UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Estudios moleculares e histoquímicos para el mejoramiento en Caña de Azúcar (Saccharum spp.)

Tesis presentada para optar al título de Doctora de la Universidad de Buenos Aires en el área de Ciencias Biológicas

Licenciada Catalina Molina

Director de Tesis: Dr. Alberto Acevedo.

Codirectora de Tesis: Dra. Norma Beatríz Paniego.

Consejera de Estudios: Viviana Confalonieri.

Lugar de trabajo: Instituto de Suelos, Centro Nacional de Investigaciones Agropecuarias, Instituto Nacional de Tecnología Agropecuaria (CNIA-INTA)

Feche de defensa: Buenos Aires 19 septiembre de 2024

Estudios moleculares e histoquímicos para el mejoramiento en Caña de Azúcar (Saccharum spp.)

Resumen

El cultivo de caña de azúcar es de gran importancia económica en Argentina y en el mundo. Su naturaleza poliploide y aneuploide dificulta el mejoramiento genético tradicional. El uso de marcadores moleculares podría aumentar el conocimiento de los genotipos utilizados para los cruzamientos y acelerar los ciclos del mejoramiento. El objetivo de este trabajo es desarrollar estrategias y generar conocimiento para el mejoramiento de caña de azúcar mediante la aplicación de tecnologías de genotipificación masiva, así como explorar la variabilidad anatómica de entrenudos para la selección de materiales de alta biomasa. En este trabajo se estableció un protocolo de genotipificación por secuenciación en cinco genotipos de caña. El protocolo se implementó en 90 individuos de la población NA 78-724 x LCP 85-384. Se generó un panel de 2.066 SNPs que se aplicó para estudios de diversidad y mapeo genético. Se generó un mapa marco representativo de los 10 grupos de ligamiento de caña de azúcar y un mapa de haplotipos a partir de los dosajes alélicos de los loci SNP. Se caracterizó la deposición de lignina, la composición de los haces vasculares y su relación con el área del entrenudo del tallo. Paralelamente se identificaron 15 SNP vinculados a proteínas y enzimas clave en la síntesis de polisacáridos y lignina. A partir de esta tesis se generaron conocimientos y herramientas moleculares fundamentales para contribuir al entendimiento de la estructura genética de las variedades de caña de azúcar y sustentar el mejoramiento de características de interés agroindustrial.

Palabras clave: genotipificación por secuenciación · híbridos de *Saccharum* · polimorfismo de un solo nucleótido · poliploides· dosaje alélico· mapas genéticos · haces vasculares·

Molecular and histochemical studies for sugarcane (Saccharum spp.) breeding

Abstract

The sugarcane crop is of great economic importance in Argentina and around the world. Its polyploid and aneuploid nature makes traditional breeding difficult. The use of molecular markers could increase the knowledge of genotypes used for crosses and accelerate breeding cycles. The aim of this work is to develop strategies and generate knowledge for sugarcane breeding programmes by applying genotyping technologies and exploring the anatomical variability of internodes for the selection of high biomass materials. A sequencing genotyping protocol was established for five sugarcane genotypes. The protocol was performed on 90 individuals from the NA 78-724 x LCP 85-384 population. A panel of 2,066 SNPs was generated and applied for diversity and genetic mapping studies. A representative genetic map of the 10 linkage groups and a haplotype map were generated from the allele dosages of the SNP loci. Lignin deposition, vascular bundle composition, and its relationship with the stem internode area were characterised. Moreover, 15 SNPs related to key proteins and enzymes in the synthesis of polysaccharides and lignin were identified. This work has provided knowledge and basic molecular tools that contribute to the understanding of the genetic structure of sugarcane cultivars and support the the breeding of agro-industrial traits.

Keywords: genotyping by sequencing - *Saccharum* hybrids - single nucleotide polymorphism - polyploids - allelic dosage - genetic mapping - vascular bundles

Financiamiento

El presente trabajo se realizó en el Instituto de Suelos, Centro de Investigación en Recursos Naturales, y el Instituto de Agrobiotecnología y Biología Molecular Unidad Ejecutora de Doble Dependencia del Instituto Nacional de Tecnología Agropecuaria (INTA) y Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Se enmarcaron los primeros tres años en una *Beca Nivel Inicial* otorgada por el *Fondo para la Investigación Científica y Tecnológica* (FONCyT) de la *Agencia Nacional de Promoción Científica y Técnica* (julio 2018- junio 2021) con disposición Nº: DN 0919-2018. Título del trabajo: "POLIMORFISMO DE MUTACIÓN SIMPLE Y PATRONES DE PARED CELULAR EN CAÑA DE AZÚCAR." Posteriormente, se continuó con *Beca de finalización* otorgada por el CONICET (abril 2021- marzo 2024). RESOL-2021-143-APN-DIR#CONICET. Título del trabajo: "HERRAMIENTAS GENÓMICAS PARA EL MEJORAMIENTO BIOMÁSICO EN CAÑA DE AZÚCAR".

A su vez, los siguientes proyectos financiaron esta investigación:

-*PROYECTO ESTRUCTURAL INTA 2019-PE-E6-I114-001* "Caracterización de la diversidad genética de plantas, animales y microorganismos mediante herramientas de genómica aplicada".

-*PROYECTO ESTRUCTURAL INTA 2019-PE-E6-I516-001.* "Mejoramiento genético y desarrollo de ideotipos de cultivos industriales (CI) caña, maní, yerba, mandioca, stevia, quinua y te para sistemas productivos resilientes".

-*PROYECTO ESTRUCTURAL INTA 2019-PE-E7-I149-001* "Bioenergía generada en origen como aporte al desarrollo territorial".

-*PROYECTO DE INVESTIGACIÓN CIENTÍFICA Y TECNOLÓGICA (PICT) 2016-1670.* "Feno y genotipificación para selección de individuos de caña de azúcar de alto contenido en azúcar y/o buenas características biomásicas para la producción de bioetanol".

Agradecimientos

Al Fondo para la Investigación Científica y Tecnológica de la Agencia Nacional de Promoción Científica y Técnica Consejo Nacional de Investigaciones Científicas y Técnicas por las becas otorgadas para la realización de este trabajo de Tesis.

Al Instituto Nacional de Tecnología Agropecuaria (INTA), en especial al Instituto de Suelos y al Instituto de Agrobiotecnología y Biología Molecular. A sus directivos por permitirme desarrollar esta Tesis con total libertad y por facilitarme los materiales, las instalaciones y gestiones necesarias para llevar a cabo mis tareas diariamente. A todos mis compañeros por su cariño y apoyo diario.

A la Universidad de Buenos Aires por darme la posibilidad de estudiar con libertad, darme las herramientas para formarme para ser una mejor profesional y enseñarme la importancia del respeto por el otro y la pluralidad de pensamiento.

A mi director Alberto por confiar en mi criterio y guiarme en estos primeros pasos. También por sus charlas sobre la vida y por compartir sus experiencias en lo profesional y en lo personal.

A mi codirectora Norma por aceptarme y guiarme con paciencia y libertad, brindándome su contención y confianza en todo momento. También por abrirme las puertas del IABIMo y darme la oportunidad de formar parte del equipo de GBS.

A José María por transitar este largo camino juntos y ser mi compañero de debate, apoyándome y aconsejándome en cada instancia.

A la Unidad de Genómica y Bioinformática (UGB), como laboratorio por permitirme realizar todos los ensayos, ayudarme y enseñarme a trabajar de forma ordenada y sistemática. Pero más importante, a las personas que trabajan en ella por ser siempre un refugio. A Andre por abrirme las puestas desde el día 1, por integrarme y explicarme con paciencia. A Pabli por las numerosas horas de mesada compartidas, por enseñarme técnicas complicadas con simplicidad y buen humor, lo que me permitió realizar costosos ensayos de biología molecular relajada. A Vero por compartirme su conocimiento y amistad. A Sergio que me ayudó y respondió cada duda bioinformática. A Marianne por ayudarme siempre que lo necesité.

A Nati Aguirre por su paciencia, por corregirme y guiarme en este camino desde el principio. También por brindarme su cariño y contención. A Susana Marcucci Poltri por estar disponible siempre que la necesité para corregirme y brindarme su punto de vista. A Carla Filippi por sus ideas para mirar los datos desde otras perspectivas. También a Martín García por su ayuda para

generar y comprender los programas para realizar los mapas genéticos. A todos ellos por su paciencia y buena onda siempre que tuve una duda o necesité reunirme.

Al Grupo Caña de Azúcar de la EEA Famaillá por recibirme y compartir su conocimiento sobre caña de azúcar, así como de brindarme los materiales para poder realizar este trabajo. A Luis Erazzú por sus consejos y siempre estar disponible.

A Lore Setten por ayudarme en todo. Por regar y ayudarme con el invernáculo, por enseñarme a usar el microscopio, contenerme en época de pandemia y asistirme en cualquier cosa que necesitara.

Juli Irigoin y Vani Cosentino por alentarme a confiar en mí y en mí trabajo, por ayudarme a organizarme y siempre tener palabras de aliento.

A Franco Calderón, mi co-worker en las clases de GBS, por ser un gran compañero y amigo durante estos años. A Melanie que con su carisma y empuje logramos muchas cosas lindas.

Agradezco a toda mi familia, por acompañarme en mis decisiones y brindarme todo su amor. A mis papás, Gregre y Marce, por apoyarme incondicionalmente, brindándome palabras de aliento para seguir adelante en todo momento, por confiar en mí y en mis convicciones y por enseñarme lo más importante, sus valores. A mi hermana Vero, por ser mi consejera fiel, defendiéndome y acompañándome en mis decisiones y ser mi mejor amiga. A Mati por estar siempre con buen humor. A ambos por regalarnos a Vicki, que con 5 meses alegra nuestras vidas. A la Tía Mimí por las horas de teléfono dedicadas en pandemia y por ser mi confidente. A mi prima Caro por su apoyo incondicional. A Mari, Gio, Rami, Agustí y Augusto por estar presentes en mi vida y permitirnos ser parte de su familia. Por último, a mi abuelita Coqui, por estar siempre conmigo, aun cuando no puede ser físicamente, y por enseñarme a ser una mejor persona.

A mis amigas de siempre Agus G, Rochi, Anto Pon, Anto Sav, Vicki, la Colo, Agus Gon, Agus Gata, con las cuales cuento siempre y me dan su cariño y apoyo en todo momento.

A mi amiga Dani Regeiro, por brindarme su amor y apoyo incondicional. A Dani Riva, a quién quiero y admiro profundamente. A Caro Cuello por sus enseñanzas de vida. A mis compañeros de la Cátedra de Bioquímica, en especial a Dani T. por sus consejos, apoyo y confianza; a Nati por su cariño y charlas; y a Andrés por escucharme y aconsejarme en todo momento.

Por último, a Alan por llegar a mi vida en el final de este camino y brindarme su amor, contención y ayudarme a creer en mí misma.

Muchas gracias a todos los que me acompañaron todos estos años!!!

Parte de los resultados de este trabajo de Tesis fueron publicados

1. García, J. M., **Molina, C.**, Simister, R., Taibo, C. B., Setten, L., Erazzú, L. E., Gómez, L. D., & Acevedo, A. (2023). Chemical and histological characterization of internodes of sugarcane and energy-cane hybrids throughout plant development. *Industrial Crops and Products*, *199*, 116739. https://doi.org/10.1016/J.INDCROP.2023.116739
2. **Molina, C**., Aguirre, N. C., Vera, P. A., Filippi, C. V., Puebla, A. F., Marcucci Poltri, S. N., Paniego, N. B., & Acevedo, A. 2022. ddRADseq-mediated detection of genetic variants in sugarcane. *Plant Molecular Biology 2022, 1*, 1–15. https://doi.org/10.1007/S11103-022-01322-4
3. Acevedo, A., Molina, C, Garcia, J.M., Federico, M. L. y Erazzú, L. E. (2021) Current Status of Sugarcane (*Saccharum spp*.) Breeding under Sustainable Conditions in Argentina en Universitat Politècnica de València (Ed.), Velázquez-Martí, B. (Coordinador), *Optimización de los procesos de extracción de biomasa Sólida para uso energético. Trabajos de investigación 2020, Red IBEROMASA* (pp. 129-140)
4. **Molina, C**.; Vera, P.; Aguirre, N.C.; Puebla, A.F.; Paniego, N.B.; Acevedo, A. 2020. "Puesta a punto de la técnica *ddRADseq* para el desarrollo de marcadores moleculares en caña de azúcar". Libro de Resúmenes IV Reunión Conjunta de Sociedades de Biología de la República Argentina "Nuevas Evidencias y Cambios de Paradigmas en Ciencias Biológicas".
5. García, J M.; **Molina, C**.; Hallam, R.; Erazzú, L. E.; Taibo, C. B.; Gomez, L.; Acevedo, A. 2020 "Composición y sacarificación de la lignocelulosa en caña de azúcar". Libro de Resúmenes IV Reunión Conjunta de Sociedades de Biología de la República Argentina "Nuevas Evidencias y Cambios de Paradigmas en Ciencias Biológicas". 9-15 de septiembre 2020.
6. García, J. M.; **Molina C**.; Acevedo, A. 2020. Caña de azúcar en Argentina: situación actual y uso para fines bioenergéticos. Capítulo 7. Velázquez-Martí, B. (Ed.). *Optimización de los procesos de extracción de biomasa Sólida para uso energético* (pp. 171-183). España. ISBN 8412179854, 9788412179859.

Índice General

Abreviaturas, acrónimos y siglas

A adenina
ADN Ácido desoxirribonucleico o Deoxyribonucleic acid
AFLP polimorfismos de longitud de fragmentos amplificados
Ampure XP Marca comercial de perlas magnéticas para purificar ADN
ANOVA Análisis de la varianza
ApeKI Enzima de restricción ApeKI
BAC cromosomas artificiales bacterianos
BAM archivo Binary Alignment Map
BIC Criterio de Información Bayesiano de Bayesian Information Criterion
Bru1 gen de resistencia a la roya
C Citosina
C4 vía de 4 carbonos o vía de Hatch-Slack
Chip Chip o microarreglo
CICVyA Centro de Investigación en Ciencias Veterinarias y Agronómicas
CNIA Centro Nacional de Investigaciones Agropecuarias
CONICET Consejo Nacional de Investigaciones Científicas y Técnicas
CP Componentes principales
d.e. Desvío estándar
DAPC Análisis Discriminante de Componentes Principales o Discriminant Analysis of
ddRADseq Double-digest Restriction-site Associated DNA sequencing
EEA Estación Experimental Agropecuaria
EST marcadores de secuencia expresada o por la sigla en inglés expressed sequence tag
FASTA formato FASTA
Fastq formato fastq de calidad de secuencias
FS Esquemas de filtrado
G Guanina
Gb Giga pares de bases
GBS Genotipado por secuenciación o Genotyping by Sequencing
gff3 General Feature Format 3
GL Grupos de ligamiento
GO Gene Ontology
GWAS Estudio de Asociación de Genoma Amplio o Genome wide association study
ha hectárea
HC cobertura de secuenciación alta
IABiMo Instituto de Agrobiotecnología y Biología Molecular (UEDD INTA-CONICET)
ie “id est” o “es decir”
INDEC Instituto Nacional de Estadística y Censos

INDEC Instituto nacional de estadísticas y censos

InDel Inserciones/deleciones

INTA Instituto Nacional de Tecnología Agropecuaria

INTA Instituto Nacional de Tecnología Agropecuaria

LD Desequilibrio de ligamiento

LD-kNNi algoritmo Linkage Disequilibrium k-nearest neighbor genotype

MAF Frecuencia alélica mínima o Minimum allele frequency

MAS Selección asistida por marcadores o Marker Assited Selection

Mb mega pares de bases

MboI Enzima de restricción MboI

MC secuenciación media

Mpb o Mb Mega pares de bases

msnm metros sobre el nivel del mar

n° número

NCBI Centro Nacional para la Información Biotecnológica o National Center for Biotechnology Information

Ne-70 75 pb lecturas simples recortados a 70 pb; ~cuatro millones de lecturas/individuo

NEA noreste argentino

NEB New England Biolabs

NextSeq Equipo de secuenciación de la empresa Illumina

NGS Next Generation Sequencing

No-150 Lecturas de 150 pb pareadas; ~diez millones de lecturas/individuo, denominada de aquí en más cobertura de secuenciación alta

No-150-PE/MC lecturas No-150 se submuestrearon de HC a MC

No-70-PE/HC lecturas obtenidas en No-150 se acortaron a 70 pb (manteniendo la alta cobertura de secuenciación y pareadas

No-70-PE/MC 4 millones de lecturas/individuo seleccionadas aleatoriamente de No-70-PE/HC

NOA noroeste argentino

NovaSeq Equipo de secuenciación de la empresa Illumina

pb pares de bases

PC componente principal

PCA Análisis de componentes principales

PCR Reacción en cadena de la polimerasa o Polymerase Chain Reaction

PE lecturas o secuencias pareadas Paired End

PMGCA Programa de Mejoramiento Genético de Caña de Azúcar

Principal Components

PstI Enzima de restricción PstI

QTLs loci de caracteres cuantitativos o Quantitative Trait Loci

R1 Lecturas leídas desde el comienzo del fragmento

R2 Lecturas leídas desde el final del fragmento

RAD Secuenciación de ADN asociada a los sitios de restricción o Restriction-site Associated DNA sequencing

RAPD amplificación aleatoria de ADN polimórfico

RD profundidad de lectura

RFLP polimorfismos de longitud de fragmentos de restricción

rho medida de la correlación (la asociación o interdependencia) entre dos variables aleatorias

S. Saccharrum

SAM Archivo en formato Sequence Alignment Map

SE lecturas o secuencias Single end

SNPs Polimorfismo de Nucleótido Sinple o Single Nucleotide Polymorphism

SSR Repetición de secuencia simple polimórfica o Simple sequence Repeat o microsatélites

T timina

USA Estados Unidos de América

USDA Departamento de Agricultura de los Estados Unidos de América

UTR unstranlated region o regiones no traducidas de los genes

VCF Formato de llamada variante o Variant Call Format

VEP predicción de efectos de las variantes del Ensembl Plants

vs versus

°C Grados centígrados

µm micrómetro

Introducción general

Taxonomía, origen y domesticación

La caña de azúcar (*Saccharrum sp*.) es una planta C4, que pertenece al género *Saccharum*, a la familia Poaceae (o Gramineae). El nombre caña de azúcar refiere cualquier especie de las diversas que pertenecen al género *Saccharum*. La clasificación taxonómica de estas especies es difícil, confundida por muchos años entre los géneros emparentados (Evans et al. 2022). Originalmente se consideraba que existían seis especies fundadoras: *S. spontaneum* (x=8; 2n=40-126)*, S. robustum* (x=10; 2n=60-80)*, S. officinarum* (x=10; 2n=80-120)*, S. sinense* (2n= 80-124)*, S. edule y S. barberi* (2n=116-120) (Paterson et al. 2012). En la actualidad suele excluirse a *S. edule* debido a que estudios *in situ* han revelado que es un probable híbrido entre *S. officinarum* y *S. robustum*, con lo cual se redujo el número de especies fundadoras a cinco. *S. robustum* y *S. spontaneum*, se consideran silvestres, y las restantes son cultivadas (Nayak et al. 2014; Zhang et al. 2019; Fickett et al. 2020). Los mapas genéticos sugieren que las diferencias en el número básico entre *S. officinarum* y *S. spontaneum* se deben a simples eventos de fusión o fisión (Grivet y Arruda 2002).

Se sostiene que la caña de azúcar se originó en el Pacífico Sur, pero fue ampliamente dispersada por los primeros exploradores, lo que dificulta precisar el origen exacto. *S. robustum*, es autóctona de Nueva Guinea y se encuentra en las riberas de los ríos. Por su parte, *S. spontaneum* se presume originaria de la India, pero puede encontrarse en estado silvestre desde el este y el norte de África, a través de Oriente Medio, hasta la India, China, el Sudeste Asiático y, a través del Pacífico, hasta Nueva Guinea. *S. officinarum* se considera que también es originaria de Nueva Guinea y que probablemente derivó de *S. robustum*. De esta forma, los cultivares comerciales modernos tienen genomas poliploides complejos como resultado de muchas generaciones de hibridación (Shearman et al. 2022). La caña energía posee una definición menos rigurosa que las variedades de caña de azúcar, y es aplicada a los cultivares de caña que tienen mayor biomasa, aunque menor concentración de sacarosa. Las variedades de caña energía, son una forma de caña de azúcar cultivada específicamente para su aprovechamiento como materia prima en la producción de bioenergía. Estas variedades provienen de la introgresión de genes de alta biomasa en los programas de mejoramiento de Estados Unidos de los años 70, con parentales que eran retrocruzamientos con *S. spontaneum* (Sandhu et al. 2016). Estas variedades poseen un alto potencial como fuente de lignocelulosa, lo que convierte al cultivo en una opción prometedora para la producción sostenible de biocombustibles. Se ha reportado que las estrategias fisiológicas de estas variedades son

diferentes a la observada por los cultivares de caña de azúcar, como, por ejemplo, que el desarrollo de brotes empieza antes del crecimiento de las raíces, sin embargo, poseen un desarrollo rápido y vigoroso, superando el volumen de raíces de la caña de azúcar (Abreu et al. 2020).

Acevedo et al. (2017) estudiaron las genealogías de 16 cultivares utilizados en la provincia de Tucumán (nueve argentinos y siete proveniente de Estados Unidos) y encontraron que todos ellos difieren en la proporción de especies fundadoras. También confirmaron que *S. officinarum* es el mayor contribuyente, seguido por *S. barberi*, y *S. spontaneum*. Además, LCP 85-384, el genotipo más cultivado (67,7%, según Ostengo et al. 2021) en Tucumán, tiene 3 genotipos de sorgo en su genealogía (Sorghum bicolor cv Wiley, TADA y Rex) (Acevedo et al. 2017).

Generalidades de la caña de azúcar

La caña de azúcar es un cultivo de gran importancia económica en el mundo (1.949.310.108 toneladas producidas al año) y es el cultivo azucarero más importante de Argentina (Walton 2020). La superficie plantada en Argentina con este cultivo es de aproximadamente 350.000 ha, siendo el cultivo industrial de mayor extensión en el país (INDEC, 2021). Tucumán es la provincia que posee la mayor extensión de superficie plantada con este cultivo (73%), seguida por las provincias de Salta (17%) y Jujuy (9%) (Benedetti, 2018). El porcentaje restante se divide entre las provincias de Santa Fe (0.8%) y Misiones (0.4%) (Benedetti, 2018). Los datos más recientes del Censo Nacional Agropecuario indican que Argentina registra 3200 explotaciones dedicadas al cultivo de caña de azúcar (INDEC, 2021). El sector agroindustrial vinculado a este cultivo destaca sobremanera en el ámbito social debido a que está representado por un gran número de explotaciones pequeñas y medianas que resultan un eslabón fundamental en el desarrollo de las economías regionales del NOA y NEA (INDEC, 2021).

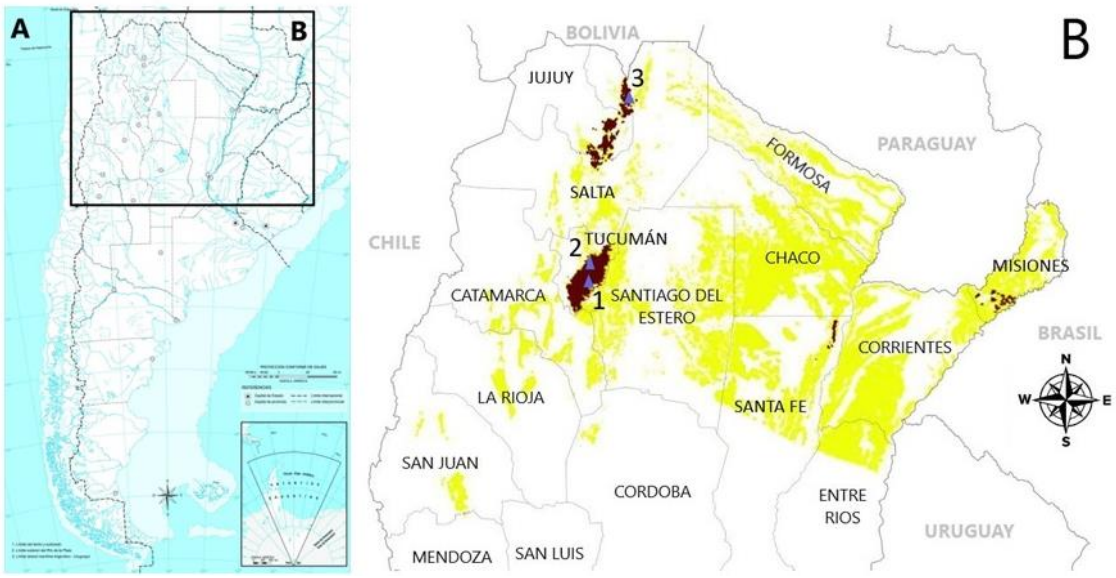

Figura i1. *Áreas actuales y potenciales de cultivo de caña de azúcar en Argentina*. A: Argentina. B: Norte de Argentina: área potencial de expansión de la caña de azúcar (amarillo), área cultivada de caña de azúcar en 2018 (marrón) y estaciones de mejoramiento (1: Estación Experimental Agropecuaria Famaillá del Instituto Nacional de Tecnología Agropecuaria, 2: Estación Experimental Obispo Colombres de la Provincia de Tucumán, 3: Chacra Experimental Agropecuaria Santa Rosa). Modificado de Benedetti (2018).

En Argentina, las condiciones climáticas conllevan a una baja fertilidad natural, y no es posible la reproducción artificial porque las plantas no llegan a la etapa de floración, o, si lo hacen, su polen no es fértil (Budeguer et al. 2021). Asimismo, la falta de sincronización floral entre los diferentes genotipos provoca momentos distintos de floración entre los cultivares, lo que aumenta las probabilidades de que haya autopolinización (García et al. 2020). La caña de azúcar se propaga vegetativamente mediante trozos de tallo que contienen yemas fértiles, lo que significa que las variedades son clones que se reproducen a partir de esquejes de tallo (Grivet y Arruda, 2002). Durante el primer año, se denomina "caña planta". Esta plantación permanece en el campo por un periodo de 4 a 5 años, con rebrotes y cosechas anuales. Cada rebrote se conoce como "caña soca", seguido del número de cosecha correspondiente (García et al. 2020). Tradicionalmente, los programas de mejoramiento se han enfocado en el incremento del contenido de azúcar en tallos por hectárea. El cambio en la forma de uso de la caña de azúcar realza la importancia de la implementación de estrategias de selección, en las que el análisis de los genotipos es la clave para revelar la variabilidad natural de su genoma.

Usos de la caña de azúcar

La caña de azúcar cumple con varios criterios que la califican como un cultivo energético. En primer lugar, presenta una alta productividad de materia seca por área y un balance

energético muy favorable al considerar toda la cadena productiva, ya que la cantidad de energía obtenida es significativa en comparación con la invertida (Acevedo et al. 2021). Además, la materia prima está disponible durante un largo período, ya que el contenido de fibra no varía considerablemente a lo largo del crecimiento, lo que la convierte en una opción económica a largo plazo (Matsuoka et al. 2014). La producción de bioetanol de primera generación a partir de la fermentación de jugos o melazas es muy atractiva para la industria energética.

En la campaña 2018/2019, la producción nacional alcanzó los 16,8 millones de toneladas de azúcar y 4,7 millones de toneladas de bioetanol 1G. En la provincia de Tucumán en el 2020 se registraron 260.800 hectáreas cultivadas, que producirán 1,4 millones de toneladas de azúcar (78% de la producción total argentina) y más de 300.000 metros cúbicos de etanol (Di Pauli et al.2021). Además, en esta provincia se encuentran 15 de los 22 ingenios azucareros del país (Ministerio de Hacienda, 2018), algunos de los cuales también producen bioetanol 1G. Los ingenios en Tucumán son: Bella Vista, Concepción, Famaillá, La Corona, La Florida, La Trinidad, Leales, Marapa, San Juan, Santa Bárbara y Santa Rosa. En Salta se ubican Alconoa S.R.L. y Bio San Isidro S.A., y en Jujuy, Bio Ledesma S.A. y Río Grande (García et al. 2020).

Otra aplicación energética de la caña de azúcar es la producción de briquetas de carbón a partir del residuo agrícola de cosecha. Después de la cosecha, quedan aproximadamente 15 toneladas (peso fresco) por hectárea de residuo agrícola en el campo (García et al. 2020). También se aprovecha el bagazo, un subproducto de la industria azucarera, para la cogeneración de energía en las calderas de los ingenios. El excedente de energía generado, equivalente a 100 MW, se integra al sistema energético argentino (García et al. 2020). El aumento del contenido de fibra en los materiales de caña energía, incorporados a los programas de mejoramiento, ofrece ventajas importantes como una mayor vigorosidad y rusticidad. Esto incrementa la resistencia a varios factores de estrés, aportando beneficios económicos y ambientales. Una mejor adaptación a regiones marginales evita la competencia por tierras destinadas a la producción de alimentos, y su baja demanda de fertilizantes y pesticidas reduce el impacto ambiental (Carvalho-Netto et al. 2014). Sin embargo, el aumento de fibra afecta negativamente la degradabilidad de la pared celular, por lo que es fundamental caracterizar los materiales para lograr un equilibrio entre el aumento de biomasa y la eficiencia de conversión en biocombustibles (García et al. 2023b).

Estructura celular del tallo de la caña de azúcar

Para comprender los procesos bioquímicos y fisiológicos que influyen en las funciones de los tejidos y órganos, así como los rasgos que integran funciones en toda la planta, es necesario dilucidar los procesos de desarrollo y regulaciones de células individuales que componen tejidos complejos. La genómica, transcriptómica, proteómica y metabolómica permiten cuantificar las funciones de las redes reguladoras de genes y proteínas en órganos, tejidos y células durante el desarrollo y en respuesta a perturbaciones ambientales (Fu et al. 2023). El resultado de esta información es la capacidad de realizar modificaciones genéticas centradas en la obtención de fenotipos de interés con mayor productividad, resistencia a estreses bióticos o abióticos, o la síntesis de nuevos compuestos (de Carvalho et al. 2019).

