Auwera G.A. *et al*., 2013; Poplin R. *et al*., 2017).

El filtrado de calidad de los SNPs se realizó con una estrategia de filtrado-duro (*hard-filtering*), con las recomendaciónes de umbrales que sugiere GATK (**Tabla 1**). La idea principal del filtrado duro es establecer unos umbrales de filtrado para diferentes parámetros y quitar los SNPs que tengan uno o más valores fuera de estos umbrales. Sin embargo, el portal de GATK (https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants) recomienda graficar los valores de los parámetros de filtrado para determinar si los valores de los parámetros necesitan afinarse y para cerciorarse de no estar quitando sitios variantes que sean reales. Para corroborar que los umbrales recomedados por GATK fuerán óptimos para nuestros datos, se graficaron cada uno de los parámetros para todas las muestras con los *scripts* de R localizados en la carpeta Gitlab bin/Variant_Calling_Genotyping/Filtering. Después de identificar los mejores valores de los parámetros, se realizó el filtrado de variantes con el *script* de shell "Obtain_matrix_SACE_from_Mexico_Mezcal.sge" y "Obtain_matrix_SACE_from_allSACE.sge" (Gitlab bin/Variant_Calling_Genotyping/SNPs_Matrices) que ejecuta la herramienta SelectVariants de GATK (van der Auwera G.A. *et al*., 2013; Poplin R. *et al*., 2017)

en ambos archivos g.vcf multi-muestra. Se realizó el filtrado de datos faltantes (*missing data*) de los SNPs con el *script* de shell "Obtain_matrix_SACE_from_Mexico_Mezcal.sge" y "Obtain_matrix_SACE_from_allSACE.sge" (Gitlab bin/Variant_Calling_Genotyping/SNPs_Matrices) que ejecuta la herramienta vcftools v0.1.14 (Danecek P. *et al*., 2011). Se quitaron SNPs que no estuvieran presente en más del 25%, 20%, 15%, 10%, 5% y 0% de los aislados en cada uno de los archivos g.vcf multi-muestra. Por último, se filtró para quedarnos exclusivamente con SNPs bialélicos con ayuda de bcftools de Samtools v1.10 (Li H. *et al*., 2009).

Se llevó a cabo un proceso de identificación de variantes, filtrado y genotipificación utilizando los archivos ID_SACE.g.vcf (archivos obtenidos del mapeo exclusivo al genoma de *S. cerevisiae*) para construir una filogenia *Neighbor-Joining,* utilizando una matriz de distancia estimada con el método de Weir & Goudet (2017), con las 1,011 cepas de *S. cerevisiae* estudiadas por Peter y colaboradores (2018). Se utilizó el *script* merge_agave_SACE_mapping.sge, que ejecuta comandos de GATKv4.1.1.0 (van der Auwera G.A. *et al*., 2013; Poplin R. *et al*., 2017), vcftools v0.1.14 (Danecek P. *et al*., 2011) y bcftools v1.9 (Li H., 2011). Después de juntar los archivos ID_SACE.g.vcf de las 146 cepas de fermentación de agave, se utilizó la herramienta vcf-merge de vcftools v0.1.14 para unir el archivo g.vcf multi-muestra de las cepas de fermentación de agave con el archivo g.vcf de las variantes identificadas por Peter y colaboradores en las 1,011 cepas de su estudio: Merge_Mezcal_SACEmap_1011.vcf.gz. Posteriormente, se realizó el filtrado duro

de variantes (**Tabla 1**) y el filtrado por datos faltantes tal y como se hizo para los archivos mencionados anteriormente.

Tabla 1. Parámetros de Filtrado de Variantes

Filtro	Significado de Filtro	Definición del filtro	Valor de filtro recomendado por GATK
QD	Quality by Depth	Confianza de la variante dividida por la cantidad de lecturas no homocigotos a la referencia	QD < 2.0
FS	Fisher Strand	Probablidad Phred de sesgo de hebra en variante	FS > 60.0
SOR	Strands Odd Ratio	Otro filtro por sesgo de hebra	SOR > 3.0
MQ	RMSMappingQuality	Raíz cuadrada del promedio de calidad de mapeo de todas las lecturas en ese sitio	MQ > 40
MQRankSum	MappingQualityRankSumTest (sólo para sitios heterocigotos)	Compara calidad de las lecturas dando soporte al alelo de referencia y alternativo	MQRankSum < -12.5
ReadPosRankSum	ReadPosRankSumTest	Evaluar variantes para identificar si hay sesgo de localización de la variante al final de la lectura.	ReadPosRankSum < -8.0

Análisis de heterocigosidad de los aislados de fermentación de agave

Se utilizó el *script* "get_snps_counts_from_vcf-stats_SACE_mezcal.sh" (Gitlab Carpeta bin/Heterozygosity_of_Strains), que ejecuta la herramienta vcf-stats de vcftools v0.1.14 (Danecek P. *et al*., 2011) para obtener diversos estadísticos de heterocigosidad (número de SNPs heterocigotos vs homocigotos y proporción de SNPs heterocigotos en cada aislado) de las 137 muestras de fermentación de agave y de las 170 cepas seleccionadas del artículo de Peter y colegas (2018) que tenían un máximo de 10% de datos faltantes. El mismo *script* "get_snps_counts_from_vcf-

stats_SACE_mezcal.sh" fue utilizado para obtener los datos necesarios de heterocigosidad del *output* de vcf-stats. Posteriormente, se utilizaron los *scripts* de R "SNPs_counts_Mezcal_SACE.R", "SNPs_count_1011.R", "Histograms_Table1.Rmd" y "Histograms_Table2.Rmd" (Gitlab bin/Heterozygosity_of_Strains) (R Core Team, 2020) con la librería ggplot2 (Wickham H., 2016) para hacer histogramas y gráficos de dispersión de las proporciones de heterocigosidad de las muestras.

Análisis filogenéticos de las cepas de fermentación de agave y de las cepas de fermentación de agave con las cepas analizadas por Peter y colaboradores (2018)

Se llevaron a cabo análisis de máxima verosimilitud de las cepas de fermentación de agave y de las cepas de fermentación de agave con las cepas de Peter y colaboradores (2018). Se utilizaron los *scripts* "RaXML_ Phylogenetic_ Trees_ Only_ Mezcal_ SACE.sge" y "RaXML_Phylogenetic_Trees_allSACE_1000bs.sge" (Gitlab bin/Phylogenetic_Analysis) que ejecutan el programa RaXML v8.2.12 (Stamakatis A., 2014). En el caso del análisis que sólo incluye a las 137 cepas de fermentación de agave y las nueve cepas mexicanas reportadas por Peter y colaboradores (2018), se utilizó una matriz con 238,293 SNPs bialélicos del genoma de *S. cerevisiae* que sólo permite un máximo de 10% de datos faltantes (*missing data*). Este análisis se llevó a cabo con 100 rondas de *bootstrap*. El modelo de sustitución para esta inferencia filogenética fue elegido con ayuda del *script* Substitution_model_MexicanAgave_RaXML.sge (Gitlab bin/Phylogenetic_Analysis)

y el programa JModelTestv2.1.10 (Darriba D. *et al*., 2018; Guindon G. & Gascuel O., 2003). Se observó que el mejor modelo de sustitución soportado por RaXML era el GTRGAMMA.

Por otro lado, un análisis que incluye las 137 cepas de fermentación de agave y 170 cepas analizadas por Peter y colaboradores (2018) se realizó con una matriz de 970,686 SNPs bialélicos y un máximo de 10% de datos faltantes (*missing data*). Este análisis se llevó a cabo con 1000 rondas de *bootstrap,* ya que el número de cepas en este caso era mayor. Con base en el análisis anterior, el modelo de sustitución seleccionado para esta inferencia filogenética fue también GTRGAMMA. Los diseños y visualizaciones de los árboles filogenéticos (trabajo realizado por Iván Sedeño) se llevaron a cabo con los *scripts* de R "treeAll.R", "treeMex.R" y "treesColors.R" (Gitlab bin/Phylogenetic_Analysis), que ejecuta el paquete de Bioconductor ggtree v2.3.4.993 (Yu G. *et al*., 2017).

Adicionalmente, se realizó un análisis filogenético con el algoritmo *Neighbor-Joining* a partir de una matriz de distancias genéticas construida con el método de Weir & Goudet (2017) y utilizando los mapeos al genoma de *S. cerevisiae*, en el que se incluyeron a las 1,011 cepas de *S. cerevisiae* del artículo de Peter y colegas (2018) y las 141 cepas de fermentación de agave. El número de SNPs bialélicos empleados para la construcción de este árbol fue 1,353,760, y se permitió un máximo de 15% de datos faltantes en cada sitio (Merge_Mezcal_SACEmap_1011_bcftools_onlySNPs_noMiseq_firstfilterlow_withoutfltSNPs_nomissing15_biallelic.vcf). Es importante mencionar que se removieron 5 cepas secuenciadas con Illumina MiSeq (**Tabla 2-Anexo),** por la baja calidad de la secuenciación. Se utilizó el programa NJ_tree_Agave_1011.R que ejecuta la librería

ape (Paradis E. & Schliep K., 2019) para calcular la matriz de disimilitudes a partir del archivo Merge_Mezcal_SACEmap_1011_bcftools_onlySNPs_noMiseq_firstfilterlow_withoutfltSNPs_nomissing15_biallelic.vcf.vcf. Posteriormente, se ejecuta el comando bionj de SNPRelate (Zheng X. *et al*., 2012) para construir el árbol *Neighbor-Joining* a partir de la matriz recién mencionada.

Remoción de cepas con alto contenido de datos faltantes y poco porcentaje de lecturas mapeadas al genoma de *S. cerevisiae*

En los siguientes análisis se decidió remover un total de 13 cepas, de las cuales 12 son híbridos interespecie *S. cerevisiae* y *S. paradoxus*. La otra cepa tenía solamente 84% de lecturas mapeadas al genoma de *S. cerevisiae* S288C. Se utilizaron dos *scripts* de Rmd "Missingness_Report_Mezcal_SACE_Samples.Rmd" y "Missingness_Report_All_SACEs.Rmd" (Gitlab bin/Removal_13_strains_high_pct_missingdata) que ejecutan la librería ggplot2 (Wickham H., 2016) para graficar el porcentaje de datos faltantes (*missing data*) por cepa. Después de evaluar las gráficas generadas por estos *scripts*, se decidió remover 13 cepas (descritas en **Tablas 3** y **4** del Material Suplementario) con los scripts "Remove_13_strains_Mezcal_lot_missing_data.sge" y "Remove_13_allSACE_strains_lot_missing_data.sge" (Gitlab bin/Removal_13_strains_high_pct_missingdata)**.**

Análisis de componentes principales de las cepas de fermentación de agave y las cepas de fermentación de agave con las cepas analizadas por Peter y colaboradores (2018)

Se llevaron a cabo dos análisis de componentes principales: uno con exclusivamente 124 cepas de fermentación de agave más nueve cepas de agave reportadas en el artículo de Peter y colaboradores (2018) y otras con 124 aislados de agave más los 170 aislados de *S. cerevisiae* del artículo de Peter y colegas (2018).

El análisis con las 124 cepas de fermentación de agave se llevó a cabo con el *script* de R "PCA_Mezcal_color_Mezcal_regions.R" (Carpeta bin/PCA) que ejecuta las librerías ade4 (Bougeard S. & Dray S., 2018) y factoextra (Kassambara A. & Mundt F., 2020). Como *input,* se utilizó una matriz de 223,013 SNPs bialélicos sin datos faltantes. La librería factoextra requiere un archivo fasta como *input,* por lo que se convirtió el archivo .vcf de la matriz a fasta con el script de Python "Convert_Agave_strains_and_allSACEs_nomissing_phylip_to_fasta.py" (Gitlab bin/PCA) que ejecuta la librería SeqIO de Biopython (Chapman B.A & Chang J.T., 2000; Cock P.A. *et al*., 2009). El script de R "PCA_Mezcal_color_Mezcal_regions.R" (Gitlab bin/PCA) se encargó de graficar el componente principal 1 vs componente principal 2, componente principal 1 vs componente principal 3 y componente principal 2 vs componente principal 3. Además, el *script* se encargó de colorear las cepas de acuerdo con la región donde fueron muestreadas.

Por otro lado, el análisis con las 124 cepas de fermentación más las 170 cepas analizadas por Peter y colegas (2018) se llevó a cabo de forma similar al anterior con una matriz de 831,175 SNPs bialélicos sin datos faltantes. Las cepas de fermentación de agave están representadas por triángulos rojos mientras que las cepas evaluadas en el trabajo de Peter y colaboradores (2018) están representadas por círculos negros.

Prueba de Mantel de cepas de fermentación de agave

Con las distancias geográficas de los puntos de colecta entre 103 cepas de fermentación de agave, se realizó una prueba de Mantel. La prueba de Mantel estima la correlación entre una matriz de distancias genética y una matriz de distancias geográficas (Diniz Filho J.A. *et al*., 2013). Sin embargo, se decidió remover seis cepas del análisis, porque son aislados que no pertenecen al clado monofilético de las cepas de fermentación de agave y su presencia podía meter ruido al análisis. Esta prueba se realizó con el *script* de R “Mantel_test.R” (Gitlab bin/Mantel_Test) que ejecuta la librería ape (Paradis E. & Schliep K., 2019) para estimar las distancias genéticas de las cepas y la librería vegan (Oksanen J. *et al*., 2019) para llevar a cabo la prueba de Mantel.

Análisis de estructura poblacional

Se realizó un análisis de estructura poblacional con el programa ADMIXTURE (Alexander D.H. *et al*., 2009). Al igual que en análisis anteriores, se llevó a cabo una corrida de ADMIXTURE que sólo incluyera a las cepas de fermentación de agave y

otro que incluyera, además, a las cepas de distintas partes del mundo artículo de Peter y colegas (2018). En el caso del análisis con sólo las cepas de agave, se utilizó la matriz de 119,317 SNPs bialélicos que sí incluye a los 13 aislados que se habían removido para el PCA. Esta matriz incluye 146 cepas de fermentación de agave (137 secuenciadas por nosotros + 9 de Peter y colaboradores). El análisis se llevó a cabo con el *script* "ADMIXTURE_Mezcal.sge" (Gitlab bin/ADMIXTURE) que ejecuta el programa ADMIXTURE para *K*=2-17. Adicionalmente, se activó el *flag* –cv para correr el *cross-validation error algorithm* que sirve para estimar los mejores valores de *K* que expliquen los datos. El gráfico de *K*=12 se realizó con el *script* de R "ADMIXTURE_Mezcal_Plots.R" y el paquete pophelper v2.3.1 (Francis, R.M., 2017) (Gitlab bin/ADMIXTURE). Las imágenes de los gráficos de barras para todas las *K* y las imágenes del error de validación cruzada pueden verse en el **Material Suplementario.**

El análisis de ADMIXTURE con las cepas de fermentación de agave y las cepas del resto del mundo fue llevado a cabo con la matriz de 831,175 SNPs bialélicos que no incluyen a las 13 cepas removidas para el PCA. En total, esta matriz tiene 294 aislados: 124 cepas de fermentación de agave secuenciadas por nosotros y 170 de Peter y colaboradores (2018). El análisis se llevó a cabo con el *script* "ADMIXTURE_allSACE.sge" (Gitlab bin/ADMIXTURE) que ejecuta el programa ADMIXTURE para *K*=2-17. Adicionalmente, se activó el *flag* –cv para correr el *cross-validation error algorithm,* que sirve para evaluar los valores de *K* que mejor expliquen los datos. El gráfico de *K*=12 se realizó con el *script* de R "ADMIXTURE_allSACE_Plots.R" y el paquete pophelper v2.3.1 (Francis, R.M., 2017) (Gitlab bin/ADMIXTURE). Las imágenes de los gráficos de barras para todas

las *K* y las imágenes del error de validación cruzada pueden verse en el **Material Suplementario.**

Análisis bioinformático de ploidías y aneuploidías de las cepas de fermentación de agave

Se utilizó el método reportado de Fay y colegas (2019) para estimar bioinformáticamente la ploidía de las cepas de fermentación de agave. Este método consiste en obtener el número de lecturas que dan soporte al alelo alternativo y al alelo de la referencia en los SNPs heterocigotos y graficarlos. Se esperaría observar diferentes patrones de nubes de puntos de acuerdo con la ploidía de la cepa. Los aislados diploides exhibirían una única nube de puntos ya que ~50% de las lecturas en los SNPs heterocigotos darían soporte al alelo alternativo y el otro 50% al alelo de referencia (**Figura 6A**). En el caso de las cepas triploides, habría dos nubes de puntos, ya que ~66% de las lecturas darían soporte al alelo alternativo y 33% al alelo de referencia o ~33% de las lecturas darían soporte al alelo alternativo y 66% al alelo de referencia (Fay J.C. *et al.,* 2019) (**Figura 6B**). Por último, los aislados tetraploides presentarían un perfil con 3 nubes de puntos porque las lecturas que dan soporte al alelo alternativo en los SNPs heterocigotos podrían estar presentes al 25%, 50% o 75% (Fay J.C. *et al.,* 2019) (**Figura 6C**).

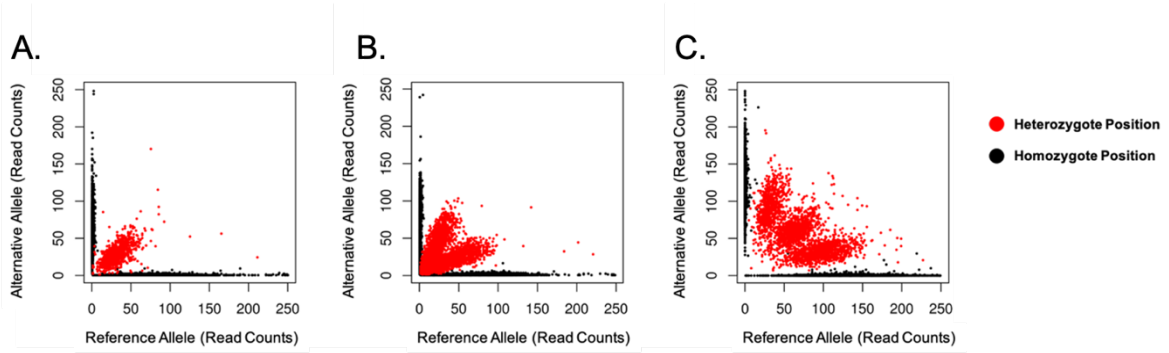


Figura 6. **Perfiles de ploidía de cepas diploides, triploides y tetraploides.** Gráficos de dispersión de las ploidías de una cepa diploide, triploide y tetraploide. El perfil del **Panel A** corresponde a una cepa diploide (50%-50% de lecturas dando soporte a alelo alternativo y alelo de referencia). El perfil del **Panel B** corresponde a una cepa triploide (66%-33% de lecturas dando soporte a alelo alternativo y alelo de referencia o viceversa). El perfil del **Panel B** corresponde a una cepa triploide (25%-%50-75% de lecturas dando soporte a alelo alternativo). Imagen tomada y modificada del artículo de Fay y colaboradores (Fay J.C. *et al.*, 2019).

En nuestro estudio, el método de Fay y colaboradores (2019) fue modificado para desglosar la ploidía de los 16 cromosomas de cada cepa. Esto nos permitiría observar, en el caso de cepas heterocigotas, si hay aneuploidías en cromosomas específicos. Los datos de las lecturas de los SNPs heterocigotos fueron obtenidos de los archivos .vcf de cada cepa sin filtrar. El análisis se realizó con los archivos sin filtrar, para ver todos los SNPs heterocigotos generados artificialmente por la presencia del genoma de *S. paradoxus* en los híbridos obtenidos de un mapeo de lecturas que se hizo exclusivamente hacia el genoma de *S. cerevisiae* S288C. Esta labor se realizó con el *script* de shell “obtain_parameter_values_gt.SNP_g_vcf.sh” (Gitlab bin/Ploidy).

Posteriormente, se utilizaron distintos programas de R Markdown “ID_plot_reads_counts_ploidy_per_chromosome.Rmd” (Gitlab Carpeta bin/Ploidy) para graficar las ploidías de los 16 cromosomas de cada cepa. Cada *script* de R

Markdown arrojaba un reporte HTML donde se desglosan los gráficos de ploidía de cada aislado (Gitlab figures/Ploidy).

En aras de complementar los análisis de ploidía hechos con el método de Fay y colaboradores (2019), se graficaron las coberturas normalizadas de los 16 cromosomas de cada uno de los aislados de fermentación de agave para determinar la presencia de aneuploidías en cada una de las cepas. Los gráficos se hicieron con el *script* de R “above_coverage” escrito por Iván Sedeño (Gitlab bin/Ploidy).

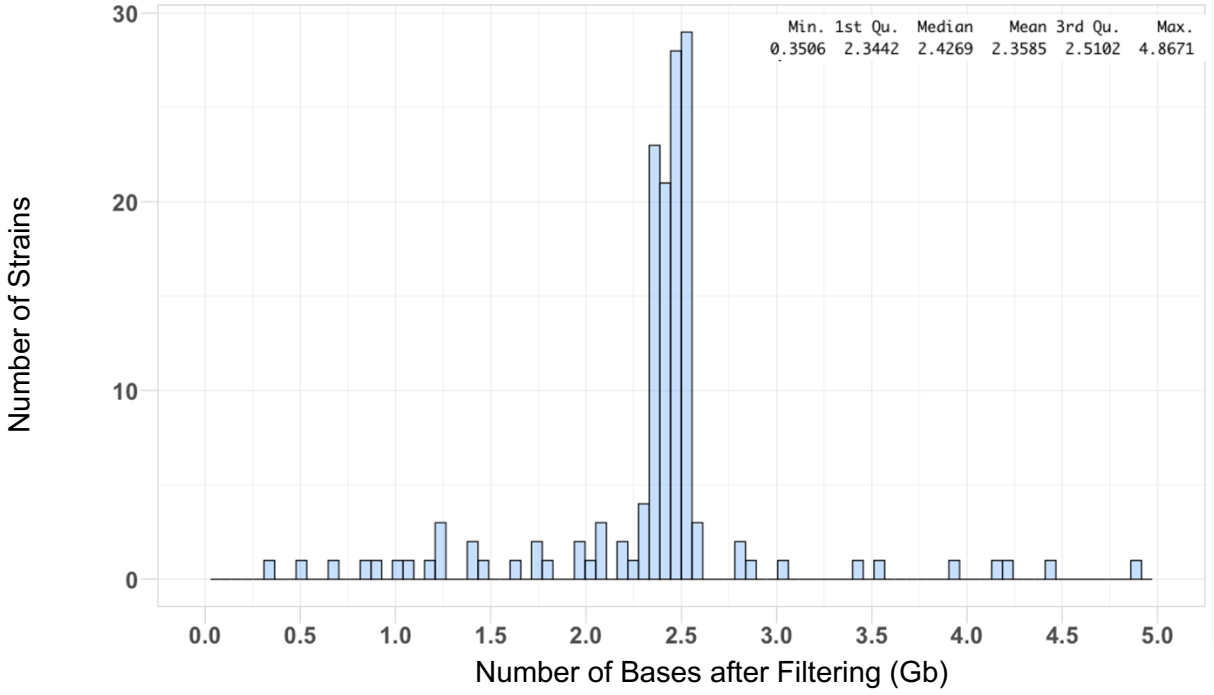
RESULTADOS

Se secuenciaron 137 aislados de fermentación de agave con la tecnología BGI-Seq y se descargaron lecturas de 170 cepas de *S. cerevisiae* del artículo de Peter y colaboradores (2018)

En total, se secuenciaron 137 cepas identificadas como levaduras pertenecientes al género *Saccharomyces* (**Tabla 1-Anexo**). Como se puede observar en la **Tabla1-Anexo**, 118 aislados fueron secuenciados con la tecnología de BGI (DNBSeq™) y 19 fueron secuenciados con plataformas de Illumina (cinco con Illumina MiSeq y 14 con Illumina NextSeq). Se decidió utilizar aislados secuenciados con ambas plataformas porque las tecnologías tienen similitudes y sus resultados son comparables (Korostin D. *et al*., 2020; Mak S.S.T. *et al*., 2017) y, salvo en algunos casos secuenciados con Illumina MiSeq, las coberturas de los aislados de Illumina resultaron en valores bastante aceptables. A estos 137 aislados se agregaron las secuencias de nueve muestras de *S. cerevisiae* previamente descritas en el artículo de Peter y colaboradores (2018) para dar un total de 146 cepas de agave. Por otro lado, al momento de realizar los análisis de nuestros aislados con respecto a cepas de otras regiones del mundo, se añadieron secuencias de 170 *S. cerevisiae* de diferentes lugares del planeta (**Tabla 2-Anexo**), que representan a los principales clados de *S. cerevisiae* descritos hasta ahora (Peter J. *et al.*, 2018). Es importante recalcar que las nueve cepas mexicanas descritas en el trabajo de Peter y colegas (2018) están incluidas en estas 170 *S. cerevisiae* (**Tabla 2-Anexo**).

Se limpiaron las lecturas de cada una de las cepas y se obtuvieron distribuciones de bases secuenciadas con una media de 2.36 Gb y 3.65 Gb para los aislados de agave (**Figura 7A**), y los tomados del trabajo de Peter y colegas (2018), respectivamente (**Figura 7B**).

A. Histograma of Bases after Filtering for Agave Strains (N=146)



B.

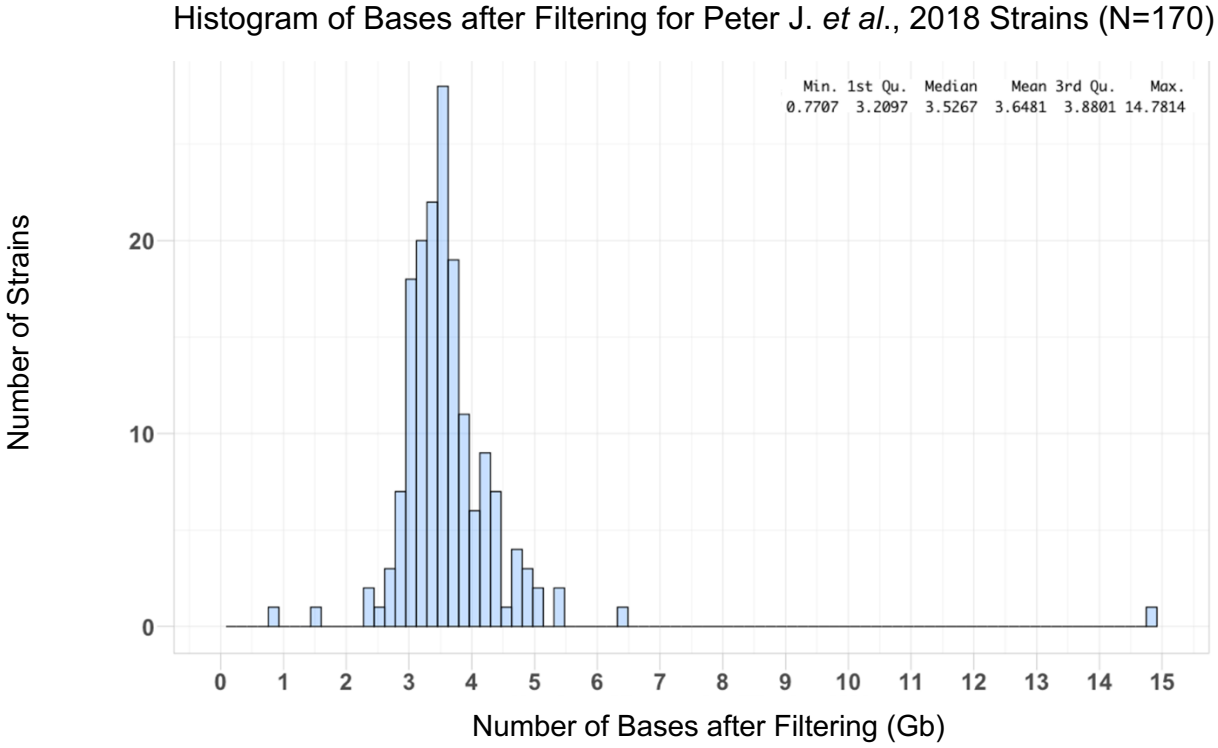


Figura 7. **Número de bases secuenciadas después de filtrado para cepas mexicanas de agave y cepas de otros lados del mundo.** Histograma de frecuencias del número de bases obtenidas (mínimo Q20 de calidad) después del filtrado de calidad de las lecturas. Las medias de las bases secuenciadas son 2.35 Gb para las cepas de agave y 3.63 Gb para las cepas del trabajo de Peter y colaboradores (2018).