El segmento cosechable de la planta de caña de azúcar es su tallo, una estructura compuesta por una sucesión de nudos y entrenudos cilíndricos alternados. Los entrenudos, que son las porciones del tallo situadas entre dos nudos, exhiben una disposición que contribuye a la morfología general de la planta. Dentro de la región central, denominada médula, se encuentran células del parénquima, que se caracterizan por su notable tamaño en comparación con las células de los tejidos circundantes (Glyn et al. 2004). La actividad de división y expansión celular en este tejido específico desempeña un papel fundamental en el aumento dimensional, tanto en sentido radial como longitudinal (Moore et al. 2014).

El sistema vascular, que incluye el xilema y el floema, resulta crucial para el transporte de nutrientes y agua (Carlquist et al. 2011). El xilema de las monocotiledóneas exhibe características únicas, como elementos vasculares con paredes con un patrón de engrosamiento en forma escalonado, que se forma debido a la deposición de materiales como lignina y otros compuestos estructurales durante el desarrollo de la célula. Esto proporciona resistencia y soporte a los elementos del xilema, permitiendo así una eficiente conducción de agua y nutrientes a través de la planta. Este tipo de estructura es característico de las monocotiledóneas y contribuye a su adaptación a diferentes condiciones ambientales. La disposición ordenada de los haces vasculares en los tallos indica una organización específica que favorece su crecimiento general (Fukuda et al. 2018).

Las paredes celulares de las plantas evolucionaron para evitar el ataque de patógenos, asegurar la rigidez de la planta y reducir la pérdida de agua (Moore et al. 2014). Presenta una estructura compleja compuesta por polisacáridos, proteínas, enzimas, polímeros fenólicos y otros elementos secretados por la célula. Su matriz, con la capacidad de lignificarse, se compone principalmente de hemicelulosas, pectinas y glucoproteínas estructurales. Las hemicelulosas, son polisacáridos no celulósicos que recubren y cristalizan las fibrillas de celulosa, formando

cadenas cohesionantes o actuando como recubrimiento resbaladizo. La combinación de microfibrillas mediante hemicelulosas forma macrofibrillas sólidas. Las pectinas y las hemicelulosas (homogalacturonanos, rhamnogalacturonanos, arabinanos, galactanos) contribuyen a la unión de microfibrillas y mantienen la hidratación en paredes celulares jóvenes (Ferreira et al. 2016). Las sustancias incrustantes como lignina, cutina y suberina aportan rigidez e impermeabilidad. También están presentes los mucílagos, ricos en polisacáridos. Los compuestos pécticos y proteínas estructurales, ricas en serina e hidroxiprolina, vinculadas con azúcares, contribuyen a la fuerza mecánica de la pared. Aunque el papel exacto de las proteínas estructurales no se comprende completamente, se sugiere que actúan como elementos de unión, formando cadenas que conectan otros componentes (Chang et al. 2022).

Generalidades de las técnicas genómicas

El estudio inicial de los genomas de las plantas se llevó a cabo mediante marcadores moleculares de bajo rendimiento, como los polimorfismos de longitud de fragmentos de restricción (RFLP), o amplificados (AFLP), amplificación aleatoria de ADN polimórfico (RAPD), y marcadores de secuencia expresada (o EST por la sigla en inglés expressed sequence tag). Esto proporcionó información sobre la organización y el contenido de los genomas, así como sobre la funcionalidad de los genes y del ADN no codificante, pudiendo determinar su ubicación precisa (Thirugnanasambandam et al. 2018). Estos marcadores se caracterizan por su precisión, reproducibilidad y, en algunos casos, neutralidad. Los ejemplos en la caña de azúcar involucran proyectos que han identificado la asociación de varios marcadores como RFLP, AFLP, microsatélites (SSR) y marcadores por PCR para generar mapas genéticos o utilizarlos para evaluar la presencia de un gen, como por ejemplo al gen resistente a la roya (Bru1) (Asnaghi et al. 2004; Costet et al. 2012; Balsalobre et al. 2017; Yang et al. 2018).

Los polimorfismos de un solo nucleótido (o SNPs por las siglas en inglés *Single Nucleotide Polymorphism)* se refieren a las variaciones de una sola base en la secuencia de ADN que ocurren comúnmente en sitios homólogos a lo largo del genoma. Este fenómeno es frecuente y tiene diversas aplicaciones en genética, incluyendo la identificación de alelos vinculados a enfermedades, estudios poblacionales y la creación de mapas de ligamiento de alta densidad. En el ámbito de las plantas cultivables, se ha investigado exhaustivamente la presencia y utilidad de los SNPs como marcadores moleculares. Estos marcadores son ideales debido a su herencia codominante, excelente reproducibilidad, asignación cromosómica específica y distribución a lo largo de todo el genoma. Los SNPs revelan polimorfismos que podrían pasar desapercibidos con

otros métodos, y su capacidad de automatización los convierte en herramientas ideales para un genotipado eficiente y de alto rendimiento (Parida et al. 2016).

Con la llegada de las tecnologías de NGS, los marcadores SNPs cobraron mayor relevancia por su eficiencia ya que pueden automatizarse fácilmente y proporcionar información a lo largo de todo el genoma. Debido a su abundancia en el genoma y a su bajo costo de detección, los marcadores SNPs han superado el uso de los microsatélites, aunque estos sigan utilizándose por ser altamente informativos. Se sabe que los SNPs aparecen con frecuencias que oscilan entre uno por cada 100-500 pb en los genomas de las plantas, y que la cantidad varía entre especies. Cuando un nucleótido de una accesión difiere del genoma de referencia en la misma localización nucleotídica, se detecta un SNP. En ausencia de un genoma de referencia, se detecta una variación comparando lecturas de diferentes genotipos mediante metodologías de ensamblado *de novo*. Los SNPs se han convertido en los marcadores preferidos del agro, debido a su densidad en el genoma, a su capacidad para identificar el dosaje alélico, con baja tasa de error, y un relativo bajo costo (Manimekalai et al. 2020).

Los SNPs tienen una ventaja significativa sobre otros tipos de marcadores en los poliploides como la caña de azúcar porque son codominantes, están extendidos por todo el genoma y pueden determinar el número de copias de cada alelo en un locus polimórfico dado (dosaje alélico) a través de los subgenomas (más de dos cromosomas por grupo homeólogo) (García et al. 2013). A su vez, se cree que el dosaje alélico está implicado en la variación de la expresión génica y, por tanto, en la variación fenotípica causada por la abundancia relativa de cada alelo (García et al. 2013). Aitken (2021) ha publicado recientemente una revisión exhaustiva de SNPs y otros marcadores moleculares en caña de azúcar. La implementación de SNPs y datos fenotípicos ayuda a identificar las correlaciones entre estas variantes y los loci de rasgos cuantitativos, asociaciones que pueden ser potencialmente útiles para futuros programas de mejoramiento de la caña de azúcar (You et al. 2019; Yadav et al. 2021; Mahadevaiah et al. 2021).

La introducción de enfoques como genotipificación por secuenciación (GBS) y otras tecnologías NGS ha dado lugar a un crecimiento significativo en el número de especies secuenciadas, en particular las especies no modelo, así como el número de marcadores disponibles para el análisis (Koseva et al. 2017). En comparación con las tecnologías anteriores, permiten construir mapas de ligamiento genético más densos, valiosos en escenarios donde los sistemas alternativos de marcadores de alta densidad son inviables (costosos de desarrollar y probar). Las técnicas GBS aprovechan las bibliotecas de representación reducida de las plataformas Illumina y pueden utilizarse sin un genoma de referencia para la detección de SNPs (Balsalobre et al. 2017).

Generalmente, los errores de genotipado causados por una baja cobertura de secuenciación se eliminan mediante filtrado, por ejemplo, marcando como ausentes los genotipos que poseen una profundidad de lectura inferior a algún valor umbral o descartando variantes mediante criterios de calidad establecidos (Blischak et al. 2017). Sin embargo, esto requiere una secuenciación más profunda, lo que se traduce en un menor número de individuos secuenciados y un menor número de loci utilizados a un costo determinado, pudiendo dejar una parte considerable de los datos originales inutilizables. Se han desarrollado varias técnicas para imputar genotipos faltantes y corregir genotipos erróneos en datos de secuenciación genómica de baja profundidad de lectura (Fu 2014; Hundia et al. 2022); pero estos algoritmos están diseñados exclusivamente para poblaciones consanguíneas. Otro problema es la presencia de errores de secuenciación, lo que lleva a inflar las distancias genéticas si no se tiene en cuenta. A diferencia de los errores producidos por una baja profundidad de lectura, los errores de secuenciación pueden hacer que los homocigotas se etiqueten como heterocigotas (Bilton et al. 2018).

El mejoramiento vegetal ha incorporado, en los últimos años, la utilización de la selección asistida por marcadores moleculares, fundamentalmente a través de técnicas de genotipado como los microarreglos de SNPs o GBS (Manimekalai et al. 2020). Al comparar los dos procesos, cada uno ofrece ventajas y desventajas para su aplicación en la caña de azúcar. Los microarreglos de SNPs son más fáciles de usar, siendo más sencilla también la interpretación de los datos, sin necesidad de un análisis bioinformático significativo. Estos son una herramienta muy utilizada en biología molecular para estudiar los genomas, ya que contienen miles o incluso millones de sondas de ADN dispuestas en una disposición ordenada diseñadas a medida por los investigadores. Estos microarreglos tienen una amplia gama de usos, como pueden ser los estudios de diversidad genética en caña de azúcar. Los marcadores SNPs de los poliploides que se representan en el microarreglo se determinan mediante la selección de variantes, el dosaje de alelos y la validación dependiendo de la calidad. De esta forma, se desarrolló un microarreglo de SNPs (Axiom Sugarcane 100K) utilizado para crear mapas genéticos de caña de azúcar e identificar QTL vinculados a la resistencia al virus de la hoja amarilla (You et al. 2019). Sin embargo, son de elevado costo inicial, al ser diseñados con sondas específicas tienen capacidad limitada para adaptarse a nuevas variedades y dependen mucho de la información genómica disponible (Manimekalai et al. 2020). Para que un microarreglo de SNPs tenga una representación efectiva en caña de azúcar, es necesario una cantidad significativa de datos de secuenciación debido a la ploidía y al tamaño de su genoma. Cuando se dispone de un microarreglo SNPs de alta densidad o de un microarreglo específico para un rasgo, el análisis de

datos y el genotipado pueden llevarse a cabo sin ambigüedades. A pesar de esta ventaja, los SNPs de los microarreglos representan una muestra sesgada de todos los SNPs que segregan en la población de interés, mientras que con GBS no existe esta limitación. Se ha demostrado que GBS es un mejor enfoque para especies no modelo, ya que no requiere genoma de referencia, aunque tiene limitaciones computacionales cuando se analizan conjuntos de datos muy grandes. Por otro lado, el enfoque GBS requiere el uso de una rutina bioinformática especializada para analizar los resultados para la detección de variantes dentro de la población secuenciada (Rochette et al. 2019).

Análisis multivariados aplicados a datos genómicos

Para el análisis de datos genómicos, existen cuatro métodos principales que permiten inferir relaciones filogenéticas y entender la evolución molecular: la estimación de la parsimonia, los métodos de distancia, la máxima verosimilitud y el análisis bayesiano (Palacio et al. 2020). Los métodos de estimación de la parsimonia buscan la simplicidad en la explicación de la evolución, minimizando el número de cambios evolutivos necesarios para explicar las diferencias observadas en las secuencias. Esto se logra mediante la identificación de características compartidas entre grupos de organismos (Smith 2019). Los métodos de distancia, como el método de neighbor-joining, calculan distancias evolutivas entre secuencias y construyen árboles filogenéticos basados en estas distancias. Aunque son rápidos computacionalmente y útiles para grandes conjuntos de datos, su validez filogenética puede ser cuestionada debido a las limitaciones en la interpretación de las distancias evolutivas (Palacio et al. 2020). La máxima verosimilitud y el análisis bayesiano son métodos probabilísticos que modelan la evolución molecular y buscan el árbol filogenético que maximiza la probabilidad de observar los datos. La máxima verosimilitud asume que los datos observados son más probables bajo un conjunto particular de parámetros del árbol y del modelo evolutivo, mientras que el análisis bayesiano incorpora información a priori sobre los parámetros y actualiza estas creencias en función de los datos observados (Felsenstein 2004; Chen et al. 2014).

Mapas de ligamiento

Los mapas de ligamiento se utilizan para analizar loci de rasgos cuantitativos (QTL), comprender los procesos evolutivos, evaluar la colinealidad del genoma y comprender los procesos meióticos. La construcción de estos mapas se basa en la detección de eventos de recombinación entre sitios genómicos y la cuantificación de las fracciones de recombinación de

a pares de loci. Este fenómeno es fácilmente apreciable en organismos diploides, donde las cromátidas hermanas intercambian segmentos después de una duplicación homóloga. Al comparar la composición cromosómica de parentales y progenies, la presencia de marcadores informativos, como los SNPs, permite determinar la fracción de recombinación entre pares de loci. Para organismos poliploides, la generación de mapas de ligamiento es más desafiante. A diferencia de los diploides, donde un marcador bialélico proporciona información completa para distinguir entre los dos alelos presentes en un par de homólogos, en los poliploides se deben considerar las proporciones de dosis alélicas. Estas limitaciones dificultan la evaluación de los eventos de recombinación durante la meiosis, especialmente a medida que aumenta el nivel de ploidía. Para obtener información multialélica en especies poliploides, es necesario tener en cuenta las frecuencias de recombinación, estimar las configuraciones de fase y reconstruir los haplotipos tanto de los progenitores como de los individuos de la población. (Wu et al. 2002; Margarido et al. 2007; Serang et al. 2012)

El limitado número de divisiones meióticas desde los primeros cruzamientos artificiales que condujeron a los cultivares modernos (entre 15-20 genotipos) ofreció escasas oportunidades para la recombinación de cromosomas de las especies fundadoras (Yadav et al. 2020). Además, se ha hipotetizado que la escasez de cromosomas recombinantes se debe al reducido número de individuos fundadores de *S. officinarum* y *S. spontaneum*. Como resultado, persiste un alto desequilibrio de ligamiento entre los cultivares modernos y una base genética estrecha en el germoplasma moderno de caña de azúcar, como se ha confirmado en una muestra de cultivares de la Isla Mauricio, donde ciertos haplotipos cromosómicos permanecen significativamente conservados en regiones de hasta 10 centiMorgans. Este descubrimiento podría facilitar la identificación y localización de genes relacionados con características de interés. En los programas de mejoramiento modernos, se sigue utilizando ampliamente material genético relativamente antiguo (>50 años) en diseños de cruzamientos para crear nuevas variedades, razón por la cual ha habido pocas oportunidades para la recombinación cromosómica a partir de los fundadores originales (Yadav et al. 2020).

Aunque la poliploidía complica el mapeo, el alto desequilibrio de ligamiento, en comparación con otros genomas, puede ofrecer algunas ventajas, como la posibilidad de trabajar con una baja densidad de marcadores para la construcción de mapas genéticos. Sin embargo, el desarrollo de un mapa saturado de baja densidad para la caña de azúcar requiere un esfuerzo considerablemente mayor que para un diploide. La cobertura de marcadores en los cultivares actuales es desigual, con los cromosomas de *S. spontaneum* siendo más densamente cubiertos que los de *S. officinarum* (Grivet y Arruda 2002). Se han propuesto nuevos modelos

teóricos para una extracción más precisa y exhaustiva de la información para el mapeo en especies poliploides complejas (Mollinari et al. 2019).

Genética del cultivo

El genoma monoploide de la caña de azúcar (cultivar R570) posee un tamaño que oscila entre 760 y 926 Mb, que es casi el doble del tamaño del genoma del arroz (389 Mb) y similar al del sorgo (760 Mb) (Garsmeur et al. 2018). A través de un estudio por citometría de flujo se determinó que la caña de azúcar tiene un gran genoma de casi 3,80-10,96 pg. (Manimekalai et al. 2020). Además, es un cultivo aneuploide, donde cada individuo exhibe pérdidas variables de cromosomas, lo que resulta en una de las principales razones por las cuales el número total de cromosomas varía entre los distintos genotipos (Vieira et al. 2018; Thirugnanasambandam et al. 2018). Con 10 grupos de ligamiento homólogos, el tamaño del genoma total estimado de los híbridos de caña de azúcar es de 10 Gb (Xiong et al. 2023). A pesar de que durante el apareamiento cromosómico en la meiosis se observan generalmente bivalentes, existe un patrón complicado y desequilibrado dentro de algunos grupos homólogos (Xiong et al. 2023).

En el género *Saccharum*, todas las especies son poliploides y han experimentado, al menos, dos duplicaciones completas del genoma desde su divergencia de un antepasado común con el sorgo, hace aproximadamente 8 millones de años, para convertirse en octoploides. A pesar de que cada octoploide posee ocho genomas, la identificación de cada genoma individual es compleja debido a que cada cromosoma puede aparearse y recombinarse en bivalentes o multivalentes con cualquiera de los otros siete cromosomas homeólogos durante la meiosis, generando un mosaico de los ocho genomas (Grivet y Arruda 2002; Vieira et al. 2018). Las características distintivas de los genomas de las variedades cultivadas incluyen un alto nivel de ploidía, aneuploidía, el origen al menos biespecífico de los cromosomas proveniente de las diversas especies fundadoras con sus diferencias estructurales, y la presencia de recombinantes cromosómicos interespecíficos (Grivet y Arruda 2002).

En los cultivos poliploides, la detección precisa de SNPs polimorfismos de un solo nucleótido (SNPs) es más compleja debido al mapeo ambiguo de loci homeólogos muy similares, lo que aumenta la tasa de falsos positivos de los SNP detectados (Fig. i2). Parida et al. 2016, detectaron que la frecuencia de SNPs por secuenciación de amplicones en genotipos de caña de azúcar (1 SNP/118,1 pb) fue inferior en comparación con un análisis *in silico* (1 SNP /130,1 pb). Alrededor del 61,6% de los SNPs predichos *in silico* en genes del metabolismo de los hidratos de

carbono, y el 45,2% en genes de resistencia a enfermedades en *S. officinarum,* fueron validados por secuenciación de amplicones PCR clonados (Parida et al. 2016).

Figura i2: *Ejemplo de SNPs en genotipos poliploides.* La barra azul representa la secuencia consenso de referencia. Las barras verdes representan una secuencia derivada de un subgenoma. Las barras naranjas representan la secuencia alternativa derivada de otro subgenoma. Las bases en rojo son variaciones entre subgenomas de un genotipo y, en este caso, son los SNP de anclaje. Las bases en amarillo son los verdaderos SNP entre genotipos. (Adaptado de Clevenger y Ozias-Akins 2015)

En un microarreglo de SNPs desarrollado para caña de azúcar lograron recuperar una media de 10000 SNPs por cromosoma y una densidad media de un SNPs cada 6832 pb (You et al. 2019). Estudios en el transcriptoma de las especies fundadoras revelaron que se encuentra una menor densidad de SNPs en la *S. officinarum* domesticada en comparación con las especies silvestres. En el transcriptoma de la hoja, *S. officinarum* mostró una media de 32,3 SNPs por kb, un 8% menos que *S. spontaneum* y *S. robustum*, que mostraron 34,9 y 35,1 SNPs por kb, respectivamente (Arro et al. 2016). En el transcriptoma del tallo, *S. officinarum* tenía una media de 45,8 SNPs por Kb, un 11% menos que *S. spontaneum*, que mostró 52,0 SNPs por Kb. Estos SNPs fueron 97-98 % bialélicos y 2-3 % multialélicos. Esto afirma que los SNPs bialélicos son las variantes SNPs más abundantes incluso en la caña de azúcar. *Saccharum* se caracteriza por la prevalencia de alelos de frecuencia intermedia en el contexto de una reproducción predominantemente asexual (Arro et al. 2016).

En las regiones de exones, se observó una densidad menor de SNPs, con una tasa de 1 SNP por cada 82 pares de bases, en comparación con las regiones de intrones, donde se registró una tasa de 1 SNP por cada 55 pares de bases. Se identificaron en total de 170544 SNPs no sinónimos y 156816 SNPs sinónimos en las regiones de exones. Además, se encontraron 3519 SNPs que

generan codones stop prematuros, 92 sitios de SNPs que ocasionan mutación del codón de inicio y 1558 sitios de empalme. Los SNPs que no se encontraron en estas categorías, 112803 (10,3%), se localizan en regiones intergénicas, mientras que 147.958 (13,4%) se encuentran en las regiones UTR (Song et al. 2016).

Genomas de referencia disponibles para *Saccharum sp*.

Las especies estrechamente emparentadas con genomas bien ensamblados y anotados suelen utilizarse para ayudar a ensamblar genomas complejos, como es el caso de las especies del género *Saccharum* (Manimekalai et al. 2020). El genoma de *Sorghum bicolor* ha sido utilizado como modelo de referencia en estudios comparativos debido a su alta afinidad con la caña de azúcar. En 2018, se publicó un genoma ensamblado a nivel de cromosomas del híbrido R570, originado a partir de la sintenia con el sorgo (Garsmeur et al. 2018). Para ello, se utilizaron clones de cromosomas artificiales bacterianos (BAC), que fueron mapeados a esta especie emparentada, para reconstruir los cromosomas haploides. Esta referencia representa muy bien la parte rica en genes ya que, además de la anotación en pares de bases (formato FASTA), posee la anotación funcional (formato gff). Además, este genoma monoploide posee un tamaño razonable (530 Mb), lo cual representa una ventaja para la demanda computacional que estos análisis requieren (Garsmeur et al. 2018).

Diferentes proyectos de secuenciación publicaron genomas de variedades de caña de azúcar, a partir del 2017, pero de forma parcial (contig o scaffold) (Manimekalai et al. 2020). Sin embargo, recientemente fueron publicados varios genomas del género *Saccharum* a nivel de cromosomas, utilizando secuenciación de nueva generación (NGS). El híbrido de caña de azúcar CC 01-1940 fue secuenciado utilizando tecnologías NGS (lecturas largas PacBio, lecturas cortas Illumina y lecturas HI-C) (Trujillo-Montenegro et al. 2021). Las especies fundadoras de los híbridos actuales *S. officinarum* (acceso al código del ensamblado del GenBank: GCA_020631735.1; BioProject: PRJNA744175) y *S. spontaneum* Np-X (Zhang et al. 2022; acceso al código del ensamblado del GenBank: GCA_022457205.1) fueron secuenciadas combinando lecturas largas PacBio y lecturas cortas Illumina.

En marzo del 2024 finalmente se publicó un nuevo ensamblado del genoma del cultivar R570 que incluye un ensamblado "primario" de 67 pseudocromosomas organizados en 10 grupos homólogos. Esto representa 5,04 Gb de secuencias, en la que se predijeron 194,593 modelos de genes codificadores de proteínas. Además, se construyó un ensamblado "alternativo" de 3,7 Gb de contigs casi idénticos al ensamblado primario pero que contienen suficientes variaciones como para ser ensamblados por separado. La mitad del genoma consiste

en múltiples copias (2-4x) de segmentos idénticos o casi idénticos que se colapsaron en el ensamblado. Considerando estos segmentos colapsados, el tamaño total del ensamblado del genoma representa 9,3 Gb (Healey et al. 2024).

Mejoramiento genético en caña de azúcar en el país y el mundo

Dentro del complejo *Saccharum* existe una gran variabilidad natural. Dado que los caracteres fenotípicos están sujetos a la influencia de factores medioambientales, el análisis del perfil molecular emerge como una herramienta fundamental para la identificación de las combinaciones óptimas entre accesiones en los cruzamientos, así como para la estimación precisa de la diversidad genética (Mahadevaiah et al. 2021). En colecciones de germoplasma de caña de azúcar, se han empleado diversos tipos de marcadores moleculares, tales como RFLP, el RAPD, el AFLP y SSRs (Perera et al. 2023).

Los principales objetivos de los programas de mejoramiento de caña de azúcar actuales son aumentar el rendimiento total de la biomasa, contenido de azúcar, la resistencia a las enfermedades e insectos, tolerancia al frío y la capacidad de macollaje (Xiong et al. 2023). Los progenitores son seleccionados entonces considerando estas características fenotípicas fundamentales. Para potenciar el contenido de azúcar en los híbridos interespecíficos de caña de azúcar, se utiliza un método conocido como "nobilización". Este proceso implica realizar entre 3 y 6 generaciones de retrocruzamiento, lo que permite estabilizar el genoma y recuperar la característica deseada, como puede ser el alto contenido de azúcar en el tallo (Xiong et al. 2023).

El contenido de fibra es otro rasgo importante considerado en los programas de mejoramiento para aumentar el rendimiento de biomasa (Leal et al. 2013). Sin embargo, como estos programas se han centrado principalmente en el contenido de azúcar, la fibra ha sido valorada negativamente en algunos índices de selección (Wei et al. 2008). Esto puede explicar por qué los cultivares y germoplasmas de élite actuales acumulan mucha sacarosa, pero no logran buenos resultados en la acumulación total de biomasa. Además, las variedades comerciales de caña de azúcar tienen una base genética relativamente estrecha, lo que podría limitar las ganancias de rendimiento en cada nueva variedad comercial (Thiappan Selvi y Nair 2009).

La alta variabilidad genética propia de la caña de azúcar podría utilizarse para desarrollar nuevos genotipos con rasgos industriales con mejores aptitudes, como el aumento del contenido de azúcar y la mejora de la digestibilidad del bagazo. Sin embargo, el mejoramiento

convencional requiere años de pruebas, tanto en invernaderos como en campos experimentales antes de que un nuevo híbrido pueda ser liberado. En este contexto, conocer los perfiles moleculares de las accesiones utilizadas para los cruzamientos y, además, la utilización de marcadores moleculares para asistir a los programas de mejoramiento permite acortar los ciclos de mejoramiento.

Por su parte, los objetivos de los programas de mejoramiento en caña energía se centran en lograr una gran capacidad para generar biomasa, lo que respalda la producción de energía renovable para contrarrestar el impacto del cambio climático. Estas variedades también se caracterizan por su resistencia tanto a factores bióticos como abióticos, lo que permite una producción más eficiente con menor necesidad de insumos como fertilizantes, pesticidas y energía. Esto es especialmente beneficioso en tierras de menor valor agronómico (menos productivas), que presentan limitaciones como menor fertilidad, escasez de agua, temperaturas extremas y suelos salinos.

En Argentina, el Instituto Nacional de Tecnología Agropecuaria (INTA) cuenta con un Programa de Mejoramiento Genético de Caña de Azúcar (PMGCA) que se desarrolla en la Estación Experimental Famaillá en Tucumán. El mismo tiene como objetivo desarrollar variedades de caña de azúcar para Argentina con resistencia/tolerancia a enfermedades y plagas, y que tengan un alto rendimiento tanto en biomasa como en azúcar (Sopena et al. 2015; García et al. 2023a). El PMGCA del INTA se enfoca especialmente en desarrollar híbridos de caña que produzcan grandes cantidades tanto de caña por hectárea como de sacarosa por hectárea (Sopena et al. 2015; García et al. 2023a). Además, se emplean rigurosos estándares fitosanitarios para seleccionar híbridos con resistencia o tolerancia a las principales enfermedades que afectan a los cultivos de caña (Rago et al. 2012), siguiendo una tradición agronómica y fitosanitaria que ha perdurado en los programas de mejoramiento genético de la caña de azúcar en todo el mundo durante más de un siglo. Sin embargo, dada la importancia creciente de la caña de azúcar como cultivo bioenergético, especialmente por su alta producción de biomasa (Santchurn et al. 2019), se han incorporado nuevos criterios de selección centrados en estos objetivos. Entre ellos, se destaca la medición de los niveles de fibra en el tallo, lo que permite evaluar la aptitud de los genotipos de caña para diferentes usos energéticos derivados de la biomasa lignocelulósica. El PMGCA del INTA también ha adoptado esta perspectiva para desarrollar variedades de caña de azúcar versátiles, capaces de utilizarse para producir azúcar y generar

bioenergía, contribuyendo a la reducción del uso de combustibles fósiles y a la mitigación de los efectos del cambio climático (García et al. 2023b).

A lo largo de los años, se han utilizado diversos métodos basados en múltiples disciplinas para apoyar al PMGCA. El uso de herramientas genómicas y metagenómicas permitió desarrollar marcadores moleculares que favorecen a reducir los tiempos de selección (Di Pauli et al. 2017; Molina et al. 2022), estimar por primera vez la composición microbiana de los suelos bajo distintos sistemas de manejo (Fontana et al. 2019) y generar nueva variabilidad genética mediante la mutagénesis in vitro (Di Pauli et al. 2021). La detección y caracterización de los principales patógenos que afectan al cultivo facilitó el descarte temprano de clones en el PMGCA (Fontana et al. 2013; 2018). Además, se logró la detección simultánea de roya marrón y anaranjada (Di Pauli et al. 2015), lo que mejoró las medidas de prevención de estas enfermedades.

Hipótesis

1. Las metodologías de genotipificación de genoma completo, como ser ddRADseq, permiten identificar variantes genómicas de tipo SNPs en genomas complejos, como el de la caña de azúcar;

2. En cruzamientos artificiales de caña de azúcar, el comportamiento genético de la poliploidía genera variabilidad adicional por sí misma. Esta variabilidad sería mayor en la progenie de cruzamientos entre híbridos comerciales que entre los respectivos progenitores y puede caracterizarse mediante la metodología ddRADseq;

3. Existe variabilidad anatómica entre los híbridos LCP85-384 e INTA05-3116, la cual puede ser visualizadas mediante técnicas histoquímicas.

Objetivo General

El objetivo del presente trabajo consiste en desarrollar estrategias y generar conocimiento útil para los programas de mejoramiento de caña de azúcar a partir de la aplicación de nuevas tecnologías de genotipificación de alto rendimiento en poblaciones de mejoramiento de caña de azúcar disponibles en INTA, así como explorar la variabilidad anatómica de entrenudos de híbridos contrastantes como variable informativa para la selección de buenas propiedades biomásicas.