Las lecturas de secuenciación de los aislados de fermentación de agave revelaron presencia de introgresiones e híbridos *S. cerevisiae*-*S. paradoxus*

El mapeo hacia un archivo con varios genomas de referencia - denominado mapeo competitivo - fue realizado con la finalidad de alinear el mayor número de lecturas con una calidad aceptable, ya que las especies del complejo *Saccharomyces* son proclives a intercambiar y compartir fragmentos genómicos a través de procesos como la hibridación y la introgresión (D' Angiolo M. *et al*., 2020; Marsit S. *et al*., 2017). Por ejemplo, las cepas de *S. cerevisiae* (**Anexo Tabla 2**) aisladas del

Alpechin (**Figura 8A)**, Guyana Francesa (**Figura 8C)** y bioetanol brasileño (**Figura 8D)** tienen un alto contenido de introgresiones genómicas provenientes de *S. paradoxus* (D' Angiolo M. *et* al., 2020; Peter J. *et al*., 2018; Pontes A. *et al*., 2019), lo cual dificultaría el mapeo, debido a que varias lecturas correspondientes a los fragmentos de *S. paradoxus* alinearían deficientemente con el genoma de *S. cerevisiae*, aumentando el número de polimorfismos.

En total, se identificaron 124 cepas de *S. cerevisiae* en los tanques de fermentación de agave, y se agregaron nueve cepas descritas por Peter y colaboradores (2018) para tener un total de 133 *S. cerevisiae* de México. Las 133 cepas de *S. cerevisiae* tienen presencia de introgresiones genómicas de *S. paradoxus,* lo cual se puede constatar con el porcentaje de lecturas mapeadas con mejor calidad (2-5%) al genoma de *S. paradoxus* YPS138 (**Tabla 1-Anexo, Figura 7B)**. De igual manera, se identificaron 13 cepas híbridas interespecies (*S. cerevisiae-S. paradoxus*) con diferentes porcentajes de contribuciones parentales de *S. paradoxus* (**Tabla 1-Anexo, Figura 8F).**

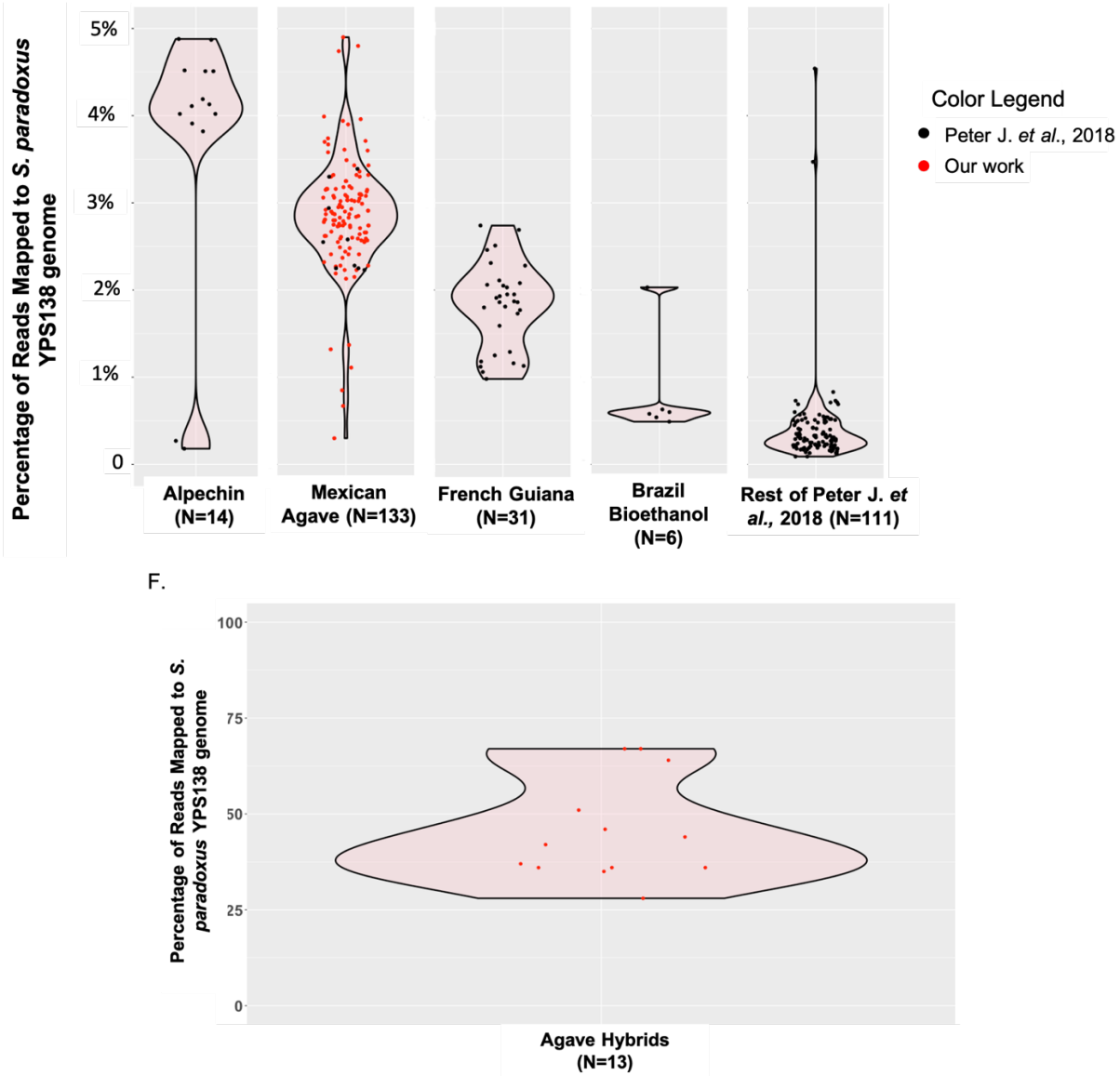
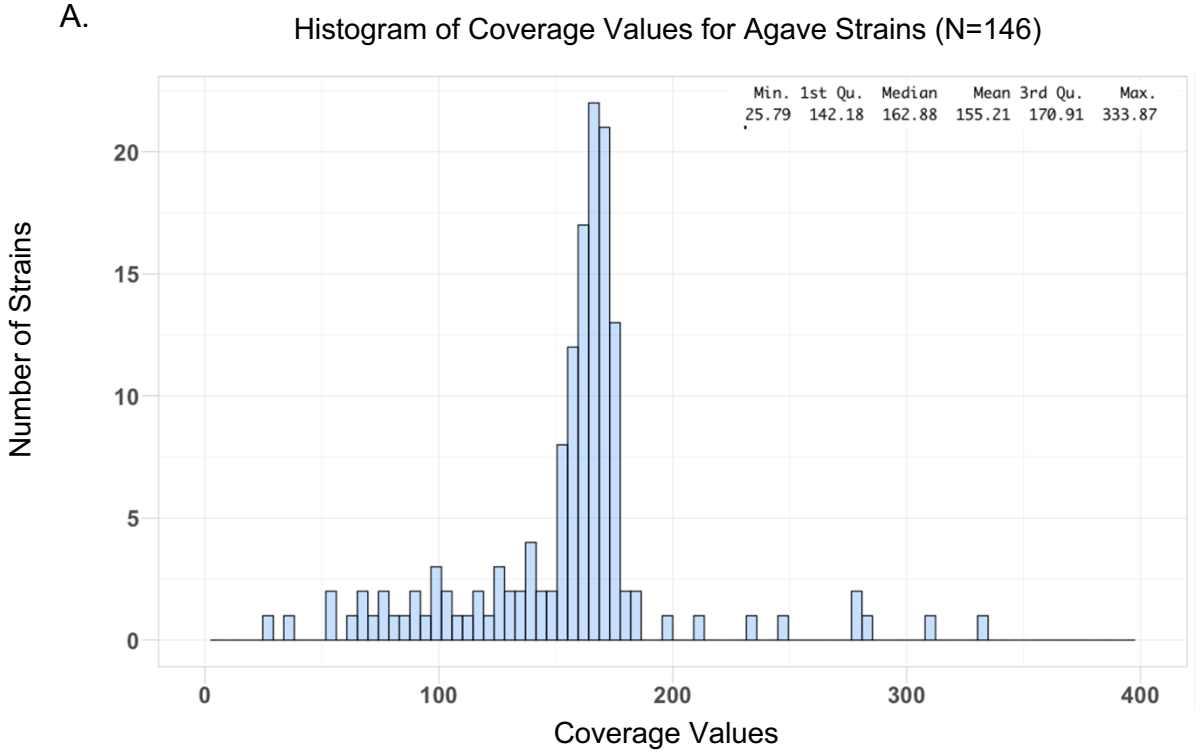


Figura 8. **Presencia de introgresiones de *S. paradoxus* en cepas de *S. cerevisiae* involucradas en la fermentación de agave.** Porcentajes de lecturas mapeadas al genoma de *S. paradoxus* YPS138 de cepas de *S. cerevisiae* pertenecientes a los grupos ***Alpechin* (A), *Mexican* (B), French *Guiana Human* (C), *Brazilian Bioethanol* (D)** y ***Resto de Cepas del trabajo de Peter* y colegas (2018) (E)** y **Cepas híbridas de agave (F).** Cada punto rojo representa el porcentaje de lecturas mapeadas al genoma de *S. paradoxus* YPS138 de cada una de las cepas en la gráfica. N: número de cepas en gráfica.

Después de mapear, se obtuvieron las medianas de las coberturas de cada una de las cepas. Se estimó una cobertura promedio de 155.21X para las cepas de agave (**Figura 9A**) y de 258.92X para las cepas evaluadas por Peter y colaboradores (2018) (**Figura 9B**).

A.



B.

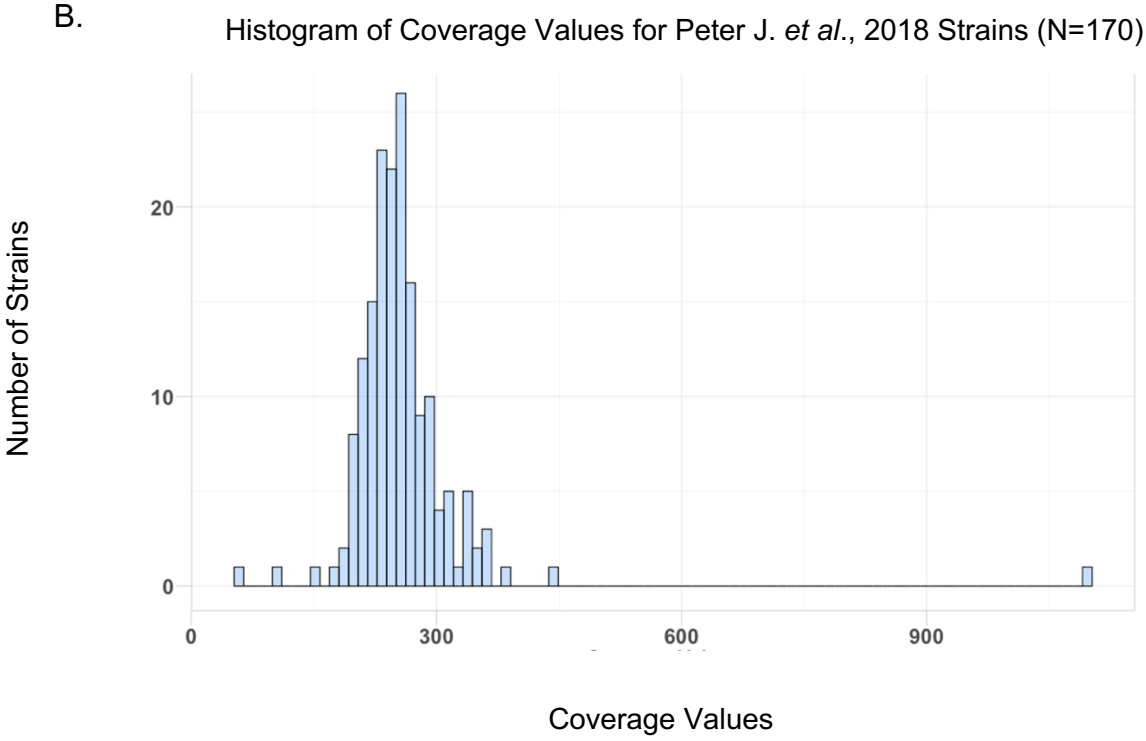


Figura 9. **Coberturas de cepas mexicanas de agave y de cepas de otras partes del mundo**. Histograma de frecuencias de las coberturas reportadas para las cepas de agave (A) y para las cepas del trabajo de Peter y colaboradores (2018) (B). Las coberturas promedio son 155.92 para los aislados de agave y 258.92 para los aislados de Peter y colaboradores (2018).

Identificación de variantes y genotipificación de las cepas mexicanas y de las del mundo

Se identificaron SNPs en las cepas mexicanas (**Anexo Tabla 1)** y de Peter y colaboradores (2018) (**Anexo Tabla 2**), con la finalidad de generar matrices de variantes que fueron utilizadas para llevar a cabo diferentes análisis de diversidad genética. Después de hacer la identificación de variantes, se obtuvieron varios archivos *genomic variant call format* (g.vcf), que contienen la información de los sitios variantes y no variantes de cada una de las muestras. Cada archivo g.vcf fue procesado de tal manera que solamente se quedaran los SNPs identificados en los

16 cromosomas del genoma de *S. cerevisiae* S288C. Posteriormente, se juntaron diferentes archivos g.vcf para generar dos matrices diferentes de SNPs:

- Matriz de SNPs de 137 cepas de México (**Anexo Tabla 1**).
- Matriz de SNPs de 137 cepas de México (**Anexo Tabla 1**), más 170 cepas del resto de mundo (**Anexo Tabla 2**).

Los SNPs de baja calidad fueron filtrados con la ayuda de una estrategia de filtrado duro (*hard filtering)* recomendada por los desarrolladores de GATK (https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants) (ver **Metodología**). Los desarrolladores de GATK recomiendan observar la distribución de los diferentes valores de los parámetros de filtrado en los SNPs de cada cepa para determinar cuáles son los mejores valores para el conjunto de datos con el que se está trabajando (**Figura 10**). Después de observar los gráficos de densidad de los diferentes parámetros de filtrado de cada una de las cepas (**ver repositorio Gitlab del proyecto**), se tomó la decisión de utilizar los parámetros de filtrado originalmente propuestos por GATK (**Tabla 1**), porque la mayoría de los SNPs llamados tenían valores asociados a una calidad aceptable para este tipo de variantes. De hecho, la **Figura 10** nos muestra un ejemplo de cómo se ven los perfiles de los parámetros de filtrado de todas las cepas utilizadas en este estudio, donde la mayor densidad de SNPs tiene un valor de parámetro de filtrado aceptable (consultar https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants) el cual indicaría que la variante tiene una alta probabilidad de ser real. Por ejemplo, el gráfico de *QualitybyDepth* (QD) de la **Figura 10** presenta una media ligeramente arriba de 30,

lo cual es mucho a dos, lo que significa que esta medida de calidad en nuestros está muy por arriba del umbral. Por lo tanto, sólo se estarían filtrando variantes localizadas en la cola del extremo izquierdo de la distribución.

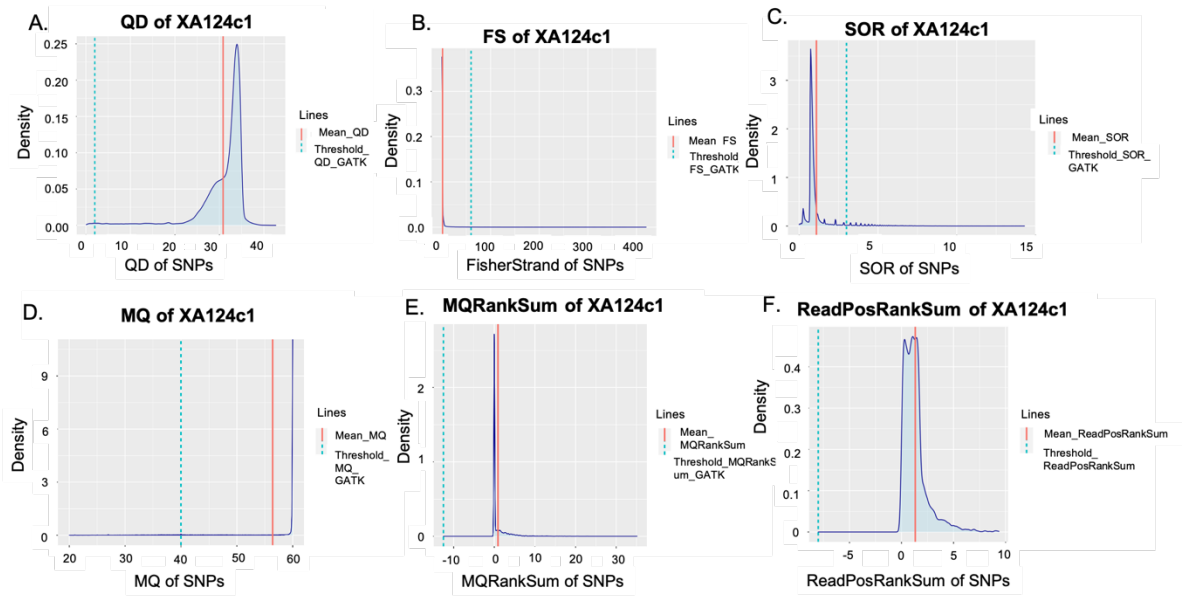


Figura 10. **Filtrado por calidad de los SNPs identificados en cada cepa.** Gráficas de densidad de los parámetros de filtrado para la cepa mexicana XA124c1. La línea roja en cada una de las gráficas representa el valor de la media de ese parámetro para esa cepa mientras que la línea azul punteada representa el valor recomendado de GATK para el umbral de filtrado (revisar **Tabla 1)**. Este es un ejemplo de las distribuciones de los parámetros de filtrado en un aislado en particular. Se puede ver que las medias (líneas roja) de las distribuciones se encuentra cercanas a valores óptimos de filtrado (revisar **Tabla 1** y **Métodos**) *QD.-Quality by Depth, FS.- Fisher Strand, SOR.- Strand Odds Ratio, MQ.- RMSMappingQuality*,MQRankSum.- *Mapping Quality Rank Sum Test, ReadPosRank Sum.- Read Position Rank Sum Test.*

Después de filtrar ambas matrices de SNPs por calidad, se procedió a evaluar la cantidad de información faltante (*missing data*), con el objetivo de remover los sitios que presentaran un cierto porcentaje de información faltante. La información faltante se define como la proporción de muestras donde una cierta variante no fue llamada, debido a falta de cobertura en ese sitio. La **Tabla 2** nos muestra como va cambiando el número de SNPs en cada matriz, dependiendo del umbral de información faltante permitido. Finalmente, se aplicó un filtro para retener exclusivamente SNPs bialélicos en las matrices (**Tabla 2**).

Tabla 2

Número de SNPs de *S. cerevisiae* (alineamiento contra archivo concatenado de genomas de ref. -> sólo SNPs del genoma de *S. cerevisiae*)				Número de SNPs de *S. cerevisiae* (alineamiento solo contra genoma de *S. cerevisiae*)	
Filtro de Calidad	Datos faltantes (min % of muestras con datos)	137 cepas mexicanas de nuestro trabajo + 9 cepas mexicanas de Peter *et al.* (2018) N=146	137 cepas mexicanas de nuestro trabajo + 170 cepas del mundo de Peter *et al.* (2018) N=307	Datos faltantes (min % de muestras con datos)	1,002 cepas de Peter *et al* + 141 cepas de fermentación de agave (incluyendo nueve de Peter *et al.* (2018)) N=1,143
OFF	0%	294,971	1,172,482	0%	2,370,738
ON	0%	260,487 (88.3%)	1,082,759 (92.3%)	0%	2,370,738
ON	75%	246,659 (83.6%)	1,047,477 (89.3%)	75%	1,572,726
ON	80%	245,964 (83.4%)	1,042,297 (88.9%)	80%	1,557,976
ON	85%	244,815 (83.0%)	1,035,868 (88.3%)	85%	1,529,943
ON	90%	243,118 (82.4%)	1,027,840 (87.7%)	85%+biallelic	1,353,760
ON	90% + biallelic	238,293 (80.1%)	970,686 (82.8%)	90%	384,463
ON	95%	239,468 (81.2%)	1,014,734 (86.5%)	90%+biallelic	256,585
ON	95% +biallelic	235,100 (79.7%)	959,596 (81.8%)	95%	377,530
ON	100%	121,169 (41.1%)	478,714 (40.8%)	95%+biallelic	252,149
ON	100% + biallelic	119,317 (40.5%)	455,579 (38.9%)	100%	188,107

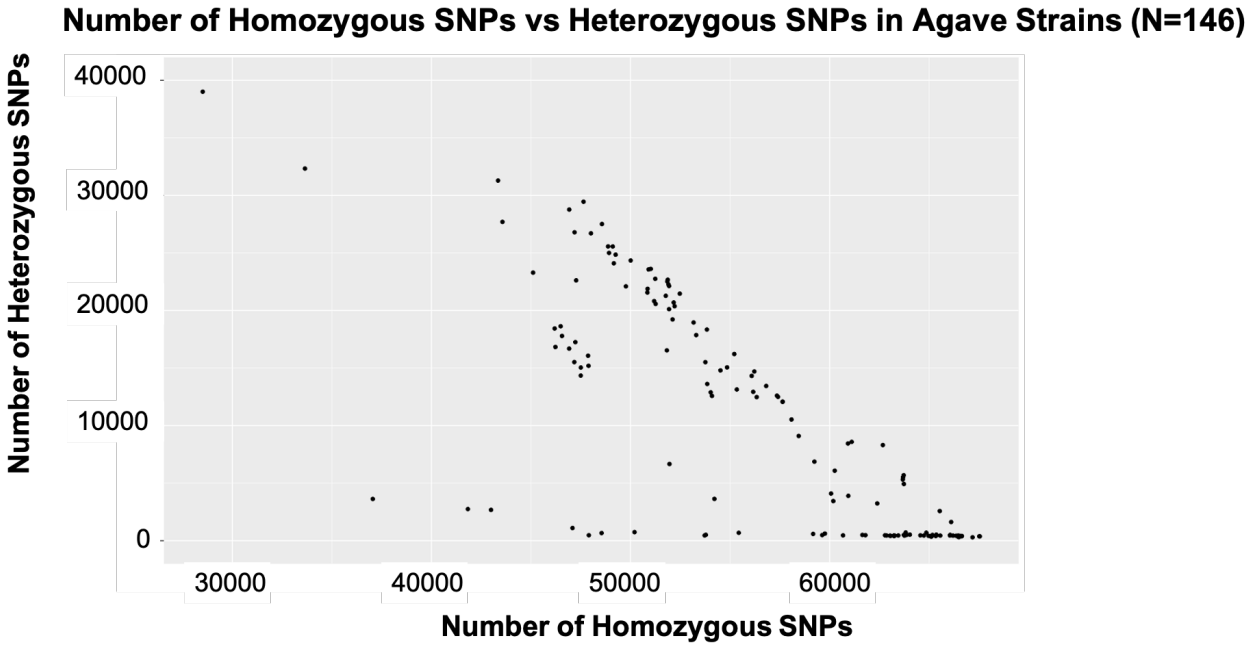
Nota: El rectángulo rojo señala las matrices de SNPs utilizadas para análisis filogenéticos, PCA y análisis de estructura poblacional. El rectángulo azul marino señala a la matriz utilizada crear el árbol filogenético de NJ de las 1,002 cepas analizadas por Peter y colegas (2018) *y 141 cepas de fermentación de agave.*

Los análisis de inferencia filogenética, PCA, prueba de Mantel y ADMIXTURE mostraron que el grupo de cepas de *S. cerevisiae* de fermentación de agave forman un grupo monofilético

Después de obtener las matrices de SNPs necesarias para los análisis de diversidad genética, se evaluó el nivel de heterocigosidad de las cepas mexicanas (**Tabla 1-Anexo**) y de las cepas del resto del mundo (**Tabla 2-Anexo**), utilizando la matriz de SNPs bialélicos con un máximo de 10% de datos faltantes (rectángulo rojo de **Tabla**

2). Se observó un número considerable de cepas mexicanas (N=60) de *S. cerevisiae* (**Tabla 1-Anexo**) con una proporción baja (< 5%) de SNPs heterocigotos, mientras que el resto de los aislados de agave mostraron niveles intermedios u altos de heterocigosidad (**Figura 11A y 12A**). Por otro lado, las cepas del resto del mundo (**Tabla 2-Anexo)** mostraron un nivel intermedio de heterocigosidad, aunque también se observó un alto número de cepas con alto contenido de SNPs homocigotos (**Figura 11B y 12B**). En el caso de los aislados de varias partes del mundo, dos cepas de China (**Grupos CHNI y CHNII en Tabla 2-anexo**) y una de Taiwán (***Taiwanese* en Tabla 2-anexo**), tienen un alto número de SNPs homocigotos, lo que sugiere que son aislados con genomas muy diferentes al genoma de referencia de *S. cerevisiae* S288C y, por lo tanto, provoca que el eje de las X de la **Figura 11B** se desplace más hacia la derecha, en contraste con lo observado en la **Figura 11A**. Esta alta presencia de SNPs homocigotos en los aislados del Este Asiático puede explicarse porque son las cepas más divergentes en términos genéticos, lo cual es congruente con la hipótesis de un origen de *S. cerevisiae* en esta región (Duan S.F. *et al*.,2018; Peter J. *et al*., 2018).

A.



B.

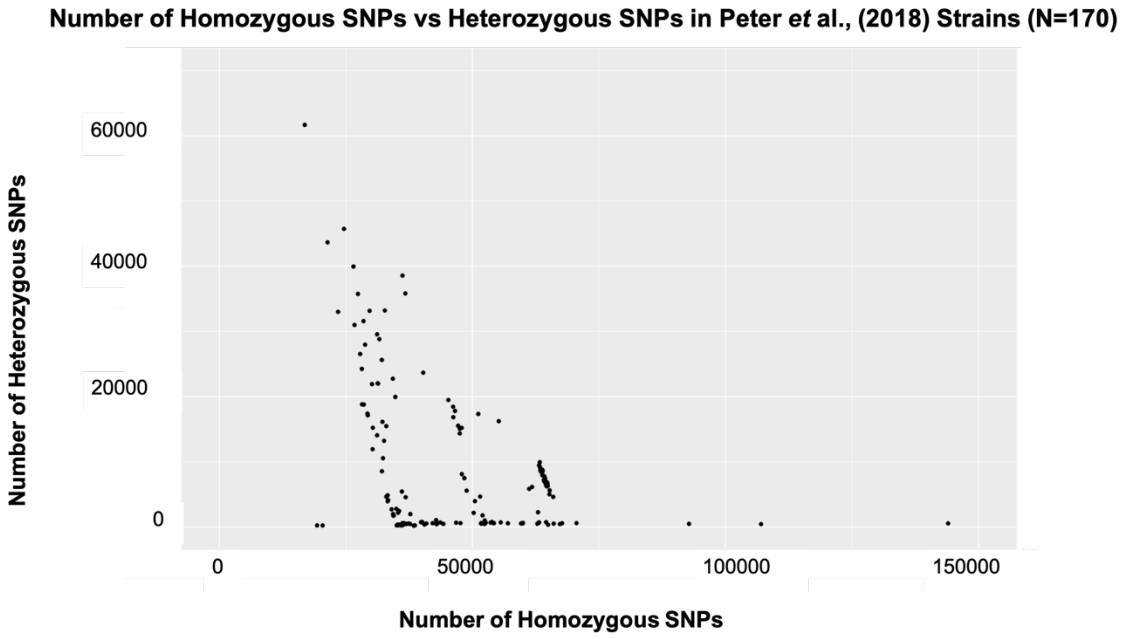
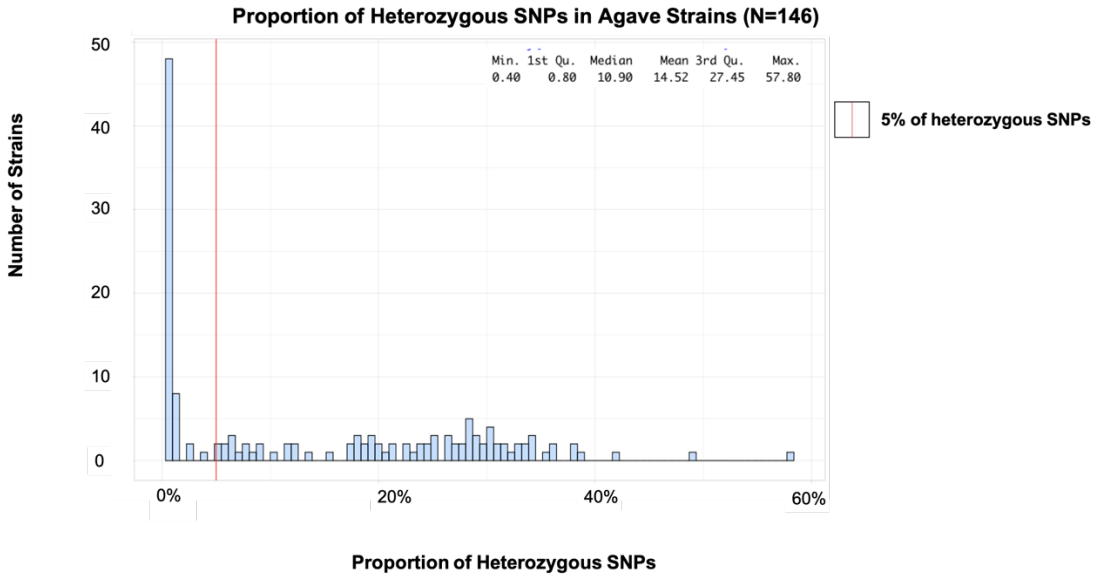


Figura 11. **Gráficas de cantidad de SNPs homocigotos vs SNPs heterocigotos de los aislados de fermentación de agave (Panel A) y los aislados del Peter y colaboradores (2018) (Panel B).** Estos gráficos de puntos comparan el número de SNPs heterocigotos y SNPs homocigotos de cada una de las muestras (cada punto negro es un aislado). El **Panel A** muestra el comparativo de las cepas de fermentación de agave (N=146) mientras que el **Panel B** muestra el comparativo de las cepas de Peter y colaboradores (2018) (N=170). Entre una cepa (punto negro) esté más a la derecha sobre el eje X y más abajo en el eje Y, significa que es una cepa altamente homocigota. Por otro lado, si una cepa se encuentra más sobre la izquierda del eje X y más arriba en el eje Y, significa que es una cepa altamente heterocigota.

A.



B.

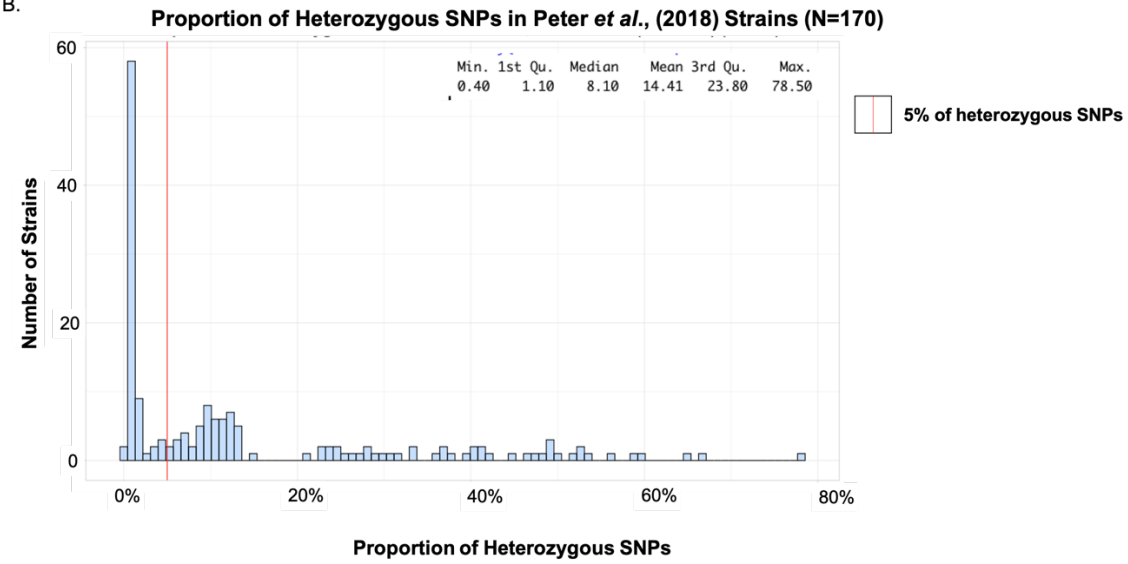
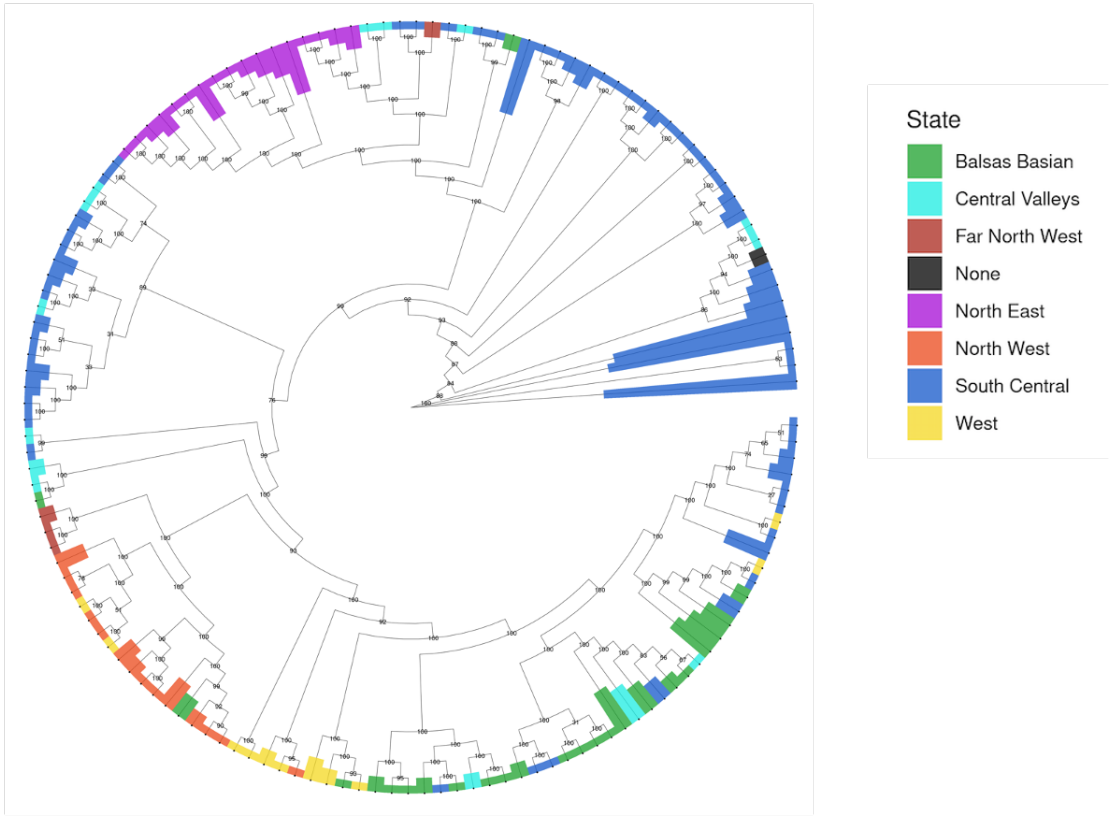


Figura 12. **Histograma de frecuencias de las proporciones (%) de SNPs que son heterocigotos en las cepas de fermentación de agave (Panel A) y en las cepas de Peter y colegas (2018)**. Estos histogramas nos muestran la cantidad de cepas heterocigotas y homocigotas que hay en las cepas de fermentación de agave (A) y en las cepas de otras partes del mundo de Peter y colegas (2018). Se pone una línea roja para delimitar el % con el que se determina si una cepa es heterocigota, es decir, si una cepa tiene más del 5% de SNPs heterocigotos, esa cepa es heterocigota (mismo criterio utilizado en el trabajo de Peter y colegas (2018)). Por lo tanto, todas las cepas al lado derecho de la línea roja son heterocigotas mientras que todas las del lado izquierdo son homocigotas.
Nota: La proporción de sitios heterocigotos se obtuvo dividiendo el número de SNPs heterocigotos entre el número total de SNPs en cada aislado.

Análisis filogenéticos de cepas de fermentación de agave

Las matrices de SNPs del genoma de *S. cerevisiae* con máximo 10% de datos faltantes y sólo SNPs bialélicos (rectángulo rojo en **Tabla 2**) fueron utilizadas para realizar análisis de inferencia filogenética con métodos de máxima verosimilitud (Stamakatis, A., 2014). Primero, se construyó un árbol filogenético (**Figura 13**) utilizando 238,293 SNPs bialélicos del genoma nuclear de las cepas mexicanas de la **Tabla 1-Anexo,** y se pudo constatar que las cepas se agrupan mayormente en subclados de acuerdo con la región geográfica donde fueron aisladas, como se muestra claramente en la **Figura 13A,** donde se ven varias cepas formando grupos de un solo color (misma región de aislamiento). Una tendencia similar de correlación entre ubicación geográfica y formación de clados es observada cuando se toman en cuenta los estados de México donde fueron recolectadas las muestras (**Figura 13B**).

A.


State
Balsas Basian
Central Valleys
Far North West
None
North East
North West
South Central
West

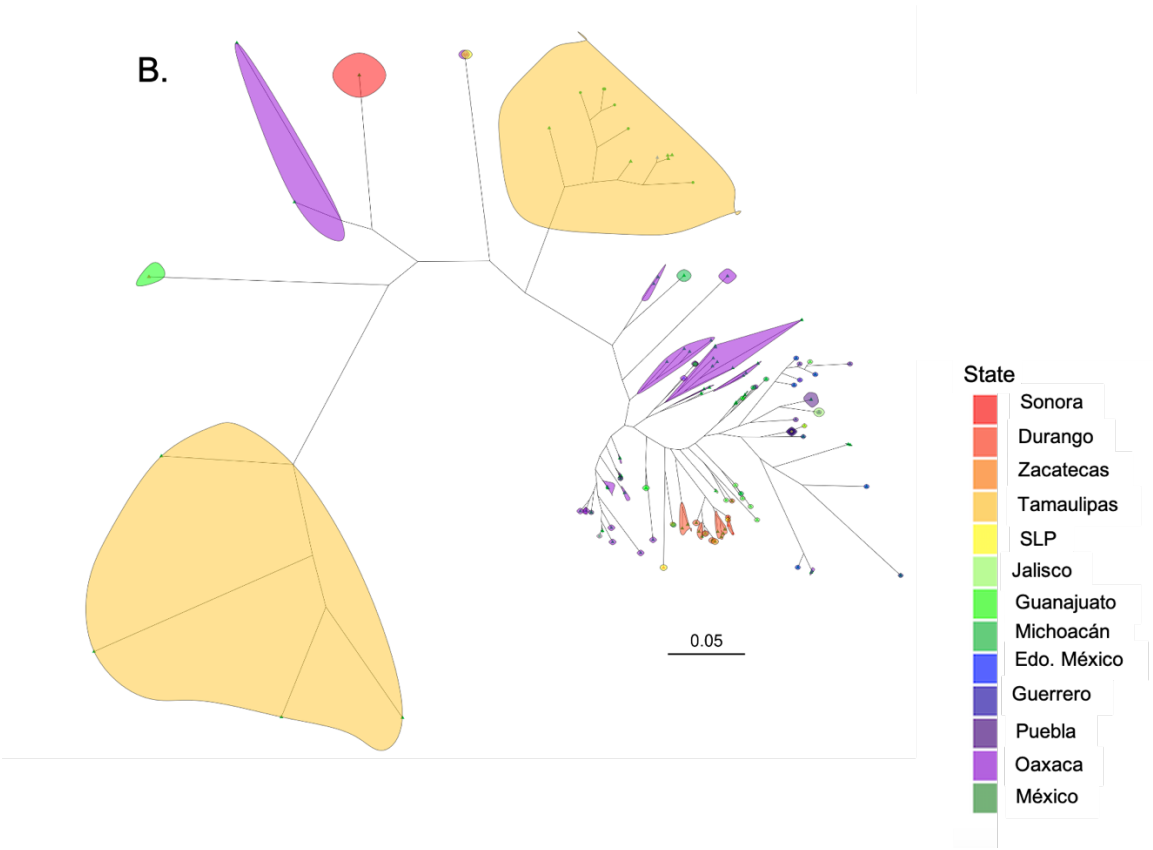


Figura 13. **Filogenia de las cepas mexicanas de fermentación de agave**. Topología no enraizada de las cepas mexicanas de *S. cerevisiae* (**Tabla 1-Anexo**) construido con el programa RaXML v8.2.12 el método de máxima verosimilitud a partir de 238,293 SNPs bialélicos (sub-genoma de *S. cerevisiae*), permitiendo un máximo de 10% de datos faltantes (*misssing data*) y con 100 rondas de bootstrap. El **panel A** muestra, por simplificación, el árbol con el tamaño de ramas iguales y los diferentes valores de bootstrap para las ramas (cladograma). Se colorearon las hojas de las ramas de acuerdo con la región geográfica donde fueron aisladas. El **panel B** muestra el mismo árbol con otro diseño de topología y con las ramas coloreadas de acuerdo con el estado donde se aislaron las cepas.

Posteriormente, se decidió construir un árbol filogenético con 970,686 SNPs bialélicos del genoma nuclear de las cepas mexicanas (**Tabla 1-Anexo**) junto con las cepas del resto del mundo, para visualizar la relación que tienen las cepas mexicanas de *S. cerevisiae* involucradas en la fermentación de agave con aislados de distintos orígenes y lugares (**Figura 14**).

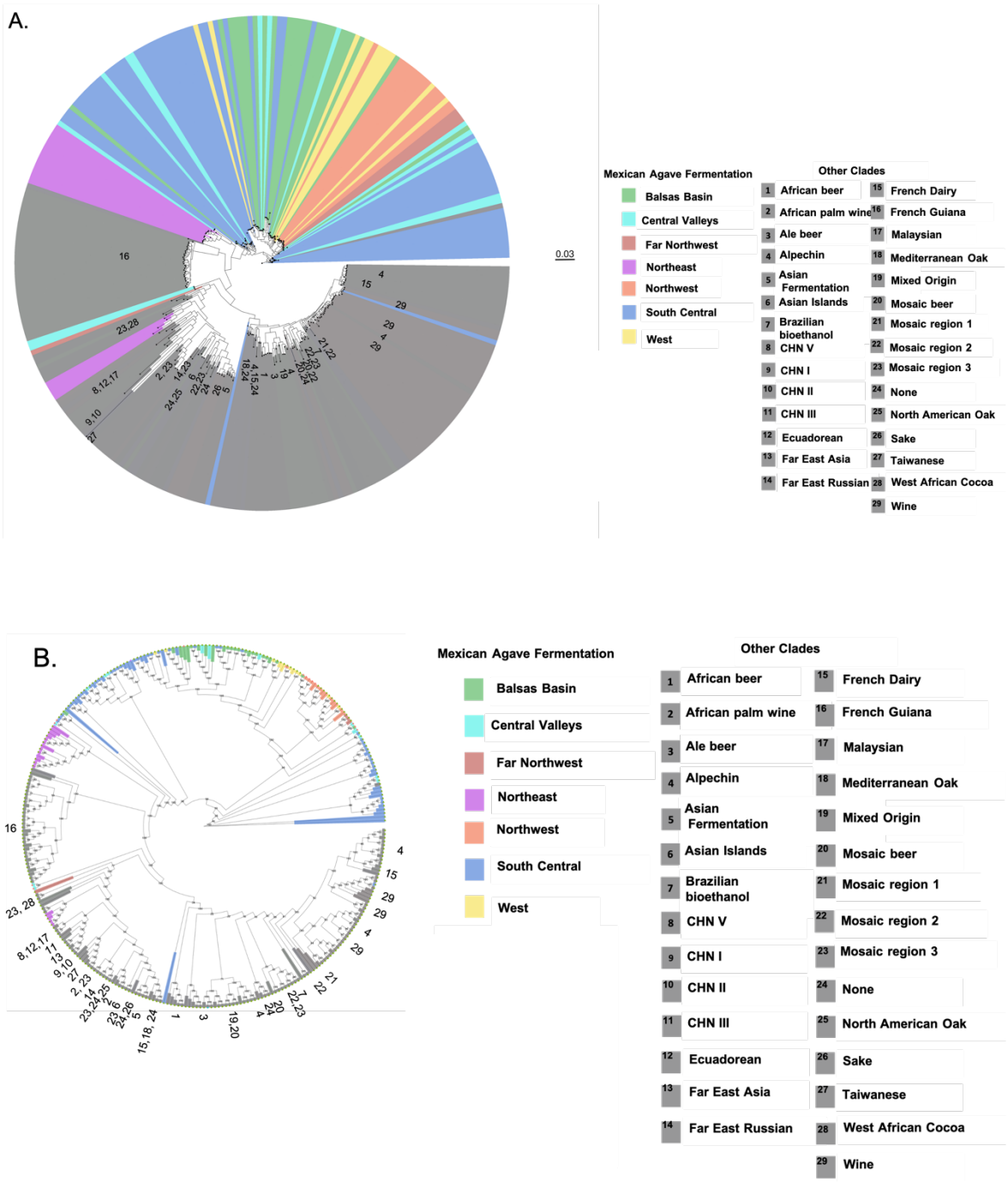


Figura 14. **Filogenia de las cepas mexicanas de agave más las cepas del resto del mundo**. Topología no enraizada de las cepas mexicanas de *S. cerevisiae* (**Tabla 1-Anexo**) y las cepas del resto del mundo (**Tabla 2-Anexo**) construido con el método de máxima verosimilitud y el programa RaXML v8.2.12 a partir de 970,686 SNPs bialélicos (sub-genoma de *S. cerevisiae*), permitiendo un máximo de 10% de datos faltantes (*misssing data*) y con 1000 rondas de bootstrap. El **Panel A** muestra el árbol con variación en la longitud de ramas y con las cepas mexicanas coloreadas de acuerdo con la región donde fueron muestreadas. El **Panel B** muestra, por simplificación, el mismo árbol pero con las longitudes iguales en las ramas, las cepas mexicanas coloreadas de acuerdo con la región donde fueron muestreadas y con los valores de bootstrap de las ramas.

La filogenia de la **Figura 14** muestra que las cepas de fermentación de agave de nuestro estudio y las nueve del estudio de Peter y colegas (2018) conforman un grupo monofilético de cepas genéticamente distintas a las del resto del mundo. En total, 128 de 137 cepas secuenciadas por nosotros conforman el grupo monofilético principal. Las otras nueve cepas están incrustadas en otros clados, lo que sugeriría que tienen un origen diferente al grupo principal.

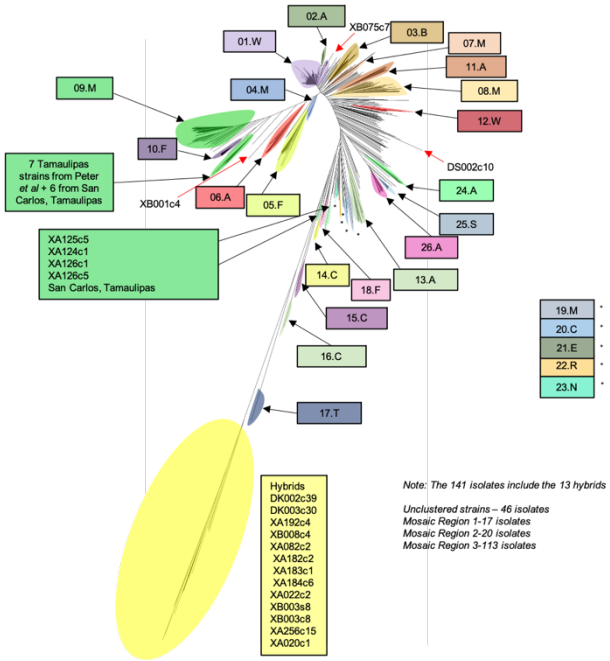
Previamente, Peter y colaboradores (2018) analizaron siete cepas de *S. cerevisiae* mexicanas (Tamaulipas) de fermentación de agave que clasificaron como ***Mexican Agave*** y, de igual manera, añadieron dos aislados de agave (JS497c1 -> Raicilla, Jalisco y JS109 -> Mosto de varias especies de agave, México) que no agruparon en el grupo recién mencionado. Nuestro análisis filogenético muestra que el muestreo centrado principalmente en el estado de Tamaulipas no detectó la totalidad de la diversidad genética existente dentro de las aislados de *S. cerevisiae* de fementaciones de agave. Las cepas de Tamaulipas forman dos subclados: un subclado hermano del grupo de ***French Guiana*** (clado morado arriba del clado de ***French Guiana*** **en Figura 14A**) y otro sumamente alejado del grupo monofilético principal. El subclado hermano de ***French Guiana*** incluye las siete cepas de Tamaulipas originalmente reportadas por el artículo de los 1,011 genomas (Peter J., *et al*., 2018), más seis cepas originarias de San Carlos en el mismo estado secuenciadas por nosotros. El otro grupo de Tamaulipas está conformado por los aislados XA126c5, XA124c1, XA126c1 y XA125c5 y se agrupa con cepas norteamericanas del grupo ***Mosaic Region*** **3.** El grupo ***Mosaic Region*** **3** es una subpoblación de cepas con orígenes geográficos y ecológicos diversos, que presenta un alto nivel de mosaicismo en sus genomas y, en consecuencia, los

aislados no pueden ser agrupados en un clado específico (Peter J. *et al*., 2018-material suplementario). Los genomas de ***Mosaic Region 3*** presentan componentes de los grupos de vino (***Wine/European***), sake y bioetanol, entre otros (Peter J. *et al*., 2018-material suplementario). Además, llama la atención que estas cuatro cepas provienen de un solo palenque, y que son muy divergentes al resto de cepas de fermentación de agave. Esto sugiere que los productores probablemente sí inoculen con alguna cepa industrial (probablemente tequilera) los tanques de fermentación en este palenque. Por otro lado, la cepa JS497c1 de Jalisco y la cepa JS109c1 de mosto de diferentes especies de agave están incluidas en el grupo monofilético principal, al lado de aislados de diferentes regiones de México.

Sin contar las cuatro cepas de Tamaulipas agrupadas con aislados de ***Mosaic Region* 3,** cinco cepas se encontraron dentro de otros clados alejados del grupo monofilético principal, como la cepa XB075c7 que, a diferencia de otras cepas de Oaxaca que forman parte del clado monofilético, se agrupó con unas cepas europeas de vino (***Wine/European***). Por último, cepas como la DS002c10 (Sonora), la XB008c4 (cepa híbrida de Guanajuato), la XA082c2 (cepa híbrida de Guanajuato) y la XB001c4 (Matatlán, Oaxaca) también se posicionaron fuera del grupo monofilético principal.

Finalmente, se decidió realizar un análisis filogenético con el algoritmo *Neighbor-Joining* que incluyera a las 1,011 cepas evaluadas por Peter y colaboradores (2018) y 132 aislados de fermentación de agave (se quitaron 5 aislados secuenciados con Illumina MiSeq por tener baja calidad). La **Figura 15** reconfirma que se está describiendo a un nuevo clado de *S. cerevisiae* que había sido pobremente estudiado. De igual manera, se puede observar que el subclado

de Tamaulipas con 13 cepas (seis secuenciadas por nosotros + siete de Peter y colegas) es divergente al resto de aislados de fermentación de agave (sombreado verde al lado derecho del grupo ***French Guiana Human***). Por último, los 13 híbridos interespecie *S. cerevisiae*-*S. paradoxus* están representadas por las ramas más largas del árbol, lo cual es esperado debido a que 33%-66% del genoma pertenece a una especie diferente a *S. cerevisiae.*



Abbreviation	Clade	Number of Isolates
01.W	Wine/European	456
02.A	Alpechin	17
03.B	Brazilian Bioethanol	35
04.M	Mediterranean Oak	8
05.F	French Dairy	32
06.A	African Beer	20
07.M	Mosaic Beer	21
08.M	Mixed Origin	72
09.M	Mexican Agave	141*
10.F	French Guiana Human	31
11.A	Ale Beer	18
12.W	West African Cocoa	13
13.A	African Palm Wine	28
14.C	CHNIII	2
15.C	CHNII	2
16.C	CHNI	1
17.T	Taiwanese	3
18.F	Far East Asia	9
19.M	Malaysian	6
20.C	CHNV	2
21.E	Ecuadorean	10
22.R	Far East Russian	4
23.N	North American Oak	13
24.A	Asian Islands	11
25.S	Sake	47
26.A	Asian Fermentation	39

Figura 15 **Árbol NJ (distancias genéticas de Weir & Goudet, 2017) de las 1,011 cepas de *S. cerevisiae* de Peter y colaboradores (2018)y 132 cepas de fermentación de agave**. Topología no enraizada de las cepas mexicanas de *S. cerevisiae* (**Tabla 1-Anexo**) y 1,011 cepas del resto del mundo (**Tabla 2-Anexo**) construido con el método de *Neighbor-Joining* utilizando la librería SNPRelate y el paquete ape de R a partir de 1,353,760 SNPs bialélicos (mapeos al genoma de referencia de *S. cerevisiae*) y permitiendo un máximo de 15% de datos faltantes (*misssing data*). La tabla de la derecha enumera a los grupos de aislados de *S. cerevisiae* tal y como se describe en el artículo de Peter y colaboradores (2018). Los sombreados en las ramas del árbol coinciden con el sombreado en cada fila de la tabla.

Análisis de componentes principales de aislados de fermentación de agave

En aras de complementar los análisis de inferencia fiilogenética se procedió a realizar unos análisis de componentes principales (PCA) en los que se buscó evaluar si los patrones de agrupamiento eran congruentes con los observados en la filogenia. Sin embargo, fue necesario remover 13 cepas (ver **Tablas 3 y 4-Anexo, Figuras Suplementarias 1A-1B** y **Metodología**) que tenían un alto contenido de datos faltantes y poco porcentaje de mapeo de lecturas al genoma de *S. cerevisiae* S288C. Doce de estas 13 cepas son híbridos interespecie y, por lo tanto, una proporción significativa de lecturas mapeó al genoma de *S. paradoxus*, lo que generó una muy alta proporción de datos faltantes en los SNPs identificados en el genoma de *S. cerevisiae*. Después de quitar las 13 cepas y todos los sitios con datos faltantes, se construyeron dos gráficos de PCA: uno con 223,013 SNPs bialélicos de los genomas nucleares de las cepas mexicanas (**Figura 16A**) (**Tabla 1-Anexo**), y otro con 831,175 SNPs bialélicos del genoma nuclear de las cepas mexicanas (**Tabla 1-Anexo**) y las cepas del resto del mundo (**Figura 16B**) (**Tabla 2-Anexo**).