Objetivos específicos

1. Poner a punto un protocolo de identificación de marcadores moleculares de tipo polimorfismo de un sólo nucleótido (SNP) usando la metodología ddRADseq (genotipificación de genoma completo) en caña de azúcar para identificar variantes genómicas potencialmente útiles y aplicables en el programa de mejoramiento del INTA;
2. Evaluar la eficiencia de los paneles de SNPs en la caracterización genética de híbridos de caña de azúcar y el mapeo genético de poblaciones del programa de mejoramiento genético a través del análisis de paneles de SNPs;
3. Predecir el efecto funcional de los SNPs presentes en genes y regiones genómicas candidato para caracteres agroindustriales;
4. Analizar de manera comparativa la anatomía del tallo en clones contrastantes para el contenido de azúcar y fibra: LCP 85-384, híbrido comercial de alto contenido de azúcar e INTA 05-3116, biotipo de caña energía de alto contenido de fibra.

Capítulo 1. Puesta a punto e implementación de un protocolo de genotipificación ddRADseq en caña de azúcar, y su evaluación en variedades del programa de mejoramiento del INTA.

Introducción

En los últimos años, los avances en biología molecular, especialmente, en el desarrollo de la secuenciación de ADN de alto rendimiento asistida por la bioinformática, representan una poderosa herramienta para la identificación de rasgos de interés agronómico en especies cultivables (de Carvalho et al. 2019). En caña de azúcar, la aplicación de este tipo de tecnología se ha centrado en el descubrimiento de nuevos genes y regiones conservadas y su asociación con rutas metabólicas asociadas a azúcares y a las vías de biosíntesis de la pared celular o de sus componentes, como la celulosa, hemicelulosa y lignina (Calviño et al. 2008; Vermerris et al. 2011). Explorar todo el genoma a través de técnicas de secuenciación masiva puede ayudar a la comprensión de la organización y estructura del genoma (de Carvalho et al. 2019) y esta información puede ser útil en futuros programas de mejoramiento.

Como se mencionó en la introducción general, la complejidad del genoma de la caña de azúcar es una de las más importantes dentro de las especies cultivadas, debido al alto nivel de poliploidía, su tamaño de aproximadamente 10 Gb y las aneuploidías (Yang et al. 2017b). La aneuploidía afecta directamente al apareamiento de los cromosomas durante la meiosis y genera una variación adicional por ganancia o pérdida de material genético (Glyn et al. 2004; Vieira et al. 2018). Para estudiar las características de los genotipos de caña de azúcar, es conveniente reducir el alto nivel de complejidad de su genoma para obtener datos de polimorfismo de alta calidad (Balsalobre et al. 2017).

La secuenciación de ADN asociada a la digestión con dos enzimas de restricción (ddRADseq), es una de las tecnologías incluidas dentro de la familia que se denominó genéricamente genotipificación por secuenciación (GBS) (Peterson et al. 2012). Este protocolo implica el uso de dos enzimas de restricción diferentes y la selección de un tamaño de fragmento específico. Este método permite la identificación y genotipificación simultánea de miles de loci distribuidos por todo el genoma, proporcionando perspectiva genómica de los genotipos analizados. También proporciona una alta reproducibilidad, una gran profundidad de lecturas por loci y la selección de polimorfismos de un solo nucleótido (SNP) distribuidos por todo el genoma (Aguirre et al.

2019). Para los cultivos poliploides, la detección de SNPs verdaderos es más desafiante que para los cultivos diploides debido a la presencia de subgenomas dentro de un mismo individuo (resultado de la interacción inter o intraespecífica). Sin embargo, muchas herramientas desarrolladas para los diploides también pueden aplicarse a los alopoliploides ya que el comportamiento en la segregación es similar (Bourke et al. 2018). No obstante, el control de calidad y el filtrado son pasos importantes en el análisis bioinformático para detectar los SNPs falsos positivos.

Un gran interrogante al comienzo de un proyecto de llamado de variantes suele ser plantearse un proceso basado en el alineamiento a un genoma de referencia, o bien, realizar un análisis *de novo*. La principal diferencia entre estas opciones radica en que, en la primera estrategia, las lecturas provenientes del secuenciador se mapean a un genoma usado como "base" o referencia para poder ordenarlas y referenciarlas a una posición específica. Esto resulta beneficioso para el posterior análisis de los SNPs obtenidos. En la segunda estrategia, se utilizan las propias lecturas para generar un catálogo y, ese catálogo es utilizado a modo de referencia para encontrar las variantes entre las muestras (Rochette et al. 2017). En estos casos, las regiones identificadas no se pueden referenciar a una posición y cada una se analiza de manera independiente. En general, suelen generar más datos, pero muchas veces esto va de la mano de errores por falsos positivo, por lo que los posteriores pasos de filtrado por calidad se vuelven claves. El software Stacks puede realizar ambas estrategias (Catchen et al. 2011; Catchen et al. 2013; Rochette et al. 2017).

La rutina planteada por el programa Stacks puede resumirse de la siguiente manera: (1) Las lecturas crudas son por un lado demultiplexadas, es decir, ordenadas en carpetas/directorios para cada una de las muestras según un código de identificación que le fue asignado en la etapa de laboratorio (preparación de la biblioteca), y en la misma instrucción, filtradas de acuerdo con diferentes criterios como la calidad, las bases indeterminadas y la longitud de la lectura (módulo process_radtags); (2) las lecturas de cada individuo son agrupadas en loci putativos, y se identifican los sitios de nucleótidos polimórficos (módulo ustacks); (3) los loci putativos son agrupados a través de los individuos y de los catálogos de RAD-loci, mientras que se identifican los SNPs, con alelos de referencia y alternativo (módulo cstacks); (4) los loci putativos de cada individuo se cotejan con el catálogo (módulo sstacks); (5) se realiza el llamado de variantes de todos los loci (módulo gstacks); y (6) los resultados se pueden escribir en formato VCF para su posterior análisis (módulo populations) (Rochette et al. 2019).

En este tipo de protocolos, es muy común que se incurra en errores durante la preparación de la biblioteca genómica o en el análisis bioinformático, por lo que etapas de evaluación de la

calidad permiten reducir los artefactos técnicos antes de que los datos sean procesados y se infieran conclusiones biológicas erróneas (O'Leary et al. 2018). Hay ciertos parámetros que deben ser considerados durante la etapa de filtrado, sin embargo, para cada set de datos se definirán específicamente los pasos empleados y los umbrales aplicados. Es así como, en general, se trabaja con el 20% de datos perdidos, es decir, que exista un 20% de indeterminaciones para un cierto locus, un determinado valor de frecuencia de alelo minoritario (MAF) y rangos mínimos y máximos de profundidad de lecturas por locus, para asegurar que dicho SNP sea verdadero (Clevenger et al. 2015). Los valores de estos parámetros pueden variar según las características de la especie con la que se esté trabajando, el número de muestras del ensayo y si se trata de una población natural o biparental (O'Leary et al. 2018).

En el presente capítulo se analizan las técnicas empleadas para la aplicación del protocolo ddRADseq en caña de azúcar, con el fin de identificar variantes genómicas de tipo SNPs en variedades utilizadas en el programa de mejoramiento de caña de azúcar de INTA. Además, se evaluaron en un número reducido de muestras los resultados de lecturas obtenidas en dos plataformas de secuenciación de Illumina, para comprobar la eficiencia de la técnica en el descubrimiento de variantes genéticas a nivel de genoma completo. Concierne también a este capítulo, el establecimiento del flujo de trabajo bioinformático para el descubrimiento de variantes SNPs. La comparación de los resultados de las estrategias con y sin genoma de referencia en el llamado de variantes y la exploración de varios esquemas de filtrado para identificar los parámetros que equilibren la tasa de datos perdidos con la precisión del SNP detectado.

Materiales y métodos

Material vegetal

En este capítulo se seleccionaron cinco accesiones de caña de azúcar (*Saccharum spp*.) (los híbridos comerciales: NA 78-724, NA 56-30, LCP 85-384 y HOCP 92-665, y el biotipo de alta fibra INTA 05-3116), todos ellos provenientes de la colección del Banco de Germoplasma de la Estación Experimental Agropecuaria Famaillá, del Instituto Nacional de Tecnología Agropecuaria, Famaillá (27°03'S, 65°25'O, 363 msnm), Tucumán, Argentina. Se seleccionaron estas muestras debido a las siguientes características: 1) El biotipo de alta fibra INTA 05-3116 fue desarrollado para maximizar la producción de energía a partir de la biomasa total (García et al. 2022); 2) Los híbridos comerciales NA 78-724, NA 56-30 (de origen argentino), LCP 85-384 y HOCP 92-665 (de origen estadounidense, USDA), son utilizados activamente como progenitores

en el programa de mejoramiento de caña de azúcar por sus características industriales y agronómicas (Acevedo et al. 2017; García et al. 2022); 3) LCP 85-384 es el cultivar predominantemente plantado en Argentina (Acevedo et al. 2017; García et al. 2022); 4) El coeficiente de parentesco (COP) de NA 56-30 con LCP 85-384 es 0,133, y 0,156 con HOCP 92-665. El COP entre LCP 85-384 y HOCP 92-665 es de 0,225 (Acevedo et al. 2017); y 5) Los cinco genotipos seleccionados fueron evaluados como potenciales materias primas para la producción de etanol lignocelulósico (Kane et al. 2021; García et al. 2023a).

Además, se seleccionó una de las especies fundadoras de los genotipos comerciales (*S. spontaneum*) y tres genotipos de sorgo (cv Wiley, Rex y MN1054) que se encontraron dentro de la genealogía del cultivar comercial LCP 85-384 (Acevedo et al. 2017).

Bibliotecas ddRADseq

Se utilizaron hojas de tres réplicas por accesión de caña de azúcar para la extracción y la purificación de ADN de alta calidad, utilizando un kit de extracción de ADN genómico vegetal EasyPure (TransGen Biotech). La calidad del ADN se comprobó mediante Nanodrop (Thermo Fisher Scientific, Waltham, MA, EE.UU.) y electroforesis en gel de agarosa (1% de agarosa en Buffer TAE). El ADN se cuantificó por fluorometría utilizando Qubit 2.0 (Thermo Fisher Scientific). El protocolo ddRADseq 1 optimizado para *Eucalyptus dunnii* (Aguirre et al. 2019) se modificó para caña de azúcar, utilizando la combinación de enzimas de restricción PstI/MboI, en el paso de digestión de ADN, teniendo la primera una enzima con un sitio de reconocimiento de seis pares de bases mientras que la segunda reconoce 4 pares de bases, siendo ambas parcialmente sensibles a metilación. El tamaño de los fragmentos de ADN de la biblioteca final osciló entre 400 y 600 pb, representando 260-460 + 140 pb de barcodes, según la evaluación de un equipo automatizado para la evaluación de fragmentos (Analizador de fragmentos, Agilent Technologies, Santa Clara, USA). Todos los experimentos se realizaron en la Unidad de Genómica, IABIMO-Instituto de Biotecnología INTA, Argentina. La biblioteca se secuenció utilizando dos plataformas Illumina (Illumina Inc., San Diego, CA, EE.UU.): **(1)** secuenciador Illumina Nextseq500 a partir de ahora denominada Ne-70 (75 pb lecturas simples recortados a 70 pb; ~cuatro millones de lecturas/individuo, denominada de aquí en más cobertura de secuenciación media (MC)), y **(2)** secuenciador Illumina Novaseq6000, desde aquí denominada No-150 (lecturas de 150 pb pareadas; ~diez millones de lecturas/individuo, denominada de aquí en más cobertura de secuenciación alta (HC)).

Eficiencia de la longitud del fragmento de lectura, la profundidad de secuenciación y las lecturas simples frente a lecturas pareadas

Para evaluar la eficiencia de las plataformas de secuenciación en términos de longitud del fragmento de lectura, profundidad de secuenciación y lecturas simples frente a lecturas leídas de ambos extremos del fragmento o pareadas: (a) las lecturas obtenidas en No-150 se acortaron a 70 pb (manteniendo la alta cobertura de secuenciación y siendo pareadas, desde aquí denominado No-70-PE/HC) y se compararon con No-150; (b) Se seleccionaron aleatoriamente 4 millones de lecturas/individuo entre las lecturas No-70-PE/HC ya acortadas, para analizar el efecto del número de lecturas por individuo (No-70-PE/MC); (c) las No-70-PE/MC seleccionadas se compararon con las Ne-70 originales para analizar las lecturas pareadas frente a las simples; (d) finalmente, las lecturas obtenidas en No-150 se submuestrearon de HC a MC (No-150-PE/MC).

Construcción de paneles de SNPs de caña de azúcar

La calidad de la corrida en el secuenciador se comprobó visualmente utilizando FastQC (Andrews 2010) para cada muestra por separado y, se utilizó el programa MultiQC para condensar la información (Ewels et al. 2016). Para el llamado de variantes alélicas se utilizó el software Stacks. Este software es un sistema de código abierto compuesto por varios programas interconectados, inicialmente creado para el ensamblado de secuencias RAD *de novo* para la generación de mapas genéticos (Catchen et al. 2011); pero posteriormente fue ampliado por su versatilidad en estudios de organismos con o sin genoma de referencia. El software permite ejecutar un conjunto de programas, gracias a que fue diseñado en el lenguaje de programación Perl. Dado que posee distintos módulos, es posible aplicarlo a diversos escenarios. La limpieza por calidad se realizó mediante el módulo "process_radtags" del paquete Stacks v 2 (Catchen et al. 2011; Rochette et al. 2019). Se descartaron aquellas lecturas cuya calidad media dentro de una ventana cayera por debajo de una puntuación Phred de 28 (Ewing y Green 1998), o tuvieran nucleótidos no identificados, secuencias de adaptadores o sitios de corte de enzimas de restricción deficientes (Catchen et al. 2011). Se siguieron dos estrategias para el llamado de SNPs: (1) utilizando un genoma de referencia y (2) *de novo*.

Para la primera estrategia, las lecturas limpias se mapearon con Bowtie2 usando los parámetros por defecto (Langmead y Salzberg 2012) y el genoma de referencia del cultivar de caña de azúcar monoploide R570 (Garsmeur et al. 2018). La identificación de SNPs se realizó utilizando la rutina "ref_map.pl" del paquete Stacks v 2 (parámetros por defecto). En la segunda estrategia, se utilizó la rutina "denovo_map.pl" del Stacks para el descubrimiento de SNPs sin

referencia. Para ambas estrategias, se obtuvieron los correspondientes paneles de SNPs en formato Variant Call Format (VCF), utilizando el módulo "Population" implementado en Stacks. Se calcularon las estadísticas de los paneles con datos crudos para determinar los subsiguientes umbrales de filtrado.

Evaluación de genomas de referencia de *Saccharum sp*. disponibles en bases de datos (NCBI-Genomes) para el llamado de variantes de caña de azúcar

Los cuatro genomas de referencia para *Saccharum sp.,* según la base de datos del NCBI, más el genoma del *Sorghum bicolor* cv BTx623, fueron evaluados para el llamado de variantes dentro del protocolo ddRADseq desarrollado y analizado en este trabajo de tesis. Para ello, se trabajó con las lecturas provenientes de la secuenciación de dos cultivares comerciales (NA 78-724, LCP 85-384), una de las especies fundadoras de los genotipos comerciales (*S. spontaneum*) y tres genotipos de sorgo (cv Wiley, Rex y MN1054) que se encontraron dentro de la genealogía del cultivar comercial LCP 85-384 (Acevedo et al. 2017). A estas muestras se le aplicó el protocolo ddRADseq desarrollado para este trabajo de tesis. La biblioteca se secuenció a través de la plataforma Novaseq6000 de Illumina (Illumina Inc, San Diego, CA, EE.UU.). Las lecturas pareadas obtenidas fueron procesadas para que todas tuvieran el mismo largo (242 pb). La cantidad de lecturas por muestra se estableció en ~4 millones de lecturas/individuo. Para cada genoma, se mapearon las lecturas utilizando el Bowtie2. Estos resultados fueron usados para la identificación de SNPs utilizando la rutina "ref_map.pl" del paquete Stacks v 2.

Filtrado de paneles SNP

El filtrado de los marcadores se realizó utilizando VCFtools (Danecek et al. 2011). Se aplicó un filtro inicial y global de 0% de datos perdidos a todos los conjuntos de datos para garantizar SNPs de buena calidad y confiables. Se utilizó una frecuencia de alelo minoritario (MAF) > 0,1, comúnmente aplicada a los diploides, como umbral conservador, mientras que se utilizó un MAF mínimo de 0,3, como se recomienda para los alopoliploides (Clevenger et al. 2015). Se evaluaron tres rangos diferentes de profundidad de lectura por locus, con umbrales mínimos y máximos:

1- 10x-36x: siendo 10x la recomendación habitual para diploides y 36x como umbral máximo (Ravinet y Meier 2020).

2- 30x-60x: siendo 30x la recomendación habitual para alopoliploides (Clevenger y Ozias-Akins 2015) y 60x como umbral máximo.

3- 56x-200x: siendo 56x el umbral descripto por Yang et al (2017b) para la caña de azúcar y 200x como umbral máximo.

Seis esquemas de filtrado (FS) resultaron de la combinación de los diferentes parámetros de MAF y profundidad de lectura (Fig. 1.1).

Fig. 1.1 Esquemas de filtrado (FS) que combinan la cobertura de la frecuencia alélica menor (MAF) y la profundidad de lectura (RD).

Análisis de clústeres jerárquicos y análisis de componentes principales

Se utilizó el paquete vcfR para introducir los archivos con formato VCF en el entorno R para su posterior análisis (Knaus BJ y Grünwald 2017; R Core Team 2019). La densidad de SNP por cromosoma para Ne-70 con FS1 y No-150 con FS5 se determinó con el paquete CMplot (Yin et al. 2021). El análisis de correlación de la longitud del cromosoma y el número de SNPs por cromosoma se realizó utilizando las funciones "cor" y "cor.mtest" (paquete corrplot) (Wei y Simko 2021). Los coeficientes de correlación se consideraron significativos si el valor p era <0,05 (*). Las distancias genéticas (alelos compartidos) entre las muestras de Ne-70 con FS1 y No-150 con FS5 se calcularon con el paquete pegas (Paradis 2010) y se graficaron con el paquete Stats de R (R Core Team 2019). El método de aglomeración utilizado fue el "promedio" ya que tenía la mayor correlación cofenética (mejor modelo de agrupación para ambos conjuntos de datos) (Borcard et al. 2018). El análisis de componentes principales (PCA) sobre los mismos paneles se calculó con Tassel y se graficó con el paquete ggplot2 de R para analizar el conjunto completo de SNPs de cada panel, y para representar las relaciones entre las muestras (Bradbury et al. 2007, Wickham 2016).

Predicción del efecto funcional de los SNPs

Se utilizó el sistema de predicción de efectos funcionales de las variantes (VEP) de Ensembl Plants (versión 51) para predecir el efecto funcional del SNP de Ne-70 con FS1 y No-150 con FS5, siguiendo las instrucciones descriptas en Ensembl/plant-scripts: Scripting analyses of genomes

in Ensembl Plants (Yates et al. 2022), con el genoma de referencia y su anotación, en formatos FASTA y gff3.

Evaluación de la robustez del protocolo

Para evaluar la robustez del protocolo, se creó una nueva biblioteca de las muestras NA 78-724 y LCP 85-384 a partir de una nueva extracción de ADN utilizando el protocolo descripto anteriormente y se secuenció utilizando la plataforma Illumina Novaseq6000 (150 bp pareadas; ~cuatro millones de lecturas, denominada cobertura de secuenciación media (MC)). Además, se realizó un submuestreo aleatorio de las lecturas No-150 para obtener un conjunto de datos de cobertura de secuenciación media para las muestras NA 78-724 y LCP 85-384. De este modo, se crearon dos conjuntos de datos en experimentos independientes, que son réplicas para la evaluación de la robustez del protocolo.

El protocolo se evaluó comparando dos réplicas de los híbridos comerciales NA 78-724 y LCP 85-384. El llamado de variantes se realizó como se ha descripto anteriormente. El archivo VCF se filtró evitando los datos faltantes y se probaron tres rangos de profundidad de lectura por locus diferentes (a. RD= 10x-36x; b. RD= 30x-60x; y c. RD= 56x-200x). Utilizando los paneles de SNPs, se calcularon las distancias alélicas entre las muestras, junto con las matrices cofenéticas asociadas, como se ha descripto anteriormente.

Resultados

Construcción y secuenciación de bibliotecas ddRADseq para caña de azúcar

El kit EasyPure (TransGen Biotech) utilizado para realizar la extracción de ADN del material vegetal fue efectivo para los genotipos de caña de azúcar evaluados. Se observaron los índices de absorbancia (A260/A280 y A260/A230) y las cantidades obtenidas de ADN para cada muestra (Apéndice 1 Tabla 1.1), determinantes de la pureza, y una banda intensa de alto peso molecular (ADN genómico) en el gel para cada muestra (Apéndice 1 Figura 1.1.).

El protocolo ddRADseq para pocas muestras implementado en este capítulo permitió la generación de una biblioteca de muestras agrupadas con buena calidad y cantidad. Las enzimas de restricción elegidas, PstI/MboI, fueron una combinación exitosa para la digestión del ADN de caña de azúcar (Fig. 1.2 a). La selección del tamaño de los fragmentos de ADN de interés de forma manual fue eficiente (Fig. 1.2 b).

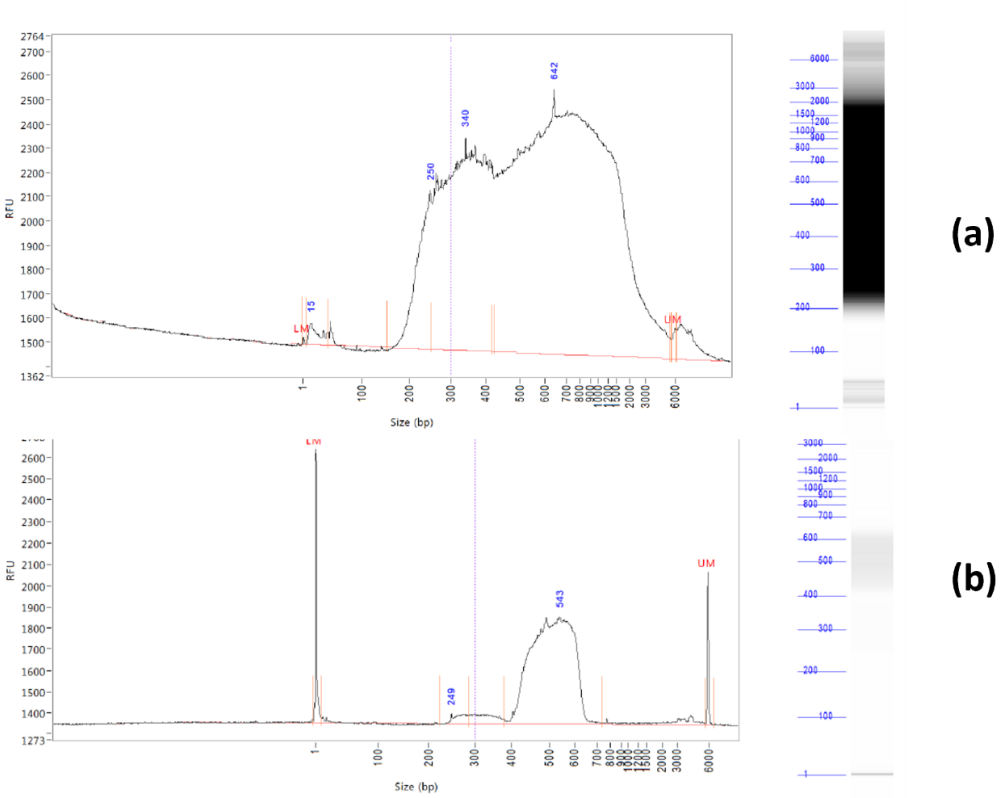

Figure 1.2. *Visualización de la corrida en el Analizador de fragmentos*. (a) Digestiones *in vitro* de ADN genómico de híbridos de caña de azúcar con *PstI-MboI* y (b) de la biblioteca genómica (400-600 pb).

Se evaluaron los resultados de dos plataformas de secuenciación a través la comparación del rendimiento obtenido utilizando el protocolo ddRADseq. Se consideraron el contenido de información obtenido, la mejor relación costo-beneficio y diferentes esquemas de filtrado. Además, la biblioteca cumplió con las expectativas de rendimiento esperada para las plataformas Illumina, lo cual se comprobó visualmente a través de los programas FastQC y MultiQC (Fig. 1.3). Las secuencias fueron de buena calidad, con un valor promedio por base de 35 según la codificación de Pherd para ambas plataformas (Fig. 1.3 b, d), que va disminuyendo hacia el final del fragmento (Fig. 1.3 a, c), lo cual es esperable por la tecnología de secuenciación. Se puede observar en la Figura 1.3 c que las lecturas leídas desde el comienzo (línea verde) y desde el final del fragmento (línea naranja) tuvieron calidades similares.

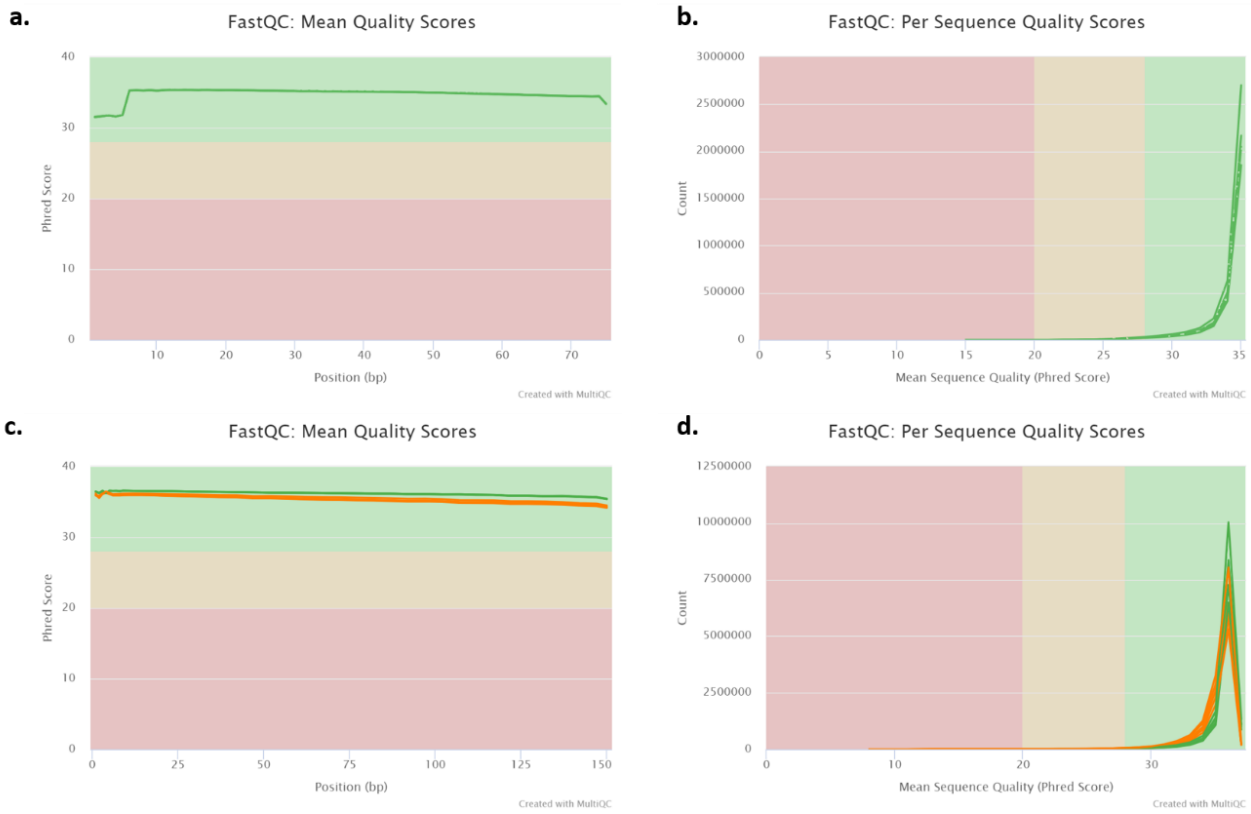

Figura 1.3. *Control de calidad visualizado con FastQC y MultiQC.* Valor de calidad promedio por base de a. Ne-70 y c. No-150; Número de lecturas con valor de calidad promedio de b. Ne-70 y d. No-150. Línea verde: lecturas leídas desde el comienzo del fragmento; Línea naranja: lecturas leídas desde el final del fragmento.

En la Tabla 1 se muestran el número de lecturas y su alineación con el genoma de referencia para cada muestra, para las corridas Ne-70 y No-150. El número de lecturas de No-150 fue en promedio más de siete veces el número detectado en Ne-70 (Tabla 1.1, Apéndice 1 Tabla 1.2), superando la relación esperada de 2,5:1. Dado que se utilizó la misma biblioteca en ambos ensayos, las proporciones diferenciales podrían deberse a la naturaleza de la muestra, el proceso de carga del secuenciador y la distribución de la solución en la celda de flujo. Las muestras NA 78-724 y NA 56-30 representaron el mayor y el menor número de lecturas, respectivamente, en ambos conjuntos de datos (Tabla 1.1). La tasa de alineamiento global de las secuencias con el genoma monoploide de la caña de azúcar fue mayor en Ne-70 que en No-150. Estas tasas, que fueron similares en todas las muestras de la primera plataforma, promediaron un 54%, y contrastaron con las tasas de la segunda plataforma (~42%), con la excepción de la muestra NA 78-724, que reportó un 50% (Tabla 1.1).