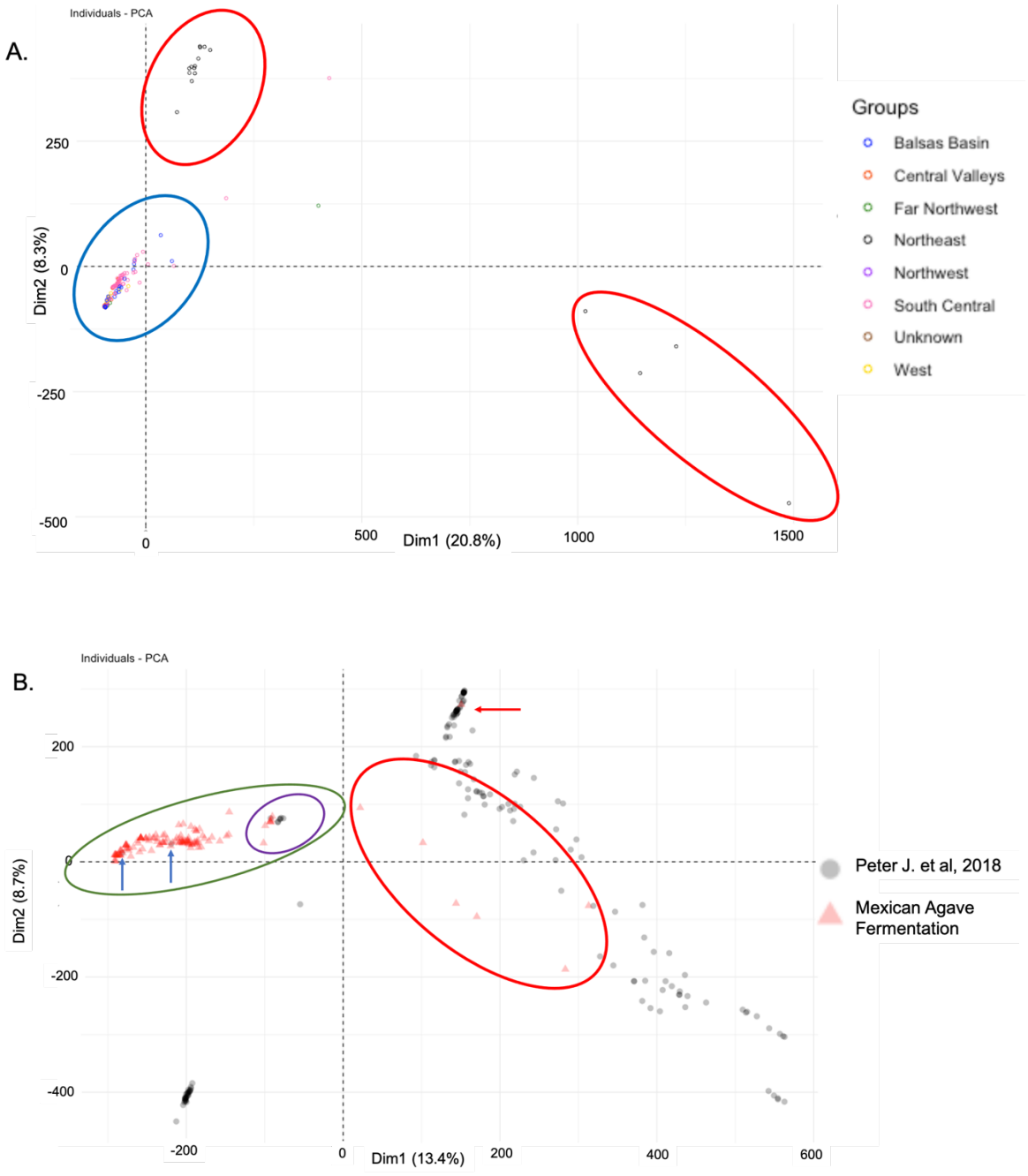


Figura 16. **Análisis de componentes principales de las cepas mexicanas (Panel A) y de las cepas mexicanas más las cepas del resto del mundo (Panel B).** Gráficos de PCA (PC1 vs PC2) de las cepas mexicanas de *S. cerevisiae* de fermentación de agave (**Panel A**) hecho a partir de 223,013 SNPs bialélicos y de las cepas mexicanas de *S. cerevisiae* de fermentación de agave mas las cepas del resto del mundo de Peter y colaboradores (2018) hecho a partir de 831,175 SNPs bialélicos (**Panel B**). En el caso del **Panel A**, las cepas están coloreadas de acuerdo con las regiones de México donde fueron muestreadas. El óvalo azul del **Panel A** encierra a las cepas pertenecientes al clado monofilético principal de aislados de fermentación de agave y los óvalos rojos encierran a las cepas de este estudio y de Peter y colaboradores (2018) provenientes de Tamaulipas. En el **Panel B,** las cepas mexicanas de fermentación de agave y los aislados de Peter y colegas (2018) están representados por triángulos rojos y círculos negros, respectivamente. El óvalo verde encierra a las cepas del clado monofilético principal de las cepas de fermentación de agave y las flechas azules señalan dos cepas del estudio de Peter y colaboradores (2018), la flecha roja señala a la cepa XB075c7 que se agrupa con las cepas de **Wine/European,** el óvalo morado encierra a los aislados provenientes de Tamaulipas (nuestro estudio y artículo de Peter y colegas, 2018) y, por último, el óvalo rojo encierra a aislados que no pertenecen al clado monofilético principal.

Los componentes principales 1 y 2 del PCA de las cepas mexicanas (**Figura 16A)** muestran que la mayoría de las *S. cerevisiae* de fermentación de agave (óvalo azul en **Figura 16A**) se juntan en un solo grupo principal de aislados, sugiriendo un posible origen común. Sin embargo, las cepas de la región Noreste (Tamaulipas) cuentan una historia diferente, porque hay dos grupos (óvalos rojos) de estos aislados que están alejados del grupo principal. Un conjunto de cepas del Noreste (óvalo rojo superior) está conformado por las siete cepas de Tamaulipas reportadas por Peter y colegas. (2018) y seis cepas de San Carlos, Tamaulipas secuenciadas por nuestro grupo de trabajo. Las siete cepas habían sido clasificadas como ***Mexican agave*** en el artículo de los 1,011 genomas (Peter J. *et al*., 2018). Las otras dos cepas mexicanas descritas en el artículo de Peter y colaboradores están incluidas en el grupo principal de cepas (óvalo azul de **Figura 16A**). La otra subpoblación de aislados de Tamaulipas (óvalo rojo inferior de **Figura 16A**) está formada por las mismas cepas (XA126c5, XA124c1, XA126c1 y XA125c5) que se agrupan en un clado distante al grupo monofilético principal en el árbol filogenético (**Figura 14A-B, Figura 15**). Este comportamiento de agrupamiento también fue observado cuando se graficó PC1 vs PC3 (**Figura suplementaria 2A**) y PC2 vs PC3 (**Figura suplementaria 2B**) con los mismos aislados mexicanos de agave.

Análisis de componentes principales de cepas de agave y del resto del mundo

El gráfico de los componentes principales 1 y 2 de las cepas mexicanas y del resto del mundo (**Figura 16B**) también confirmó lo que ya había mostrado el análisis

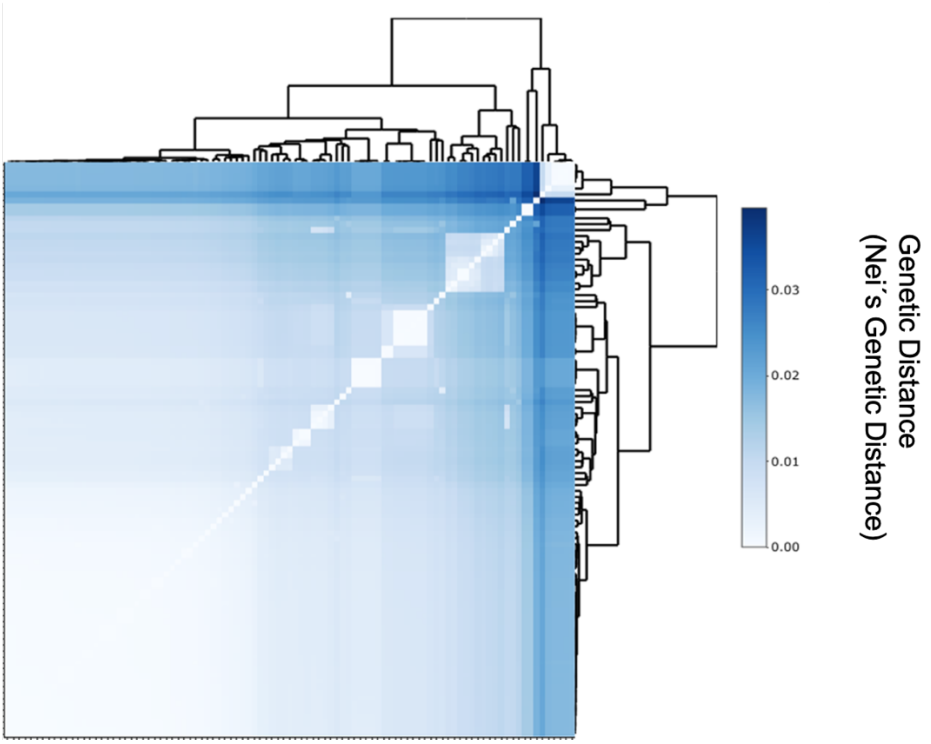
filogenético. Las cepas de *S. cerevisiae* involucradas en la fermentación de agave (triángulos rojos dentro de óvalo verde) conforman un grupo de cepas genéticamente distintas al resto de aislados de otras partes del mundo. Adicionalmente, este grupo de cepas está proyectado a lo largo del componente principal 1 (Dim1), lo que indicaría que este grupo tiene una gran diversidad genética. Las otras dos cepas reportadas por Peter y colaboradores (2018) (JS497c1 -> Raicilla, Jalisco y JS109-Mosto de varias especies de agave, México) están en el grupo principal de aislados de agave (flechas azules en **Figura 16B**). Este dato sugiere que estas dos cepas son parte de un continuo de diversidad genética presente en el grupo de las *S. cerevisiae* involucradas en la fermentación de agave. De igual manera, se observó que los aislados XA126c5, XA124c1, XA126c1 y XA125c5 de Tamaulipas (óvalo rojo en **Figura 16B**) están proyectadas lejos del grupo principal de cepas de agave (óvalo verde), lo que apunta a un origen distinto de esta subpoblación de un palenque en particular de Tamaulipas. Por último, la cepa XB075c7 de Oaxaca, al igual que en el análisis filogenético, aparece incrustada en medio del grupo de aislados de vino (flecha roja en **Figura 16B**). Esto se podría explicar por un probable evento de flujo génico (hibridación) entre alguna cepa de fermentación de vino (o alguna cercana a este clado) y un aislado de fermentación de agave. Cabe mencionar que estos comportamientos de agrupamiento para las cepas mexicanas de agave y del resto del mundo también son observados cuando se grafica PC1 vs. PC3 (**Figura suplementaria 3A**) y PC2 vs. PC3 (**Figura suplementaria 3B**).

Prueba de Mantel de cepas de fermentación de agave

Como se mencionó anteriormente, el análisis filogenético de las cepas mexicanas de fermentación de agave parece indicar que existe una correlación entre la agrupación de las cepas en clados y su ubicación geográfica. Para evaluar mejor este patrón, se decidió realizar una prueba de Mantel, la cual se encarga de estimar la correlación que existe entre una matriz de distancias genéticas y una matriz de distancias geográficas (Diniz Filho J.A. *et al*., 2013), ya que se tienen datos de distancias geográficas (vuelo de pájaro) de 103 cepas de agave secuenciadas por nosotros (**Tabla Mantel Test-Anexo**), las cuales fueron utilizadas para construir la matriz de distancias geográficas (QGIS.org,2021). La matriz de distancias genéticas (**Figura 17A**) se estimó a partir de la matriz con 223,013 SNPs de los genomas nucleares de aislados mexicanos de agave, de donde ya habían sido removidas las 13 cepas por alto contenido de datos faltantes (ver **Tablas 3 y 4-Anexo, Figuras Suplementarias 1A-1B** y **Metodología**). Adicionalmente, se quitaron las cepas XA126c5, XA124c1, XA126c1, XA125c5, XB001c4 y XB075c7 (**Tabla 1-Anexo**) porque no forman parte del clado monofilético principal (**Figura 14**) y eso introducía un sesgo por valores atípicos (*outlier bias*). En total, se contó con información de 97 cepas de agave para el análisis. Se corrió la prueba de Mantel con ambas matrices listas y depuradas. Este análisis arrojó un estadístico r de Mantel de 0.4453 (**Figura 17B**) con una significancia estadística de 0.001. Entre más cercano esté el resultado de r a 1, significa que la correlación entre distancias geográficas y genéticas es más fuerte. Este resultado resultó significativo, lo cual indica que el aislamiento genético y la distancia geográfica entre las distintas localidades donde se muestrearon las

cepas de fermentación de agave sí estarían influyendo en su diferenciación genética. Adicionalmente, este dato sugiere que los procesos neutrales, como la deriva génica estarían participando en el moldeo de la variación genética presente en los diferentes subclados de los aislados de fermentación de agave. De igual manera, la correlación positiva entre las distancias genéticas y geográficas podría explicarse por adaptaciones locales a condiciones ambientales (clima, suelo, altitud, entre otros) de cada región. Estas adaptaciones locales, a su vez, influirían en las variaciones a nivel de frecuencias alélicas en cada región.

A.



B.

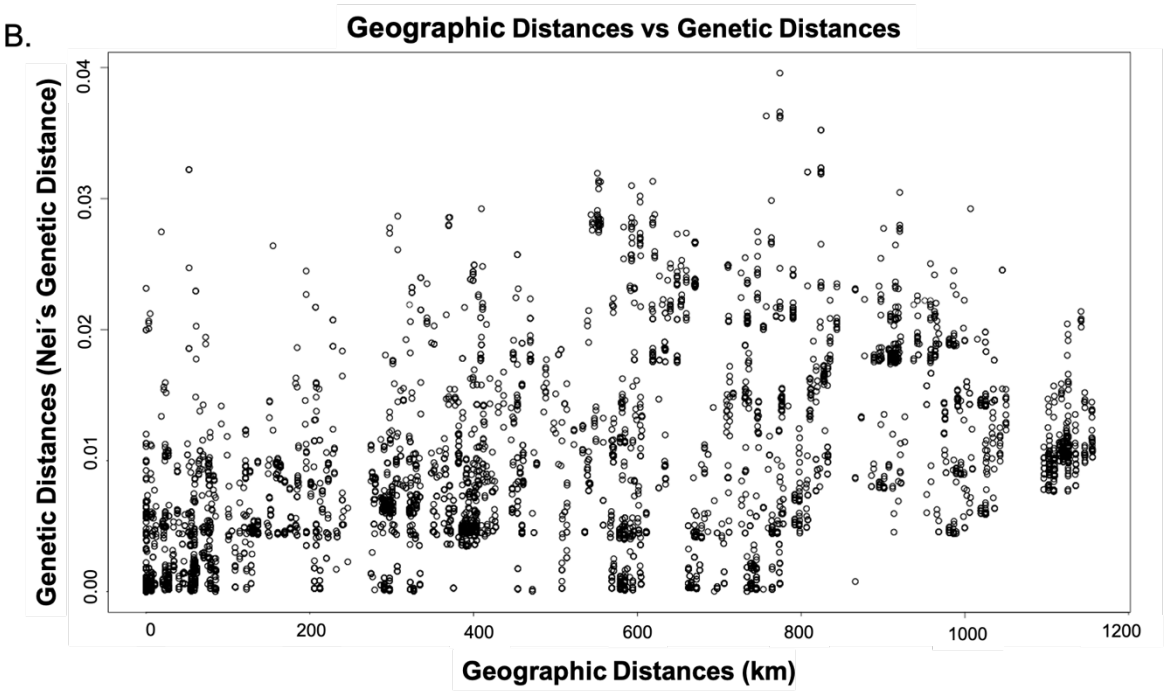


Figura 17. **Prueba de Mantel en un subconjunto de cepas de fermentación de agave. Panel A.** *Heatmap* de distancias genéticas estimadas con el método de Tamura y Nei (1993) de las 97 cepas usadas para llevar a cabo la prueba de Mantel. **Panel B.** Gráfico de dispersión entre las distancias geográficas (eje X) y las distancias genéticas (eje Y) de las 97 cepas utilizadas para llevar a cabo la prueba de Mantel. Se observa una correlación entre la distancia genética y la distancia geográfica de las cepas.

Análisis de estructura poblacional

Para finalizar los análisis de diversidad genética en las cepas de *S. cerevisiae* involucradas en la fermentación de agave, se realizó un estudio de estructura poblacional con el software ADMIXTURE (Alexander D.H. *et al*., 2009). Al igual que con las filogenias y el PCA, se realizaron dos análisis: uno con la matriz de 119,317 SNPs bialélicos sin datos faltantes de las cepas mexicanas de agave que sí incluyen a las 13 cepas que habían sido removidas para los PCA y la Prueba de Mantel (**Tabla 1-Anexo**), y otro con la matriz de 831,175 SNPs bialélicos de cepas mexicanas más los aislados del resto del mundo que no incluyen los 13 aislados previamente eliminados para los PCA y la prueba de Mantel (**Tabla 2-Anexo**).

Análisis de estructura poblacional de cepas mexicanas de agave

En el caso del análisis de ADMIXTURE de las cepas mexicanas, se decidió incluir los subgenomas de *S. cerevisiae* de las 13 cepas que originalmente habían sido removidas porque 12 de ellas son híbridas interespecie y resulta de interés observar cómo sus subgenomas de *S. cerevisiae* se estructuran poblacionalmente. Se utilizó el algoritmo de *cross-validation error* incluido en ADMIXTURE (Alexander D.H. *et al*., 2009) para probar valores de *K*=2 a 17 (**Figura Suplementaria 4A** *K*=2-17) y estimar cuáles serían los valores de *K* idóneos de acuerdo con nuestro set de datos. **La figura suplementaria 4A** muestra que los valores de *K*=6 y *K*=12 son los valores de *K* que más minimizan el valor del error de validación cruzada. Después de una inspección visual de las dos gráficas, se determinó que el gráfico de *K*=12 era el que más congruentemente agrupaba a las cepas en los diferentes grupos. La gráfica de ADMIXTURE (**Figura 18**) muestra tendencias similares a lo observado en el análisis filogenético y en el PCA, por lo que se constata que Peter y colegas (2018) no habían abarcado la mayor parte de la diversidad genómica presente en el grupo de aislados de fermentación de agave, porque siete de las nueves cepas incluidas en su estudio provenían exclusivamente de Tamaulipas. Esto puede observarse cuando las cepas del Noreste se juntan en dos clados aparte con poco mosaicismo genético (**Figura 18-Grupos 9 y 12**). Por ejemplo, se observa que en general hay una estructura genética poblacional bien definida con algunos casos de mosaicismo genómico, principalmente en los grupos 1, 4, 6 y 7. En el caso del grupo 1, tenemos un número considerable de aislados originarios de varios palenques de Michoacán;

el grupo 4 cuenta con cepas de Guerrero en su mayor parte; los grupos 6 y 7 están conformados por cepas de diversos palenques de Oaxaca.

Por otro lado, también se observó que las cepas identificadas como híbridos interespecie se agruparon en grupos particulares de la **Figura 18**: 2,8,10 y 11. El grupo 2 muestra un cierto nivel de mosaicismo en los subgenomas de *S. cerevisiae* de estos híbridos. Por otro lado, los subgenomas de *S. cerevisiae* de los híbridos presentes en el grupo 8,10 y 11 presentan una estructura poblacional bien definida sin rasgos de mosaicismo, lo que sugeriría que las *S. cerevisiae* parentales de estos híbridos estuvieron aisladas y no tuvieron flujo génico con otras *S. cerevisiae* del país y, probablemente, se tengan descendientes cercanos de las *S. cerevisiae* parentales de los híbridos interespecie. Es probable que los eventos de hibridación hayan ocurrido recientemente; sin embargo, se necesitaría comparar a los diferentes híbridos con sus posibles cepas parentales (o descendientes directos de los mismos) para revisar cuántas mutaciones han acumulado. Esto permitiría tener una noción más clara del tiempo que ha transcurrido desde que ocurrieron los eventos de hibridación.

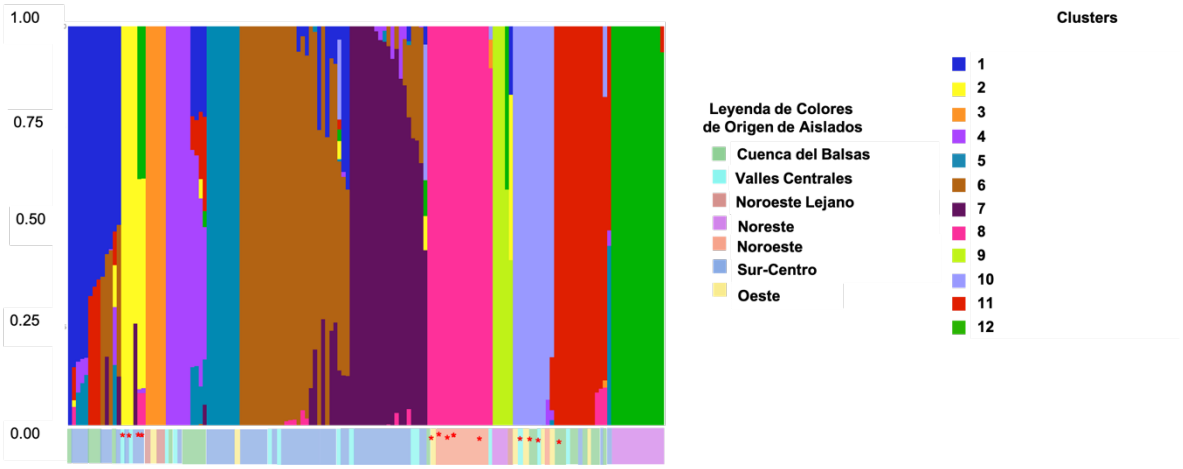


Figura 18. **Estructura poblacional de las cepas de fermentación de agave**. Gráfico de ADMIXTURE con *K*=12 (146 cepas de *S. cerevisiae* o híbridos interespecies *S. cerevisiae-S. paradoxus*) llevado a cabo con 119,317 SNPs bialélicos del genoma nuclear de *S. cerevisiae*. Los sombreados de colores en la parte inferior del gráfico representan las regiones de origen de los aislados mientras que los colores de cada *cluster* vienen señalizados en la leyenda con título "**Clusters**". Los asteriscos rojos marcan a los híbridos interespecie *S. cerevisiae-S. paradoxus.*

Análisis de estructura poblacional de las cepas de agave y las cepas del mundo

Por otro lado, el gráfico de ADMIXTURE (**Figura 19**) de las cepas mexicanas de agave junto con los 170 aislados de varias partes del mundo (Peter J. *et al*., 2018) también se llevó a cabo evaluando *K*=2-17 (**Figura suplementaria 4B**) y, con el algoritmo de *cross-validation error*, se observó que el mejor valor de *K* para este análisis de ADMIXTURE era *K*=12 (**Figura 19**). El gráfico de ADMIXTURE muestra una estructura poblacional definida, con algunos ejemplos de mosaicismo genómico tanto en las cepas de fermentación de agave como en cepas de otras partes del mundo. Las cepas del clado monofilético principal (**Figura 14**) se encuentran distribuidas en los grupos 1,3,4 y 9 respectivamente. La agrupación de los aislados de estos tres grupos correlaciona con su distribución geográfica, lo cual ya había sido descrito en los análisis filogenéticos y en el PCA. En el caso de las siete cepas de Tamaulipas clasificadas como ***Mexican agave*** por Peter y colaboradores (2018) (**Tabla 2 -Anexo**), éstas se agruparon en el grupo 1 (azul) del gráfico (**Figura 18**), junto con otras seis cepas de San Carlos, Tamaulipas, secuenciadas por nosotros. Este patrón complementa lo ya observado en el análisis filogenético y en el PCA, donde se observó que estas cepas se juntan en un subclado y en un grupo alejado del *cluster* principal de aislados, respectivamente.

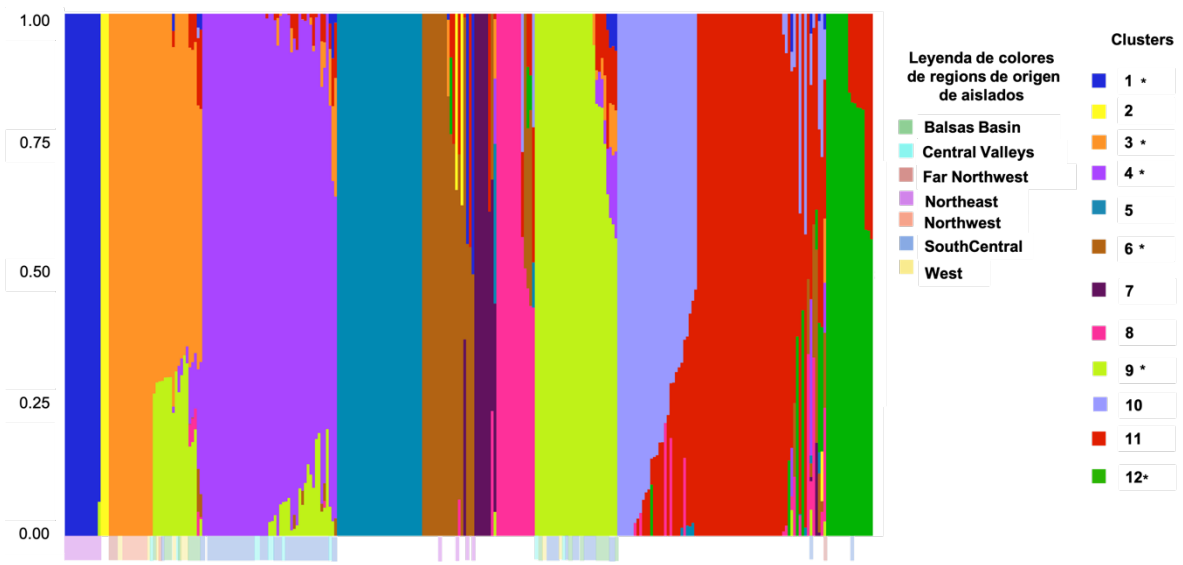


Figura 19. **Estructura poblacional de las cepas de fermentación de agave más las cepas del resto del mundo.** Gráfico de ADMIXTURE con *K*=12 (294 cepas de *S. cerevisiae* -> 124 cepas de agave secuenciadas por nosotros +170 cepas de diversos orígenes analizadas por Peter y colaboradores) llevado a cabo con 831,175 SNPs bialélicos del genoma nuclear de *S. cerevisiae*. Las cepas se posicionaron en los 12 grupos de la siguiente manera: **Grupo 1**.- 7 cepas del clado ***Mexican agave*** más cepas de fermentación de agave de San Carlos, Tamaulipas, **Grupo 2**.- Cepas de los clados ***CHNI, CHNII y Taiwanese***, **Grupo 3**.- Cepas de fermentación de agave de Jalisco, Durango, Zacatecas, Michoacán, Guanajuato y Oaxaca. **Grupo 4**.- Cepas de fermentación de agave de Oaxaca, Puebla y Guanajuato. **Grupo 5**.- Cepas del clado ***French Guiana Human***, **Grupo 6**.- Cepas de los clados ***CHNIII, CHNV, Malaysian, Far East Russian, North American Oak, Ecuadorean, Mosaic Region 3, Far East Asia*** **y** ***Unclustered*** y cuatro cepas de fermentación de agave de San Carlos, Tamaulipas. **Grupo 7**.- Cepas de los clados ***African Palm Wine***, ***Mosaic Region 3 y Unclustered*****. Grupo 8**. Cepas de los clados ***Asian Fermentation, Sake, Asian Islands, Mosaic Region 3 y Unclustered*****. Grupo 9**.- Cepas de fermentación de agave del Edo.Mex, Jalisco, Puebla, Oaxaca y Guerrero. **Grupo 10**.- Cepas de los clados ***Ale Beer, Mixed Origin, Alpechin, Mosaic Region 3, Mosaic Beer, French Dairy, Brazilian Bioethanol y Unclustered.*** **Grupo 11**.- Cepas de los clados ***Wine/European,Alpechin, French Dairy, Mosaic Region 1-3, West African Cocoa y Unclustered*** y cepa de fermentación de agave de Oaxaca y Sonora. **Grupo 12**.- Cepas de los clados ***Mosaic Region 2, Mediterranean Oak, African Beer y Wine/European*** y una cepa de Oaxaca. *Nota El color del sombreado debajo del gráfico de barras corresponde a la región a la que pertenece cada aislado.*

Por otro lado, las cepas JS497c1 (Raicilla, Jalisco) y JS109c1 (Mosto de varias especies de agave de México) se posicionaron dentro de los grupos 3 (**Figura 19-naranja**) y 4 (**Figura 19-Morado**), junto con otras cepas del clado monofilético principal (**Figura 13**), sugiriendo que estas cepas forman parte del grupo principal de aislados de *S. cerevisiae* involucrados en la fermentación de agave. Las cepas XA126c5, XA124c1, XA126c1, XA125c5 de San Carlos, Tamaulipas siguieron la misma tendencia de agruparse en un *cluster* independiente (**Figura 18 Grupo 6-Café**) y llama la atención que los genomas de estos aislados presentan un nivel de mosaicismo considerable. Además, estas cepas se agruparon con aislados

pertenecientes a los clados ***Malaysian, CHNIII, North American Oak, Far East Russian, Ecuadorean, Mosaic Region 3 y Far East Asia.*** La presencia de mosaicismo en los genomas de estas cepas y el hecho de que se posicionen en un grupo diferente con los clados recién mencionados sugiere que estos aislados tienen un origen distinto al del grupo monofilético original. Por último, las cepas XB001c4 (Matatlán, Oaxaca), XB075c7 (Matatlán, Oaxaca) y DS002c10 (Huasabas, Sonora) se posicionaron en los grupos 11 (**Figura 19-Rojo**) y 12 (**Figura 19-Verde**) donde se ubican aislados de los grupos ***Wine/European, Alpechin, Mosaic Regions 1-3, African beer*** y ***Mediterranean Oak***. La XB075c7 forma parte del clado ***Wine/European*** en el árbol filogenético (**Figura 18**) y esta cepa presente en el grupo 12 (**Figura 18-Verde**) tiene un alto componente del grupo 11 (**Figura 18-Rojo**) que es donde se encuentra la mayor cantidad de cepas "puras" del clado ***Wine/European***.

Las cepas de fermentación de agave presentaron una gama amplia de ploidías y aneuploidías en sus genomas

Varios artículos han mostrado que las cepas de *S. cerevisiae* de diferentes lugares y orígenes pueden presentar una gama amplia de ploidías y aneuploidías que, muchas veces, juegan un papel importante en la adaptación de los aislados a distintos medios (Duan S.F. *et al*., 2018; Legras J.L. *et al*., 2018; Peter J. *et al.*, 2018). Por ejemplo, Peter y colaboradores (2018) estimaron las ploidías a través de citometría de flujo y observaron que las cepas presentaban un continuo de ploidías desde 1N hasta 5N. De igual manera, Duan y colegas (2018) mostraron que las

cepas del Oriente Lejano también presentaban distintos números de ploidías, aunque la mayoría de éstas eran diploides.

Por lo tanto, se decidió hacer un análisis de estimación de ploidías de las 146 cepas (*S. cerevisiae* e híbridos *S. cerevisiae*-*S. paradoxus*) de fermentación de agave (137 secuenciadas por nosotros y las nueve reportados por Peter y colaboradores). Si bien es necesario hacer análisis de citometría de flujo y de electroforesis de *clap homogenous electrical field* (CHEF) para tener una noción más certera de la ploidías y aneuploidías de una cepa en particular, se decidió utilizar una estrategia bioinformática (**ver Metodología**) para observar las ploidías y aneuploidías de las cepas de fermentación de agave. La estrategia fue diseñada por Fay y colegas (2019) e involucra las proporciones de lecturas de secuenciación que dan soporte al alelo alternativo y al alelo de referencia en los SNPs heterocigotos (**ver Metodología**). Algunas de las desventajas de este método son su poca capacidad de detectar ploidías en cepas haploides y/o homocigotas (Fay J.C. *et al*., 2019) y de detectar tetraploides donde haya dos cromosomas homocigotos a la misma variante y otros dos cromosomas homocigotos a otra variante (autotetraploides). Para complementar el análisis realizado con el método de Fay y colegas (2019), se decidió hacer un análisis de cobertura normalizada (figura elaborada por el integrante de nuestro grupo Iván Sedeño) de cada uno de los cromosomas. Esta estrategia también nos permite observar con detenimiento si existen aneuploidías en algún cromosoma de un aislado.

Se identificaron 65 cepas diploides (**Ej: Figura 19A**), 25 cepas triploides (**Ej: Figura 20B**), una posible cepa tetraploide y 55 cepas cuyas ploidías no fueron estimadas, porque son cepas con una baja proporción de heterocigosidad y/o son

haploides. De igual manera, llamó la atención que 21 cepas de *S. cerevisiae* presentan una aneuploidía de aumento en número de copias en el brazo derecho del cromosoma 9 (**Figura 21A y 21B**). El aumento del número de copias en el brazo derecho en el cromosoma 9 está presente en más del 10% de las cepas muestreadas, y esto sugiere que la selección natural podría estar promoviendo la duplicación de esta región genómica. Sin embargo, es necesario primero delimitar con exactitud esta región genómica y hacer otro tipo de análisis de selección (D de Tajima, CLR score, H de Fu y Way, entre otros) para comprobar esta aseveración.

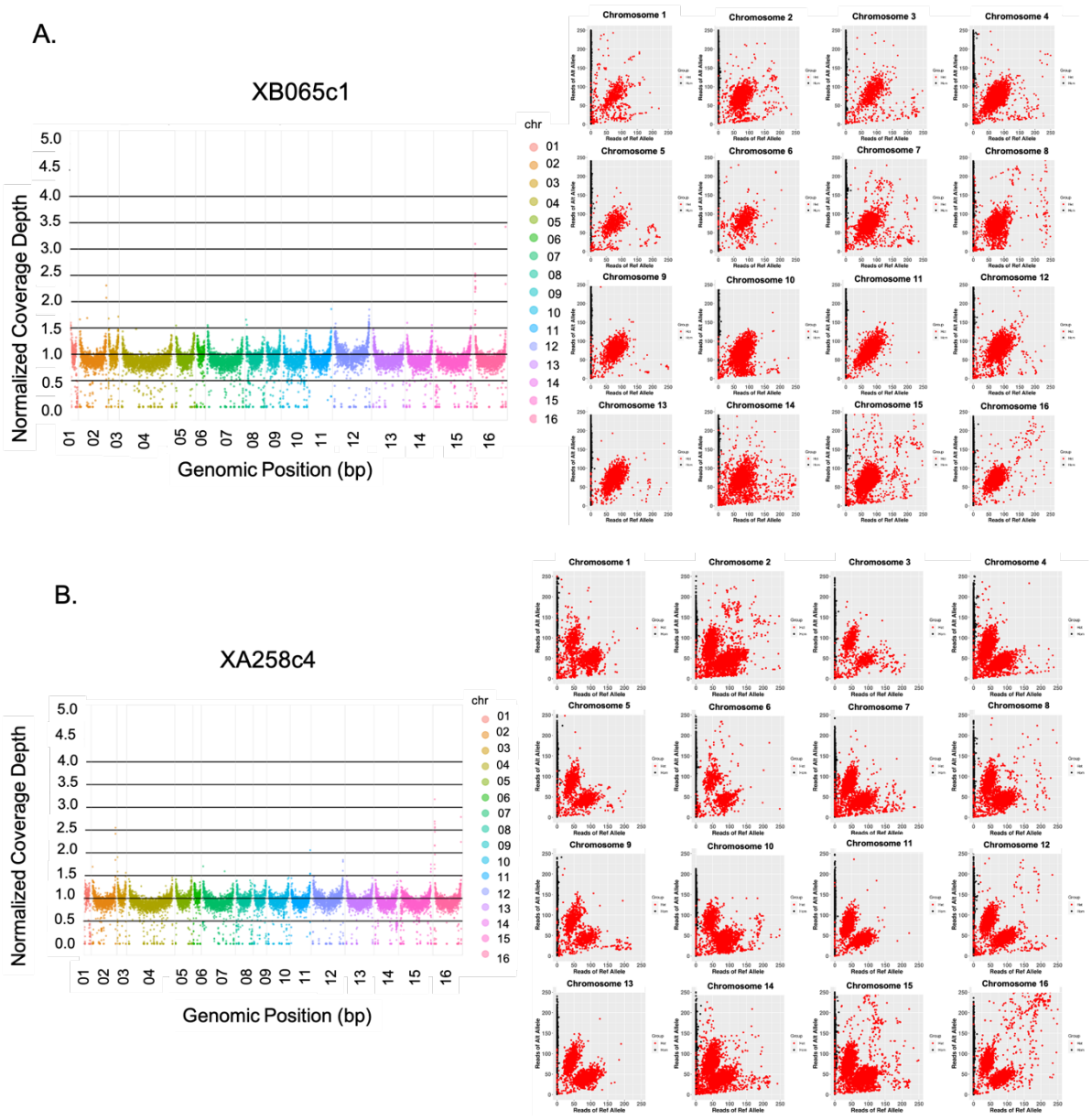


Figura 20. **Plodías de cepas euploides de fermentación de agave.** Gráficas de cobertura normalizada y conteos de lecturas heterocigotas de cepas euploides. El **Panel A** muestra el gráfico de y de conteos de lecturas heterocigotas de la cepa diploide XB065c1 (Santa Catarina. Oaxaca) mientras que el **Panel B** muestra el gráfico de cobertura y de conteos de lecturas heterocigotas de la cepa triploide euploide XA258c4 (Matatlán, Oaxaca). Las gráficas de cobertura fueron realizadas por Iván Sedeño.

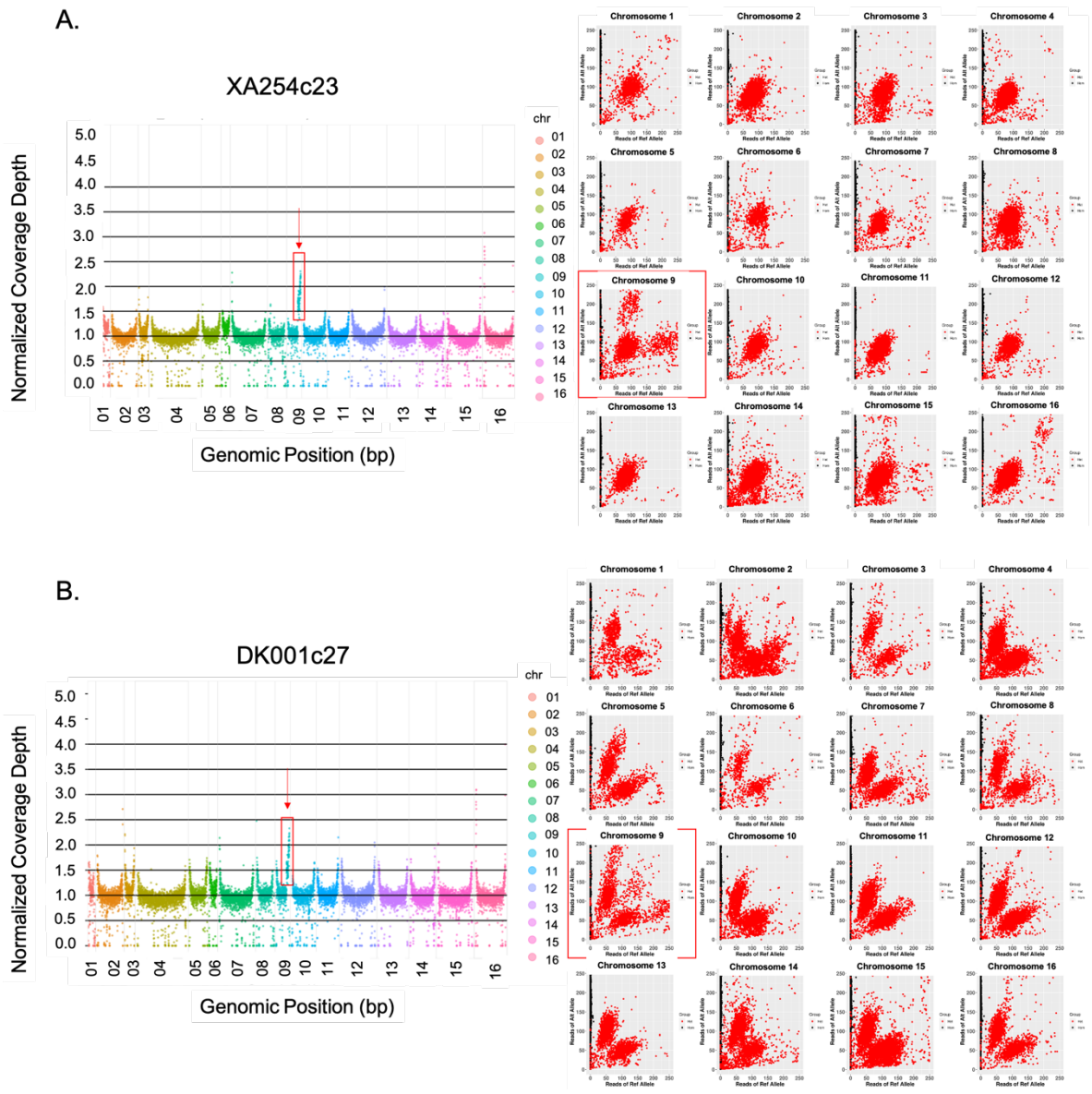


Figura 21. **Aneuploidías en brazo derecho de cromosoma 9 en cepas de fermentación de agave.** Gráficos de cobertura normalizada y conteos de lecturas heterocigotas de cepas con aneuploidía (aumento en número de copias) en el brazo derecho de cromosoma 9. La aneuploidía está marcada con un rectángulo rojo y una flecha en las gráficas de cobertura normalizada mientras que el cromosoma 9 está encerrada en un cuadro rojo en las gráficas de conteo de lecturas heterocigotas. El **Panel A** muestra el gráfico de cobertura y de conteos de lecturas heterocigotas de la cepa diploide aneuploide XA254c23 (Piedras de Lumbre, Michoacán) mientras que el **Panel B** muestra el gráfico de cobertura (Sedeño, Iván) y de conteos de lecturas heterocigotas de la cepa triploide aneuploide DK001c27 (Santiago Matatlán, Oaxaca). Las gráficas de cobertura fueron realizadas por Iván Sedeño.

Adicionalmente, se identificaron otras aneuploidías recurrentes en las cepas de fermentación de agave. Por ejemplo, se observó la presencia de una duplicación en el cromosoma 12 en 10 cepas (**Figura 22A**) de fermentación de agave y posibles deleciones que involucran una región cercana al centrómero del cromosoma 14 en cinco cepas (**Figura 22B)**. En la **Tabla Suplementaria 5**, se describen las observaciones con respecto a aneuploidías en otros cromosomas y regiones con base en los gráficos de cobertura normalizada y el método de Fay y colegas (2019).

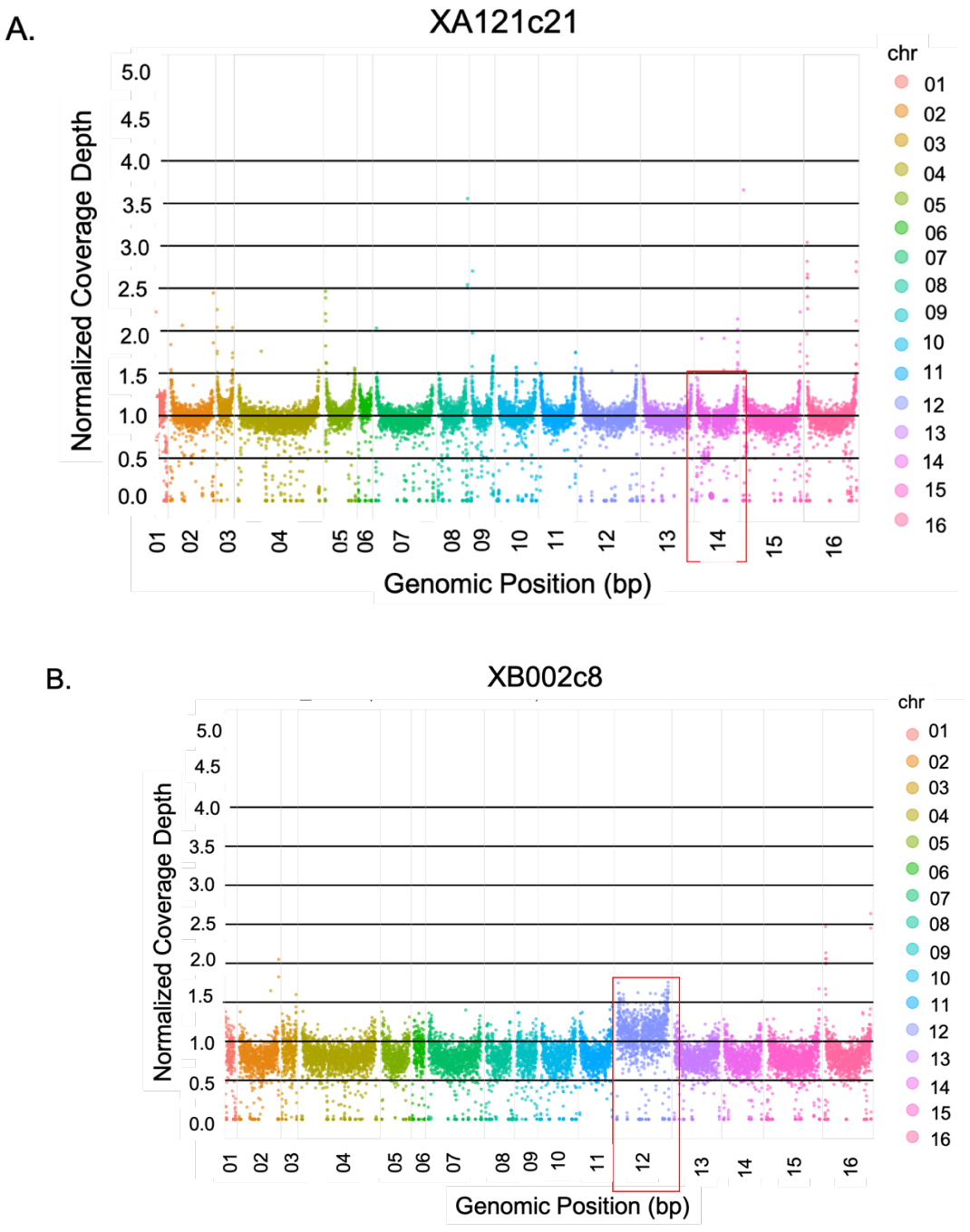


Figura 22. **Aneuploidías recurrentes en cromosoma 12 (duplicación) y cromosoma 14 (deleción).** Gráficos de cobertura normalizada de duplicación del cromosoma 12 (**Panel A**) de la cepa XB002c8 (Guanajuato) y de posible deleción de una región del cromosoma 14 de la cepa XA121c21 (San Carlos, Tamaulipas). La duplicación del cromosoma 12 y la deleción en el cromosoma 14 están marcadas por un rectángulo rojo en sus respectivos gráficos. Imágenes realizadas por Iván Sedeño.

DISCUSIÓN

S. cerevisiae en fermentaciones de agave para producción de mezcal

En este trabajo, se muestreatron 137 cepas de fermentación de agave de siete regiones mezcaleras del país (Aguirre X., *et al*., 2006). De esos 137 aislados, 118 fueron secuenciados con la ayuda de la tecnología DNBSeq™ de BGI y 19 con plataformas de Illumina. En general, se obtuvieron valores de cobertura buenos (**Tabla 1-Anexo**, **Figura 9A**) Además, el precio de secuenciación es 30% menor al de otras plataformas como Illumina. La única desventaja de este método estuvo relacionado con un error por parte de BGI al momento de diseñar las librerías, ya que el tamaño del inserto fue demasiado pequeño. Sin embargo, se comunicó este error a BGI y se comprometieron a no repetir el error.

En recientes años, diversos autores han comenzado a publicar trabajos donde se describen algunas características importantes de la estructura poblacional de distintos subclados de *S. cerevisiae* de alrededor del mundo (Almeida P. *et al*., 2015; Bigey F. *et al*., 2020 Barbosa R. *et al*., 2016; Gallone B. *et* al., 2016; Gallone B. *et* al., 2019; Legras J.L. *et al*., 2018; Peter J. *et al*., 2018; Pontes A. *et al*., 2019). Dichos estudios han descrito características importantes de estas subpoblaciones de levadura, como son su estructura filogenética y poblacional, genes bajo selección y sus procesos demográficos. Adicionalmente, se han comparado a las cepas estudiadas con aislados de otras regiones del mundo con ayuda de herramientas como análisis de componentes principales (PCA) (Legras J.L. *et al*., 2018; Peter J. *et al*., 2018).

En el caso de aislados de levadura involucrados en fermentación de agave, ningún estudio había evaluado la estructura poblacional usando aislados provenientes de las diferentes regiones productoras de mezcal. Flores Berrios E.P. *et al*., (2005) describieron la diversidad genética de algunos aislados de agave con AFLP, que fueron comparadas con los aislados de levaduras de fermentación de vino. Kichmayr y colegas (2017), con análisis de espectometría de masas en proteínas totales que en dos palenques oaxaqueños de mezcal, encontraron que *S. cerevisiae* era la especie más prevalente en distintos tiempos (Kichmayr M.R. *et al*., 2017). Por último, el análisis de 1,011 genomas de *S. cerevisiae* de Peter y colaboradores (2018) incluyó nueve cepas mexicanas de fermentación de agave; siete aisladas de Tamaulipas que fueron agrupadas en un clado denominado ***Mexican Agave***.

Esta tesis evaluó de manera más amplia de la diversidad genética presente en las levaduras de fermentación de agave con la finalidad de observar si características como las diferentes condiciones geográficas o las especies de agave utilizadas en la fermentación pueden estar influyendo en la estructura poblacional de estas cepas. Adicionalmente, las cepas de agave fueron comparadas con aislados pertenecientes a clados de otras partes del mundo. Los resultados de los diferentes análisis realizados en este trabajo muestran que las cepas de *S. cerevisiae* involucradas en la fermentación de agave en México conforman una población genéticamente única y distinta al resto de clados de *S. cerevisiae* de otras regiones del mundo. De igual manera, se observó que la agrupación de las cepas de fermentación de agave en distintos grupos o subclados está influenciada tanto por su ubicación geográfica como por la deriva génica.

Introgresión con *S. paradoxus*

Se evalúo la cantidad de lecturas secuenciadas de cepas de *S. cerevisiae* que mapearon con mayor calidad al genoma de *S. paradoxus*. Un porcentaje considerable de lecturas con mejor mapeo a *S. paradoxus* sugeriría que las cepas de *S. cerevisiae* de fermentación de agave tienen una presencia considerable de introgresiones genómicas de *S. paradoxus*. También se compararon los porcentajes de lecturas mapeadas con buena calidad a *S. paradoxus* YPS138 de los aislados de *S. cerevisiae* de Alpechin (N=14), fermentación de agave (N=133), Guyana Francesa (N=31), bioetanol de Brasil (N=6) y el resto de las cepas analizadas por Peter y colegas (2018) (N=111) (**Figura 8**) no pertenecientes a los grupos recién mencionados. Estos grupos de aislados fueron evaluados porque diversos reportes (D'Angiolo M. *et* al., 2020; Peter J. *et al*., 2018; Pontes A. *et al*., 2019) han descrito que tienen una cantidad considerable de introgresiones provenientes de *S. paradoxus.* La **Figura 8** muestra que un número considerable de cepas mexicanas de fermentación de agave tienen 2-4% de lecturas mapeadas con mejor calidad al genoma de *S. paradoxus*. De hecho, las cepas mexicanas de fermentación de agave son el segundo grupo con mayor cantidad de segmentos genómicos provenientes de *S. paradoxus*, sólo por debajo de las cepas del ***Alpechin***. Previamente, se había descrito que la presencia de introgresiones en cepas naturales (i.e., no modificadas genéticamente) de *S. cerevisiae* contribuía de manera importante en la generación de variación fenotípica y en la adaptación a ciertos nichos (Peter J. & Schacherer J., 2016).

Una de las explicaciones más plausibles para el origen de las introgresiones en levaduras es la hibridación recurrente entre especies del género *Saccharomyces* (Dujon B.A. & Louis E.J., 2017). Esta hibridación ocurre porque son débiles las barreras precigóticas de las especies de este género. Sin embago, las barreras pos-cigóticas son más fuertes y promueven el aislamiento reproductivo entre dos especies *Saccharomyces* diferentes (Marsit S. *et al*., 2017). Por ejemplo, muchas veces las esporas resultantes de este proceso son estériles, debido a que la alta divergencia nucleotídica entre las especies inhibe la recombinación meiótica en los gametos o por incompatibildades genéticas entre genes nucleares y mitocondriales (incompatibilidades Bateson-Dobzhansky-Müller) (D'Angiolo M. *et al*., 2020; Marsit S. *et al*., 2017). Sin embargo, algunas esporas pueden ser viables y posteriormente retrocruzarse con alguna de las dos cepas parentales (D'Angiolo M. *et al*., 2020, Dujon B.A. & Louis E.J., 2017). Posteriormente, es teoricamente más fácil para la cepa realizar retrocruzas con la cepa parental, al grado de disminuir el contenido genómico de uno de sus progenitores, hasta quedarse sólo con algunos fragmentos del genoma de uno de los padres (Dujon B.A. & Louis E.J., 2017). Sin embargo, la presencia de introgresiones de otras especies en *S. cerevisiae* no es común en la mayoría de las subpoblaciones, por lo que se infiere que las introgresiones podrían estar jugando un rol adaptativo en ciertos clados de levaduras (Peter J. *et al*., 2018).

En el caso de las cepas de fermentación donde se ha encontrado un alto contenido de introgresiones de *S. paradoxus*, una ruta propuesta para la población de ***Alpechin*** implica una hibridación ancestral de ambas especies, donde ambos genomas coexisten en la levadura (D'Angiolo M. *et al*., 2020). Posteriormente, la inestabilidad genómica derivado de la coexistencia de genomas muy diferentes

resultaría en la pérdida de heterocigosidad después de muchas generaciones de división mitótica, lo cual ayudaría a que el híbrido recupere su fertilidad y pueda retrocruzarse con la cepa parental de *S. cerevisiae* (D'Angiolo M. *et al*., 2020). D´Angiolo y colegas *(*2020) pudieron demostrar esta ruta experimentalmente gracias a que aislaron un híbrido ancestral *S. cerevisiae* - *S. paradoxus* de la población de ***Alpechin*** (D'Angiolo M. *et al*., 2020)**.** En nuestro caso, se identificaron 13 cepas híbridas *S. cerevisiae* - *S. paradoxus* que son buenos candidatos de ser descendientes directos del/de los híbrido(s) ancestral(es) que dio(eron) lugar a las introgresiones de los aislados de fermentación de agave.

Finalmente, es necesario hacer hincapié en la importancia que tiene la hibridación interespecie entre levaduras del género *Saccharomyces*, ya que pone en entredicho la definición biológica de especie. Estas levaduras híbridas han desarrollado estrategias para lidiar con genomas divergentes y, eventualmente, dar lugar a progenie viable. Las cepas de fermentación de agave tienen alta presencia de introgresiones, y alrededor del 10% son híbridos interespecie, resaltando la prevelencia de estos eventos en ambientes antropogénicos.

Heterocigosis en las cepas mezcaleras de *S. cerevisiae*

Se evaluaron los niveles de heterocigosidad de las cepas de fermentación de agave. Se observó que 60 de 146 aislados de agave tienen un porcentaje menor a 5% de SNPs heterocigotos en sus genomas (**Tabla 1-Anexo**). Esto sugiere que, si bien hay muchas cepas altamente homocigotas, la mayoría de las cepas cuentan con

una alta proporción de sitios heterocigotos en su genoma (~5-58% de SNPs heterocigotos en cada cepa clasificada como heterocigota).

Una probable explicación de esta tendencia estaría relacionada con la alta diversidad nucleotídica presente en *S. cerevisiae* (Peter J. & Schacherer J., 2016). Por ejemplo, Peter y colegas (2018) clasificaron a 509 aislados de los 1,011 como heterocigotos. Ellos, al igual que nosotros, clasificaron a las cepas como heterocigotas si el 5% o más de los SNPs identificados en cada aislado eran heterocigotos. Es importante recalcar que en cepas de *S. cerevisiae* involucradas en actividades humanas, estudios previos han reportado una alta proporción de SNPs heterocigotos (Borneman A.R. *et al*., 2011,Duan S.F. *et al*., 2018; Gallone B. *et al*., 2016; Magwene P.M. *et al*., 2011). Se sabe que *S. cerevisiae* es una especie de levadura altamente asexual y endógama con ciclos sexuales ocasionales (Fischer G. *et al*., 2020; Peter J. *et al*., 2018). En consecuencia, muchos de los linajes de *S. cerevisiae*, especialmente aquellos de origen natural, tienen un nivel bajo de diversidad genética (Peter J. & Schacherer J., 2016). Sin embargo, se sabe que los ambientes altamente estresantes como las fermentaciones pueden aumentar la tasa de exogamia (*outcrossing*) al crear espacios masivos de apareamiento donde coexisten diferentes cepas de una misma especie o diferentes especies o al promover la dispersión de esporas por insectos vectores (Magwene P.M. *et al*., 2011). Además, los eventos de reproducción sexual, a pesar de ser muy raros en estas levaduras, influyen de manera considerabe en la arquitectura y evolución genómica de los linajes de *S. cerevisiae* asexuales facultativos y les

permite adaptarse a ambiente altamente cambiantes en intervalos de tiempo relativamente cortos (Magwene P.M. *et al*., 2011).