Tabla 1.1. *Número de lecturas obtenidas con dos plataformas de secuenciación*. Tasa de alineamiento total respecto al genoma de caña de azúcar monoploide del cultivar R570 (Garsmeur et al. 2018), número de lecturas alineadas en un solo sitio, y número de lecturas alineadas en más de un sitio. Ne-70 = Lecturas simples de 75 pb recortada a 70 pb, MC; No-150 = Lectura pareadas de 150 pb, HC.

Plataforma de secuenciación	*Accesión*	*N° lecturas*	*Lecturas alineadas en un solo sitio*	*Lecturas alineadas en más de un sitio*	*Tasa de alineamiento total / N° lecturas*	*Lecturas alineadas en un solo sitio / N° lecturas*
Ne-70	INTA 05-3116	3.055.298	1.182.532	430.432	0,528	0,387
	NA 78-724	3.759.023	1.499.488	569.465	0,55	0,399
	LCP 85-384	3.307.885	1.315.870	462.541	0,538	0,398
	HOCP 92-665	2.826.753	1.098.629	435.747	0,543	0,389
	NA 56-30	2.535.384	1.006.667	368.096	0,542	0,397
No-150	INTA 05-3116	10.504.705	4.162.735	2.208.809	0,415	0,396
	NA 78-724	13.056.991	5.696.204	2.484.897	0,504	0,436
	LCP 85-384	10.810.451	4.375.030	2.314.485	0,427	0,405
	HOCP 92-665	9.302.776	3.714.998	2.100.697	0,433	0,399
	NA 56-30	8.269.031	3.356.088	1.794.476	0,435	0,406

Evaluación de genomas de referencia de *Saccharum sp*. disponibles en bases de datos (NCBI-Genomes)

Cinco genomas de referencia para *Saccharum sp.,* según la base de datos del NCBI, fueron evaluadas para ser utilizados durante este trabajo de tesis. Las muestras utilizadas en este ensayo fueron dos cultivares comerciales (NA 78-724, LCP 85-384), una de las especies fundadoras de los genotipos comerciales (*S. spontaneum*) y tres genotipos de sorgo (cv Wiley, Rex y MN1054). Cómo se puede observar en la Figura 1.4, al utilizar el genoma de referencia de caña de azúcar del cultivar híbrido R570, el mapeo general disminuye respecto a los resultados obtenidos previamente, dado que aquí se trata de lecturas más largas, de 242 pares de bases. Para los cultivares comerciales se obtuvo un alineamiento del ~30%, para el *s. spontaneum* un porcentaje similar (~27%), mientras que para los sorgos fue menor (~22%). Al evaluar el alineamiento contra Sorgo bicolor BTx623 como referencia, se obtienen porcentajes muy por debajo para las cañas de azúcar comerciales y la especie salvaje (~13%), mientras que para los sorgos el porcentaje aumenta considerablemente (~95%), lo que era de esperar ya que hablamos de la misma especie. Respecto a las otras referencias evaluadas, con el híbrido de caña de azúcar CC 01-1940 y *S. spontaneum* Np-X, ambas arrojaron tasas de alineamiento similares, siendo ~46-49% para las cañas y ~35% para los sorgos. Por último, el ensamblado de otra de las especies fundadoras de los cultivares comerciales, *S. officinarum,* dio como resultado un menor

porcentaje de alineamiento (~30% para las cañas de azúcar, ~27% para el genotipo de *S. spontaneum* evaluado y ~20% para los sorgos).

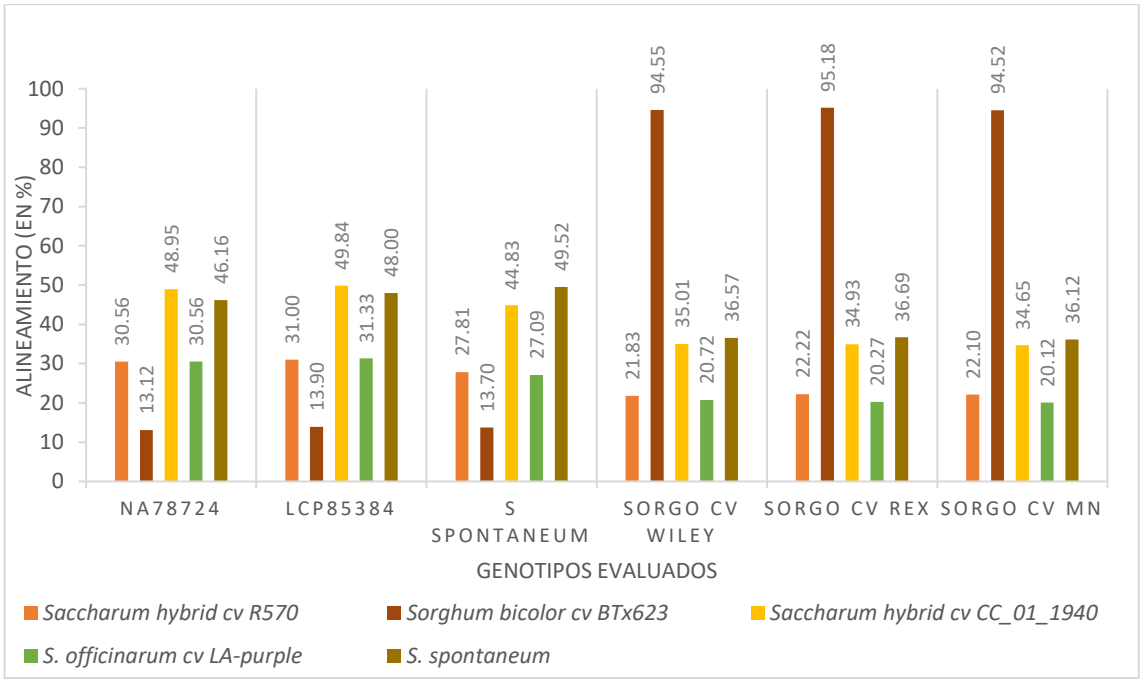

Figura 1.4. *Porcentaje de alineamiento entre las muestras evaluadas y los genomas de referencia.* Cultivares comerciales de caña de azúcar: NA 78-724, LCP 85-384; especies fundadora: *S. spontaneum*; cultivares de sorgo: Wiley, Rex y MN1054.

Paneles de marcadores SNPs para caña de azúcar

Se obtuvieron paneles de SNPs para los cinco genotipos de caña de azúcar secuenciados en cada plataforma Illumina (Ne-70 y No-150), a partir del llamado de variantes polimóficas guiada por referencia y *de novo*. Para el primer set de datos, se recuperaron 188.199 SNPs usando la estrategia *de novo* y 86.051 SNPs utilizando el alineamiento con una referencia, mientras que en el segundo se identificaron 2.831.922 SNPs para el análisis *de novo* y 460.890 SNPs en el usando el genoma monoploide de la caña de azúcar (Tabla 1.2). La estrategia *de novo* permitió identificar una cantidad de SNPs 2,1 y 6,1 veces mayor que la basada en genoma de referencia en Ne-70 y No-150, respectivamente. Además, se recuperaron 15 veces más SNPs con el análisis *de novo* en el No-150 en comparación al Ne-70, mientras que usando la referencia se recuperaron 5,4 veces más SNPs.

Tabla 1.2. *Número de SNPs por panel obtenido*. Los paneles de SNPs crudos se obtuvieron de los cinco genotipos de caña de azúcar analizados en cada plataforma (Ne-70 y No-150) y de dos estrategias de llamado de SNPs (guiada por referencia y *de novo*). Las simulaciones aplicadas a la corrida No-150, es decir, el submuestreo de HC a MC (No-150-PE/MC), más las lecturas acotadas a 70bp (No-70-PE/HC) y el posterior submuestreo de HC a MC (No-70-PE/MC) también se analizaron con referencia guiada y *de novo*. Por último, se aplicaron los seis FS a cada uno de los paneles de SNPs crudos.

FS= esquemas de filtrado; MAF= frecuencia de alelo minoritario; RD= profundidad de lectura por locus; MC: cobertura de secuenciación media; HC: cobertura de secuenciación alta; SR: lecturas simples; PE: lecturas pareadas; Ne: Plataforma de secuenciación Nextseq500; No: Plataforma de secuenciación Novaseq6000; 70: Lectura de 70 pb de longitud; 150: Lectura de 150 pb de longitud; Ne-70 = Lecturas simples de 75 pb recortada a 70 pb, MC; No-150 = Lectura pareadas de 150 pb, HC; No-70-PE/MC= Lectura de 70 pb pareadas, MC; No-70-PE/HC= Lectura de pb pareadas, HC, No-150-PE/MC= Lectura de 150 pb pareadas, MC. El tipo de letra normal corresponde a los resultados experimentales, mientras que la fuente en cursiva representa los resultados simulados. Los números en negrita representan los SNPs que se tuvieron en cuenta para continuar el análisis.

Plataforma de secuenciación y simulaciones	Estrategia	Datos crudos	FS 1	FS 2	FS 3	FS 4	FS 5	FS 6
			MAF>0,1	MAF>0,3	MAF>0,1	MAF>0,3	MAF>0,1	MAF>0,3
			RD= 10x-36x		RD= 30x-60x		RD= 56x-200x	
Ne-70	referencia	86.051	**9.094**	6.237	4.246	3.111	2.122	1.619
	de novo	188.199	14.36	9.206	3.222	2.193	723	555
No-70-PE/MC	referencia	174.585	16.464	10.31	9.954	6.921	6.719	4.964
	de novo	363.919	7.205	4.011	3.282	2.012	1.195	852
No-70-PE/HC	referencia	351.864	15115	8770	17689	11570	24799	17280
	de novo	767.203	28734	15947	17290	10461	11433	7469
No-150-PE/MC	referencia	127.556	13.227	9.571	3.595	3.047	1.883	1.712
	de novo	218.241	23.054	15.807	6.204	4.785	1.713	1.409
No-150	referencia	460.89	23.174	13.403	28.081	18.29	**47.964**	33.4
	de novo	2.831.922	30.167	16.925	15.568	9.102	5.43	3.392

Después de recortar las secuencias de No-150 a 70 pb (No-70-PE/HC) y de reducir el número de lecturas de HC a MC (No-70-PE/MC), también se realizó el llamado de polimorfismos utilizando la estrategia con referencia y *de novo*. De esta forma, se obtuvieron otros cuatro paneles de SNPs. No-70-PE/HC se comparó con los paneles obtenidos a partir de las lecturas de No-150 para evaluar el efecto de la longitud de las lecturas en el número de SNPs identificados. El número de SNPs identificados en No-150 con la estrategia *de novo* fue 3,7 y con referencia 1,3 veces más que el número de SNPs identificados en No-70-PE/HC, respectivamente. En el análisis sin referencia, se detectó el menor número de SNPs en las primeras 70 pb (767.203 SNPs en No-70-PE/HC de 2.831.922 SNPs en No-150), mientras que usando la referencia la mayoría de los SNPs se localizaron en las 70 primeras posiciones de los fragmentos (351.864 SNPs en No-70-PE/HC de 460.890 SNPs en No-150) (Tabla 1.2).

Se evaluó también el efecto del número de lecturas por individuo en el descubrimiento de SNPs. Para ello, se comparó el panel No-70-PE/HC habiendo seleccionado de forma aleatoria cuatro millones de lecturas por individuo, considerado como cobertura de secuenciación media

(No-70-PE/MC). Los SNPs identificados en No-70-PE/HC (351.864 SNPs utilizando la búsqueda con referencia y 767.203 SNPs usando el análisis *de novo*) duplicaron la cantidad de los identificados en No-70-PE/MC (174.585 SNPs con referencia y 363.919 SNPs sin referencia), y esto fue así para ambas estrategias de llamado de variantes (Tabla 1.2). Esto sugiere además que el número de SNPs no es directamente proporcional al número de lecturas por individuo. Como se predijo, al comparar No-150-PE/MC con No-150, fue con esta última que se recuperaron más SNPs.

Por último, para estudiar el efecto de las lecturas pareadas frente a las simples en el número de SNPs identificados, se comparó el Ne-70 original con el panel No-70-PE/MC. Como se esperaba, el número de SNPs identificados en No-70-PE/MC fue casi el doble del número de SNPs identificados en Ne-70, para ambas estrategias de llamado de variantes (Tabla 1.2).

Paneles de SNPs filtrados

Los paneles de SNPs generados a partir de los conjuntos de datos de Ne-70 y No-150, sumando las tres simulaciones, se filtraron en función de la calidad, los datos faltantes, rangos de profundidad de las lecturas y el MAF, como se muestra en la Fig. 1.1. Se generaron sesenta paneles de SNPs a partir de seis esquemas de filtrado aplicados (FS). Las matrices Ne-70, No-70-PE/MC y No-150-PE/MC mostraron una tendencia decreciente de SNPs desde FS1 hasta FS6 con ambas estrategias (con y sin referencia), lo que indica que los SNPs retenidos disminuyeron con la profundidad de lectura a medida que menos SNPs cumplían las condiciones restrictivas impuestas por las profundidades de lectura 56x-200x (Tabla 1.2). Con No-150 la tendencia de SNPs filtrados fue similar al de Ne-70 en la estrategia sin referencia, pero No-150 se recuperó un mayor número de SNPs en el análisis con referencia con una profundidad de lectura de 56x-200x (FS5 y FS6) (Tabla 1.2). Esta tendencia también se observó en la simulación No-70-PE/HC, pero en este panel el número de SNPs fue menor debido al menor tamaño de los fragmentos. A esta profundidad, los alelos que presentan dosaje único podrían recuperarse en un locus heterocigota, mientras que a una profundidad menor no es posible distinguirlos o se pueden confundir con errores de secuenciación. Los SNPs retenidos dentro de cada rango de profundidad de lectura evaluado fueron mayores con MAF >0,1 para ambas plataformas. Dado que la aplicación de FS1 en Ne-70 y FS5 en No-150 retuvo el mayor número de SNPs, se seleccionaron ambos esquemas de filtrado para las comparaciones posteriores.

Para visualizar la distribución de SNPs a lo largo de cada cromosoma se representó gráficamente la densidad de SNPs (Fig. 1.5). Los datos crudos de Ne-70 mostraron regiones con más de 289 SNPs repartidos en todos cromosomas, y las regiones distales mostraron una densidad de SNPs relativamente alta, es decir, más de 181 SNPs (Fig. 1.5 a). La aplicación del FS1

sobre la plataforma Ne-70 disminuyó drásticamente la densidad de SNPs en todos los cromosomas (Fig. 1.5 b), siendo esta reducción consistente con el número de SNPs retenidos en el FS1 en comparación con los datos crudos (Tabla 1.2). Los datos crudos de No-150 mostraron regiones con más de 1621 SNPs repartidos en todos cromosomas (Fig. 1.5 c). La aplicación del FS5 sobre la plataforma No-150 reveló una menor distribución de SNPs en las regiones pericentroméricas en comparación con las distales (Fig. 1. 5 d).

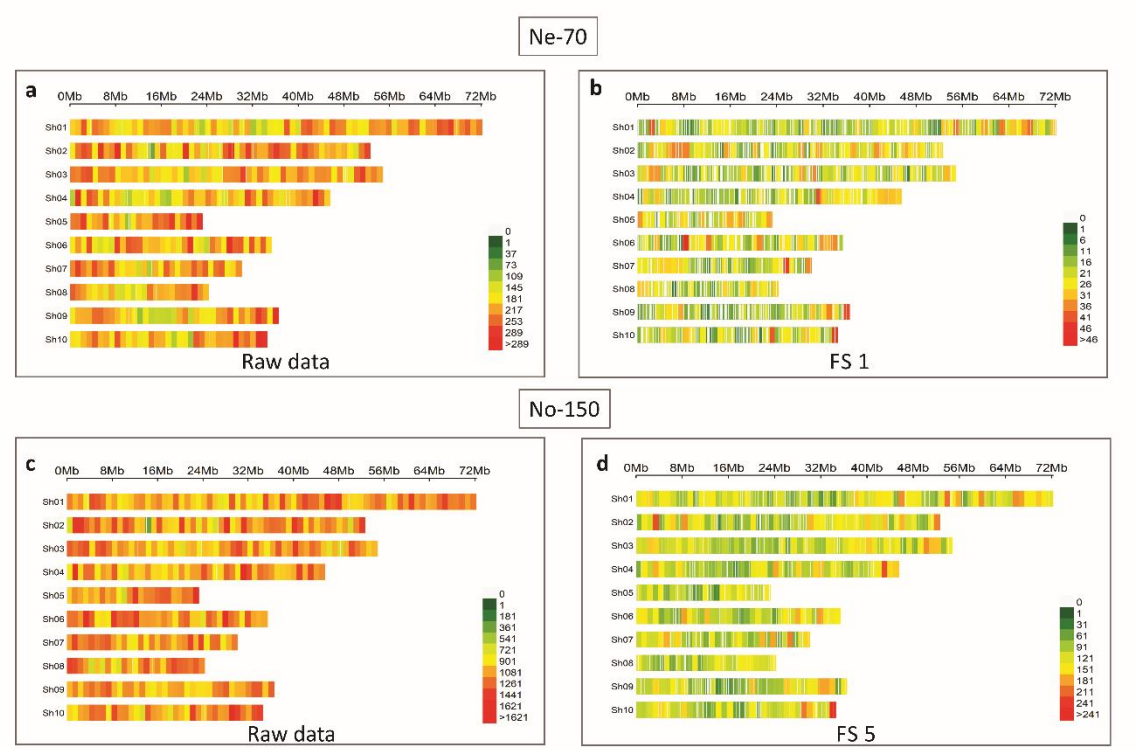

Figura 1.5. *Densidad de SNPs por cromosoma.* Distribución de SNPs en 10 cromosomas de la caña de azúcar cv R570. a. Datos crudos de Ne-70; b. Ne-70 con FS1; c. Datos crudos de No-150; d. No-150 con FS5. La escala de colores de la leyenda indica el número de marcadores dentro de una ventana de un Mbp

La longitud del cromosoma y el número de SNPs por cromosoma se correlacionaron positivamente en ambos paneles de SNPs, Ne-70 con FS1 (rho= 0,94; valor p < 2,2e-16) y No-150 con FS5 (rho= 0,96; valor p < 2,2e-16), (Fig. 1.6). De esta forma, el cromosoma 1, que es el que presenta mayor número de bases, representó 1545 SNPs en el primer panel (Tabla 1.3) mientras que 8850 en el segundo (Tabla 1.3). Por el contrario, los cromosomas 5 y 8, los más cortos en pares de bases, mostraron 573 y 526 SNPs respectivamente en Ne-70 con FS1, mientras que 2534 y 2702 SNPs respectivamente en No-150 con FS5 (Tabla 1.3).

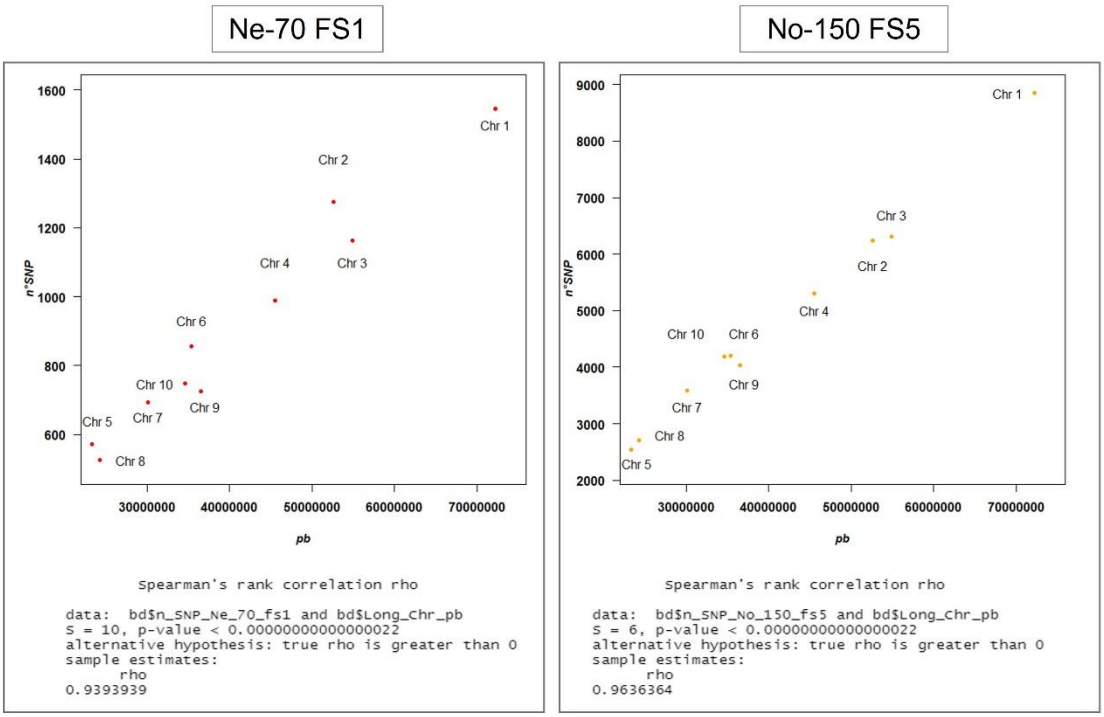

Figura 1.6. Análisis de correlación de Spearman entre el número de SNPs (n° NPs) detectados por cromosoma y la longitud del cromosoma (pb) para Ne-70 con FS1 y No-150 con FS5.

Tabla 1.3. *Relación entre el largo de cada cromosoma y el número de SNPs*. Largo de cada cromosoma de caña de azúcar y número de SNPs por cromosoma en Ne-70 con FS 1 y No-150 con FS 5. pb= pares de bases.

Cromosoma	Largo del cromosoma (pb)	n° SNPs Ne-70 FS1	n° SNPs No-150 FS5
1	72.286.729	1.545	8.850
2	52.638.177	1.275	6.246
3	54.887.512	1.163	6.312
4	45.576.573	989	5.305
5	23.282.433	573	2.534
6	35.411.752	855	4.196
7	30.096.520	694	3.592
8	24.320.556	526	2.702
9	36.579.524	726	4.033
10	34.651.920	748	4.194

Análisis jerárquico de clústeres y análisis de componentes principales

Los paneles filtrados, Ne-70 FS1 y No-150 FS5, con el análisis basado en el genoma de referencia, fueron analizados de acuerdo con dos métodos multivariados diferentes, uno basado en el agrupamiento jerárquico de clústeres y el otro en el ordenamiento reducido en el espacio, a través de un análisis de componentes principales (PCA). El análisis de agrupamiento de los paneles de SNPs reveló dos grandes grupos dentro de los dendrogramas (Fig. 1.7 a, b). En ambos dendrogramas, el híbrido INTA 05-3116 (biotipo de alta fibra) formó un solo clúster (Fig. 1.7 a, b). El segundo clúster, lo conformaron los cuatro híbridos comerciales: NA 78-724, LCP 85-384, HOCP 92-665 y NA 56-30; sin embargo, el agrupamiento entre los genotipos comerciales en cada dendrograma fue diferente (Fig. 1.7 a, b).

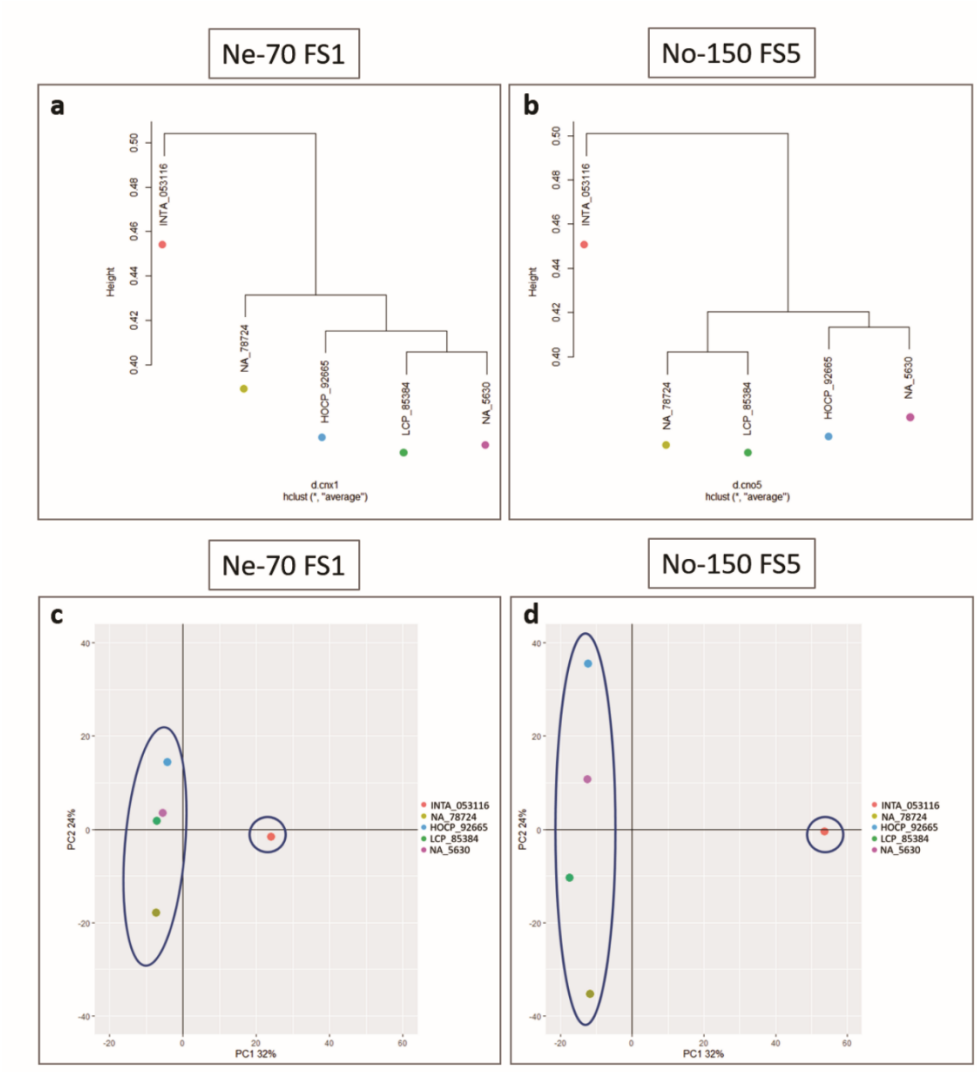

Figura 1.7. *Relación entre cultivares de caña de azúcar basada en marcadores SNPs*. Distancia promedio de agrupamiento en (a) Ne-70 con FS1 y (b) No-150 con FS5. Análisis de componentes principales en (c) Ne-70 con FS1 y (d) No-150 con FS5. PC, componente principal. Agrupados por líneas sólidas azules se identifican los dos grandes grupos que se forman a partir de la ubicación de las muestras en el espacio.

El PCA permitió representar gráficamente las relaciones entre los cinco genotipos de caña de azúcar (Fig. 1.7 c, d). El gráfico de PCA, explicó el 56% de la varianza total, en ambos paneles de SNPs, distribuidos en un primer (PC1) y el segundo (PC2) componentes principales. El PC1 contribuyó al 32% de la varianza total y distinguió dos grupos de genotipos (agrupados por líneas sólidas en la Fig. 1.7 c, d). El grupo 1 estaba compuesto por el INTA 05-3116, mientras que el grupo 2 lo conforman los cuatro genotipos comerciales. El PC2 representó el 24% de la varianza total, separando LCP 85-384, NA 56-30 y HOCP 92-665 (Fig. 1.7 c, cuadrante superior izquierdo) de NA 78-724 (Fig. 1.7 c, cuadrante inferior izquierdo) para Ne-70 FS1, mientras que NA 56-30 y HOCP 92-665 (Fig. 1.7 d, cuadrante superior izquierdo) de NA 78-724 y LCP 85-384 (Fig. 1.7 d, cuadrante inferior izquierdo) para No-150 FS5. En conjunto, el análisis de agrupamiento jerárquico y el PCA revelaron patrones similares de agrupación y ordenación en el espacio reducido de genotipos analizados.

Predicción del efecto funcional de los SNPs

Para predecir las posibles implicancias funcionales de los SNPs en los cinco genotipos de caña de azúcar estudiados en este capítulo se utilizó el sistema de predicción de efectos de las variantes (VEP) del Ensembl Plants (Fig. 1.8). Del total de variantes procesadas en Ne-70 FS1, las variantes genéticas ubicadas río arriba y río abajo de los genes alcanzaron el 60,6% (8.415 SNPs), representando más de la mitad de las variantes detectadas (Fig. 1.8 a). Las variantes intergénicas e intrónicas alcanzaron 2.070 (14,9%) y 1.314 SNPs (9,5%), respectivamente (Fig. 1.8 a, Apéndice 1 Tabla 1.3). Casi el 13% de las variantes (1.776 SNPs) correspondieron a regiones codificantes (Fig. 1.8 a, Apéndice 1 Tabla 1.3). De este último porcentaje de SNPs, 894 causaron cambios en los aminoácidos (variante con cambio de sentido, 50,3%), 868 no causaron ningún cambio resultante en el aminoácido codificado (variantes sinónimas, 48,9%), 5 evitaron el inicio de la transcripción (codón de inicio perdido, 0,3%), 5 provocaron una ganancia (0,3%) y 4 una pérdida de codones de finalización (0,2%) (Fig. 1.8 b, Apéndice 1 Tabla 1.3).