Diferenciación geográfica y estructura en *S. cerevisiae*

i) Analisis filogenéticos

El análisis de inferencia filogenética de los 146 aislados de *S. cerevisiae* sugiere que la mayoría de los aislados de *S. cerevisiae* involucrados en la fermentación de agave se agrupa en subclados de acuerdo con su ubicación geográfica, inclusive, cepas de diferentes palenques, pero de la misma región geográfica, tienden a agruparse juntas (**Figura 13A, 13B**). Una observación interesante es la agrupación de cuatro cepas de Tamaulipas (XA126c5, XA124c1, XA126c1 y XA125c5) en un subclado aparte del grupo principal de cepas de agave (**Figura 13B - > Subclado inferior sombreado de naranja**). Esto significa que estas cepas son muy distintas al resto de cepas de agave. Algo similar ocurre con las siete cepas del grupo ***Mexican agave*** (Tamaulipas) evaluadas por Peter y colaboradores (2018), ya que forman otro subclado junto con las cepas XA121c18, XA121c6, XA121c21, XA126c7, XA121c14 y XA121c20 de San Carlos, Tamaulipas (**Figura 13B - > Subclado superior sombreado de naranja**). Este subclado tampoco se encuentra dentro del grupo principal de cepas de fermentación de agave, sugiriendo, probablemente, que las cepas de Tamaulipas han estado aisladas y no han tenido flujo génico con el resto de las *S. cerevisiae* de otras regiones del país o tienen un origen diferente al resto.

Posteriormente, se decidió incorporar a 170 cepas analizadas por Peter y colaboradores (2018) para determinar, a través de una filogenia de máxima

verosimilitud, si el grupo de fermentación de agave, tal como lo describió Peter J. *et al*., (2018), formaba un grupo monofilético o si compartía origen en común con otras subpoblaciones de *S. cerevisiae.* Cuando se incorporaron las 170 cepas del artículo de los 1,011 genomas (incluyendo a las nueve cepas de agave evaluadas por ellos) (Peter J. *et al*., 2018), el árbol filogenético muestra que la mayoría de las cepas de fermentación de agave (N=137) tiene un origen monofilético (**Figura 14A, 14B**). Algunas de las 137 cepas del clado monofilético se agrupan de acuerdo con las regiones geográficas de dónde fueron muestreadas.

Dentro del grupo monofilético, hay un subclado conformado por siete cepas tamaulipecas evaluadas por Peter y colegas (2018) y seis aislados de San Carlos, Tamaulipas, muestreadas por nosotros (**Figura 14A - > subclado morado contiguo al clado 16 de Guyana Francesa**). De manera interesante, este subclado de Tamaulipas es el más cercano al grupo de ***French Guiana*****,** el cual había sido descrito previamente por Peter y colaboradores (2018) como el más cercano al clado ***Mexican Agave*** . Sin embargo, el análisis de esta tesis nos muestra que el artículo de los 1,011 genomas de *S. cerevisiae* (Peter J. *et al*., 2018) denominó como ***Mexican Agave*** a un conjunto de cepas que solamente representa sólo una fracción de la diversidad que alberga el grupo monofilético principal. De hecho, las dos cepas mexicanas sin clasificar del trabajo de Peter y colaboradores (2018) se agrupan en otros subclados dentro del grupo monofilético, sugiriendo que la diversidad de las cepas de fermentación de agave es mayor de lo que originalmente se creía. Por último, nueve aislados de fermentación de agave se agruparon fuera del clado monofilético. La cepa XB075c7 de Oaxaca se agrupa con cepas del grupo de vino (***Wine***), por lo que sería interesante investigar cómo una cepa cercana a los

aislados de vino llegó a una fábrica de mezcal. Por otro lado, cuatro aislados de San Carlos, Tamaulipas (XA126c5, XA124c1, XA126c1 y XA125c5) forman un subclado alejado del grupo monofilético principal junto con dos cepas norteamericanas (JS612c1 y JS235c1) del grupo ***Mosaic Region 3***. Es necesario hacer análisis adicionales de estos cuatro aislados de Tamaulipas para determinar por qué son tan diferentes al resto.

La identificación de cepas en grupos fuera del grupo monofilético principal ha sido observado en otros estudios poblacionales de levadura. Por ejemplo, Bigey y colaboradores (2020) observaron que algunas *S. cerevisiae* de uno de los clados fermentación de pan se agrupan con cepas de cerveza, sugiriendo que probablemente ha existido flujo génico entre ambos grupos, como consecuencia de prácticas humanas. En conclusión, las cepas de fermentación de agave forman un grupo monofilético que los separa de los aislados de otras partes del mundo. Dentro de esta clado monofiletico se distinguen dos grupos: los aislados de Tamaulipas y los del resto del país.

Para poner en el contexto de los 1,011 aislados del mundo a las cepas de este trabajo, se realizó un análisis filogenético con el algoritmo *Neighbor-Joining.* Este analisis permitó apoyar el resutlado de que las cepas originalmente descritas como ***Mexican Agave*** en el artículo de Peter y colegas (2018) forman un subclado divergente dentro del grupo monofiletico de fermentación de agave. Adicionalmente, se volvió a confirmar -ya con más cepas de otras partes del mundo- que el grupo de aislados de *S. cerevisiae* involucrado en la fermentación de agave es un clado bastante distinto al de otras regiones del mundo, lo que sugiere que aún hay mucha diversidad de esta especie por evaluar. México es un país megadiverso donde

coexisten diferentes tipos de ecosistemas, climas, suelos, altitudes, fauna y flora (Sarukhan, J., 2008). Por lo tanto, era de esperarse que el muestreo de Peter y colaboradores (2018) en un sólo estado (Tamaulipas) no fuera representativo de toda la diversidad genética presente en los aislados de *S. cerevisiae* involucrados en la fermentación de agave. Diversos estudios han correlacionado la diferenciación genética de linajes y subpoblaciones de *S. cerevisiae* con el nicho ambiental, las condiciones geográficas y el nivel de asociación con el humano (Peter J. & Schacherer J., 2016). En consecuencia, la enorme diversidad de los factores recién mencionados en México son fundamentales para explicar la diferenciación genética de diferentes aislados localizados en diferentes regiones de producción de mezcal (**Figura 1**).

ii) Análisis de componentes principales

Posteriormente, se llevó a cabo un análisis de componentes principales (PCA), sin tomar en cuenta a 13 cepas (12 híbridas interespecie y una con probable contaminación) y sin datos faltantes en los SNPs. El análisis de sólo 133 cepas de fermentación de agave (124 de nosotros más nueve cepas de Peter y colaboradores) muestra que, dentro del clado monofilético principal de aislados, hay dos grupos principales de cepas de *S. cerevisiae*. El primero (**Figura 16A- óvalo azul**) agrupa a todas aquellas cepas están distribuidas en todo el país, a excepción de Tamaulipas. El segundo grupo está compuesto únicamente por las 13 cepas de Tamaulipas (**Figura 16A- óvalo rojo de la parte superior**): siete utilizadas en el análisis realizado por Peter y colegas (2018) (Illumina HiSeq2000) y nueve

secuenciadas por nosotros de San Carlos, Tamaulipas (BGI DNBSeq™). La diferenciación de este grupo de Tamaulipas con el resto de las cepas del grupo monofilético nos sugiere que, por alguna razón, las poblaciones de *S. cerevisiae* de Tamaulipas divergieron del resto, formando un subclado aparte al resto.

En otros estudios, se ha observado que algunas cepas del mismo grupo pueden variar considerablemente por diversos factores como la variedad de uva en el caso de la fermentación de vino (Schuller D. *et al*., 2012). En el caso de agave, sería importante revisar qué especies de agave se utilizan en Tamaulipas, y si hay alguna otra característica del lugar como suelo, clima, aislamiento geográfico y altitud que pueda estar coadyuvando en la diferenciación genética de estos aislados tamaulipecos de *S. cerevisiae*. Tamaulipas es un estado que, orográficamente, está aislado del resto del país por un segmento de la Sierra Madre Oriental. Resultaría interesante investigar si la divergencia genética de los aislados de esta región puede ser explicada por este hecho. Este fenómeno ya ha sido reportado en otros organismos y se debe, en parte, a que la Sierra Madre Oriental es la cadena montañosa más antigua de México (Mastretta-Yanes A. *et al*., 2015). Los linajes de especies altamente relacionadas entre sí tienden a diferenciarse genéticamente, dependiendo del lugar montañoso donde se encuentran y se han identificado linajes de especies más divergentes en lugares aislados por la Sierra Madre Oriental (Mastretta-Yanes A., *et al*., 2015).

Después de incorporar las 170 cepas de distintas regiones del mundo (Peter J. *et al.*, 2018) al PCA (**Figura 16B**), se observó que las cepas de fermentación de agave forman un grupo aislado del resto de las *S. cerevisiae* de otras partes del

mundo. Los componentes principales 1 y 2 son suficientes para separar y aislar a las cepas de fermentación de agave del resto de aislados del mundo (**Figura 16B -> óvalo verde**). Cabe destacar que, para la realización de este PCA (ver Métodos), la presencia de introgresiones de *S. paradoxus* en las cepas de agave no fue tomada en cuenta. Peter y colaboradores (2018) realizaron un análisis de PCA (Material suplementario de Peter J. *et al*., 2018 -> **Figura S7 Panel b**) tomando en cuenta los ORFs variable entre las 1,011 cepas y también mostraron que los grupos ***French Guiana Human*** y ***Mexican Agave*** forman *clusters* lejanos al resto de las cepas, lo cual no sólo se explica por el número elevado de introgresiones de *S. paradoxus* como se inferiría.

De forma interesante, el PCA complementa el resultado de la inferencia filogenética (**Figura 14A, 14B**), ya que podemos observar que 13 cepas de Tamaulipas (nueve secuenciadas por nosotros +siete del trabajo de Peter y colaboradores) están alejadas (**Figura 16B -> óvalo morado**) del grupo monofilético principal. Estos datos también sugieren que las subpoblaciones de levaduras de Tamaulipas son divergentes al resto de cepas del país. Por otro lado, las dos cepas mexicanas de agave analizadas por Peter y colegas (2018) (JS109c1 y JS497c1) están embebidas en el grupo monofilético principal (**Figura 16B -> flechas azules**), lo que indica que las cepas originalmente reportadas como ***Mexican agave*** en el artículo de los 1,011 genomas solamente conforman una pequeña porción de un clado de cepas de *S. cerevisiae* más grande y diverso. Por último, también fue interesante ver que algunos aislados aparecen afuera del grupo principal de *S. cerevisiae* de fermentación de agave y son las mismas cepas que no se agrupan en el clado monofilético en el árbol filogenético (**Figura 16A, 16B**).

iii) Prueba de Mantel

Después de observar la correlación entre los patrones de agrupación de las cepas de agave y su ubicación geográfica, se decidió llevar a cabo una prueba de Mantel para evaluar si la distancia genética de los aislados de agave está relacionada con su distancia geográfica. Los resultados de Mantel confirman que existe una correlación significativa entre el aumento de distancia geográfica y el incremento en distancia genética entre las cepas. Esto podría explicarse por dos escenarios no mutuamente excluyentes: que procesos de evolución neutral como la deriva génica estaría jugando un rol importante en la diferenciación genética de las cepas de fermentación o que procesos selectivos y adaptativos locales estarían moldeando la variación genética de las diferentes subpoblaciones de acuerdo con las características ambientales de cada región. Esto tiene sentido, ya que los productores no inoculan con alguna cepa predeterminada y la diferenciación de los aislados estaría influenciada por características propias de cada lugar. Schuller & Casal (2007) y Schuller y colaboradores (2012) reportaron que existían otros factores, como la variedad de uva, que influían más que la distancia geográfica en la diferenciación de cepas de *S. cerevisiae* involucradas en fermentación de vino en Portugal. En el caso de nuestro estudio, es necesario extender el análisis a más cepas, utilizar estadísticos como F_{ST} para evaluar diferenciación genética y revisar si otros factores como la especie de agave, el clima y/o las condiciones de fermentación están influyendo en la diferenciación de las subpoblaciones de *S. cerevisiae* de fermentación de agave.

iv) Análisis ADMIXTURE

En aras de complementar los análisis filogenéticos y el PCA, se decidió evaluar la estructura poblacional de los aislados de fermentación de agave. Primero, se hizo un análisis con 146 cepas de fermentación de agave sin datos faltantes en los SNPs y, con base en el algoritmo de *Cross-validation error* incluido en el programa ADMIXTURE y en una inspección de los gráficos de barra de ancestría, se decidió que el número de *K*=12 (**Figura 18**) era el que mejor explicaba nuestros datos. Der manera interesante, también se identificó una correlación entre la agrupación en *clusters* y el origen de las cepas. En el caso de Tamaulipas (**Figura 18 -> Grupo 12**), se observó que las siete cepas de ***Mexican Agave*** (Peter J. *et al*., 2018) y seis cepas de San Carlos, Tamaulipas son miembros de un grupo aislado, lo cual abona al resto de la evidencia sobre el aislamiento de estas *S. cerevisiae* con respecto a las otras cepas del México.

Por otro lado, el mosaicismo -diferentes componentes de ancestría en algunos aislados- presente en algunos grupos (**Figura 18 -> Grupos 1,4,6 y 7**) indicaría que han existido algunos eventos de flujo génico entre grupos de diferentes regiones del país. La presencia de *admixture* en cepas relacionadas a procesos industriales ha sido ampliamente descrita en la literatura. Por ejemplo, Tilarkaratna & Bensasson (2017) encontraron altos niveles de *admixture* en *S. cerevisiae* de hábitats relacionados íntimamente al ser humano. En este caso, identificaron mayor entrecruzamiento entre cepas de fermentación de vino que en cepas de bosques de robles (Tilarkaratna V. & Bensasson D., 2017). De igual manera, Peter y colaboradores (2018) identificaron la mayor cantidad de *admixture* en cepas de ambientes humanos. También se incluyeron los híbridos interespecie (**Figura 18 ->**

flechas rojas) para tener una noción más clara de cuáles podrían ser las *S. cerevisiae* mayormente relacionadas con el subgenoma de *S. cerevisiae* de los híbridos. Se observó que los subgenomas de *S. cerevisiae* de estos híbridos no parecieran contener un nivel alto de mosaicismo. Un ejemplo consiste en las cepas híbridas del grupo 10 (**Figura 18**), las cuales se ven homogéneas con los aislados no híbridos. Otro aspecto importante es el hecho de observar que los híbridos se encuentran distribuidos en distintos *clústers* y esto es algo que el análisis filogenético también muestra. El surgimiento de híbridos en distintos subclados localizados en regiones diferentes nos hace pensar que han existido varios eventos de hibridación a lo largo de la historia evolutiva de las *S. cerevisiae* involucradas en la fermentación de agave. Como ya se mencionó, la alta presencia de introgresiones en todo el grupo, sin importar región u origen, sugiere que los eventos de hibridación fueron recurrentes a lo largo del tiempo.

Una vez incluidas las 170 cepas de otras regiones del mundo, se realizó el análisis de ADMIXTURE con la finalidad de observar la relación que guardan los aislados mexicanos de fermentación de agave con *S. cerevisiae* de diversas regiones del planeta. Se observaron algunas tendencias. Por ejemplo, las cepas de ***French Guiana Human*** forman el grupo 5 (**Figura 19**) y parecieran ser un conjunto de *S. cerevisiae* que han estado aisladas del resto. Peter y colegas (2018) también realizaron un análisis con ADMIXTURE y en todas las *K*s evaluadas (*K*=2-17), el grupo de ***French Guiana Human*** también forma un grupo aislado del resto (Material suplementario de Peter J. *et al*., 2018 -> **Figura S8**). De manera interesante, las siete cepas del grupo ***Mexican Agave***, en la mayoría de los las *K*s evaluadas, presentan, según Peter y colegas (2018), un 50% de componente de ancestría de

French Guiana, y por ello, habían propuestos que ambos clados eran muy cercanos entre sí. Esta relación de ancentría podría explicarse por la presencia de introgresiones compartidas de *S. paradoxus*. Sin embargo, nuestro análisis no muestra que el componente de ancestría de las cepas de **French Guiana** forme parte de algunos de los grupos donde están metidos aislados de fermentación de agave. Esto podría explicarse porque no se tomó en cuenta el componente de *S. paradoxus* al momento de realizar nuestro análisis (ver **Métodos**). En nuestro estudio, las siete cepas de ***Mexican Agave*** forman un grupo aparte con otros seis aislados de San Carlos, Tamaulipas, lo que también sugiere que las subpoblaciones de *S. cerevisiae* de Tamaulipas son diferentes al resto. Por último, se puede observar que algunas cepas no pertenecientes al grupo monofilético principal (**Figura 19 -> asteriscos rojos**) de fermentación de agave están distribuidas en distintos grupos. Cinco cepas llamaron la atención: XB075c7 (Grupo 12) porque tiene componente de ancestría de vino y es la misma cepa que se agrupó con el clado de ***Wine*** en el árbol filogenético (**Figura 19A,19B**) y XA126c5, XA124c1, XA126c1 y XA125c5 porque son cuatro cepas de San Carlos, Tamaulipas que forman un grupo independiente en la **Figura 18**. Esta evidencia se suma a la descrita anteriormente con respecto a una divergencia genética marcada en estos cuatro aislados de San Carlos, Tamaulipas.

Ploidías y aneuploidías de las cepas de fermentación de agave

Se utilizó el método de Fay y colegas (2019) para estimar bioinformáticamente la ploidía de cada una de las cepas estudiadas. Como se mencionó anteriormente, este método sirve para estimar las ploidías de cepas heterocigotas, únicamente.

Además de determinar la ploidía, este método nos permite identificar aneuploidías en cromosomas o regiones genómicas específicas. La mayoría de los aislados de fermentación de agave cuya ploidía pudo determinarse por este método son diploides (**Figura 20A**) (65 de 91 cepas o 71.4%). La prevalencia de aislados diploides es esperada debido a que la diploidía es una condición que parece conferir una ventaja adaptativa a *S. cerevisiae* sin importar las condiciones en las que se encuentre. De hecho, el trabajo de Peter y colegas (2018) identificó que 68.6% de las cepas son diploides, esto puede explicarse, en parte, por una ventaja que representa la diploidía para el crecimiento mitótico de las cepas, sin importar su origen (Gerstein A.C. *et al*., 2006).

Los cambios de ploidía y la presencia de aneuploidías podrían tener un efecto benéfico para la adecuación (*fitness*) de las cepas de *S. cerevisiae* en ambientes muy cambiantes con altos niveles de estrés (Gilchrist C. & Stelkens R., 2019). Por ejemplo, se ha reportado que altos contenidos de etanol propician la aparición de una duplicación total o parcial del cromosoma 3 de *S. cerevisiae* (Gilchrist C. & Stelkens R., 2019, Morard M. *et al*., 2019; Voordeckers K. *et al*., 2015). Adicionalmente, otros eventos de aneuploidización fomentados por los niveles altos de alcohol en el recipiente han sido observados en diferentes cepas industriales de fermentación (Gilchrist C. & Stelkens R., 2019; Gorter de Vries A.R. *et al*., 2017). En el caso de las cepas de fermentación de agave, éstas tienen que lidiar con un estrés adicional: el alto contenido de saponinas en el jugo de agave (Alcázar-Valle M. *et al*., 2019). Las saponinas son un compuesto utilizado por algunas plantas para inhibir el crecimiento de eucariontes patógenos, como algunas levaduras (Alcázar-Valle M. *et al*., 2019). Alcázar-Valle M. *et al*. (2019) encontraron que algunas cepas

de *S. cerevisiae* no podía crecer adecuadamente en los jugos de *Agave durangensis* y *A. salmiana* porque son las especies con mayor contenido de saponinas.

En el presente estudio, identificamos algunas aneuploidías recurrentes, como la duplicación del brazo derecho del cromosoma 9 (**Figura 21**). Resultaría interesante evaluar si la presencia de esta aneuploidía estaría incrementando la adecuación de las cepas frente al exceso de saponinas o algún otro presente en el jugo de agave cocido, sin descatar alguna otra posibilidad. Por ejemplo, Gilchrist C. & Stelkens R., (2019) evaluaron aneuploidías reportadas en ocho artículos distintos y observaron una correlación negativa entre el tamaño del cromosoma y el número de aneuploidías presente en el mismo. El cromosoma 9 es el cuarto cromosoma más chico de *S. cerevisiae* y el segundo con mayor número de aneuploidías detectadas (Gilchrist C. & Stelkens R., 2019), por lo que no se podría descartar que la aneuploidía del cromosoma 9 sea resultado de una inestabilidad inherente del mismo.

CONCLUSIONES

En este trabajo se evaluó la estructura poblacional de los aislados de *S. cerevisiae* involucrados en la fermentación de agave en distintas regiones del país. Se identificó a este clado como un grupo de levaduras genéticamente único y diferente a otras subpoblaciones de *S. cerevisiae* previamente estudiadas, como los clados de vino, sake, cerveza, pan, bioetanol, encinos, entre otros.

Por un lado, se observó que las cepas de fermentación de agave tienen algunas características en común con otros aislados involucrados en procesos industriales. Por ejemplo, se observaron incrementos en la ploidía de algunas cepas y presencia recurrente de aneuploidías, en particular la duplicación del cromosoma 9. Se ha reportado en la literatura la presencia recurrente de aneuploidías en cepas industriales de *S. cerevisiae* porque, al parecer, este fenómeno genético les permite lidiar con estreses altos y cambiantes propios de ambientes íntimamente relacionados con el hombre (Gilchrist C. & Stelkens R., 2019). En nuestro caso, pareciera que las cepas de fermentación de agave también estarían generando aneuploidías, que tendrían un rol adaptativo en un contexto de fermentación de agave. Sin embargo, es necesario realizar otros análisis (identificación de outliers, *Fst*, D de Tajima, entre otros) y experimentos para revisar si algún gen/región dentro de las aneuploidías podrían jugar un rol en la adaptación.

Por otro lado, se realizaron diversos análisis de agrupación por diversidad genética (análisis filogenético, PCA y ADMIXTURE) con la finalidad de tener una noción sobre las semejanzas/diferencias entre estas cepas y con respecto a subpoblaciones de *S. cerevisiae* de otras partes del mundo. Los tres análisis dieron

resultados consistentes sobre un origen monofilético en común para la mayoría de las cepas de fermentación de agave.

Las cepas de Tamaulipas --tanto del clado monofilético como externas -- parecieran ser diferentes al resto de aislados de otras regiones del país. Curiosamente, las *S. cerevisiae*, originalmente descritas como ***Mexican Agave*** por Peter J. *et al*. (2018), son originarias de Tamaulipas y se agruparon con otras seis cepas de San Carlos, Tamaulipas en un subclado dentro del grupo monofilético. Una tendencia similar fue observada en el PCA y los análisis de estructura poblacional con ADMIXTURE. Esta evidencia apunta a una subrepresentación (sesgo hacia la región del Noreste) importante en la evaluación original del grupo ***Mexican Agave*** llevado a cabo por Peter y colegas (2018) y este trabajo describe, por primera vez, al gradiente de diversidad genómica presente en un grupo de aislados de *S. cerevisiae* pobremente descrito con anterioridad. Seguramente, subrepresentaciones similares podrían estar ocurriendo con otros clados descritos por Peter J. *et al*., (2018). Los diferentes análisis apuntan a que México, como país megadiverso, es hogar de una población diversa (se realizó un análisis de estimación global de $\pi = 2.1 \times 10^{-3}$) de *S. cerevisiae* involucrada en fermentación de agave. Si bien los análisis realizados en este trabajo deben afinarse y mejorarse para obtener resultados aún más confiables, la evidencia presentada en esta tesis sirve como punto de arranque para afirmar que las cepas de fermentación de agave forman un grupo genético distintos de otras poblaciones de *S. cerevisiae* previamente estudiadas.

PERSPECTIVAS

Este trabajo es un punto de partida en el estudio de las levaduras involucradas en la fermentación de agave, existen implicaciones importantes de nuestra investigación:

Primero, nos permite saber que, probablemente, las características físicas y del proceso de fermentación en cada región juegan un papel importante en la diferenciación genética de estos aislados.

Segundo, se confirma que el grupo de fermentación de agave tiene un alto nivel de introgresiones de *S. paradoxus,* lo que sugiere que procesos de hibridación en el pasado habrían jugado un papel trascendental en el proceso de domesticación de este grupo de levaduras.

Tercero, se apoya la idea que las levaduras *Sacccharomyces* son un muy buen modelo para estudiar estabilidad genómica, hibridación y evolución en intervalos de tiempo relativamente cortos.

Finalmente, los estudios derivados de esta tesis servirán para comunicarle a los productores de mezcal algunas características de las levaduras que participan en la fermentación de agave, como su origen, capacidad de producción de etanol, resistencia a compuestos como las saponinas, entre otras.

Algunas perspectivas de este trabajo serían:

- Volver a correr los análisis mapeando solamente al genoma de *S. cerevisiae* y contrastar resultados. El mapeo exclusivo al genoma de *S. cerevisiae* nos permitiría identificar y tomar en cuenta a variantes que estén íntimamente

relacionados con la presencia de introgresiones provenientes de *S. paradoxus.*

- Volver a realizar el análisis filogenético de las cepas de *S. cerevisiae* involucradas en la fermentación de agave enraizando con *S. paradoxus* y corrigiendo por sesgo muestral (*ascertainment bias*). Esta corrección, recomendada por los autores de RaXML, debe hacerse porque se usan datos de SNPs (sitios únicamente variables).
- "Adelgazar" las matrices de SNPs, para filtrar por desequilibrio de ligamiento (*pruning*) y filtrar por ancestría (*relatedness*), es decir, quitar cepas que sean muy parecidas entre sí y sólo dejar un representante de cada subgrupo. Este paso es necesario para obtener resultados más confiables en el PCA (con nuestros aislados de agave más 1,011 del artículo de Peter y colegas) y el análisis de estructura poblacional por ADMIXTURE.
- El análisis de ADMIXTURE debe volverse a correr con 100 repeticiones para hacer un promedio de los resultados o elegir aquel con el mejor valor de verosimilitud.
- Estimar valores de π, F_{ST} y correr prueba de Mantel con información de más cepas. En el caso de prueba de Mantel, es necesario contar con información de más cepas, para corroborar que la correlación entre ambas matrices se mantiene y, de igual manera, se podrían llevar un cabo pruebas de Mantel seccionadas por región.
- Estimar D de Tajima, H de Fay y Wu y omega (dn/ds) para medir selección en las cepas de fermentación de agave.

Posteriormente, los resultados obtenidos de este trabajo pueden dar pie a nuevas investigaciones:

- Determinar los ORFs introgresados de *S. paradoxus* y determinar su rol en la adaptación a la fermentación de agave, así como el linaje de *S. paradoxus* de dónde provienen estas introgresiones.
- Determinar y confirmar ploidías/aneuploidías experimentalmente mediante FACS y CHEF.
- Secuenciar con tecnología de secuenciación de lecturas largas para determinar presencia de variantes estructurales en los genomas de las cepas de fermentación de agave.

LITERATURA CITADA

Aguirre, X., C. Illsley, y J. Larson. 2006. Dulce semblanza de los mezcales del Altiplano y del Balsas. México Desconocido 352:36-45.

Alcazar-Valle, M., Gschaedler, A., Gutierrez-Pulido, H., Arana-Sanchez, A., & Arellano-Plaza, M. (2019). Fermentative capabilities of native yeast strains grown on juices from different Agave species used for tequila and mezcal production. *Brazilian Journal of Microbiology*, *50*(2), 379-388.