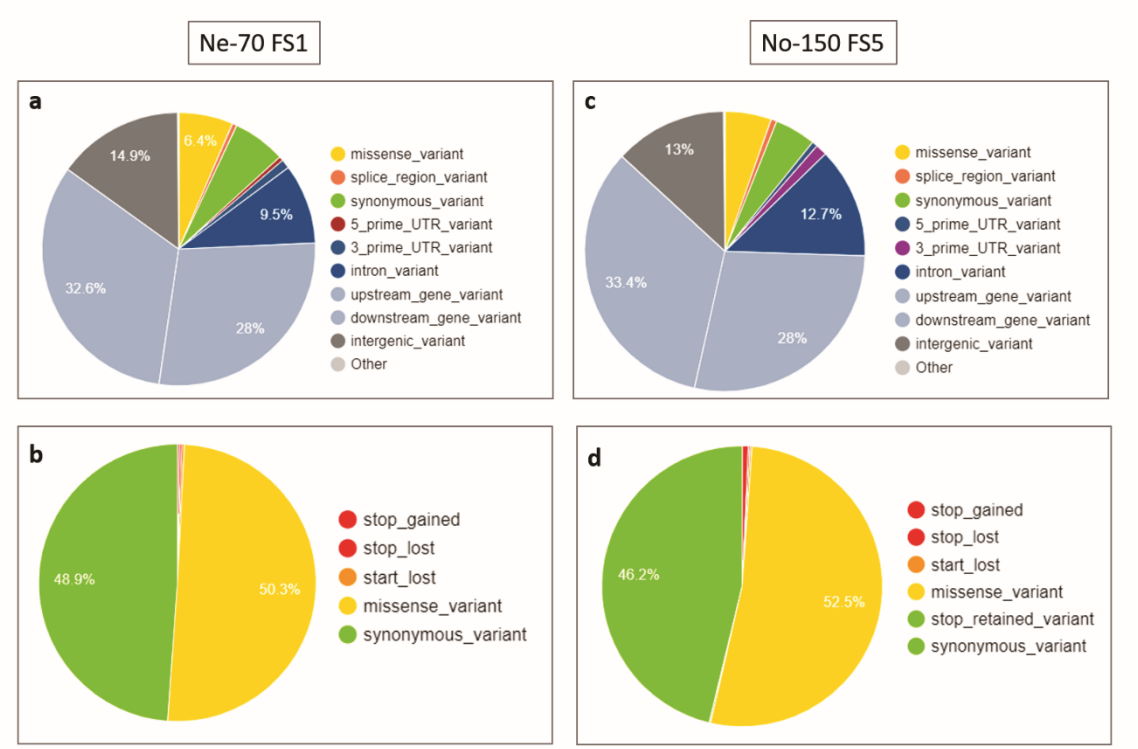

Figura 1.8. *Análisis funcional in silico de SNPs.* Porcentajes del efecto biológico de las variantes SNPs para todas las consecuencias en a. Ne-70 FS1; o c. No-150 FS5, y las consecuencias en las regiones codificantes según la localización del sitio del genoma en b. Ne-70 con FS1 o d. No-150 con FS5.

Las variantes génicas ubicadas río arriba y río abajo de los genes en No-150 FS5 lograron un porcentaje similar al de Ne-70 FS1 (46.522 SNPs, que representan el 61,4%) (Fig. 1.8 c). Las variantes intergénicas e intrónicas lograron 9.834 (13%) y 9.662 SNPs (12,7%), respectivamente (Fig. 1.8 c, Apéndice 1 Tabla 1.3). Alrededor del 10% de los SNPs se localizaron en regiones codificantes, lo que representa 7.756 variantes. De esta cantidad 1) 4.071 SNPs fueron variantes con cambio de sentido (52,5%), con posibles impactos en la funcionalidad de la proteína; 2) 3.586 SNPs fueron variantes sinónimas (46,2%); 3) 19 SNPs causaron la pérdida de codones de inicio; y 4) 80 SNPs modificaron los codones de finalización (55 se ganaron, 14 se perdieron, y en 11 codones se cambió al menos una base, pero se mantuvo el codón de finalización) (Fig. 1.8 d, Apéndice 1 Tabla 1.3).

Robustez del protocolo

En total, las dos muestras de cada hibrido comercial y sus réplicas (NA 78-724 y LCP 85-384), produjeron 77.918 SNPs. Tras filtrar los datos por datos perdidos, se retuvieron 22.744 SNPs. Los tres umbrales de profundidad de lectura probados dieron lugar a tres paneles de SNPs: a. RD= 10x-36x, donde se encontraron 14.024 SNPs; b. RD= 30x-60x, donde se encontraron 5.137

SNPs; y c. RD= 56x-200x, donde se encontraron 2.465 SNPs. Estos paneles de SNPs se utilizaron para calcular las distancias alélicas entre las muestras (Fig. 1.9). En todos los casos, las réplicas se agruparon juntas, pero a diferentes valores de distancia. La mayor distancia entre réplicas se observó en el rango de menor profundidad de lectura por locus, con valores de 0,211 para NA 78-724 y 0,161 para LCP 85-384 (Fig. 1.9 a). En el panel con profundidad de lectura medida por locus media, el valor de la distancia fue de 0,153 para NA 78-724 y de 0,112 para LCP 85-384 (Fig. 1.9 b). En el panel con profundidades de lecturas más altas, se obtuvieron valores de distancia más bajos, de 0,066 para NA 78-724 y 0,071 para LCP 85-384 (Fig. 1.9 c).

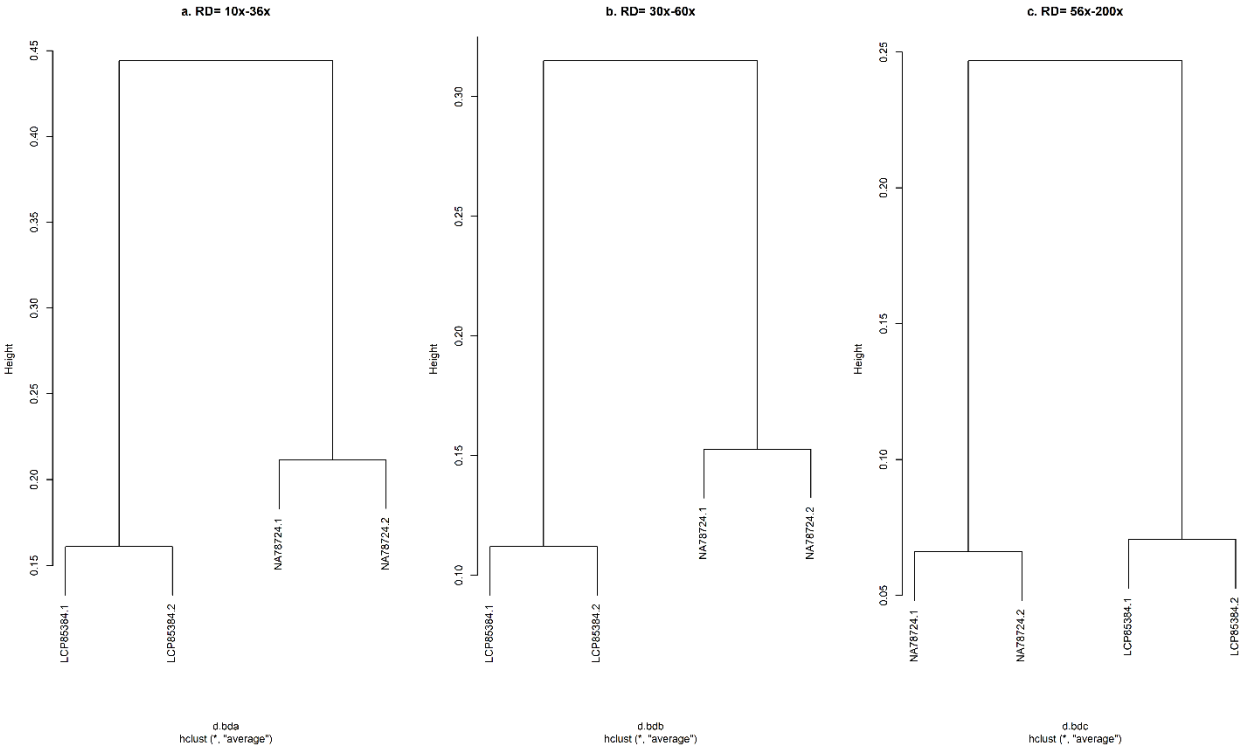

Figura 1.9. *Distancia promedio de agrupamiento en dos réplicas*. NA 78-724.1 y LCP 85-384.1 (cobertura media/individuo seleccionadas aleatoriamente del conjunto de datos No-150), y sus réplicas NA 78-724.2 y LCP 85-384.2 (lecturas de 150 pb pareadas; MC). Se probaron tres umbrales de profundidad de lectura (RD) diferentes a. RD= 10x-36x, b. RD= 30x-60x, y c. RD= 56x-200x.

Discusión

El análisis genómico de la caña de azúcar resulta complejo debido a la ganancia y la pérdida de material genético estrechamente asociadas a la naturaleza altamente poliploide y aneuploide del cultivo. La aneuploidía y la ploidía, desconocida de algunos híbridos, dificultan el genotipado de la caña de azúcar debido al múltiple número de copias alélicas (Serang et al. 2012). Además, en los poliploides la tasa de falsos positivos en la detección de SNPs aumenta debido a que el mapeo puede resultar ambiguo para loci homeólogos altamente similares, lo que se convierte en un reto para la detección precisa de SNPs (Clevenger et al. 2015). Las técnicas de GBS han

demostrado ser una forma exitosa de analizar la caña de azúcar (Balsalobre et al. 2017; Yang et al. 2017b), sin embargo, hasta el momento, el enfoque ddRADseq no se ha aplicado a este cultivo. Las simulaciones y comparaciones basadas en datos reales mostraron que, independientemente de la complejidad del genotipado de los híbridos de caña de azúcar, hay consistencia en los resultados obtenidos tras la aplicación del protocolo ddRADseq.

Las plataformas NGS actuales generan datos para analizar genes y genomas, a través de la secuenciación por síntesis (López de Heredia 2016). Las diferencias entre plataformas se basan en el rango de lecturas obtenidas por corrida, la longitud máxima de lectura de los fragmentos y los costos por corrida (Stoler and Nekrutenko 2021). Por esta razón, la comparación entre plataformas de secuenciación es crucial para decidir cómo invertir los recursos disponibles. Los resultados de este capítulo demuestran que el número de lecturas asignadas a cada genotipo siguió el mismo orden en las plataformas Ne-70 y No-150, no obstante, el número final de lecturas en esta última fue superior al detectado en la primera (Tabla 1.1). Cabe destacar que NA 78-724 y NA 56-30, dos híbridos argentinos provenientes del mismo programa de mejoramiento de caña de azúcar presentaron valores extremos para el número de lecturas en ambas plataformas de secuenciación, lo que sugiere una consistencia en el desempeño de la metodología entre ambas muestras.

La puesta a punto del protocolo ddRADseq en caña de azúcar se realizó utilizando como genoma de referencia al híbrido R570, el único genoma ensamblado a nivel de cromosoma con que se contaba al iniciar la tesis. De esta forma, la tasa de alineamiento general obtenida para cada grupo de lecturas de ADN de ambas plataformas evaluadas en la primera instancia fue de un promedio de ~48%. Varios fenómenos no excluyentes podrían ser responsables de este valor. En primer lugar, el ensamblado del genoma de referencia híbrido R570 se originó a partir de la sintenia con el sorgo, utilizando BACs basados en los alineamientos globales entre ambas, representando la parte rica en genes (Garsmeur et al. 2018). Esto indica que es poco probable que las regiones menos conservadas se alineen con otros híbridos de caña de azúcar. En segundo lugar, R570 es un cultivar de la Isla de la Reunión, ubicada en el Océano Índico, lo que sugiere que podría ser genéticamente distante de las variedades argentinas y americanas utilizados en esta investigación (Acevedo et al. 2017; Garsmeur et al. 2018). En tercer lugar, el software de alineamiento fue desarrollado para especies diploides, lo que representa un inconveniente técnico para una especie altamente poliploide y aneuploide como la caña de azúcar, más allá de las limitantes computacionales. Estas observaciones aumentan la dificultad de localizar las lecturas en la posición exacta del genoma de referencia, aunque Yang et al. (2017b) informaron que el programa Bowtie 2 era más eficaz que otros alineadores en la caña de azúcar.

Curiosamente, el biotipo de alta fibra y los cultivares de caña de azúcar mostraron tasas de alineación similares, independientemente de la plataforma utilizada (Tabla 1.1).

Luego, se llevó a cabo la evaluación de cuatro genomas de *Saccharum sp.* que fueron publicados en los últimos años, junto con el genoma del *Sorghum bicolor* cv BTx623, el cual se comparó con el R570. Como se explicó previamente en la introducción, muchos estudios sobre el genoma de la caña de azúcar continúan utilizando al sorgo bicolor como punto de referencia, considerándolo la especie diploide más relacionada (Garsmeur et al. 2018), a pesar de la disponibilidad de varios genomas de *Saccharum* en las bases de datos. La selección de las muestras permitió llevar a cabo una evaluación de los diferentes "tipos" de cañas, tanto de azúcar como de energía, así como variedades de sorgo. Los sorgos se utilizaron como "control" en la técnica, ya que, al alinear los materiales propios con el genoma de referencia, se obtuvo un alto porcentaje de coincidencia. Al menos de manera preliminar, observamos tasas de alineamiento más elevadas, aunque no se incrementó de manera significativa el número de SNPs detectados.

El software Stacks proporciona dos tipos de análisis para la recuperación de SNPs: el *de novo* y el basado en la referencia (Rochette y Catchen 2017). Como era de esperar, el análisis *de novo*, considerando los datos crudos, permitió una mayor tasa de detección de SNPs que el análisis basado en la referencia en ambas plataformas, ya que esta última rutina solo consideró las lecturas que mapearon contra el genoma de referencia. Sin embargo, los paneles obtenidos a partir del análisis con genoma de referencia dan información sobre la posición del SNP en el cromosoma, por lo que resultan más informativos. Además, como discutiremos más adelante, los filtros de calidad son esenciales para encontrar SNPs robustos.

Las simulaciones aplicadas para dilucidar el rol del tamaño del fragmento de lectura, la cobertura de la secuenciación y el impacto de las lecturas pareadas frente a las simples han demostrado que los SNPs no están distribuidos uniformemente a lo largo del fragmento. Como se esperaba, se recuperaron más SNPs en las lecturas más largas (No-150) que en las más cortas (No-70-PE/HC) para ambas estrategias de mapeo, pues que la probabilidad de encontrar una variante aumenta con el tamaño de la lectura, junto con la precisión del locus mapeado. Otros autores como Manimekalai et al. (2020) recomiendan lecturas pareadas de al menos 150 pb para distinguir los SNP homeólogos de los SNP alélicos en aloploides para el descubrimiento de variantes de alta calidad. Por su parte, el número de SNPs no es directamente proporcional al número de lecturas por individuo. En términos relativos, la cantidad de SNPs recuperados en los datos crudos en No-70-PE/MC (174.585 SNPs en el análisis con referencia y 363.919 SNPs en el *de novo*) fue mayor a la mitad de la cantidad recuperada en No-70-PE/HC (351.864 SNPs usando

la referencia y 767.203 SNPs en el análisis sin referencia), lo que sugiere que asignar una cobertura de secuenciación media (~4 millones de lecturas/individuo) sigue siendo una alternativa adecuada y rentable para aplicar ddRADseq a un mayor número de muestras.

La falta de información sobre la aplicación del protocolo ddRADseq en la caña de azúcar nos llevó a estudiar seis esquemas de filtrado (FS) compuestos por una combinación de diferentes MAF y profundidades de lectura por locus a fin de seleccionar los valores de los parámetros más adecuados para mejorar el llamado de variantes SNPs. La inclusión de filtros de mínima profundidad por locus se implementó para eliminar los falsos positivos y asegurar variantes de mayor calidad, y de máxima profundidad por locus que tienden a evitar las regiones paralógas o repetitivas (Ravinet y Meier 2020). Se observó que, con una cobertura de secuenciación media, a medida que el valor de la profundidad de lectura aumenta, el número de SNPs recuperados disminuye. Con una cobertura de secuenciación alta, esta tendencia se invirtió. La combinación de este MAF con el aumento del valor de corte mínimos y máximos para la profundidad de secuenciación por locus disminuyó los SNPs recuperados en ambas estrategias de mapeo Ne-70 y en No-150 sin referencia. Sin embargo, en No-150 analizado con genoma de referencia la combinación de este MAF con profundidades de lectura crecientes aumentó el número de SNPs recuperados. Esto podría estar relacionado con las regiones parálogas procedentes de los eventos de duplicación o con la naturaleza poliploide, donde se pueden encontrar varias copias de algunas regiones (o genes) dentro del mismo genoma que pertenecen a diferentes ancestros (subgenomas).

Los resultados obtenidos muestran que, las lecturas pareadas y de secuenciación de lecturas más largas (comparando No-150 con No-70-PE/MC, No-70-PE/HC, No-150-PE/HC y Ne-70) producen un mayor número de SNPs. Aunque la cobertura de secuenciación alta es ideal para los poliploides, en particular para la caña de azúcar, la cobertura de secuenciación media (alrededor de 4 millones y lecturas PE/individuo) proporciona suficientes variantes y, combinada con la plataforma Illumina Novaseq6000, podría ser una opción rentable para aplicar el protocolo ddRADseq a un mayor número de muestras. Por otra parte, se vio que la mejor combinación de filtros fue un MAF 0,1 y una profundidad de lectura entre 56x y 200x (para ambas estrategias de llamada de variantes).

Las regiones distales a lo largo de cada cromosoma de la caña de azúcar mostraron una densidad de SNPs relativamente alta tanto en los datos crudos como, en menor medida, en los FS1 y FS5 (Fig. 1.5) debido al número diferencial de SNPs retenidos en estos FS en comparación con los datos crudos (Tabla 1.2). Estas observaciones son consistentes con el patrón observado en el genoma de referencia del sorgo, donde las regiones distales de los cromosomas mostraron

una alta densidad de genes y las regiones pericentroméricas mostraron una baja densidad de genes y bajas tasas de recombinación (McCormick et al. 2018). En cuanto al número de SNPs por cromosoma, resultados similares fueron encontrados por You et al (2019) en donde el número máximo de SNPs se encontró en cromosoma 1 (16,66%), mientras que el número mínimo de SNPs se encontró en cromosoma 5 (4,35%). La mayoría de los SNP (91,77%) se localizaron en regiones génicas, con una media de 7,1 SNPs por gen (You et al. 2019).

Los resultados de los análisis de agrupamiento jerárquico y de componentes principales revelaron patrones similares. Es importante destacar que el biotipo de alta fibra quedó aislado respecto de los híbridos comerciales (Fig. 1.7), probablemente debido a los diferentes fondos genéticos de los híbridos comerciales respecto del biotipo de alta fibra (García et al. 2022; Kane et al. 2021). Curiosamente, NA 78-724 y NA 56-30 no compartieron el mismo grupo en ninguno de los paneles analizados, pese a pertenecer al mismo programa de mejoramiento de caña de azúcar. Sin embargo, NA 56-30 y HOCP 95-665 mostraron la distancia genética más cercana en No-150 (Fig. 1.7 b) entre las muestras ensayadas, en concordancia con las similitudes genéticas basadas en las estimaciones del coeficiente de parentesco entre ambos genotipos (Acevedo et al. 2017).

Para dilucidar el efecto funcional de los SNPs retenidos sobre genes, transcriptos, secuencias de proteínas y regiones reguladoras en los cinco genotipos de caña de azúcar ensayados, se utilizó el predictor de efecto de variantes (VEP). A grandes rasgos, se asignaron proporciones similares de SNPs a las mismas consecuencias funcionales en ambos paneles (Fig. 1.8 a, c). Esto demuestra además que, independientemente del número diferencial de SNPs detectados en cada plataforma, se asignaron proporciones similares de variantes a las mismas consecuencias. Entre los SNPs recuperados, las variantes con cambio de sentido (6,4 - 5,4%, Fig. 1.8 b, d), surgieron como variantes interesantes para tener en cuenta para análisis posteriores. Estos SNPs serían responsables del cambio de una base de la secuencia original, pero se conserva la longitud de la proteína. Más de la mitad de las variantes detectadas se localizaron dentro de regiones reguladoras, lo que sugiere que podrían alterar la funcionalidad biológica del gen (Fig. 1.8 a, c). Por último, en las variantes no codificantes o que afectan a genes no codificantes, el impacto es difícil de predecir o no hay evidencia de impacto en absoluto.

Aguirre et al. (2019) evaluaron la robustez del protocolo utilizando réplicas en *Eucaliptus dunnii*. En este trabajo, se realizó una comparación similar utilizando paneles de SNPs obtenidas de experimentos ddRADseq independientes para dos genotipos. Los paneles se generaron con tres rangos de profundidades de lectura por loci diferentes y excluyendo los datos perdidos. En todos los casos, las dos réplicas se agruparon con diferentes valores de distancia dependiendo

de la profundidad por locus analizada. Las diferencias observadas podrían estar relacionadas con la capacidad de detectar la heterocigosidad de un locus con baja frecuencia, como en SNPs de dosaje único, que son comunes en la caña de azúcar y requieren una alta profundidad de secuenciación por locus para ser detectados. Estos resultados son consistentes con los de Yang et al. (2017b), e indican que el uso de una profundidad de 56 lecturas por locus permite la selección de SNPs de alta calidad. Esto se debe a que con esta profundidad es posible obtener dos lecturas para alelos de baja frecuencia en los loci de dosaje único e identificar más del 90% de los SNPs potenciales incluso en una especie dodecaploide.

Capítulo 2. Generación de paneles de SNP en una población biparental y evaluación de su eficiencia en la construcción de mapas genéticos.

Introducción

Los programas de mejoramiento actuales buscan, además de mejorar ciertas características de interés, que pueden ser tanto agronómicas como de calidad del producto final, disminuir los tiempos del proceso. Las técnicas de biología molecular contribuyen a lograr ambos objetivos. El desarrollo de protocolos especie específicos es una de las principales ventajas de las técnicas de GBS, lo que permite caracterizar el germoplasma a través de SNPs específicos representativos de cada población. El desarrollo de paneles de SNPs potencia al programa al brindar las herramientas necesarias para un análisis exhaustivo de los genotipos y sus relaciones genéticas, así como de los potenciales cruzamientos y de las asociaciones fenotipo-genotipo.

A diferencia del capítulo anterior, en este capítulo, el llamado de variantes alélicas se realizó a nivel de una población de mejoramiento utilizando el mismo software Stacks v2 (Catchen et al. 2011; Rochette et al. 2019). Dentro del módulo que permite identificar SNPs previo al alineamiento con genoma de referencia, el Stacks permite especificar si se trabaja con una población natural o una población biparental, y el algoritmo a aplicar es ligeramente distinto en un caso o el otro (Rochette et al. 2019). De esta manera, los loci se procesan de manera diferente. El primer paso analiza las variantes considerando la profundidad de lectura de cada sitio polimórfico en el conjunto de individuos del grupo. Posteriormente, se revisa las variantes de cada individuo en particular. Así, la agrupación de los individuos puede resultar en un número variable de variantes dependiendo de cómo se realice esta agrupación (Rochette et al. 2019). Este enfoque puede ser especialmente interesante al analizar una población biparental en comparación con la agrupación de los mismos individuos sin relaciones de parentesco en un genoma complejo, como el de la caña de azúcar. A su vez, permite evaluar solo los SNPs que difieren en los progenitores y observar su comportamiento en la progenie, o comparar todos los SNPs que puedan aparecer dentro de los individuos. Esto resulta relevante debido a que el genoma altamente heterocigota y poliploide de la caña de azúcar representa un cuello de botella para el análisis de la heredabilidad de los rasgos de interés (Singh Sandhu et al. 2022).

La construcción de mapas genéticos se fundamenta en identificar los eventos de recombinación entre sitios genómicos y calcular las tasas de recombinación entre pares de loci.

En el caso de especies poliploides, para obtener datos multialélicos es crucial considerar las frecuencias de recombinación, realizar estimaciones de configuraciones de fase y reconstruir los haplotipos de los progenitores e individuos de la población (Margarido et al. 2007). Los grandes bloques de desequilibrio de ligamiento (LD), causados por los largos ciclos de autofecundación en caña de azúcar, pueden ser útiles para localizar marcadores vinculados a genes que controlan las características agronómicas claves. Estudiar la descendencia F1 de dos parentales heterocigóticos o incluso la población de descendientes obtenidas por autofecundación de un determinado cultivar, puede ser desafiante debido al genoma altamente heterocigota de la caña de azúcar y a la fuerte depresión por endogamia para varias características (Xiong et al. 2023). Sin embargo, existen varias alternativas de software que permiten hacer las estimaciones de LD basándose en datos de SNPs (Yang et al. 2017a; Gerard et al. 2018; Gerard 2021). Esta información resulta muy útil para la selección de SNPs que permitan la realización de mapas genéticos.

En estudios previos, distintos grupos de trabajo han desarrollado mapas genéticos para caña de azúcar, utilizando distintas metodologías y tipos de marcadores. Raboin et al. (2008) trabajo con 72 cultivares modernos, se estudiaron 1537 polimorfismos del tipo AFLP encontrando una mayor frecuencia de bloques de LD en los primeros 5 cM, disminuyendo progresivamente dentro de una ventana de 0-30 cM. Luego del análisis estadístico para evaluar la asociación de los marcadores, se conservaron 119 AFLPs por bloque de LD. Sin embargo, los autores encontraron dificultades metodológicas debido la presencia de múltiples copias del genoma que dificultaron el reconocimiento de marcadores ligados (Raboin et al. 2008). Recientemente, otros autores trabajando con una nueva generación de marcadores basados en secuenciación aplicaron diferentes estrategias para construir mapas genéticos obteniendo versiones fragmentadas de los mismos. Balsalobre et al. (2017) generaron un mapa, usando 933 SNPs obtenidos por GBS, con 223 grupos de ligamiento (LGs), 48 de los cuales se pudieron agrupar posteriormente en 18 grupos homólogos (HGs). También, utilizando técnicas de GBS, Yang et al. (2017c) construyeron mapas genéticos de alta densidad para los híbridos de caña de azúcar CP 95-1039 y CP 88-1762. Para el primer cultivar, utilizaron 2453 marcadores que se agruparon en 162 LGs con una longitud total del mapa de 4224,4 cM. Para el segundo cultivar, lograron obtener un mapa con 2154 marcadores que se agruparon en 142 LGs con una longitud total del mapa de 4373,2 cM (Yang et al. 2017c). En otro estudio, se utilizó el microarreglo Affymetrix Axiom de 100K para genotipar 469 clones derivados del cruzamiento de las variedades Green German e IND81-146, y una población autofecundada de CP80-1827. Se obtuvieron mapas genéticos con 3482 marcadores SNPs para Green German, con un tamaño de

3336 cM y 138 LGs, un mapa para IND81-146 con 1513 SNPs que abarcan 2615 cM y 90 LGs, mientras que la población de CP80-1827 con 536 SNP presentó una extensión de 3651 cM y 102 LGs (You et al. 2019).

Para especies del género *Saccharum*, resulta conveniente utilizar poblaciones F1 para construir mapas de ligamiento para uno o ambos progenitores. Esto se debe a que la alta heterocigosis, la depresión por endogamia y su propagación clonal dificultan la utilización de poblaciones F2, líneas endógamas recombinantes, líneas endógamas por retrocruzamiento y líneas de doble haploide (DH), como sí es posible en otros cultivos (Yang et al. 2018). La construcción de un mapa genético en una población de hermanos completos puede resumirse en cinco pasos básicos:

i. estimación de las fracciones de recombinación por pares y pruebas estadísticas asociadas;
ii. separación de marcadores en grupos de ligamiento;
iii. ordenamiento de marcadores dentro de cada grupo de ligamiento;
iv. cálculo de las probabilidades de la fase parental y actualización de la fracción de recombinación
v. si el orden es óptimo, el mapa está completo; de lo contrario, se vuelve al paso iii.

Durante la meiosis, las posibles configuraciones resultantes del apareamiento cromosómico son diversas, predominando los bivalentes en muchos casos. Sin embargo, la complejidad de estas configuraciones aumenta exponencialmente con el nivel de ploidía. Si se considera "m" como el nivel de ploidía, es posible identificar hasta "m" alelos distintos para un locus en un organismo. Asimismo, si algunos de estos alelos no pueden ser distinguidos entre sí, es necesario considerar la cantidad de copias de cada variante alélica específica, lo que se conoce como dosaje alélico. El dosaje alélico de los SNPs puede estimarse utilizando las proporciones de dos alelos de un mismo locus (la relación entre la abundancia de sus dos formas alélicas) (Mollinari et al. 2019).

En estudios genéticos de especies con altos niveles de ploidía se utiliza como estrategia para bajar la complejidad, marcadores de dosis única. En una población de hermanos completos, los marcadores presentan una segregación en proporciones de 1:1 (cuando solo están presentes en uno de los progenitores) o en proporciones de 1:2:1 (si se encuentran en ambos progenitores). De esta forma, es posible emplear el método de los cinco pasos mencionados anteriormente, junto con un software estándar apropiado para poblaciones diploides (Mollinari et al. 2020). No obstante, la utilización de marcadores de dosis única presenta limitaciones en la

construcción de mapas genéticos precisos. Estos enfoques submuestrean el genoma, lo que impide considerar los efectos de múltiples alelos en los modelos de mapeo de loci con características cuantitativas (QTL) u otros análisis. Además, cuando los marcadores están en fase de repulsión, resulta más difícil detectar un ligamiento estadísticamente significativo entre ellos debido a que existe una mayor probabilidad de que ocurra recombinación entre los marcadores.

Los avances en las técnicas de secuenciación han mejorado significativamente la precisión y exhaustividad de los mapas genéticos, permitiendo la generación de mapas más densos gracias a la capacidad de descubrir muchos más marcadores. Además, es posible relacionar estos marcadores con los mapas físicos, ya que se pueden referenciar a posiciones genómicas específicas (Vining et al. 2017). Las tecnologías de secuenciación de alto rendimiento han permitido la generación de mapas de ligamiento de alta densidad, ayudando al desarrollo de ensamblados genómicos para especies no modelo (Audano y Beck 2023). Sin embargo, los problemas como los errores de secuenciación y los errores de genotipado debidos a la baja profundidad pueden dar lugar a mapas genéticos inflados si no se abordan adecuadamente. Dada la complejidad del genoma de la caña de azúcar, caracterizado por extensos elementos repetitivos y copias casi idénticas difíciles de separar en cada cromosoma homo(eó)logo con lecturas cortas de NGS, se ha optado por el ensamblado de las secuencias en un único contig colapsado. Además, los métodos actuales de NGS generan contigs y scaffolds de longitud limitada, lo que resulta numerosas regiones sin información. Por tanto, los mapas de ligamiento de alta resolución se vuelven herramientas valiosas para colocar adecuadamente los scaffolds en los cromosomas y ordenar los marcadores SNPs (Yang et al. 2018). Se han desarrollado nuevas herramientas bioinformáticas como los paquetes de R OneMap, para diploides, y posteriormente el MAPpoly, para poliploides, que aplican la metodología descripta por Wu et al. (2002), basadas en los modelos ocultos de Markov, para modelar con precisión datos de secuenciación de baja cobertura, proporcionando estimaciones precisas de las fracciones de recombinación y de las distancias totales de los mapas (Margarido et al. 2007; Mollinari y García 2019). Estos avances ponen de manifiesto la viabilidad de la utilización eficaz de los datos de secuenciación para mejorar la calidad de los mapas genéticos sin necesidad de un filtrado exhaustivo de genotipos potencialmente erróneos.

En este capítulo se aplicó el protocolo ddRADseq desarrollado en el capítulo anterior a una población biparental con el objetivo de investigar el potencial de los marcadores SNP desarrollados para la construcción de mapas genéticos. Se analizaron estrategias que tienden a reducir la complejidad genética y otras que captan la complejidad de la presencia de múltiples

alelos para cada gen. Se realizó la evaluación de dosaje alélico y se evaluó la estructura de la población para evaluar la constitución genética de las progenies respecto a sus parentales.

Materiales y métodos

Material vegetal

Para este estudio se analizaron las progenies (90 individuos) de un cruzamiento biparental de los genotipos comerciales de caña de azúcar (*Saccharum spp*.) NA 78-724 (de origen argentino) x LCP 85-384 (de origen estadounidense (Xiong et al. 2023)) provenientes de la Estación Experimental Agropecuaria Famaillá, del Instituto Nacional de Tecnología Agropecuaria, Famaillá (27°03'S, 65°25'O, 363 msnm), Tucumán, Argentina. Estos materiales se usan como parentales en los programas de mejoramiento de caña de azúcar por sus características industriales y agronómicas (Acevedo et al. 2017; García et al. 2022)

Se sembraron trozos de tallos con yemas viables en macetas de 4 litros, para obtener un clon por genotipo. Las plantas crecieron en invernáculo en el Instituto de Suelos del Centro Nacional de Investigaciones Agropecuarias (CNIA-INTA) hasta el momento de realizar los ensayos.

Construcción y secuenciación de bibliotecas ddRADseq para caña de azúcar

Se utilizaron hojas de tres réplicas por accesión de caña de azúcar para la extracción de ADN de alta calidad. La extracción de ADN genómico vegetal, la evaluación de su calidad y cantidad se realizó de la misma forma que se explicó en la sección Materiales y métodos del capítulo 1. Se construyeron dos bibliotecas de 48 muestras cada una. Para ello, se utilizó el protocolo ddRADseq 2 optimizado para *Eucalyptus dunnii* (Aguirre et al. 2019). En este caso, como se mencionó en la sección Materiales y métodos del capítulo 1 se modificó este protocolo para caña de azúcar utilizando la combinación de enzimas de restricción PstI/MboI en el paso de digestión de ADN. El tamaño de los fragmentos de ADN seleccionado fue de 260-460 pb + 140 pb de barcodes (Molina et al. 2022). Todos los experimentos se realizaron en la Unidad de Genómica, IABIMO-Instituto de Biotecnología INTA, Argentina. La biblioteca se secuenció utilizando la plataforma Novaseq6000 de Illumina (Illumina Inc, San Diego, CA, EE. UU.). Las lecturas obtenidas fueron de 250 pb pareadas; ~cuatro millones de lecturas/individuo.

Paneles de SNPs de caña de azúcar

El flujo de trabajo bioinformático usado para analizar los datos fue el descripto en la sección Materiales y métodos del capítulo 1. De las lecturas obtenidas del secuenciador, se descartaron

aquellas que fueran de baja calidad o tuvieran nucleótidos no identificados, secuencias de adaptadores o sitios de corte de enzimas de restricción deficientes. Las lecturas recuperadas se mapearon contra el genoma de referencia del genotipo comercial de caña de azúcar R570 (Garsmeur et al. 2018) con el software Bowtie2 (Langmead y Salzberg 2012).

Para la identificación de SNPs, se utilizó la estrategia “con referencia”, implementada con el módulo "ref_map.pl" del paquete Stacks v2, donde es posible agrupar las muestras según su procedencia, en este caso en particular las opciones son si son poblaciones biparentales o naturales, entre otras. La diferencia fundamental entre estas opciones radica en el método de procesamiento de los loci. En el caso de poblaciones biparentales se registran los alelos de los parentales primero y luego los de las progenies, mientras que, en las poblaciones naturales se priorizan los loci con información por encima de la jerarquía de parentesco en la población. Se exploraron estas dos opciones disponibles para determinar el agrupamiento de las muestras a evaluar y de esta manera seleccionar la que mejor refleje la realidad o genere SNPs robustos. En la primera, se definieron las lecturas correspondientes a ambos parentales y a las progenies, para analizar como población biparental. En este enfoque, se comparan inicialmente los dos parentales y luego se registra el genotipo de las progenies en las regiones donde se detectan datos. En una segunda instancia, se consideraron todas las muestras como parte de una población natural, evaluando a todos los individuos sin distinción de orden o parentesco. Esta estrategia permite obtener paneles de SNPs más numerosos que permitan evaluar información que este en baja frecuencia dentro de la población analizada. Además, para cada análisis en el llamado de variantes, se estableció que el porcentaje mínimo de individuos de la población requerido para procesar un sitio nucleotídico fuese el 90% para el análisis como población biparental y 80% para el análisis como población natural, y un valor de MAF de 0,1. Los paneles de SNPs se reportaron en formato VCF. Este software reporta al alelo que se encuentra en mayor proporción en los datos analizados como alelo de referencia, mientras que llama alelo alternativo a la variante encontrada para ese locus.

Se realizó una evaluación de los paneles de marcadores a través del paquete VCFTools (Danecek et al. 2011) y se utilizó el software R para el análisis estadístico y para realizar los gráficos (R Core Team 2023). Luego de este análisis, los paneles de SNPs, en formato VCF, se filtraron para recuperar solo los loci que estuvieran en el 80% de los individuos, un MAF de > 0,1, y una profundidad de lectura RD= 10x-200x. Se tomó la determinación de bajar el umbral de profundidad de lectura a 10, respecto al usado en el capítulo 1, para poder recuperar mayor número de SNPs, teniendo en cuenta el aumento considerable del número de muestras evaluadas. Se realizó la imputación de datos faltantes a través del módulo LD KNNi del software

Tassel, para continuar los análisis posteriores. Este algoritmo de imputación utiliza el desequilibrio de ligamiento (LD) para identificar buenos predictores para cada SNP, y luego utiliza los SNPs de LD alto para identificar a los vecinos más cercanos a K. El genotipo se llama con un modo ponderado de los KNNs

Construcción de mapas genéticos

Aproximación para diploides

Para conocer la segregación de los marcadores SNPs filtrados se utilizó el paquete de R OneMap (Margarido et al. 2007). Este análisis emplea la verosimilitud multipunto mediante modelos ocultos de Markov (HMM) para estimar simultáneamente el ligamiento, las fases de ligamiento y el ordenamiento de marcadores. Se utilizó un conjunto de diferentes patrones de segregación de marcadores que fueron codificados según la notación propuesta por Wu et al. 2002. Los alelos codominantes se representaron como "a" y "b". Se utilizaron marcadores de tipo codominante para evaluar la segregación en tres tipos de cruzamientos: 'B3.7' (ab × ab), 'D1.10' (ab × aa, la madre es heterocigota) y 'D2.15' (aa × ab, el padre es el heterocigota).

Utilizando la función find_bins se determinó cuáles son los SNPs redundantes. Para evaluar si existe alguna distorsión en la segregación de los marcadores se utilizó la función segreg_test, que realiza una prueba estadística de chi-cuadrado para comprobar si un marcador específico sigue el patrón de segregación esperado según la segregación mendeliana. Por defecto, utiliza como umbral para la prueba un α global=0,05 corregido para pruebas múltiples con la corrección de Bonferroni.

Para agrupar los marcadores en grupos de ligamiento, se utilizó la función suggest_lod para calcular una puntuación LOD sugerida teniendo en cuenta que se están realizando múltiples pruebas. La fracción de recombinación asignada fue de 0,40, según las sugerencias de estudios previos para la detección de las fases de acoplamiento y repulsión (Andru et al. 2011). El mapa genético marco se construyó con la función de mapeo de Haldane, que permite incluir múltiples sobrecruzamientos entre dos marcadores. Se utilizaron las funciones make_seq y record, para definir los grupos de ligamiento y ordenar los marcadores, respectivamente. La función draw_map2 permitió obtener una visualización de cada grupo de ligamiento representado como un cromosoma y los marcadores como líneas que los cortan a una determinada distancia (en centimorgan). La calidad de los datos para la construcción del mapa marco se analizó con un mapa de calor a partir de las frecuencias de recombinación de los marcadores.

Aproximación para poliploides

Se utilizó el paquete de R updog (por las siglas en inglés de "Using Parental Data for Offspring Genotyping") (Gerard et al. 2018) para la estimación del dosaje alélico para cada SNP por individuo, con un parámetro de ploidía fijo de 8 y considerando el modelo de población F1. Este paquete permite estimar cada genotipo, basándose en la profundidad de lectura y las medias de probabilidad posterior para cada SNP de cada individuo, considerando los posibles problemas técnicos asociados con los datos de secuenciación.

Con la información del dosaje alélico para cada SNP/individuo, se utilizó el paquete de R MAPpoly (Mollinari et al. 2019) para construir mapas genéticos con una aproximación para poliploides. Este paquete fue desarrollado para autopoliploides con niveles de ploidía uniformes, y aunque este no es el caso de la caña de azúcar, ya que es alopoliploide y los niveles de ploidía son variables, se incorporó al análisis para intentar obtener más información acerca del comportamiento en la segregación de los marcadores. Se eligió un nivel de ploidía de 8 que es el valor máximo que puede manejar el paquete cuando se utilizan Modelos ocultos de Markov (HMM).

Estructura genética y diversidad poblacional mediante análisis multivariados con datos genómicos

Se utilizó el paquete Adegenet para realizar un Análisis Discriminante de los Componentes Principales (DAPC), implementado en R (Jombart 2008; Jombart y Ahmed 2011). Este método permite asignar a los individuos en grupos o clusters y visualizar de sus contribuciones a la estructura genética de la población. Para lograr esto, se emplea una estrategia que implica la transformación de datos genómicos mediante un análisis de componentes principales (PCA), como paso previo a un análisis discriminante (DA). Esta secuencia asegura que las variables utilizadas en el DA estén completamente no correlacionadas y que su cantidad sea inferior a la de los individuos analizados. Este requisito del DA no se cumple cuando se trabaja con datos de paneles de SNPs y utiliza un enfoque multivariado que no se fundamenta en el equilibrio de Hardy-Weinberg (HWE) ni supone la inexistencia de desequilibrio de ligamiento. El DAPC asigna a cada individuo un coeficiente de pertenencia a cada grupo. Se retuvieron los componentes principales (PC) y dos funciones discriminantes, que en conjunto representaban el 99% de la variación genética global. Se utilizó el algoritmo find-cluster para determinar el número ideal de conglomerados (K); este algoritmo realiza agrupaciones sucesivas K-means con valores crecientes de K. El número de conglomerados debe ser el que posea menor valor según el Criterio Bayesiano de Información (BIC). Las accesiones se dividieron en subpoblaciones

mediante el método DAPC. Los gráficos de carga destacaron las contribuciones de los SNPs a cada función discriminante.

Por último, se realizó un análisis de componentes principales (PCA) sobre el panel poliploide y se graficó con el paquete ggplot2 de R para analizar el conjunto completo de SNPs de cada panel

Resultados

Paneles de SNPs de caña de azúcar

Al igual que sucedió con el set de muestras utilizado en el capítulo 1 de esta tesis, al evaluar la progenie de los genotipos comerciales NA 78-724 x LCP 85-384, el kit EasyPure (TransGen Biotech) permitió obtener una buena calidad y cantidad de ADN genómico. En el Apéndice 2 Tabla 2.1 se observan los índices de absorbancia (A260/A280 y A260/A230) obtenidos en un Nanodrop (Thermo Fisher Scientific, Waltham, MA, EE.UU.), y la cuantificación de ADN por fluorometría medida en un Qubit 2.0 (Thermo Fisher Scientific). El protocolo de ddRADseq para muchas muestras, implementado en este capítulo, permitió generar dos bibliotecas de 48 muestras agrupadas de buena calidad y cantidad. Nuevamente, las enzimas de restricción elegidas, PstI/MboI, fueron una combinación exitosa para el paso de digestión.

Solamente el 10,8% de las lecturas fueron descartadas por no pasar alguno de los criterios de calidad establecidos (presentar errores en el sitio de corte de las enzimas de restricción, barcodes indeterminados). Las lecturas leídas desde el comienzo (R1) y desde el final (R2) del fragmento fueron de buena calidad, con un valor promedio de 35 según la codificación de Phred por base. El valor medio de las lecturas R1 fue de 4,73 ± 2,33 millones de lecturas y 4,70 ± 2,24 millones de lecturas para las R2 (Fig. 2.1). Esto demuestra que para cada muestra los valores de la R1 fueron similares a los de su par R2, pero entre muestras hubo mucha variación, lo que se puede apreciar en el valor del desvío estándar. En ambos conjuntos de muestras se incluyeron los parentales. Por esta razón, se obtuvo un mayor número de lecturas para estos dos genotipos, lo que se puede observar en los puntos que se encuentran por encima de ambos gráficos de cajas y bigotes. Los otros puntos que aparecen como outliers o fuera de tipo corresponden a las muestras 99, 125 y 163 (con 12.4, 13.3 y 10.3 millones de lecturas, respectivamente, tanto para las R1 como para las R2). Se lograron resultados similares a los ensayos anteriores al alinear con el genoma de referencia del cultivar comercial R570, con valores de mapeo de lecturas que oscilaron entre el 30% y el 40% de alineamiento.

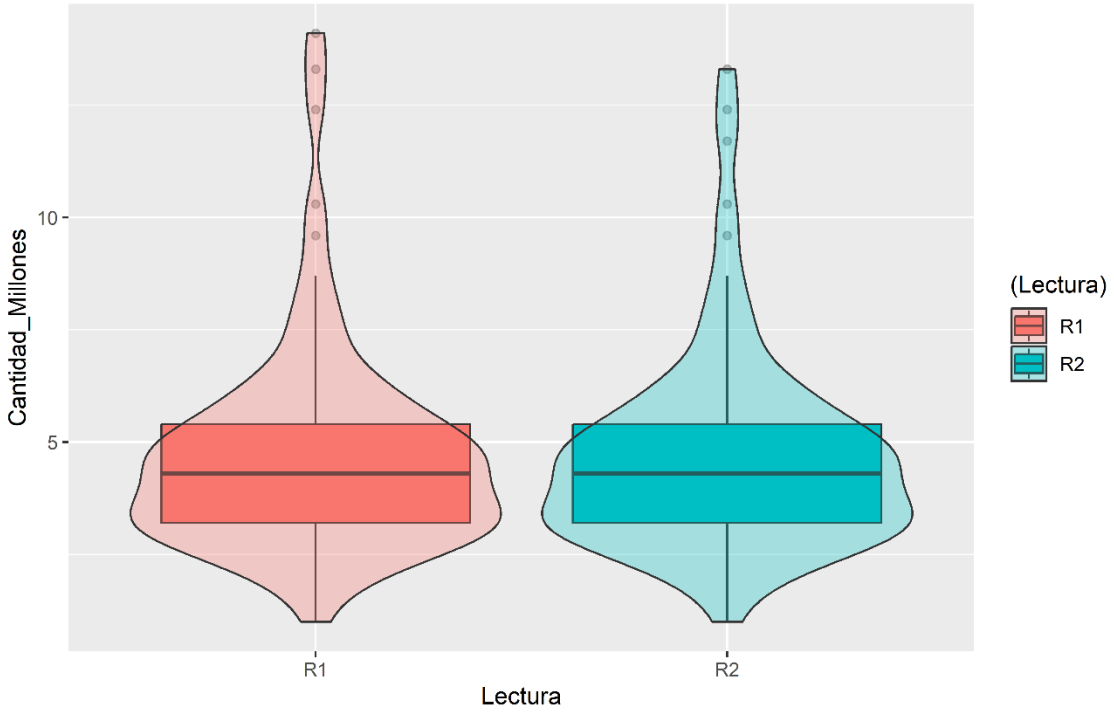

Figura 2.1. *Secuencias obtenidas para la población biparental.* Combinación gráficos violín y de cajas y bigotes para el número de lecturas leídas desde el comienzo y desde el final del fragmento; R1: lecturas leídas desde el comienzo del fragmento; R2: lecturas leídas desde el final del fragmento.

Llamado de variantes SNPs

El llamado de las variantes se realizó utilizando el mapeo con el genoma de referencia, bajo dos estrategias de agrupamiento de las muestras (biparental, natural), para evaluar aquella que ofreciera resultados más precisos en cuanto a la identificación de SNPs.

Selección de estrategias para identificación de variantes SNPs: panel A, una población biparental de caña de azúcar

La implementación de la estrategia con referencia para una población biparental, utilizada con el paquete Stacks (v2), permitió la identificación de un panel de SNPs en el cual las variantes identificadas están presentes en ambos parentales y en al menos 81 progenies (esto es porque el software toma como un grupo a los parentales y como otro grupo a las progenies). De esta forma, el panel quedó conformado por 24443 SNPs. La profundidad media de cada una de las variantes, esto es esencialmente el número de lecturas que se han mapeado en esta posición, tuvo mucha variación, con un mínimo de 1 y 2106 lecturas por sitio, con una mediana de 10,50 y media de 26,77 lecturas por sitio. Las regiones con mayor densidad de datos (Fig. 2.2) fueron en valores menores a 25.

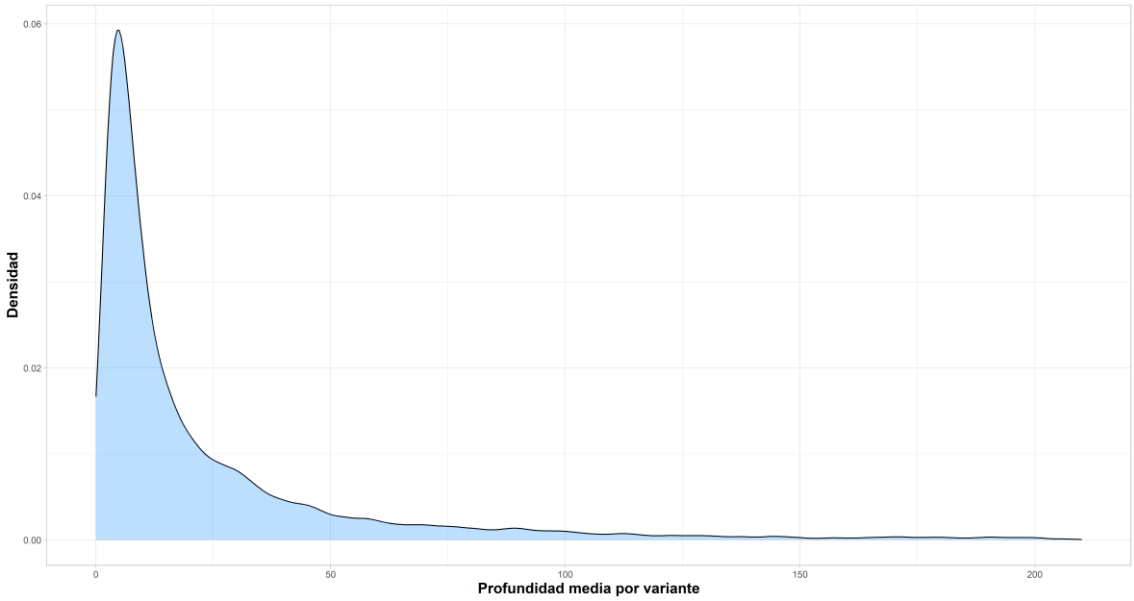

Figura 2.2. *Profundidad media de lecturas por variante.* Distribución de la profundidad media de las lecturas.

Al evaluar la profundidad media de lecturas por individuo, también se obtuvo mucha variación (Fig. 2.3). Este parámetro varió entre 27 y 294 lecturas promedio por sitio por individuo evaluado. La mediana y la media de la profundidad de lectura en todas las muestras fueron 72 y 85, respectivamente.

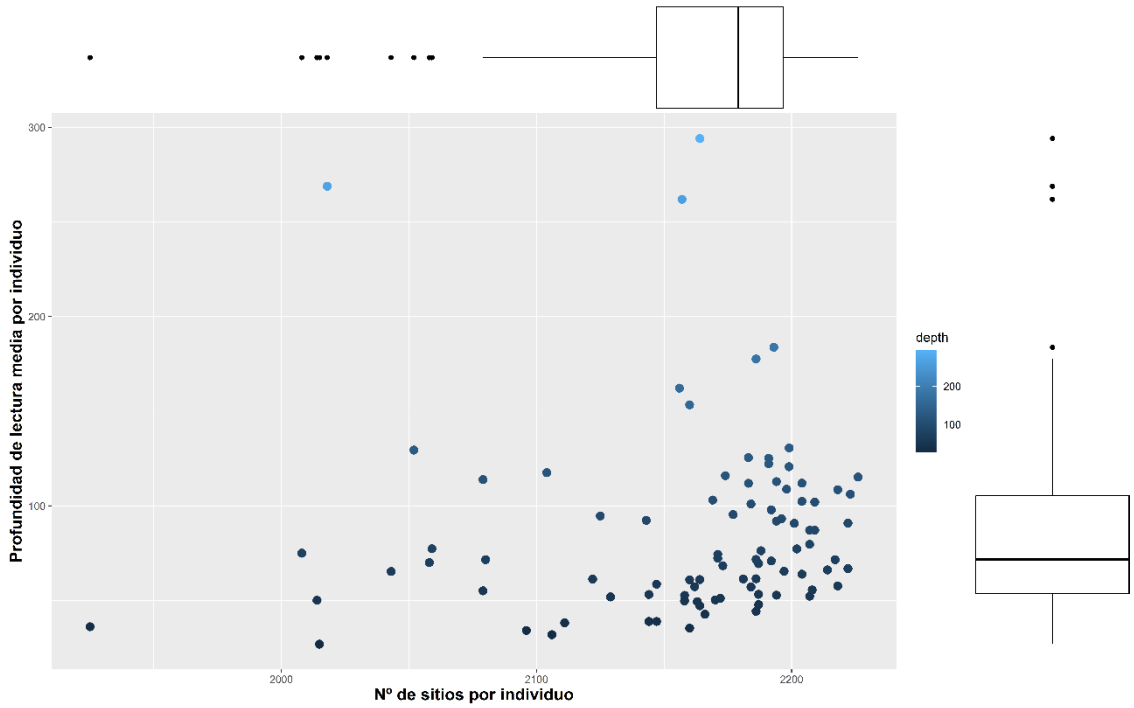

Figura 2.3. *Profundidad media de lecturas de las variantes y número de sitios evaluados por individuo.* Relación entre la profundidad media de lecturas de las variantes y el número total de variantes por individuo. La intensidad del color azul refleja el valor de la profundidad media de lecturas de las variantes, siendo más oscuro para los valores más bajos, y más claros para los más altos. Los diagramas de cajas y bigotes de la profundidad media de lecturas de

las variantes y del número de sitios para cada individuo permiten ver la distribución de los datos de estas variables por separado.

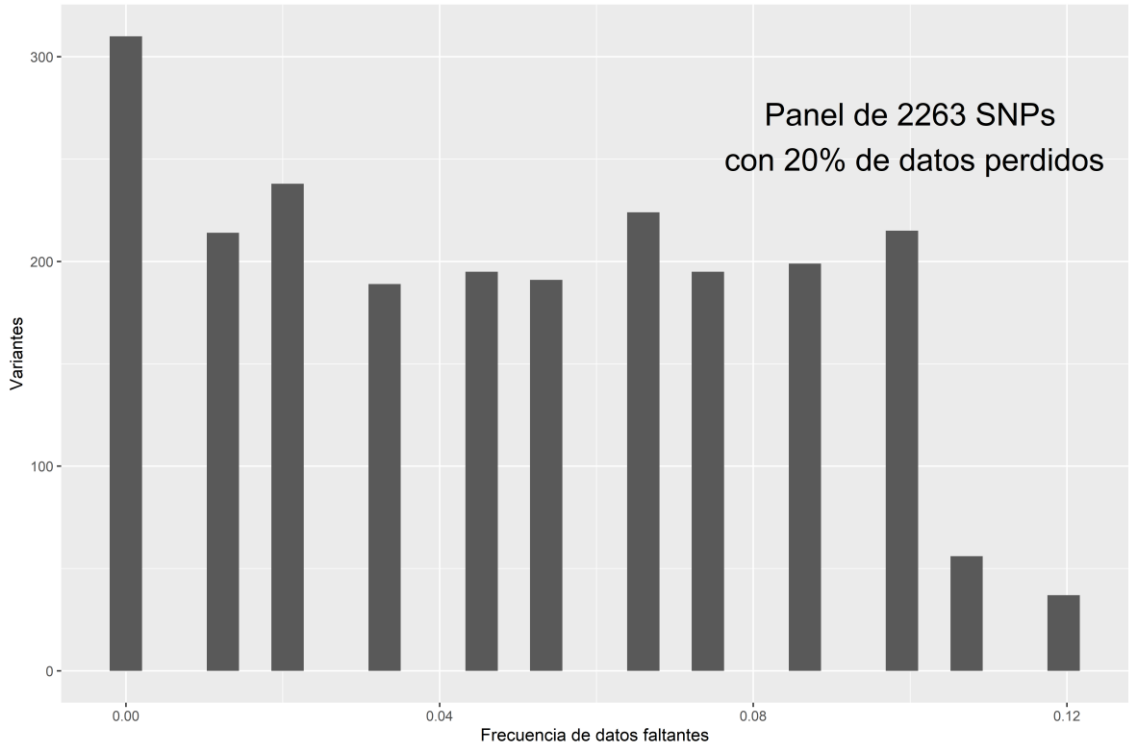

Figura 2.4. *Frecuencia de datos faltantes por variante SNP.* Número de variantes según la frecuencia de datos perdidos, seleccionando solo los SNPs que están presentes en el 80% de los individuos.

Dado que el análisis fue realizado para una población biparental, primero se buscaron los SNPs en los parentales y, luego, en las progenies. Además, como se mencionó anteriormente, se obtuvo mayor número de lecturas para los parentales que para la mayoría de las progenies. Por lo tanto, los parentales tuvieron muy pocos datos perdidos, mientras que en las progenies hubo mayor número de datos perdidos por cuestiones propias de la metodología aplicada. Esto resultó en que el número de SNPs disminuya de 24.443 SNPs a 2.263 SNP al aplicar el filtro de tolerancia de 20% de datos perdidos o datos presentes en el 80% de los individuos de la población. De esta forma, se puede observar que un pequeño número de sitios en todos los individuos tuvieron una baja frecuencia de datos faltantes. (Fig. 2.4).

La población de mapeo se creó mediante el cruce de progenitores poliploides que eran heterocigotas. En este sentido, la distribución de frecuencias alélicas para el panel de SNPs presentó una mayor densidad en frecuencias 0,5 (Fig. 2.5). Esto indica que hay mucha heterocigosis, lo cual es esperable de una población F1 proveniente de parentales heterocigóticos. Sin embargo, también se observaron loci en frecuencias diferentes a 0,5 (Fig.

2.5), que podrían asociarse a los largos ciclos de reproducción y a la depresión por endogamia que se ha reportado en estudios previos o a la segregación que no sigue un patrón bivalente (Xiong et al. 2023).

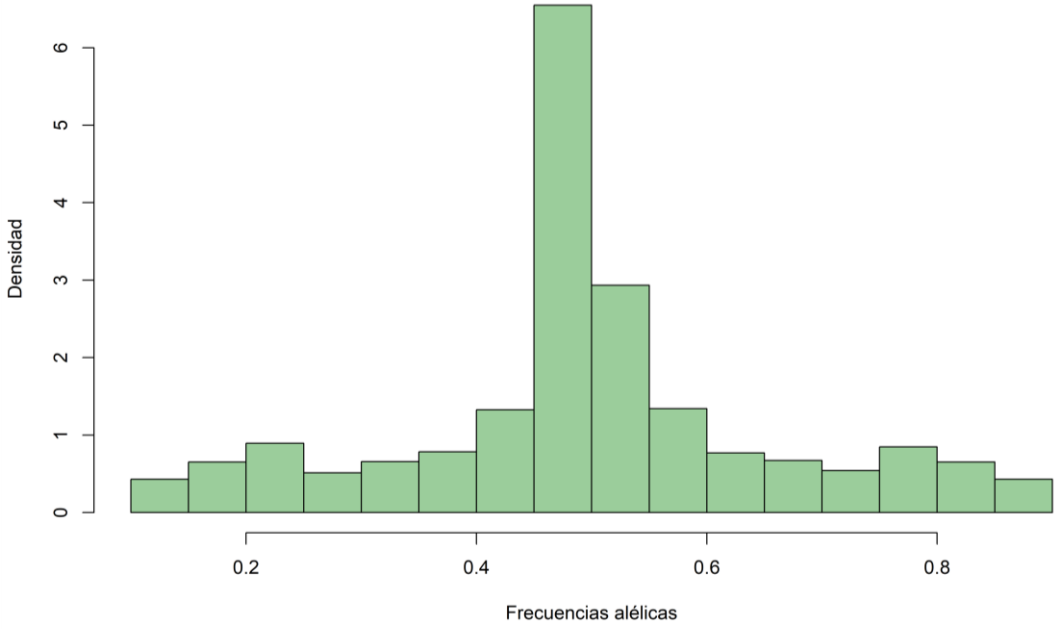

Figura 2.5. *Distribución de frecuencias alélicas.* Histograma de frecuencias alélicas.