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

Almeida, P., Barbosa, R., Zalar, P., Imanishi, Y., Shimizu, K., Turchetti, B., Legras, J. L., Serra, M., Dequin, S., Couloux, A., Guy, J., Bensasson, D., Gonçalves, P., & Sampaio, J. P. (2015). A population genomics insight into the Mediterranean origins of wine yeast domestication. *Molecular ecology*, *24*(21), 5412-5427.

Alsammar, H., & Delneri, D. (2020). An update on the diversity, ecology and biogeography of the Saccharomyces genus. *FEMS yeast research*, *20*(3), foaa013.

Barbosa, R., Almeida, P., Safar, S. V., Santos, R. O., Morais, P. B., Nielly-Thibault, L., Leducq, J. B., Landry, C. R., Gonçalves, P., Rosa, C. R., & Sampaio, J. P. (2016). Evidence of natural hybridization in Brazilian wild lineages of Saccharomyces cerevisiae. *Genome Biology and Evolution*, *8*(2), 317-329.

Bigey, F., Segond, D., Friedrich, A., Guezenec, S., Bourgais, A., Huyghe, L., Agier, N., Nidelet, T., & Sicard, D. (2020). Evidence for two main domestication trajectories in Saccharomyces cerevisiae linked to distinct bread-making processes. *Current Biology*.

Borneman, A. R., Desany, B. A., Riches, D., Affourtit, J. P., Forgan, A. H., Pretorius, I. S., Egholm, M., & Chambers, P. J. (2011). Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of Saccharomyces cerevisiae. *PLoS Genet*, *7*(2), e1001287.

Bougeard, S., & Dray, S. (2018). Supervised multiblock analysis in R with the ade4 package. *Journal of statistical software*, *86*(1), 1-17.

Chapman, B., & Chang, J. (2000). Biopython: Python tools for computational biology. *ACM Sigbio Newsletter*, *20*(2), 15-19.

Charron, G., Marsit, S., Hénault, M., Martin, H., & Landry, C. R. (2019). Spontaneous whole-genome duplication restores fertility in interspecific hybrids. *Nature communications*, *10*(1), 1-10.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., Hoon M. J. L., & De Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422-1423.

D'Angiolo, M., De Chiara, M., Yue, J. X., Irizar, A., Stenberg, S., Persson, K., Llored, A., Barré, B., Schacherer, J., Marangoni, R., Gilson, E., Warringer, J., & Liti, G. (2020). A yeast living ancestor reveals the origin of genomic introgressions. *Nature*, *587*(7834), 420-425.

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., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158.

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature methods*, *9*(8), 772-772.

De Chiara, M., Friedrich, A., Barré, B., Breitenbach, M., Schacherer, J., & Liti, G. (2020). Discordant evolution of mitochondrial and nuclear yeast genomes at population level. *BMC biology*, *18*, 1-15.

De la Torre González, F. J., Avendaño, D. O. G., Mathis, A. C. G., & Kirchmayr, M. R. (2018). Evaluation of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for differentiation of Pichia kluyveri strains isolated from traditional fermentation processes. *Rapid Communications in Mass Spectrometry*, *32*(17), 1514-1520.

de Vries, A. R. G., Pronk, J. T., & Daran, J. M. G. (2017). Industrial relevance of chromosomal copy number variation in Saccharomyces yeasts. *Applied and Environmental Microbiology*, *83*(11).

Di Paola, M., Meriggi, N., & Cavalieri, D. (2020). Applications of wild isolates of Saccharomyces yeast for industrial fermentation: the gut of social insects as niche for yeast hybrids' production. *Frontiers in Microbiology*, *11*.

Diniz-Filho, J. A. F., Soares, T. N., Lima, J. S., Dobrovolski, R., Landeiro, V. L., Telles, M. P. D. C., Rangel, T. F., & Bini, L. M. (2013). Mantel test in population genetics. *Genetics and molecular biology*, *36*(4), 475-485.

Dori-Bachash, M., Shema, E., & Tirosh, I. (2011). Coupled evolution of transcription and mRNA degradation. *PLoS Biol*, *9*(7), e1001106.

Duan, S. F., Han, P. J., Wang, Q. M., Liu, W. Q., Shi, J. Y., Li, K., Zhang, X. L., & Bai, F. Y. (2018). The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. *Nature communications*, *9*(1), 1-13.

Dujon, B. A., & Louis, E. J. (2017). Genome diversity and evolution in the budding yeasts (Saccharomycotina). *Genetics*, *206*(2), 717-750.

Fay, J. C., Liu, P., Ong, G. T., Dunham, M. J., Cromie, G. A., Jeffery, E. W., Ludlow, C. L., & Dudley, A. M. (2019). A polyploid admixed origin of beer yeasts derived from European and Asian wine populations. *PLoS biology*, *17*(3), e3000147.

Fischer, G., Liti, G., & Llorente, B. (2021). The budding yeast life cycle: More complex than anticipated? *Yeast*, *38*(1), 5-11.

Francis, R. M. (2017). pophelper: an R package and web app to analyse and visualize population structure. *Molecular ecology resources*, *17*(1), 27-3

Flores Berrios, E. P., Alba González, J. F., Arrizon Gavino, J. P., Romano, P., Capece, A., & Gschaedler Mathis, A. (2005). The uses of AFLP for detecting DNA polymorphism, genotype identification and genetic diversity between yeasts isolated from Mexican agave-distilled beverages and from grape musts. *Letters in applied microbiology*, *41*(2), 147-152.

Gabaldón, T. (2020). Hybridization and the origin of new yeast lineages. *FEMS Yeast Research*, *20*(5), foaa040.

Gallone, B., Mertens, S., Gordon, J. L., Maere, S., Verstrepen, K. J., & Steensels, J. (2018). Origins, evolution, domestication and diversity of Saccharomyces beer yeasts. *Current opinion in biotechnology*, *49*, 148-155.

Gallone, B., Steensels, J., Mertens, S., Dzialo, M. C., Gordon, J. L., Wauters, R., Thesseling, F. A., Bellinazzo, F., Saels, V., Herrera-Malaver, B., Prahl, T., White, C., Hutzler, M., Meussdoerffer, F., Malcorps, P., Souffriau, B., Daenen L., Baele, G., Maere, S., & Verstrepen, K. J. (2019). Interspecific hybridization facilitates niche adaptation in beer yeast. *Nature Ecology & Evolution*, *3*(11), 1562-1575.

Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B., Merlevede, A., Roncoroni, M., Voordeckers, K., Miraglia, L., Teiling, C., Steffy, B., Taylor, M., Schwartz, A., Richardson, T., White, C., Baele, G., Maere, S., & Verstrepen, K. J. (2016). Domestication and divergence of Saccharomyces cerevisiae beer yeasts. *Cell*, *166*(6), 1397-1410.

Gerstein, A. C., Chun, H. J. E., Grant, A., & Otto, S. P. (2006). Genomic convergence toward diploidy in Saccharomyces cerevisiae. *PLoS genetics*, *2*(9), e145.

Giannakou, K., Cotterrell, M., & Delneri, D. (2020). Genomic adaptation of Saccharomyces species to industrial environments. *Frontiers in Genetics*, *11*.

Gilchrist, C., & Stelkens, R. (2019). Aneuploidy in yeast: Segregation error or adaptation mechanism? *Yeast*, *36*(9), 525-539.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Phiippsen, P., Tettelin, H., & Oliver, S. G. (1996). Life with 6000 genes. *Science*, *274*(5287), 546-567.

Gonçalves, M., Pontes, A., Almeida, P., Barbosa, R., Serra, M., Libkind, D., Hutzler, M., Gonçalves, P., & Sampaio, J. P. (2016). Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. *Current Biology*, *26*(20), 2750-2761.

Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, *52*(5), 696-704.

http://broadinstitute.github.io/picard/

Kassambara, A., & Mundt, F. (2017). Package 'factoextra'. *Extract and visualize the results of multivariate data analyses*, *76*.

Kirchmayr, M. R., Segura-García, L. E., Lappe-Oliveras, P., Moreno-Terrazas, R., de la Rosa, M., & Mathis, A. G. (2017). Impact of environmental conditions and process modifications on microbial diversity, fermentation efficiency and chemical

profile during the fermentation of Mezcal in Oaxaca. *LWT-Food Science and Technology*, *79*, 160-169.

Korostin, D., Kulemin, N., Naumov, V., Belova, V., Kwon, D., & Gorbachev, A. (2020). Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing. *Plos one*, *15*(3), e0230301.

Legras, J. L., Galeote, V., Bigey, F., Camarasa, C., Marsit, S., Nidelet, T., Sanchez, I., Couloux, A., Guy, J., Franco-Duarte, R., Marcet-Houben, M., Gabaldon, T., Schuller, D., Sampaio J. P., & Dequin, S. (2018). Adaptation of S. cerevisiae to fermented food environments reveals remarkable genome plasticity and the footprints of domestication. *Molecular biology and evolution*, *35*(7), 1712-1727.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987-2993.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Liti, G., Barton, D. B., & Louis, E. J. (2006). Sequence diversity, reproductive isolation and species concepts in Saccharomyces. *Genetics*, *174*(2), 839-850.

Liti, G., Warringer, J., & Blomberg, A. (2017). Isolation and laboratory domestication of natural yeast strains. *Cold Spring Harbor Protocols*, *2017*(8), pdb-prot089052.

Liu, L., Redden, H., & Alper, H. S. (2013). Frontiers of yeast metabolic engineering: diversifying beyond ethanol and Saccharomyces. *Current opinion in biotechnology*, *24*(6), 1023-1030.

López-Romero, J. C., Ayala-Zavala, J. F., González-Aguilar, G. A., Peña-Ramos, E. A., & González-Ríos, H. (2018). Biological activities of Agave by-products and their possible applications in food and pharmaceuticals. *Journal of the Science of Food and Agriculture*, *98*(7), 2461-2474.

Magwene, P. M., Kayıkçı, Ö., Granek, J. A., Reininga, J. M., Scholl, Z., & Murray, D. (2011). Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, *108*(5), 1987-1992.

Mak, S. S. T., Gopalakrishnan, S., Carøe, C., Geng, C., Liu, S., Sinding, M. H. S., Kuderna, L.F.K., Zhang, W., Fu, S., Vieira, F.G., Germonpré, M., Bocherens, H., Fedorov, S., Petersen, B., Sicheritz-Pontén, T., Marques-Bonet, T., Zhang, G., Jiang, H., & Gilbert, M. T. P. (2017). Comparative performance of the BGISEQ-500

vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *Gigascience*, *6*(8), gix049.

Marcet-Houben, M., & Gabaldón, T. (2015). Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biol*, *13*(8), e1002220.

Marsit, S., & Dequin, S. (2015). Diversity and adaptive evolution of Saccharomyces wine yeast: a review. *FEMS Yeast Research*, *15*(7), fov067.

Marsit, S., Leducq, J. B., Durand, É., Marchant, A., Filteau, M., & Landry, C. R. (2017). Evolutionary biology through the lens of budding yeast comparative genomics. *Nature Reviews Genetics*, *18*(10), 581-598.

Mastretta-Yanes, A., Moreno-Letelier, A., Piñero, D., Jorgensen, T. H., & Emerson, B. C. (2015). Biodiversity in the Mexican highlands and the interaction of geology, geography and climate within the Trans-Mexican Volcanic Belt. *Journal of Biogeography*, *42*(9), 1586-1600.

McGovern, P. E., Zhang, J., Tang, J., Zhang, Z., Hall, G. R., Moreau, R. A., Nuñez, A., Butrym E.D., Richards, M.P., Wang, C.S., Cheng, G., Zhao, Z., & Wang, C. (2004). Fermented beverages of pre-and proto-historic China. *Proceedings of the National Academy of Sciences*, *101*(51), 17593-17598.

Meriggi, Niccolo´., Cavalieri, D., & Stefanini, I. (2020). Saccharomyces cerevisiae–insects association: impacts, biogeography, and extent. *Frontiers in Microbiology*, *11*.

Morard, M., Macías, L. G., Adam, A. C., Lairón-Peris, M., Pérez-Torrado, R., Toft, C., & Barrio, E. (2019). Aneuploidy and ethanol tolerance in Saccharomyces cerevisiae. *Frontiers in genetics*, *10*, 82.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., & Stevens, M. H. H. (2019). vegan: Community Ecology Package. R package version 2.5-6. 2019.

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526-528.

Peter, J., & Schacherer, J. (2016). Population genomics of yeasts: towards a comprehensive view across a broad evolutionary scale. *Yeast*, *33*(3), 73-81.

Peter, J., De Chiara, M., Friedrich, A., Yue, J. X., Pflieger, D., Bergström, A., Sigwalt, A., Barré, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J.M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., Liti, G., & Schacherer, J. (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature*, *556*(7701), 339-344.

Pontes, A., Čadež, N., Gonçalves, P., & Sampaio, J. P. (2019). A quasi-domesticate relic hybrid population of Saccharomyces cerevisiae× S. paradoxus adapted to Olive Brine. *Frontiers in genetics*, *10*, 449.

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Moonshine A. L., Roazen D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek M., Gabriel S., Daly M. J., Neale B., MacArthur D. G., & Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 201178.

QGIS.org, 2021. QGIS Geographic Information System. QGIS Association. http://www.qgis.org

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Randez-Gil, F., Corcoles-Saez, I., & Prieto, J. A. (2013). Genetic and phenotypic characteristics of baker's yeast: relevance to baking. *Annual Review of Food Science and Technology*, *4*, 191-214.

Sarukhan, J. (Ed.). (2008). *Capital natural de México* (No. 333.95160972 333.95160972 C3 C37). Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO).

Schacherer, J., Shapiro, J. A., Ruderfer, D. M., & Kruglyak, L. (2009). Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. *Nature*, *458*(7236), 342-345.

Schuller, D., & Casal, M. (2007). The genetic structure of fermentative vineyard-associated Saccharomyces cerevisiae populations revealed by microsatellite analysis. *Antonie Van Leeuwenhoek*, *91*(2), 137-150.

Schuller, D., Cardoso, F., Sousa, S., Gomes, P., Gomes, A. C., Santos, M. A., & Casal, M. (2012). Genetic diversity and population structure of Saccharomyces cerevisiae strains isolated from different grape varieties and winemaking regions. *PLoS One*, *7*(2), e32507.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313.

Steensels, J., Gallone, B., Voordeckers, K., & Verstrepen, K. J. (2019). Domestication of industrial microbes. *Current biology*, *29*(10), R381-R393.

Tilakaratna, V., & Bensasson, D. (2017). Habitat predicts levels of genetic admixture in Saccharomyces cerevisiae. *G3: Genes, Genomes, Genetics*, *7*(9), 2919-2929.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, *43*(1), 11-10.

Voordeckers, K., Kominek, J., Das, A., Espinosa-Cantu, A., De Maeyer, D., Arslan, A., Van Pee, M., van der Zande, E., Meert, W., Yang, Y., Zhu, B., Marchal, K., DeLuna, A., Van Noort, V., Jelier, R., & Verstrepen, K. J. (2015). Adaptation to high ethanol reveals complex evolutionary pathways. *PLoS Genet*, *11*(11), e1005635.

Weir, B. S., & Goudet, J. (2017). A unified characterization of population structure and relatedness. *Genetics*, *206*(4), 2085-2103.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. springer.

Wolfe, K. H. (2015). Origin of the yeast whole-genome duplication. *PLoS Biol*, *13*(8), e1002221.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28-36.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326-3328.

ANEXO/MATERIAL SUPLEMENTARIO

TABLAS SUPLEMENTARIAS

Tabla 1- Anexo -> Contiene 146 cepas de fermentación de agave (137 nuestras + 9 analizadas por Peter y colaboradores)

La tabla en formato Excel puede ser descargada del siguiente enlace:

http://app.liigh.unam.mx/lmorales/Master_Thesis_JAUA/blob/master/metadata/Table_1_anexxed.xlsx

Tabla 2-Anexo -> Contiene 170 cepas analizadas por Peter y colaboradores (2018) (incluye a las nueve cepas de fermentación de agave analizadas por ellos)

La tabla en formato Excel puede ser descargada del siguiente enlace:

http://app.liigh.unam.mx/lmorales/Master_Thesis_JAUA/blob/master/metadata/Table_2_anexxed.xlsx

Tabla 3-Anexo - > 13 cepas a ser removidas de la matriz de SNPs que sólo incluye a las cepas de fermentación de agave

Strain	Geo_Region	Isolation	Hybrid	Pct mapped	Pct of Missing Data
XB003s8	SLP	Mezcal	Yes	36	28.94
DK002c39	Oaxaca	San Pedro Totolapan	Yes	33	20.3
DK003c30	SLP	Laguna Seca	Yes	33	19.33
XA256c15	Michoacan	Rio de Parras	Yes	57	11.57
XB003c8	SLP	SLP	Yes	63	8.79
XA192c4	Durango	Nombre de Dios	Yes	63	8.26
XB008c4	Guanajuato	San Luis de la Paz	Yes	52	7.49
XA184c6	Durango	Nombre de Dios	Yes	63	5.13
XA022c2	Estado de Mexico	Palmar de Guadalupe	Yes	65	4.94
XA182c2	Durango	Nombre de Dios	Yes	64	4.74
XA082c2	Guanajuato	San Felipe	Yes	53	4.43
XA183c1	Durango	Nombre de Dios	Yes	63	4.38
XB063c2	Oaxaca	Santa Catarina	No	84	3.23

Pct mapped to SACE genome: Porcentaje de lecturas mapeadas al genoma de *S. cerevisiae*

Esta tabla puede descargarse del repositorio de GitLab en el link: http://app.liigh.unam.mx/lmorales/Master_Thesis_JAUA/blob/master/metadata/Strains_to_be_removed_from_MexicanMezcal_SNP_dataset.xlsx

Tabla 4-Anexo- > 13 cepas a ser removidas de la matriz de SNPs que sólo incluye a las cepas de fermentación de agave + cepas analizadas por Peter y colaboradores (2018)

Strain	Geo_Region	Isolation	Hybrid	Pct mapped to SACE genome	Pct of Missing Data
XB003s8	SLP	Mezcal	Yes	36	27.66
DK002c39	Oaxaca	San Pedro Totolapan	Yes	33	19.5
DK003c30	SLP	Laguna Seca	Yes	33	18.7
XA256c15	Michoacan	Rio de Parras	Yes	57	10.36
XB003c8	SLP	SLP	Yes	63	7.6
XA192c4	Durango	Nombre de Dios	Yes	63	7.18
XB008c4	Guanajuato	San Luis de la Paz	Yes	52	7.02
XA184c6	Durango	Nombre de Dios	Yes	63	4.15
XA022c2	Estado de M	Palmar de Guadalupe	Yes	65	4.03
XA082c2	Guanajuato	San Felipe	Yes	53	3.84
XA182c2	Durango	Nombre de Dios	Yes	64	3.83
XA183c1	Durango	Nombre de Dios	Yes	63	3.44
XB063c2	Oaxaca	Santa Catarina	No	84	2.63

Pct mapped to SACE genome: Porcentaje de lecturas mapeadas al genoma de *S. cerevisiae*

Esta tabla puede descargarse del repositorio de GitLab en el link: http://app.liigh.unam.mx/lmorales/Master_Thesis_JAUA/blob/master/metadata/Strains_to_be_removed_from_allSACE_SNP_dataset.xlsx

Tabla 5-Anexo- > Aneuploidías presentes en las 146 cepas de fermentación de agave (137 cepas secuenciadas por nosotros + 9 reportadas por Peter y colaboradores)

La tabla en formato Excel puede ser descargada del siguiente enlace: http://app.liigh.unam.mx/lmorales/Master_Thesis_JAUA/blob/master/metadata/Ploidy_table.xlsx .

Tabla 6-Anexo -> Pasos de *pipeline* para obtener las matrices crudas de SNPs utilizadas para los análisis filogenéticos y de estructura poblacional

La tabla en formato Excel puede ser descargada del siguiente enlace: http://app.liigh.unam.mx/lmorales/Master_Thesis_JAUA/blob/master/metadata/Pipeline_Steps_to_obtain_raw_SNPs_matrices.xlsx .

Tabla 7-Anexo -> Pasos de *pipeline* para obtener las matrices filtradas a utilizar en los análisis filogenéticos

La tabla en formato Excel puede ser descargada del siguiente enlace: http://app.liigh.unam.mx/lmorales/Master_Thesis_JAUA/blob/master/metadata/Pipeline_Steps_to_filter_and_obtain_matrices_for_phylogenetic_analyses.xlsx

Tabla 8-Anexo -> Lugares de origen de cepas de fermentación de agave utilizadas en el análisis de ADMIXTURE que solamente incluye a las cepas de fermentación de agave (N=146)

Lugares de Origen de Cepas de Fermentación de Agave en Figura 16 -> 146 cepas y *K*=12	
Grupo en Figura	**Lugares de Origen**
1	• San José Chalmita, Edo. México • Morelia Pino Chico, Piedras de Lumbre, Michoacan • Matatlán, Oaxaca • Eduardo Neri, Guerrero
2	• San Luis de la Paz, San Felipe, Guanajuato • Santa Catarina, San Pedro Totolapan, Matatlán, Oaxaca • Laguna Seca, SLP
3	• Huasabas, Sonora • Las Guasimas, Jalisco
4	• Laguna Seca, SLP • Huitzuco, Chilapa, Apetlán, Guerrero • Oaxaca • San José Chalmita, Edo. México
5	• Santa Maria Xoyatla, Puebla • Las Guasimas, Jalisco
6	• San Juan del Río, Matatlán, Jayacatlán, Santiago Matatlán, Jayacatlán, San Agustín Amatengo, Oaxaca • San Felipe, Guanajuato • San Nicolás Huejapan, Puebla • Cepa de múltiples especies de agave de Peter y colegas (2018) de México
7	• San Luis Amatlán, Santa Catarina, San Agustín Amatengo, Jayacatlán, Mihuatlán, San Isidro Gusihe, Oaxaca • San Felipe, San Luis de la Paz, Guanajuato
8	• Palmar de Guadalupe, Edo. México • La Estancia, Jalisco • Nombre de Dios, Temohaya Mezquital, Yonora Mezquital, Durango • Huitzila, Zacatecas
9	• San Carlos, Tamaulipas • Huasabas, Sonora
10	• Mascota, San Juan Espanatica, Raicilla (Peter J. *et al*., 2018), Jalisco • Rio de Parra, Michoacán • SLP
11	• Palmar de Guadalupe, Edo. México • San Juan del Río, Sta. Catarina, Mihuatlán, Oaxaca • Guanajuato • Etúcuaro, Michoacán • Mascota, Jalisco
12	• San Carlos, 7 cepas de Tamaulipas de Peter y colegas (2018)

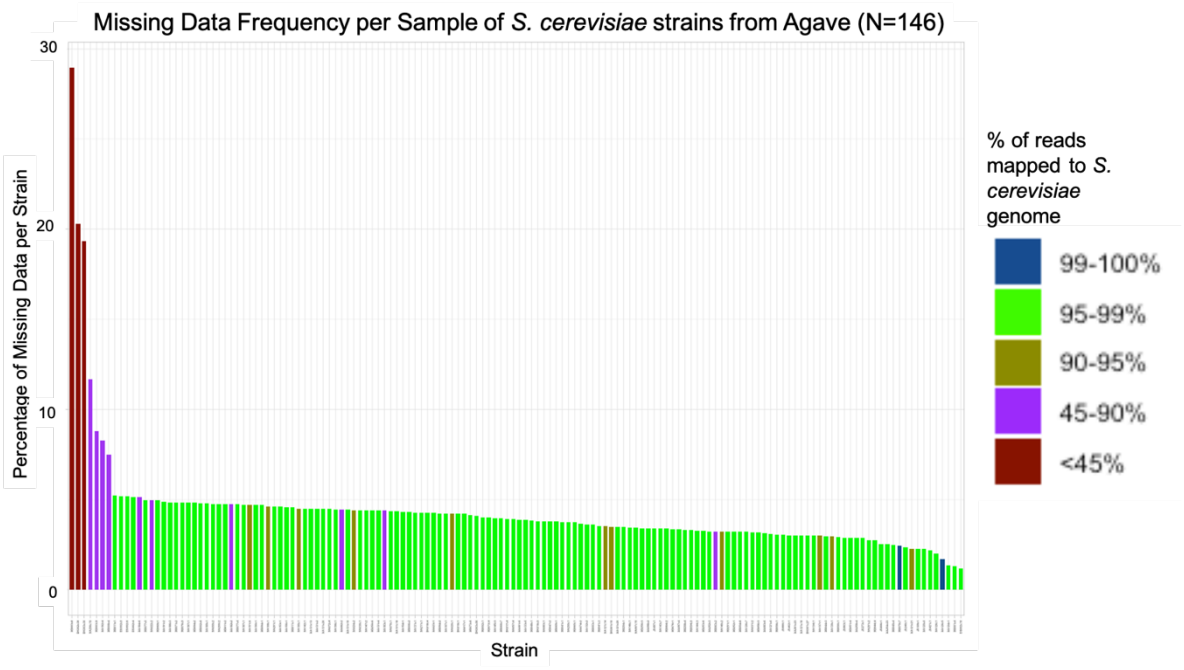
Tabla 9-Anexo -> Lugares de origen de cepas de fermentación de agave utilizadas en el análisis de ADMIXTURE que s incluye a las cepas de fermentación de agave (N=124) y las cepas analizadas de Peter y colaboradores (2018) (N=170)

Lugares de Origen de Cepas de Fermentación de Agave en Figura 17	
Grupo en Figura	**Lugares de Origen**
1	• San Carlos, Tamaulipas y 7 cepas de fermentación de Tamaulipas de Peter y colegas
2	• Clados **CHNI, CHNII** y **Taiwanese**
3	• Jalisco • Durango • Michoacán • Zacatecas • Guanajuato • Oaxaca
4	• Oaxaca • Puebla • Guanajuato
5	• Clado **French Guiana Human**
6	• Clados **CHNIII, CHNV, Malaysian, Far East Russian, North American Oak, Ecuadorean, Mosaic Region 3, Far East Asian, Unclustered** • 4 cepas de San Carlos, Tamaulipas
7	• Clados **African Palm Wine, Mosaic Region 3** y **Unclustered**
8	• Clados **Asian Fermentation, Sake, Asian Islands, Mosaic Region 3** y **Unclustered**
9	• Edo. México • Jalisco • Puebla • Oaxaca • Guerrero
10	• Clados **Ale Beer, Mixed Origin, Alpechin, Mosaic Region 3, Mosaic Beer, French Dairy, Brazilian Bioethanol** y **Unclustered**
11	• Clados **Wine/European, Alpechin, French Dairy, Mosaic Region 1-3, West African Cocoa** y **Unclustered**
12	• Clados **Mosaic Region 2, Mediterranean Oak, African Beer, Wine/European** y una cepas de Matatlán, Oaxaca

Los nombres de los clados del artículo de Peter y colaboradores (Tabla 2-Anexo) se indican en negritas.

FIGURAS SUPLEMENTARIAS

A.



B.