Con esta información, se aplicó un segundo filtro para recuperar solo los loci que presenten un MAF mayor a 0,1. Finalmente, se aplicó el filtro de profundidad de lectura de entre 10x y 200x. Con estos filtros se obtuvo un panel de 2066 SNPs. La distribución de los SNPs identificados por cromosoma de caña fue heterogénea (Fig. 2.6 y 2.7) y coherente con el tamaño de estos, como se vio también en los resultados del capítulo 1. El cromosoma 1 acumuló 388 variantes, seguido de los cromosomas 3, 2 y 4 que acumularon entre 260, 251 y 234 SNPs respectivamente. Los cromosomas con menos variantes polimórficas fueron el 5 y el 8, con 118 y 114 SNPs respectivamente (Fig. 2.6). La distribución de los SNPs en cada cromosoma siguió un patrón similar al encontrado en el capítulo 1, donde se observó mayor variabilidad en las regiones distales y en el centro menos frecuencia de variantes o baches sin información (regiones en blanco) (Fig. 2.7).

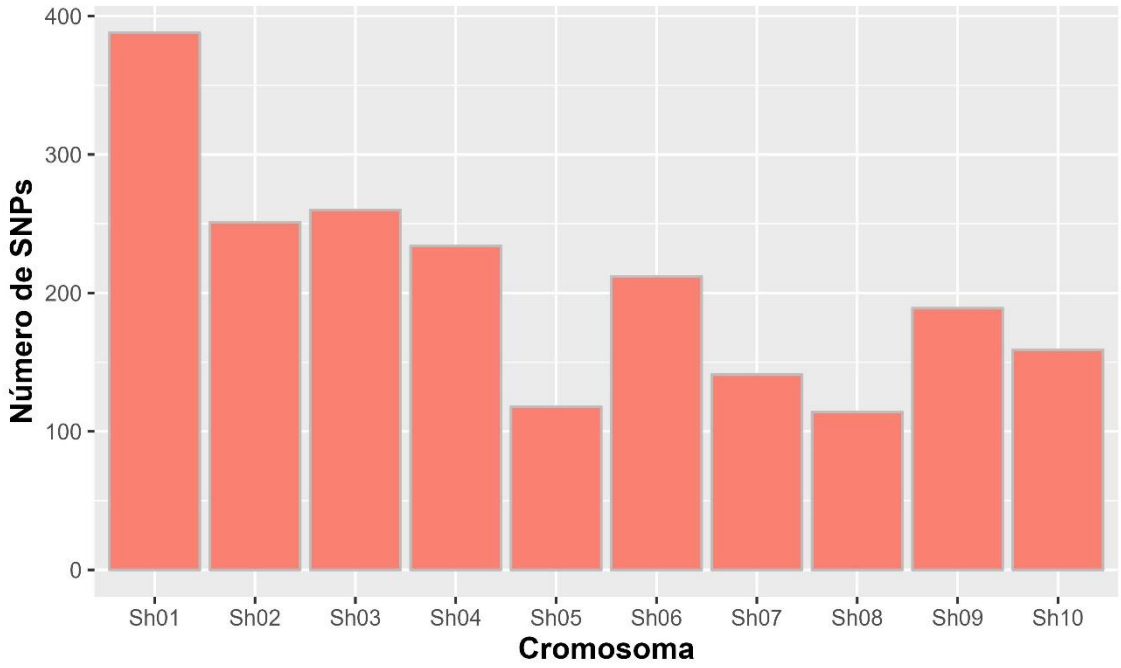

Figura 2.6. Número de v*ariantes SNPs por cromosoma.* Número de SNPs para cada uno de los 10 cromosomas de la caña de azúcar.

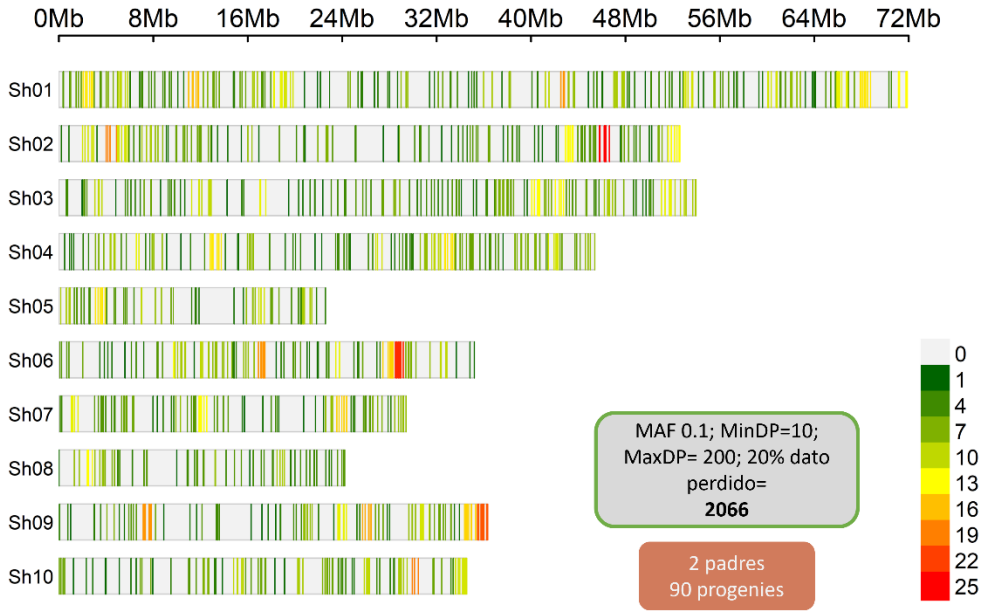

Figura 2.7. *Densidad de SNPs por cromosoma.* Distribución de SNPs en 10 cromosomas de la caña de azúcar cv R570. La escala de colores de la leyenda indica el número de marcadores dentro de una ventana de un Mb.

Selección de estrategias para identificación de variantes SNPs: panel B, población natural o priorización de loci informativos a relaciones de parentesco.

El llamado de alelos bajo el criterio de población natural permitió identificar un panel de 4372 SNPs en el cual las variantes registradas estén presentes en un mínimo de 73 muestras. Este parámetro fue más restrictivo que el utilizado en el apartado anterior, por lo que el panel generado tuvo menor proporción de datos faltantes. La profundidad media de lectura de cada variante, es decir el número de lecturas mapeadas en cada posición, osciló entre 3,6 y 2107 lecturas por sitio, con una media de 60,5 y una mediana de 35,9 lecturas por sitio respectivamente. La distribución de frecuencias alélicas se puede apreciar en la Figura 2.8a, en la que se observó un patrón similar al del Panel A, donde se registró una mayor concentración en el valor de 0,5. Posteriormente, se filtró el panel para recuperar solo los loci que tuvieran una profundidad de lectura de entre 10x y 200x, lo que resultó en el Panel B de 3907 SNPS. La cantidad de SNPs identificados por cromosoma (Fig. 2.8b) siguió el mismo patrón observado para el Panel A (Fig. 2.6).

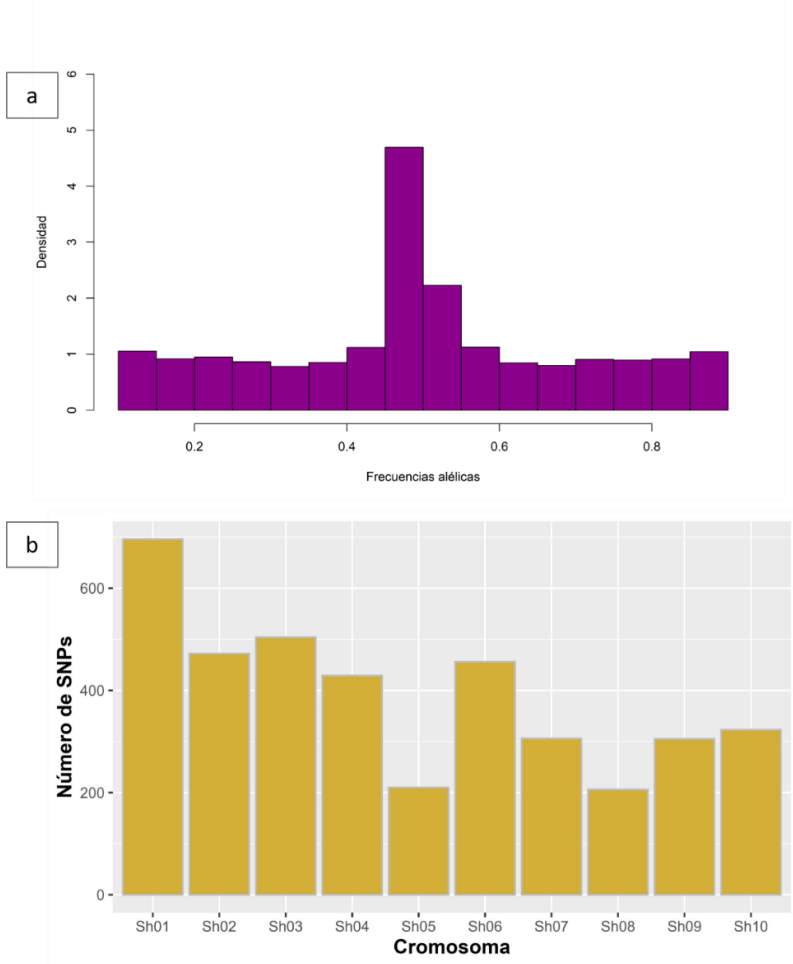

Figura 2.8. *Distribución de frecuencias alélicas y número* de v*ariantes SNPs por cromosoma.* a. Distribución de frecuencias alélicas; b. Número de SNPs para cada uno de los 10 cromosomas de la caña de azúcar.

Construcción de mapas genéticos

Se exploraron dos metodologías alternativas para la creación de mapas genéticos en organismos poliploides. Inicialmente, se evaluaron los 2066 SNPs del Panel A tratándolos como si fueran de una especie diploide (metodología más común para la creación de mapas) usando el algoritmo OneMap de R y se siguieron los criterios recomendados para tales casos. Luego, se calcularon las dosis alélicas de cada SNP y se exploró una metodología que permitiera inferir los grupos de ligamiento a partir de estos dosajes, con el fin de generar mapas para especies poliploides usando el paquete MAPpoly de R.

Aproximación para diploides

Para la construcción del mapa, utilizando el software OneMap (Margarido et al. 2007) desarrollado para especies diploides, se evaluó la segregación de los marcadores del panel (2066 SNPs) para retener aquellos que segregan de manera mendeliana. Se eliminaron las variantes del conjunto de datos que en ambos progenitores eran homocigotas para el mismo alelo, por lo que estos marcadores no se consideran informativos en poblaciones de cruzamiento biparental. De esta forma, el panel (2066 SNPs) quedó conformado por 553 marcadores SNPs. Dado que la información genotípica de algunos SNPs puede coincidir con la de otros marcadores debido a que no ha ocurrido ninguna recombinación entre ellos, su inclusión incrementa la carga computacional en la construcción del mapa y no aporta información adicional. En consecuencia, se excluyeron estos marcadores para la construcción del mapa. Por consiguiente, se agruparon los marcadores en intervalos según su información genotípica en 430 bins o grupos, con un promedio de 1,29 marcadores/grupo.

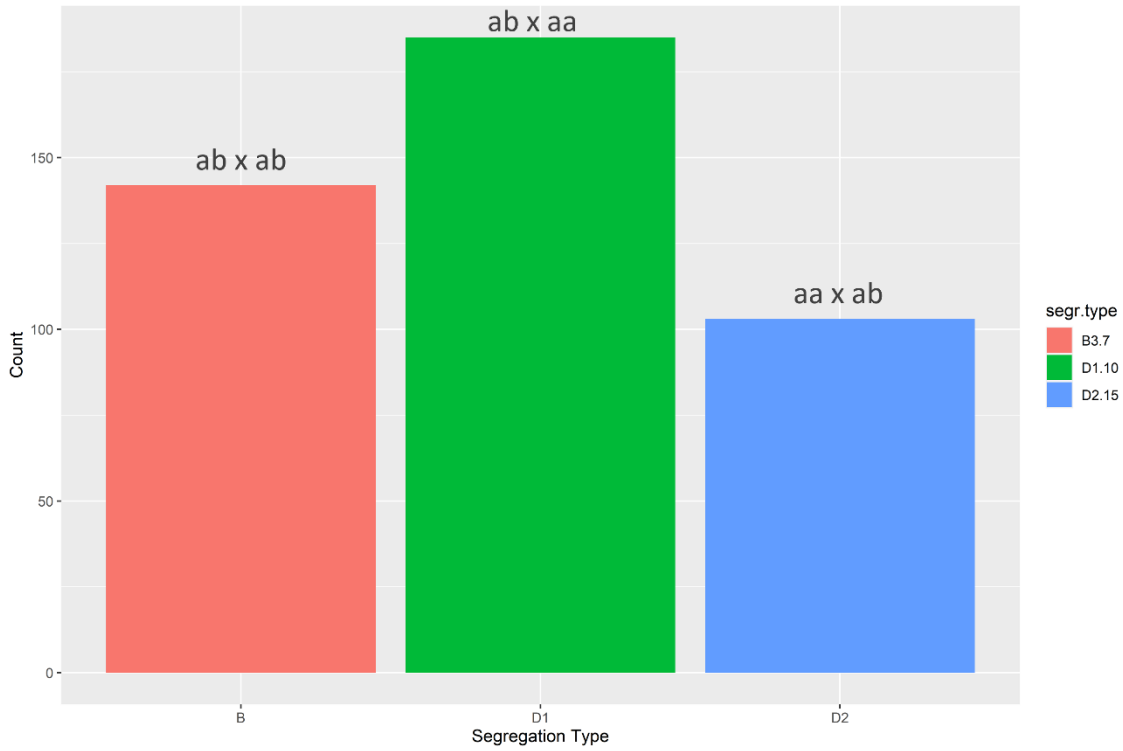

Figura 2.9. *Tipos de segregación.* B3.7= Parentales= ab×ab; segregación= 1:2:1 (a, 2ab, b). D1.10= Parentales= ab×aa; segregación= 1:1 (a,ab). D2.15= Parentales= aa×ab; segregación= 1:1 (a,ab).

Se evaluaron los tipos de segregación propuestos por Wu et al. (2002) y se observó que 142 SNPs resultaron heterocigotas para ambos progenitores (B3.7= ab × ab), mientras que 185 SNPs eran heterocigotas para el parental NA 78-724 (D1.10= ab × aa) y 103 SNPs eran heterocigotas para el parental LCP 85-384 (D2.15=aa × ab) (Fig. 2.9). Mediante la prueba de chi-cuadrado, se identificaron 183 marcadores sin distorsión en la segregación (Fig. 2.10), y estos se utilizaron para la construcción del mapa de ligamiento.

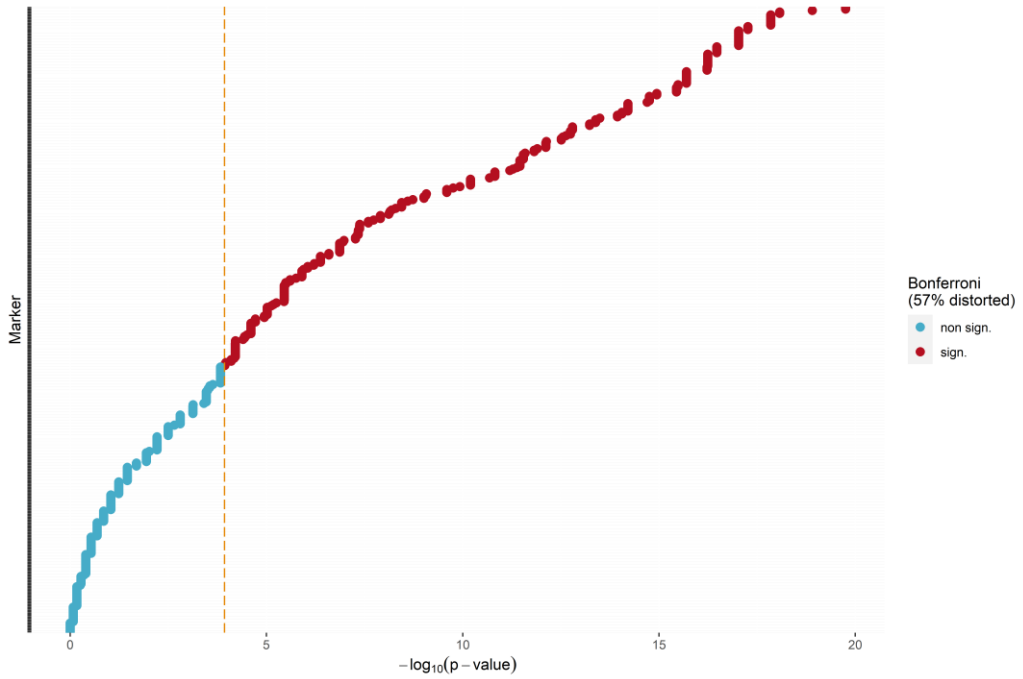

Figura 2.10. *Distorsión en la segregación mendeliana*. Non sign.= no es significativa la distorsión; sign= es significativa la distorsión.

Para la generación del mapa se emplearon 183 SNPs, con un LOD de 5,45 y una fracción de recombinación de 0,40, resultando en un mapa compuesto por 38 grupos de ligamiento (LG). De estos, 112 marcadores se asignaron a alguno de estos grupos, mientras que 71 no pudieron ser asignados. Con esta información, se elaboró un mapa genético marco que muestra cada LG con sus respectivos marcadores y la distancia entre ellos en centimorgan (cM) (Figura 2.11). Dado que las variantes se llamaron originalmente utilizando una referencia, se disponía de las posiciones físicas de los marcadores mapeados en los cromosomas para cada LG. Cuando un LG tenía más del 80% de marcadores correspondientes al mismo cromosoma de caña, se agrupaba en el grupo homólogo (HG), que se nombraba según el número de cromosoma de referencia correspondiente. El tamaño total del mapa fue de 570,1 cM. El grupo LG2, ubicado dentro del HG 1, fue el más extenso con 49,9 cM, mientras que los LGs 13, 32 y 33 fueron los más cortos, con solo 1,1 cM cada uno.

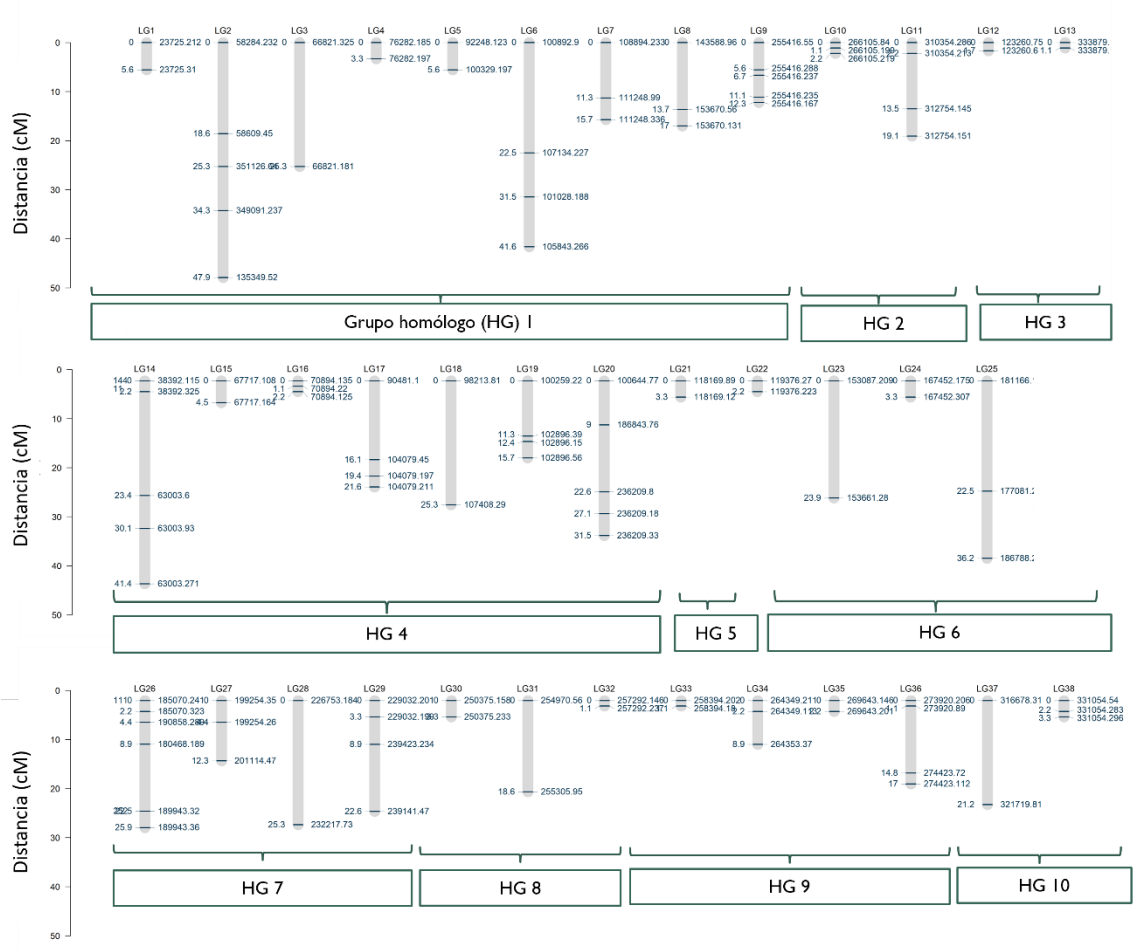

Figura 2.11. *Mapa genético marco.* Los 38 grupos de ligamiento (LGs) se representaron como cromosomas grises. Representados como líneas azules se encuentran cada uno de los 112 marcadores SNPs agrupados en 38 LGs. A la izquierda se detalla la distancia dentro de cada LG y a la derecha el nombre del SNP. Se agregó de forma manual el grupo homólogo (HG) al que pertenecía cada LG, según la posición física de cada marcador.

Para evaluar la calidad de los datos obtenidos para la construcción del mapa marco se graficó un mapa de calor a partir de las frecuencias de recombinación de los 112 marcadores (Fig. 2.12). Se puede observar que los colores más cálidos, que representan frecuencias de recombinación más bajas, se encuentran cercanas a la diagonal, formando pequeños grupos. La mayor parte de los marcadores se encuentran a distancias de recombinación más altas, aunque es posible ver varios puntos en celeste (frecuencias de recombinación entre 0,3 y 0,4).

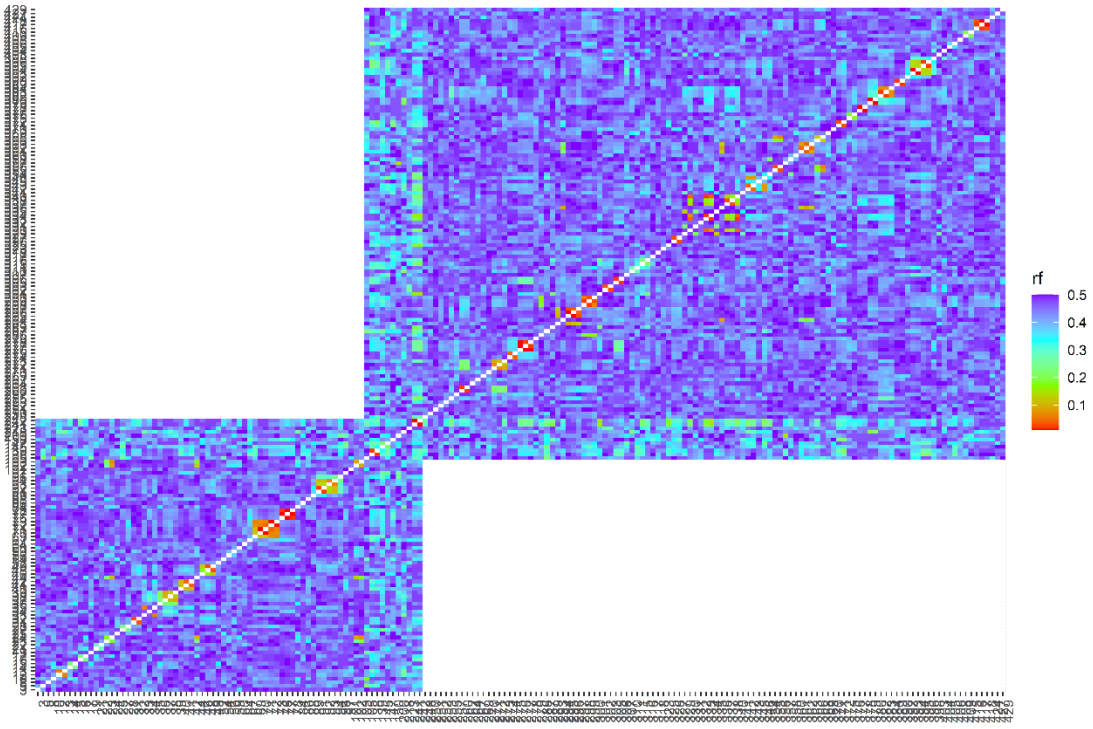

Figura 2.12. *Mapa de calor de 112 marcadores.* Cada punto representa las frecuencias de recombinaciones entre dos marcadores. En la diagonal se encuentra cada marcador consigo mismo. Los colores cálidos representan frecuencias de recombinación más bajas, mientras que los colores fríos representan frecuencias de recombinación más altas. rf= frecuencias de recombinación.

Aproximación para poliploides

La estimación de los dosajes alélicos para cada SNP por individuo del panel (2066 SNPs) se realizó con el paquete de R updog (Gerard et al. 2018). Este paquete devuelve dos conjuntos de datos. El primero engloba las características de los SNPs, incluyendo valores estimados del sesgo alélico y la tasa de error de secuenciación, así como la probabilidad del dosaje alélico de cada marcador para cada uno de los parentales (Fig. 2.13). El segundo conjunto de datos comprende las características de cada individuo respecto a cada SNP (Fig. 2.14).

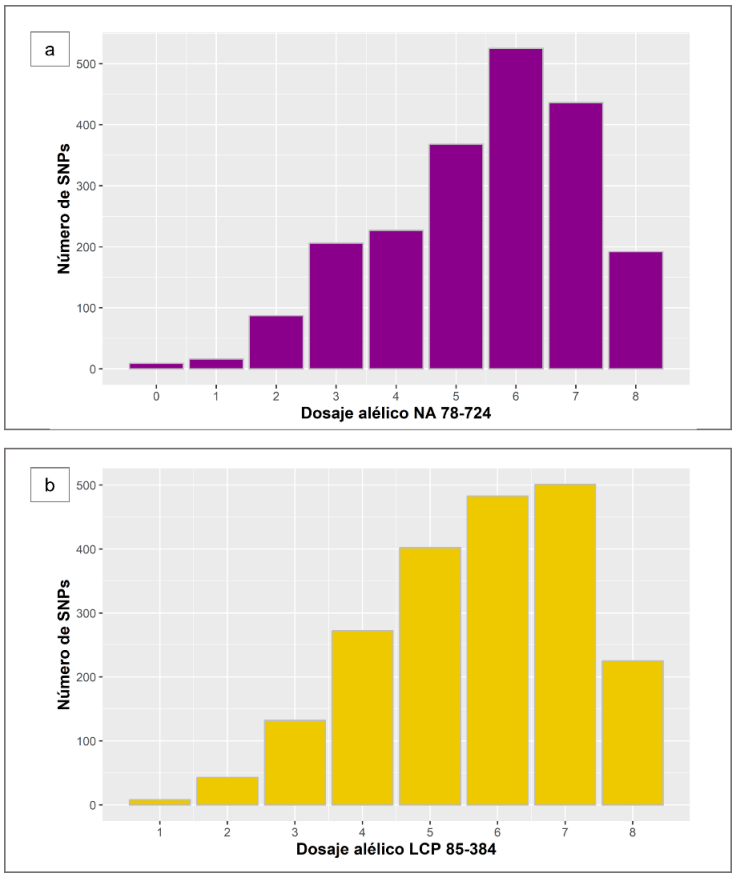

Figura 2.13. *Estimación de los dosajes alélicos de todos los marcadores para los parentales*. a. NA 78-724 y b. LCP 85-384. La estimación se realizó en base a la profundidad de lectura del alelo de referencia, respecto al número de lecturas total, asumiendo una ploidía uniforme de 8. Un marcador "7" representa siete copias del alelo de referencia y una copia del alelo alternativo.

La mayoría de los SNPs mostraron una alta probabilidad de tener 6 o 7 copias del alelo de referencia en los parentales, lo que significa que tenían 2 o 1 copias del alelo alternativo, respectivamente (por ejemplo, AAAAAAaa para el primer caso). Se observó una proporción mayor de la relación 6-2 en el NA 78-724 (Figura 2.13a) y de 7-1 en el LCP 85-384 (Figura 2.13b). Las progenies mostraron dosajes alélicos intermedios entre los parentales en algunos loci, como es esperado por la segregación mendeliana, mientras que en otros se comportaron de manera similar a uno de los parentales. El programa updog permitió tener una estimación de la correcta asignación de los genotipos de la progenie, teniendo en cuenta los errores de secuenciación, la profundidad de lectura por SNP para cada individuo y los datos faltantes. Esto permitió realizar una estimación más real de los genotipos poliploides, considerando la cantidad de copias de cada alelo para cada loci. A su vez, fue posible representar gráficamente la probabilidad de los dosajes alélicos de la población, como se muestran los ejemplos de la Figura 2.14. En la Figura 2.14a, ambos parentales tienen un dosaje alélico de 7-1, pero el padre tiene más lecturas que la madre. En la Figura 2.14b, el NA 78-724 tiene un dosaje alélico de 8, mientras que el LCP 85-384

tiene un dosaje alélico de 7. Esto permite observar los dosajes alélicos de toda la población para cada marcador.