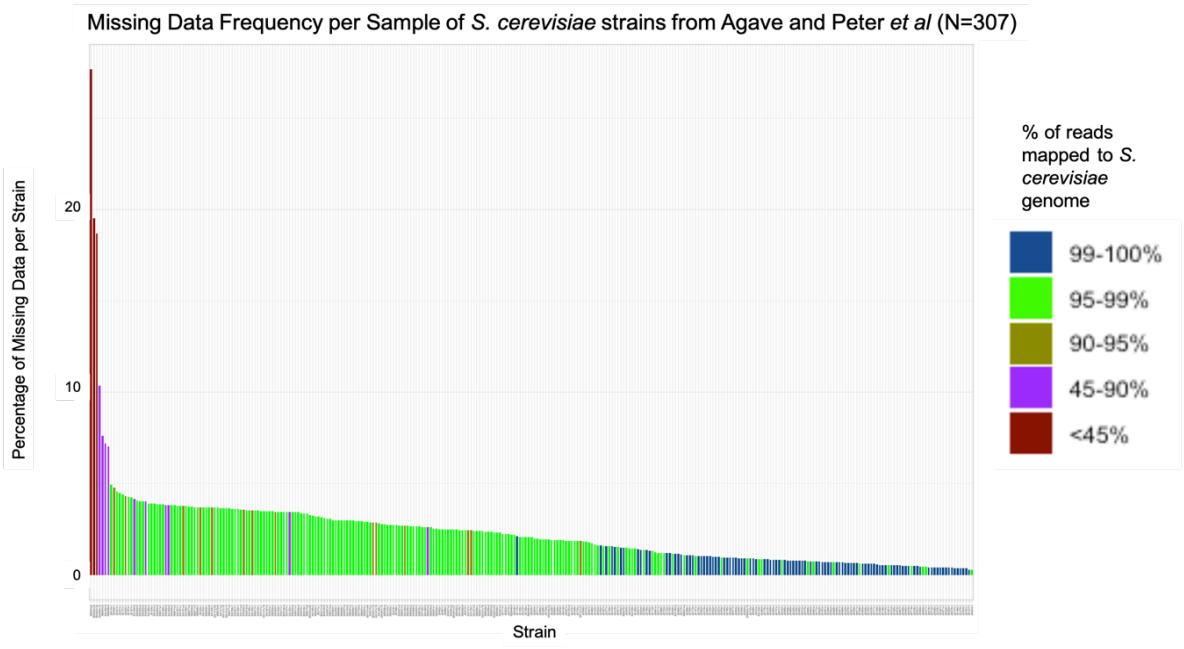
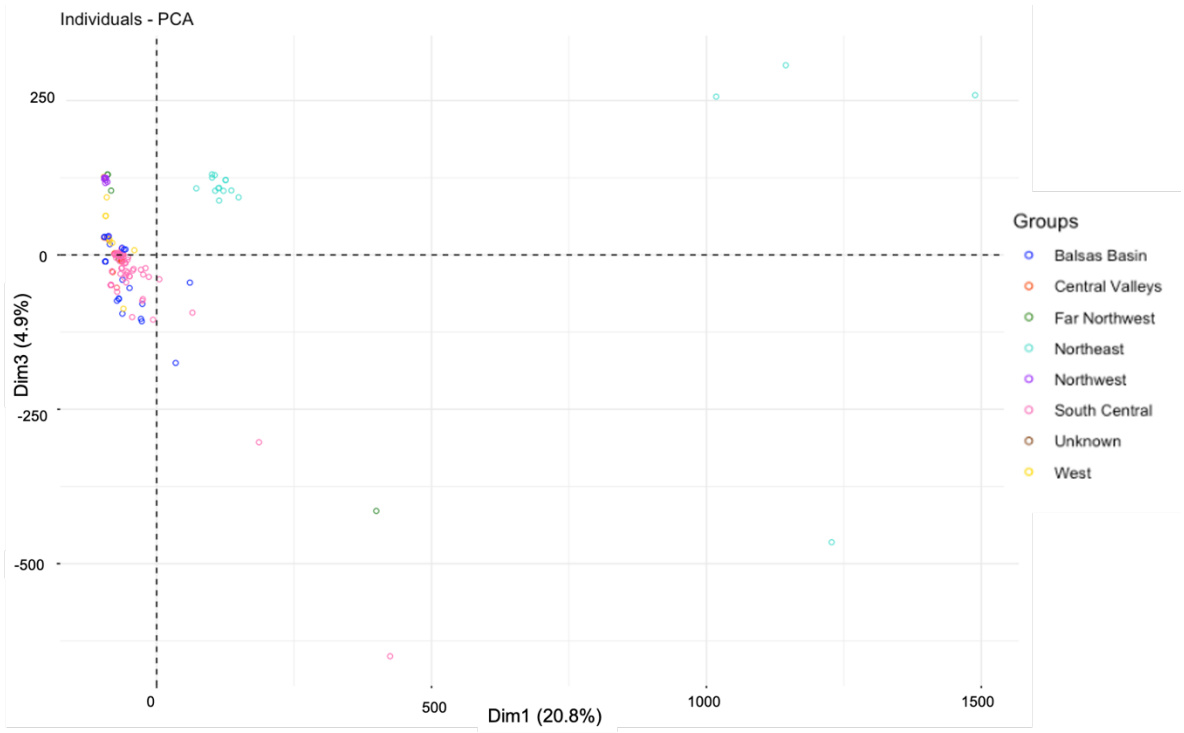


Figura suplementaria 1. **Gráfica de barras para cuantificar porcentaje de datos faltantes en cada cepa y porcentaje de lecturas mapeadas al genoma de *S. cerevisiae*.** El **Panel A** muestra la cantidad de datos faltantes por cepa de las 146 cepas de fermentación de agave. Los colores de las barras representan el intervalo del porcentaje de lecturas mapeadas al genoma de *S. cerevisiae*. Las cepas por remover (**Tabla 3-Anexo**) de la matriz de SNPs son las cepas coloreadas de rojo y morado. El **Panel B** muestra la cantidad de datos faltantes por cepa de 307 cepas (137 cepas de fermentación de agave + 170 cepas analizadas por Peter y colaboradores). Los colores de las barras representan el intervalo del porcentaje de lecturas mapeadas al genoma de *S. cerevisiae*. Las cepas por remover (**Tabla 4-Anexo**) de matriz de SNPs son las cepas coloreadas de rojo y morado.

A.



B.

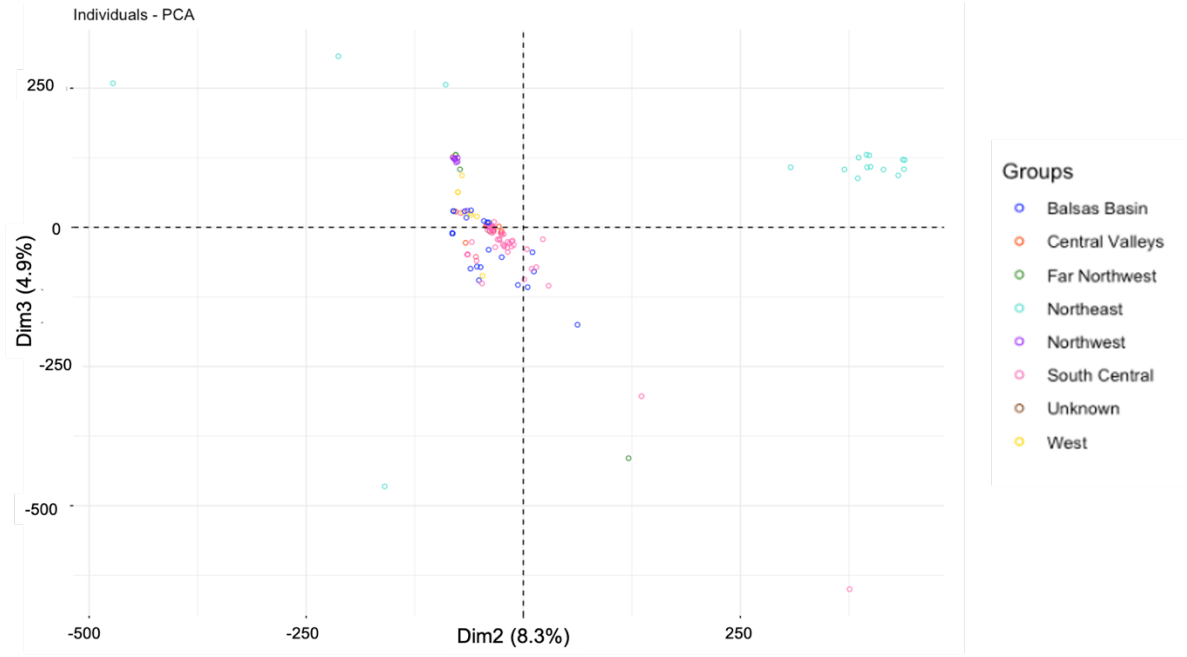
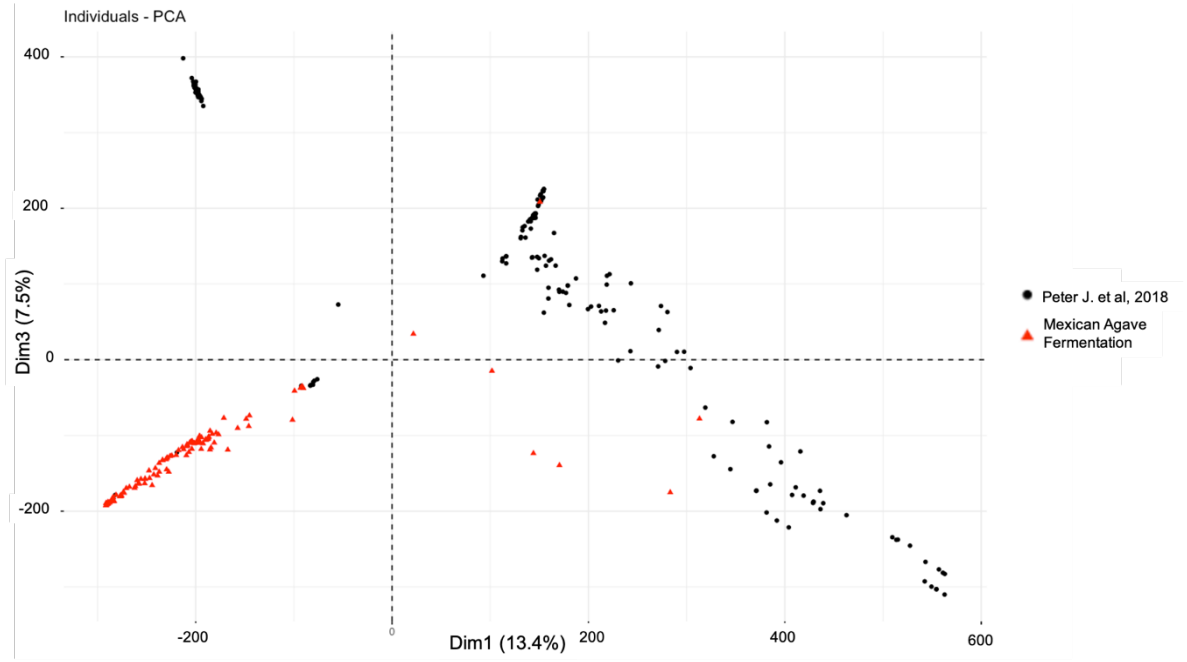


Figura suplementaria 2. **Gráficas de componentes principales PC1 vs PC3 y PC2 vs PC3 de las *S. cerevisiae* involucradas en fermentación de agave (124 cepas de fermentación de agave + 9 de fermentación de agave analizadas por Peter y colaboradores).** El **Panel A** muestra la gráfica de PC1 vs PC3 hecha a partir de 223,013 SNPs bialélicos de las cepas de fermentación de agave. Por otro lado, el **Panel B** muestra la gráfica de PC2 vs PC3 hecha a partir de los mismos 223,013 SNPs bialélicos de las cepas de fermentación de agave.

A.



B.

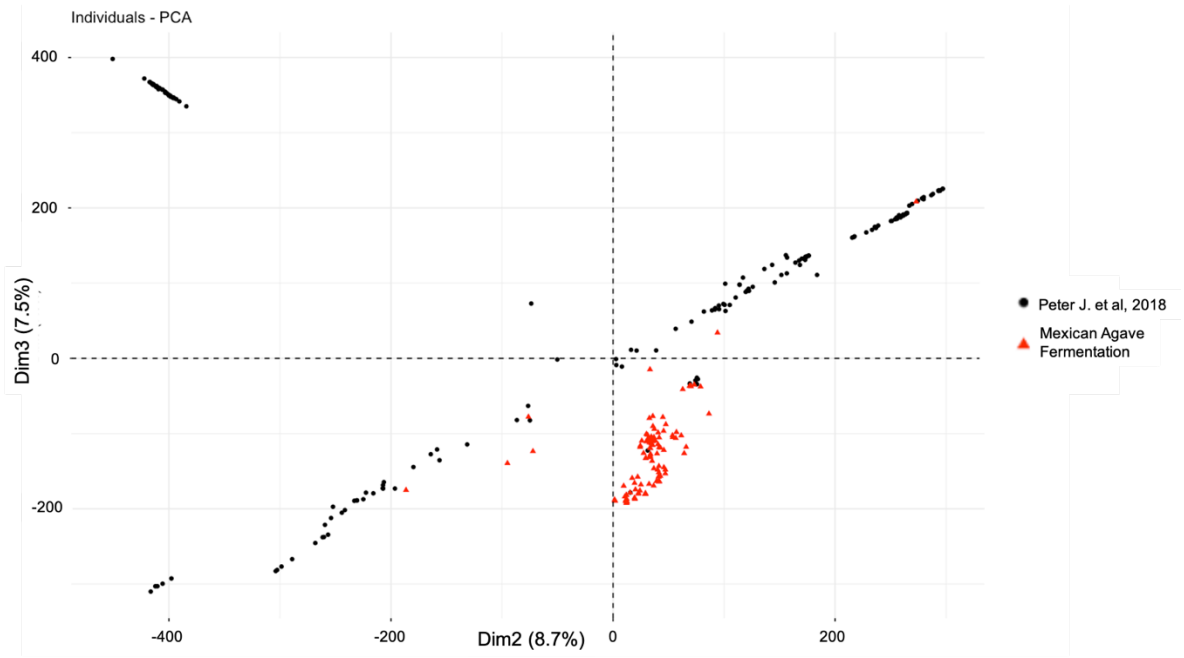
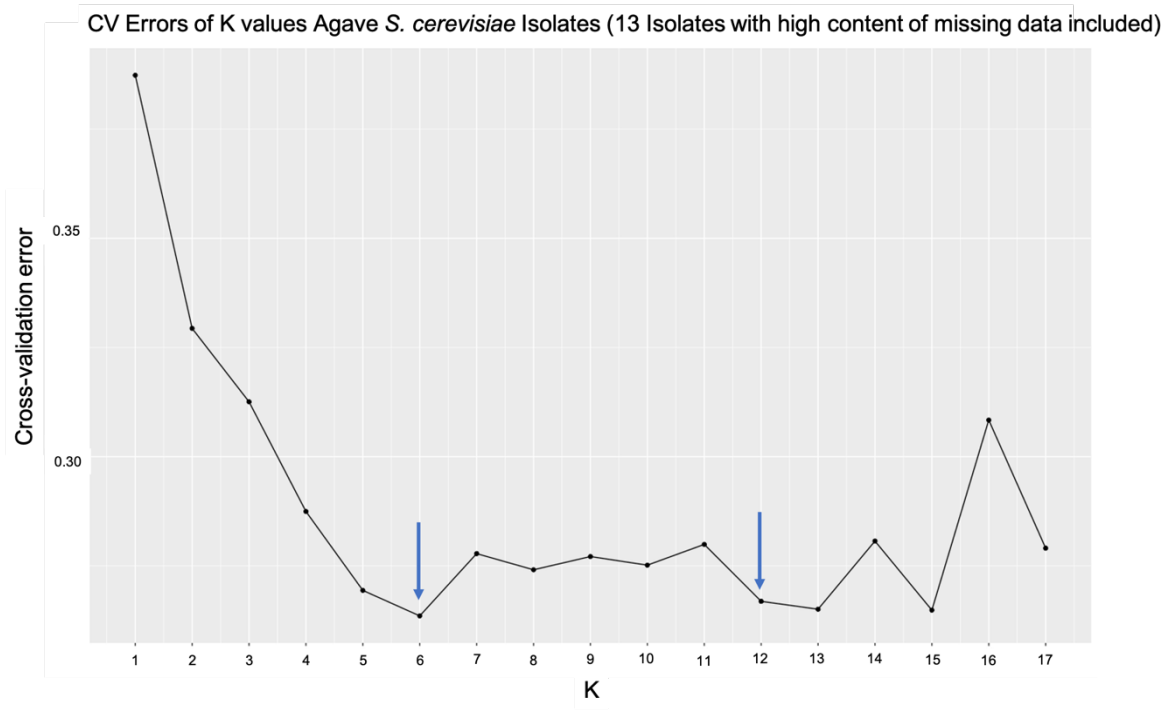


Figura suplementaria 3. **Gráficas de componentes principales PC1 vs PC3 y PC2 vs PC3 de las *S. cerevisiae* involucradas en fermentación de agave y cepas de otras partes del mundo (124 cepas de fermentación de agave + 170 cepas analizadas por Peter y colaboradores)** El **Panel A** muestra la gráfica de PC1 vs PC3 hecha a partir de 831,175 SNPs bialélicos de las cepas de fermentación de agave y las cepas de otras partes del mundo. Por otro lado, el **Panel B** muestra la gráfica de PC2 vs PC3 hecha a partir de esos mismos 831,175 SNPs bialélicos de las cepas de fermentación de agave y las cepas de otras partes del mundo.

A.



B.

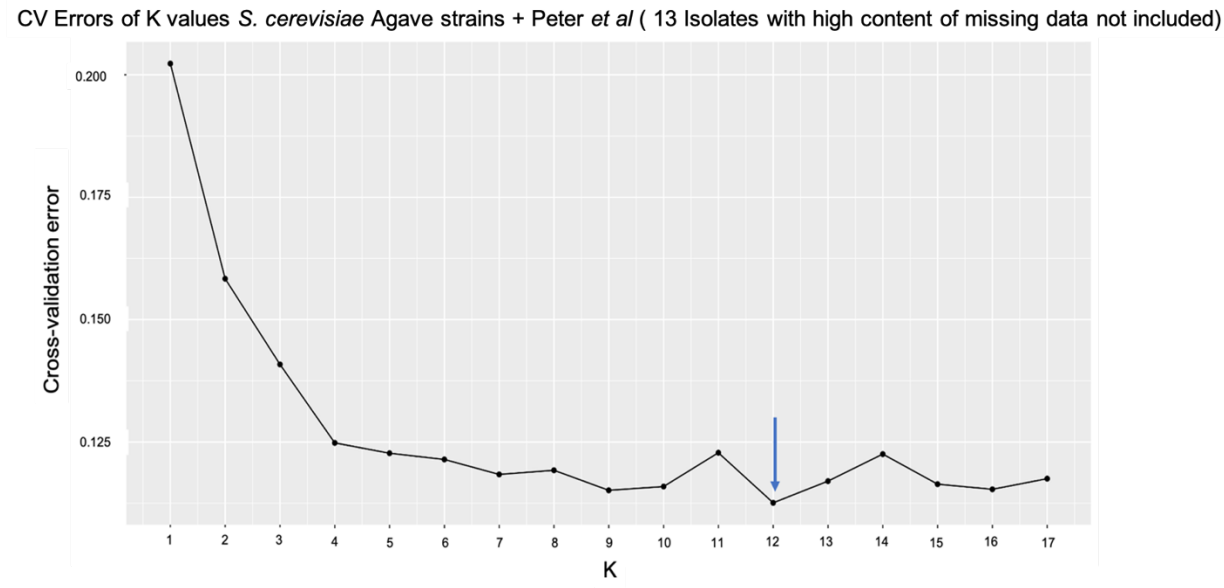


Figura suplementaria 4. **Gráficas de *Cross-Validation* (CV) *error* de los análisis de ADMIXTURE para determinar mejores valores de *K*.** El **Panel A** muestra la gráfica de CV para los valores de *K* del análisis de ADMIXTURE llevado a cabo con 119,317 SNPs bialélicos del genoma nuclear de *S. cerevisiae* de las cepas de fermentación de agave (N=146). El **Panel B** muestra la gráfica de CV para los valores de *K* del análisis de ADMIXTURE llevado a cabo con 831,175 SNPs bialélicos del genoma nuclear de *S. cerevisiae* de las cepas de fermentación de agave y las cepas analizadas por Peter y colaboradores (N=294).

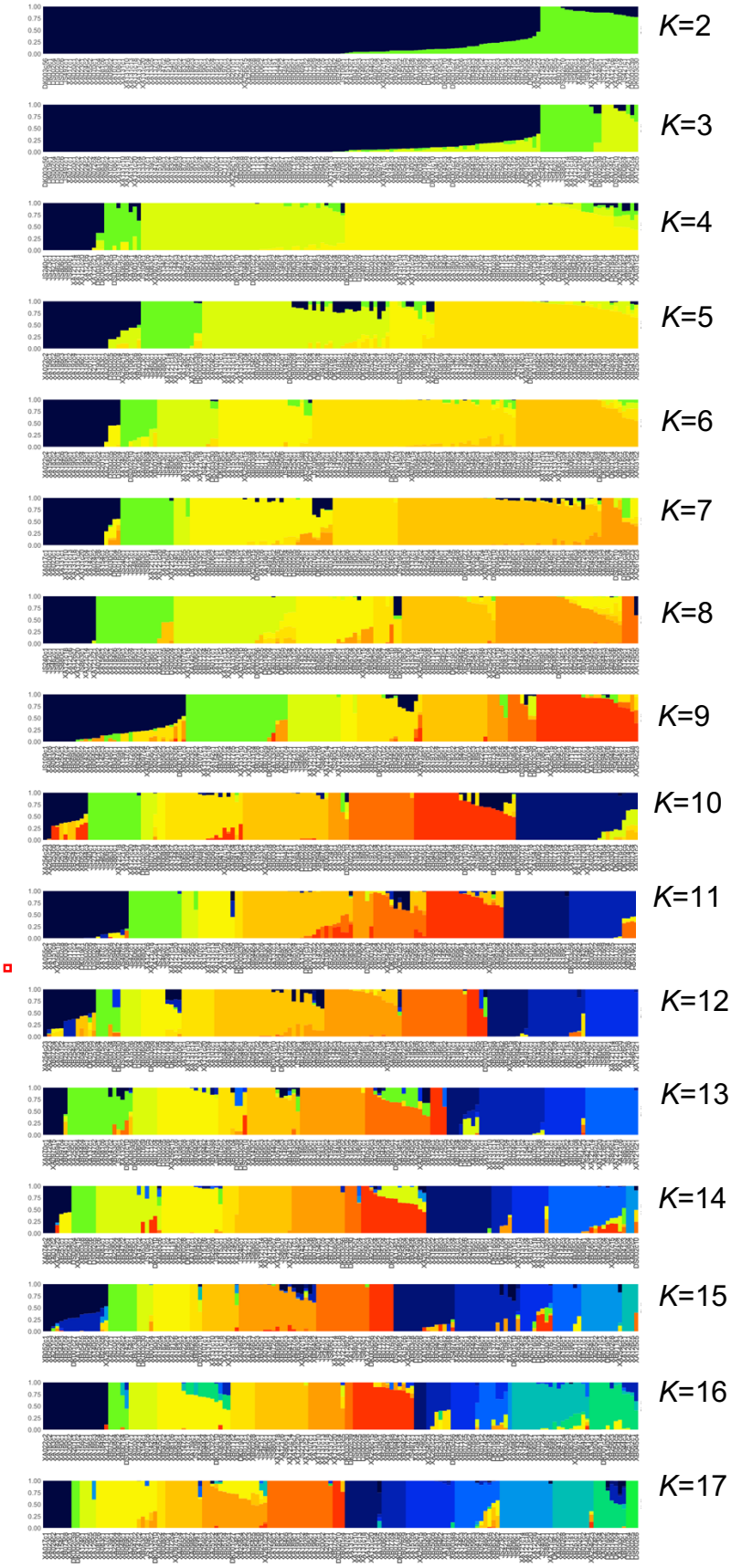

K=2
K=3
K=4
K=5
K=6
K=7
K=8
K=9
K=10
K=11
K=12
K=13
K=14
K=15
K=16
K=17

Figura suplementaria 5. **Estructura poblacional de las cepas de fermentación de agave**. Gráficas de ADMIXTURE con *K*=2-17 (146 cepas de *S. cerevisiae* o híbridos interespecies *S. cerevisiae-S. paradoxus*) llevado a cabo con 119,317 SNPs bialélicos del genoma nuclear de *S. cerevisiae*. Esta imagen fue creada con la ayuda del paquete de R pophelper v2.3.1 (Francis, R.M., 2017) y las gráficas fueron ordenadas con respecto a los *clústers* identificados en cada una de ellas, por lo tanto el orden de las cepas varía en cada gráfica. El rectángulo rojo encierra el gráfico de *K*=12, el cual se muestra en la **Figura 15**.

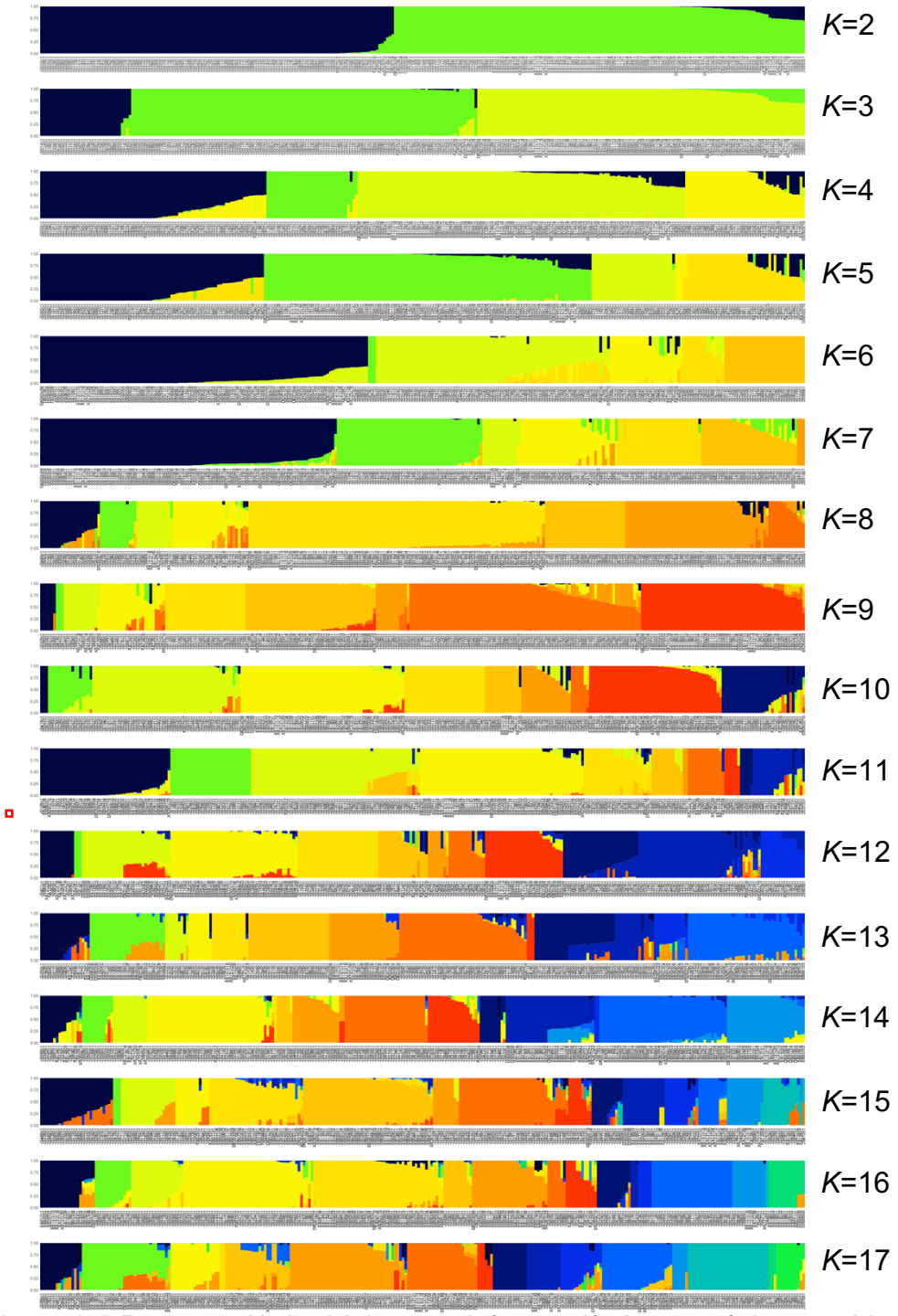


Figura suplementaria 6. **Estructura poblacional de las cepas de fermentación de agave más las cepas del resto del mundo.** Gráficas de ADMIXTURE con *K*=2-17 (294 cepas de *S. cerevisiae* -> 124 cepas de agave secuenciadas por nosotros +170 cepas de diversos orígenes de Peter y colaboradores) llevado a cabo con 831,175 SNPs bialélicos del genoma nuclear de *S. cerevisiae*. Esta imagen fue creada con la ayuda del paquete de R pophelper v2.3.1 (Francis, R.M., 2017) y las gráficas fueron ordenadas con respecto a los *clústers* identificados en cada una de ellas, por lo tanto el orden de las cepas varía en cada gráfica. El rectángulo rojo encierra el gráfico de *K*=12, el cual se muestra en la **Figura 16**.