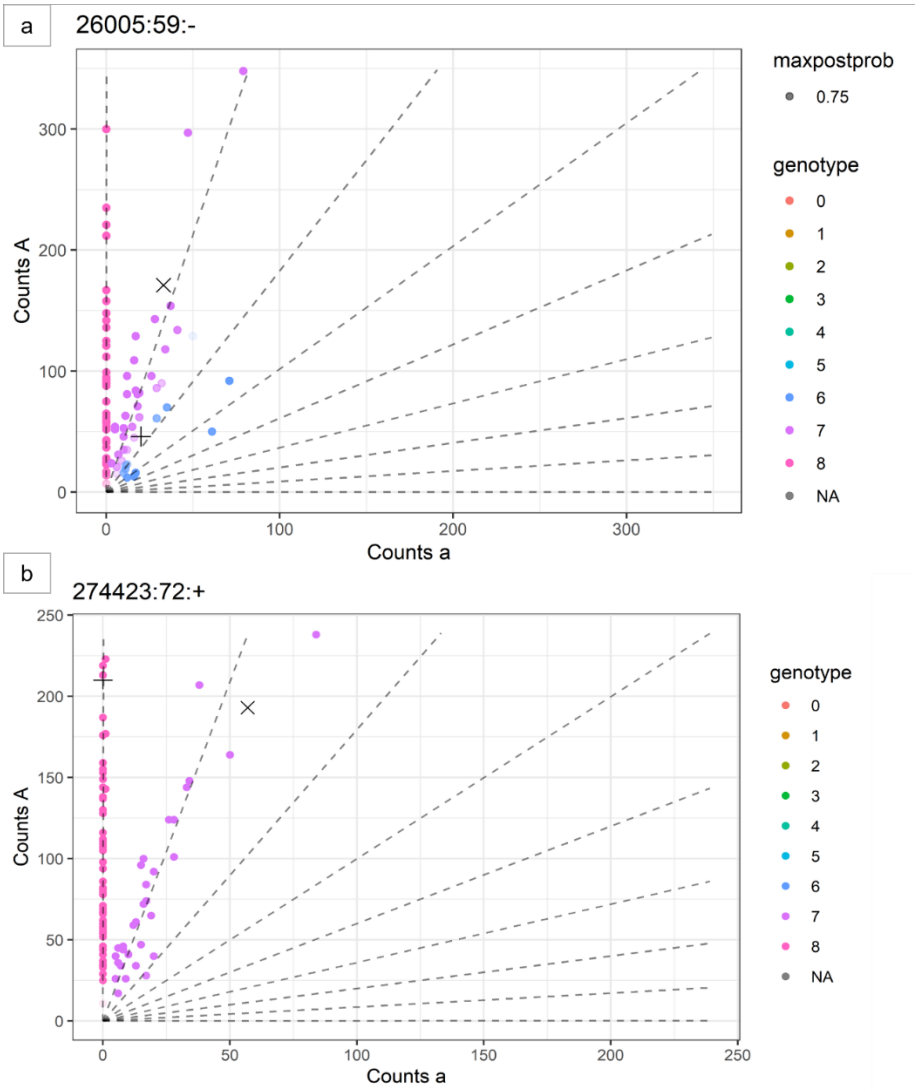

Figura 2.14. *Ejemplo del llamado de dos marcadores SNP con sus posibles dosajes alélicos utilizando updog*. a. Marcador 26005:59: -. b. Marcador 274423:72: +. El eje Y representa el número de copias del alelo de referencia (A) y el eje X representa el número de copias del alelo alternativo (a). El símbolo "+" es la ubicación del parental femenino (NA 78-724) y "x" es el parental masculino (LCP 85-384). Los individuos de la progenie se muestran como puntos codificados por colores según su genotipo con la probabilidad posterior más alta. Por ejemplo, un genotipo "7" representa siete copias del alelo de referencia y una copia del alelo alternativo (AAAAAAAa). Las líneas discontinuas indican nueve dosis posibles en un individuo octoploide. El nivel de transparencia es proporcional a la probabilidad posterior máxima. Esto equivale a la probabilidad posterior de que la estimación del genotipo sea correcta. NA= sin dato.

Para construir el mapa genético con una aproximación para poliploides se utilizó el paquete de R MAPpoly (Mollinari et al. 2019), a partir de la estimación de los genotipos, con sus dosajes alélicos, como se explicó previamente. Primero, se realizó un análisis exploratorio en el que se compararon los resultados de analizar los datos como si se tratara de una especie diploide, utilizando los mismos 2066 SNPs, pero con los dosajes alélicos asignados para aproximarse a un análisis del tipo poliploide (Fig. 2.15). En el análisis como diploide, se observa que la mayoría de los marcadores son heterocigotas, tanto en los parentales, que se puede observar en el panel de la izquierda como gráfico de barras del número de marcadores por cada combinación de dosis alélica en ambos progenitores (mayor proporción segregación 1:1), o en el panel de las progenies como color amarillo en el panel central (Fig. 2.15a). En caso del análisis como poliploide, se observa que mayoría de los marcadores son heterocigotas, con dosis alélicas muy variables, pero con mayor proporción de dosis alélicas correspondiente al alelo de referencia respecto al alternativo (segregaciones 7:7, 6:6, 6:7, 7:8). Las progenies por su parte tienen un comportamiento similar a los parentales (colores del celeste al azul) (Fig. 2.15b).

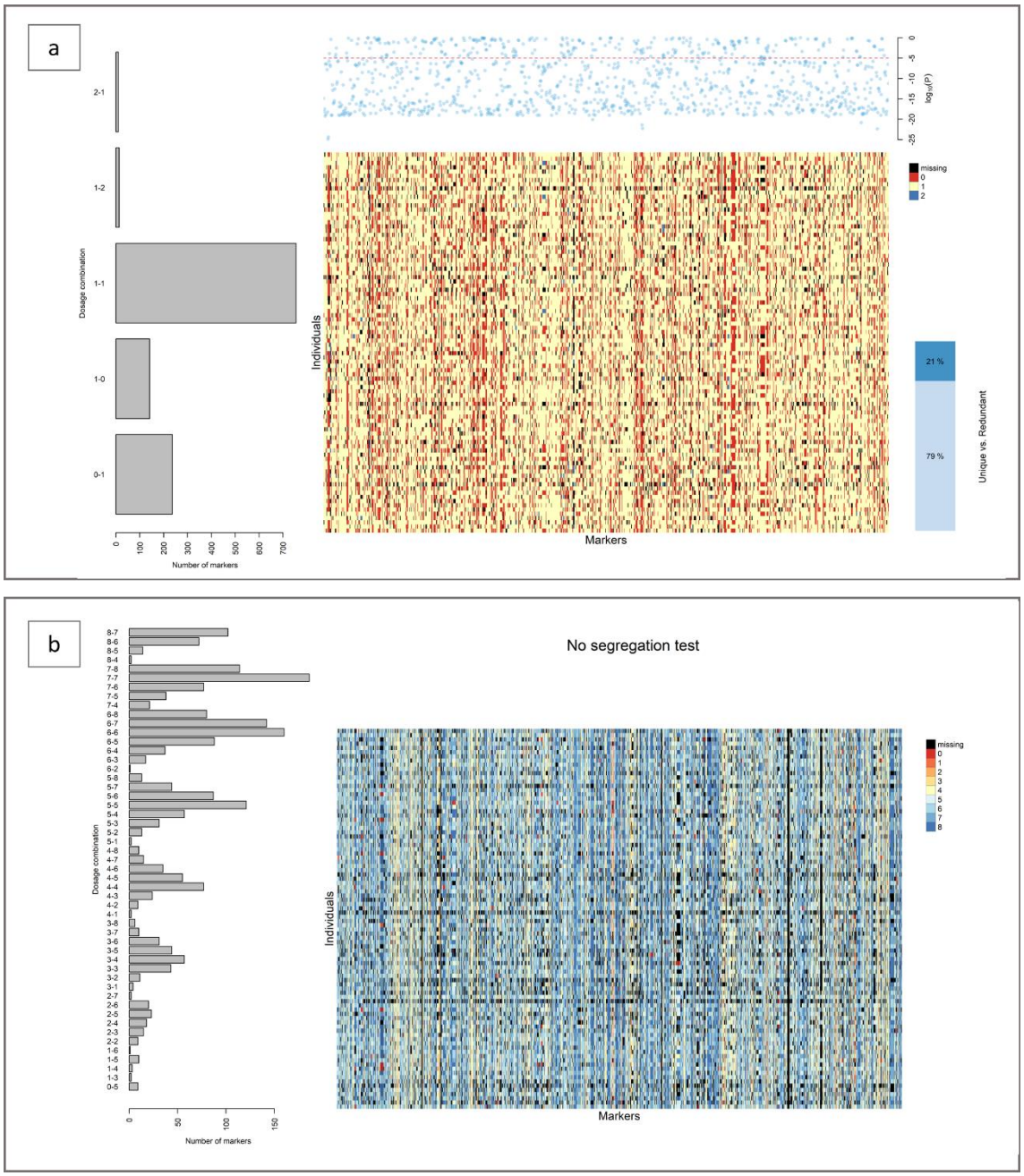

Figura 2.15. *Composición genética de cada genotipo según las estrategias utilizadas para el armado de los mapas.* Análisis como diploide con OneMap (Panel de 2066 SNPs diploidizados) y como poliploide con MAPpoly (Panel de 2066 SNPs con los dosajes alélicos estimados). a. El diagrama de barras de la izquierda es una representación gráfica del número de marcadores por cada combinación de genotipos de los progenitores (0= homocigota para el alelo de referencia; 1= heterocigotas; 2= homocigota para el alelo alternativo). Los números separados por un guion indican la dosis en los parentales NA78-724 y LCP85-384 respectivamente. En el panel superior derecho se muestra el log10 (valor p) de una prueba chi-cuadrado para los patrones de segregación esperados según la herencia mendeliana. El panel inferior derecho muestra una representación gráfica de la distribución de dosis en la población de hermanos completos. La barra azul indica la proporción de marcadores únicos vs. redundantes. b. A la izquierda se encuentra la representación gráfica del número de marcadores por cada combinación de genotipos de los progenitores, asumiendo una ploidía uniforme de 8. Los números separados por un guion indican la dosis en los parentales NA78-724 y LCP85-384 respectivamente. El panel superior muestra la leyenda "no segregation test" porque no se pudo obtener la información de los patrones de segregación esperados según la herencia mendeliana. El panel inferior derecho muestra una representación gráfica de la distribución de dosis en la población de hermanos completos.

El software MAPpoly permite una evaluación gráfica del desempeño de cada marcador (Figura 2.16), sintetizando la información y facilitando la comparación entre marcadores, así como la evaluación de la calidad. Por ejemplo, consideremos el marcador 274423:72:+. En este caso, podemos observar que el genotipo NA 78-724 tiene un dosaje alélico de 8, mientras que el parental LCP 85-384 presenta 7 copias del alelo de referencia y 1 del alelo alternativo. Además, este locus no tiene datos faltantes y la mayoría de las progenies son homocigotas para el alelo de referencia, aunque algunos muestran 1 copia del alelo alternativo. Este análisis nos permite visualizar las probabilidades de que cada marcador esté correctamente genotipado y la cantidad de genotipos que cumplen con estas características.

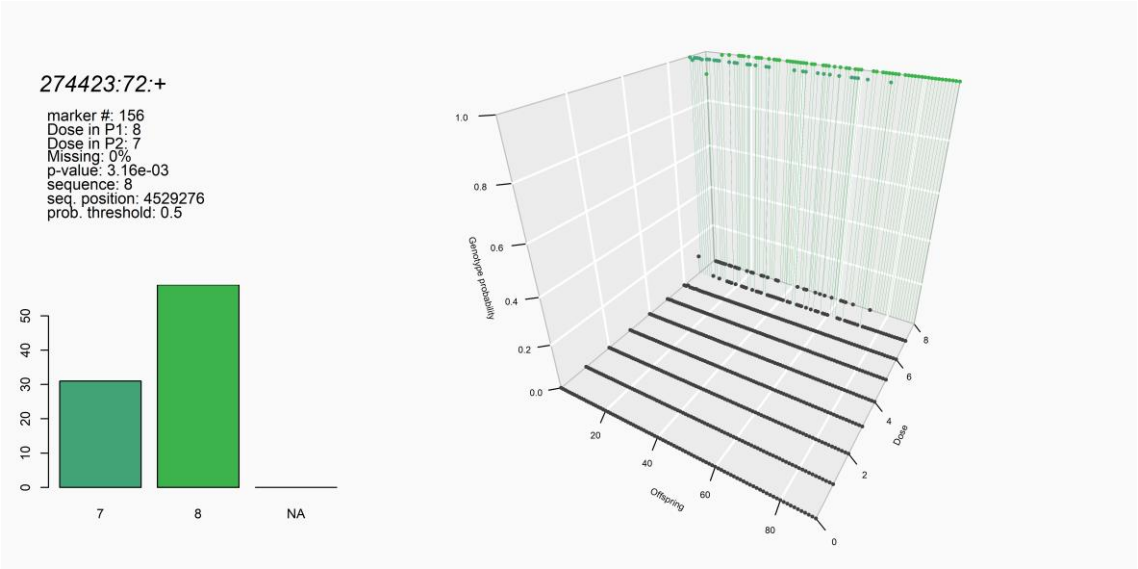

Figura 2.16. *Visualización de las probabilidades del dosaje alélico de los marcadores utilizando MAPpoly.* Ejemplo de un marcador SNP "274423:72:+" con sus posibles dosajes alélicos. Se utilizó uno de los marcadores de la Figura 2.14 para evaluar la información del marcador. El panel superior izquierdo da un resumen, informando el dosaje de cada parental, porcentaje de datos perdidos, valor p asociado, posición en el genoma de referencia y probabilidad de recombinación. En el panel inferior izquierdo se visualizan la proporción de progenies con dosajes de 7 u 8 y sin datos. Por último, como figura central se ve un gráfico de tres dimensiones en la que se ve por cada progenie (eje x) el dosaje alélico (eje z) y su probabilidad (eje y).

Se aplicó un proceso de filtrado de marcadores similar al del OneMap, por lo que el panel inicial de 2066 SNPs fue ajustado a 1252 SNPs tras eliminar los marcadores que no eran informativos y, luego, se refinó a 1196 SNPs al eliminar aquellos ubicados en la misma región genómica. Es importante aclarar que muchos de los marcadores se perdieron al no poseer la cantidad de lecturas necesarias para una evaluación del dosaje alélico. Por este motivo, los marcadores que pasaron los filtros de calidad fueron 568, distribuidos por todo el genoma, que conformarán el mapa poliploide. El algoritmo del MAPpoly es diferente al del OneMap, razón

por la cual cambió el número de marcadores final con el que se construyeron los mapas de ligamiento.

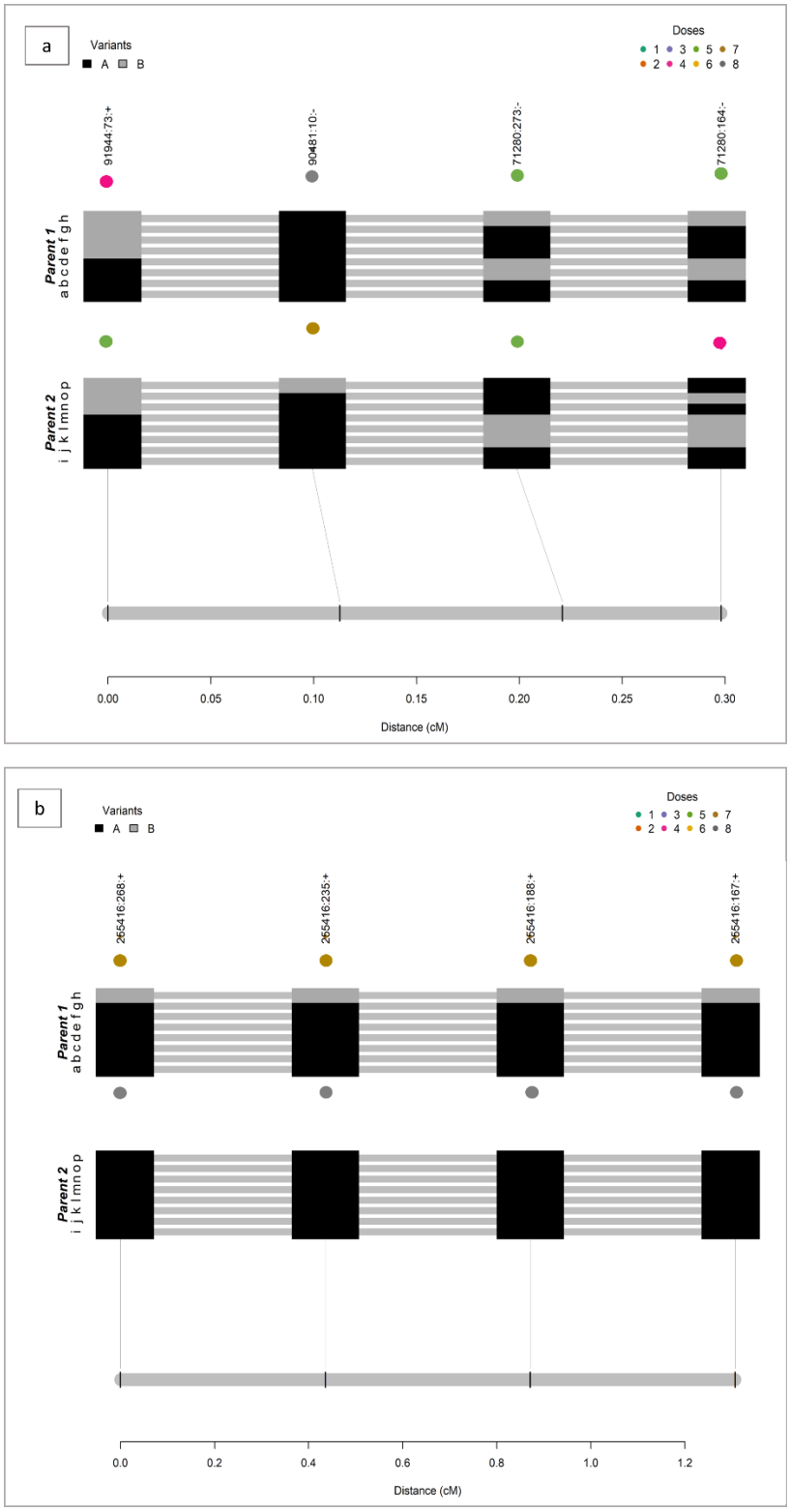

Figura 2.17. *Mapa genético de los parentales ejemplificado para los grupos de ligamiento 2 (a) y 7 (b).* Representados como líneas grises se encuentran cada una de las 8 copias del genoma. Los bloques negros o grises representan las posiciones de los marcadores, siendo el primer color la variante de referencia y la segunda la alternativa. Los puntos están codificados por colores según la dosis alélica de cada genotipo.

El mapa de ligamiento se construyó utilizando un algoritmo multipunto modelo oculto de Markov, en donde se estiman las configuraciones de las fases de ligamiento. A partir de las frecuencias de recombinación de los 568 SNPs, solo 51 se pudieron agrupar, formando grupos de marcadores muy distanciados entre sí. Debido a esto, no fue posible graficar todos los LGs en un mismo mapa. Sin embargo, se pudieron graficar los LGs por separado, mostrando los SNPs en los parentales y los dosajes alélicos representados con diferentes colores. La Apéndice Tabla 2.2 muestra el número de marcadores por cromosoma utilizados para calcular las fracciones de recombinación (Apéndice Tabla 2.2 a) y el número de marcadores por grupo homólogo (HG) (Apéndice Tabla 2.2 b). Los haplotipos de los parentales para los LG2 y 7, que poseen 4 marcadores cada uno se pueden visualizar en la Figura 2.17. El LG2 tiene una distancia de 0,30 cM, siendo la distancia física 26 Mb, mientras que el LG7 tiene 1,2 cM con una distancia física de 101 pb. Esta permite también la visualización de las copias de cada alelo para cada marcador, en cada parental. En la Fig. 2.17a hay más variación en los dosajes de los marcadores, algunos con 4 o 5 copias para el alelo de referencia, mientras que en la Fig. 2.17b es más uniforme, donde el LCP 85-384 no presenta copias para el alelo alternativo y NA 78-724 tiene un dosaje alélico de 7.

Estructura genética y diversidad poblacional mediante análisis multivariados con datos genómicos

Para conocer la diversidad genética y la estructura poblacional se realizó un Análisis Discriminante de los Componentes Principales (DAPC) sobre los 183 SNPs que no poseen distorsión en la segregación analizado con técnicas diploides (Fig. 2.10). Utilizando el valor BIC más bajo, se detectaron tres clústeres (Fig. 2.18). Se conservaron los 80 primeros PCs (99.3% de la varianza total) y dos funciones discriminantes. En la Figura 2.11 a. y b., se observa la que Función Discriminante 1 (eje horizontal) separó al clúster 1 del 2 y del tres, mientras que Función Discriminante 2 (eje vertical) permitió separar los clústeres dos y tres. El clúster 1 agrupó 43 genotipos, entre ellos el parental LCP 85-384, el grupo 2 permitió asociar a 12 individuos, entre ellos el otro parental NA 78-724, mientras que el clúster 3 agrupó a 37 líneas (Fig. 2.18a).

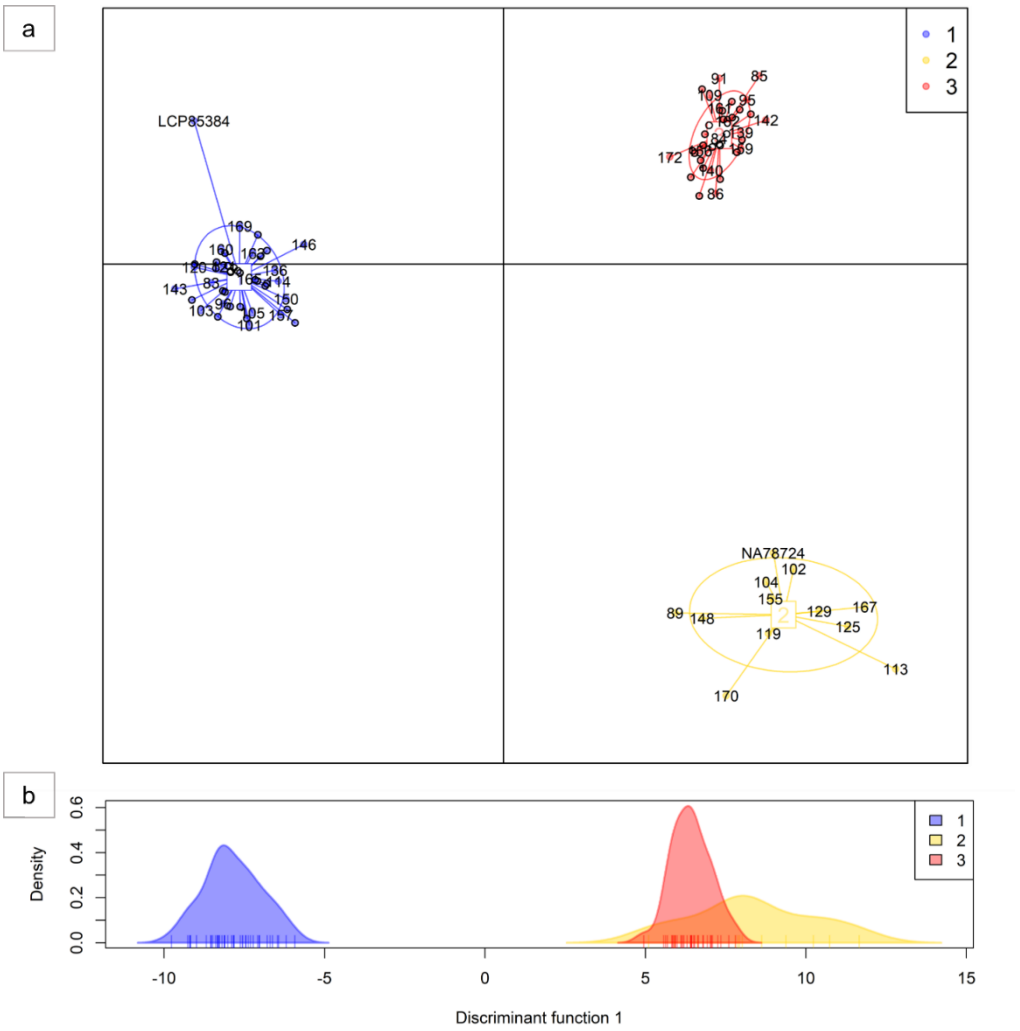

Figura 2.18. *Análisis discriminante de los componentes principales.* a. DAPC para la población biparental (dos parentales y las 90 progenies) a partir de 183 SNPs. 80 PCs y 2 funciones discriminantes fueron retenidos durante los análisis, para describir la relación entre los clusters. Los ejes representan las primeras dos Funciones Discriminantes Lineales. Cada círculo representa un clúster y cada punto representa un individuo. Los números representan las diferentes subpoblaciones identificadas por el análisis DAPC; b. Densidad individual y por clusters (3) en la primera función discriminante.

De forma análoga, se realizó un análisis de componentes principales (PCA) para el panel de 568 SNPs, codificados según sus dosajes alélicos (Fig. 2.19). Se observa que el componente principal 1 (PC1, eje horizontal) explicó el 4,5% de la varianza total, mientras que PC2 (eje vertical) permitió explicar el 4,2% de la varianza total. El parental NA 78-724, resaltado con un rectángulo amarillo en la Figura 2.19, se agrupó junto a 10 líneas, que coinciden con el agrupamiento de la Figura 2.18, con la excepción de la línea 150. El otro parental LCP 85-384, marcado con un rectángulo azul (Fig. 2.19), se ubicó más alejado de otras muestras. Un grupo de 6 progenies (94, 98, 140, 145, 153 y 164), identificadas con un rectángulo rojo en la Figura 2.19, se diferenció de los dos parentales y de las progenies restantes que se encuentran más próximas espacio.

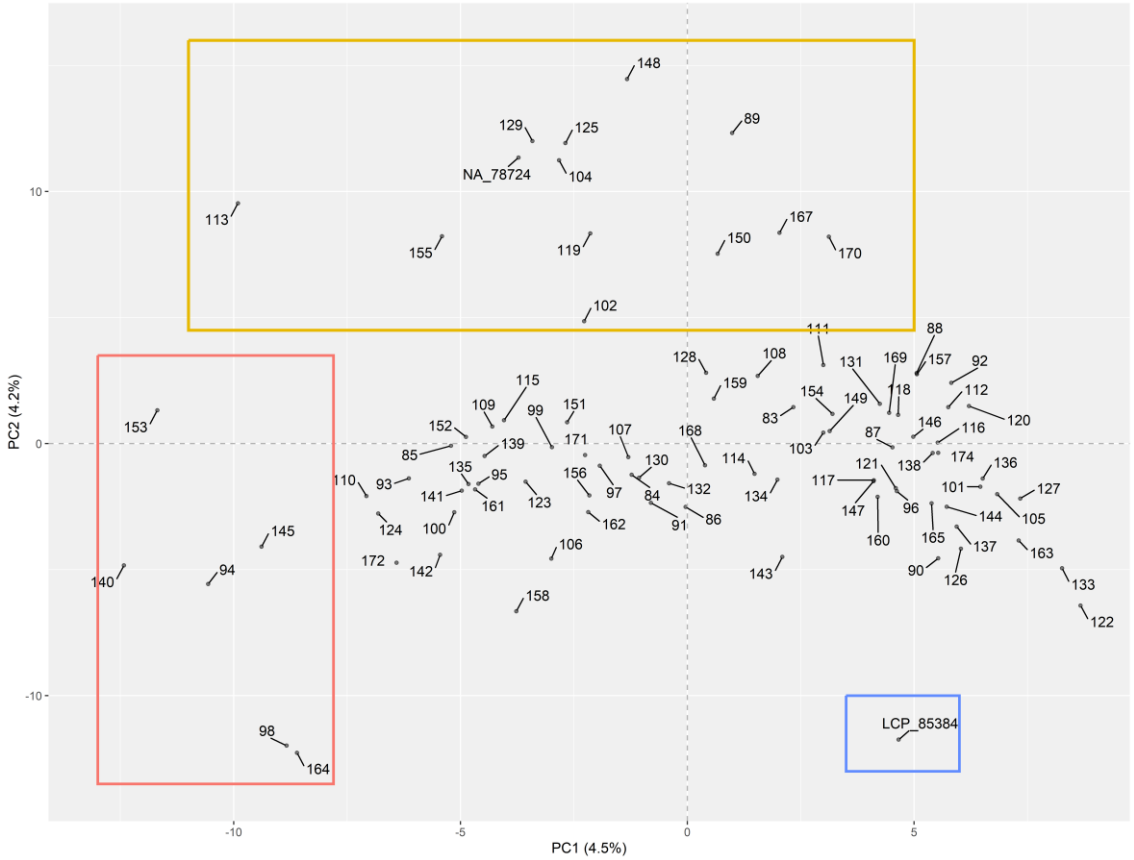

Figura 2.19. *Análisis de componentes principales.* PCA para la población biparental (dos parentales y las 90 progenies) a partir de 568 SNPs, codificados con sus dosajes alélicos. Los ejes representan los primeros dos componentes principales. Cada rectángulo representa un grupo y cada punto representa un individuo.

Discusión

La implementación del protocolo ddRADseq desarrollado para caña de azúcar en esta tesis demostró ser aplicable tanto para pocos individuos como para un escalado en el número de muestras al evaluar una población biparental de 90 individuos. El mismo permitió la obtención de datos genómicos de buena calidad que se pudieron utilizar para el desarrollo de paneles de SNPs y la caracterización de germoplasma complejo como lo es el de la caña de azúcar. Los marcadores moleculares relacionados a loci genéticos pudieron utilizarse para identificar polimorfismos entre los progenitores, y para genotipar a los individuos recombinantes dentro de la población.

Como se discutió en el capítulo anterior, las técnicas de ddRADseq presentan limitaciones inherentes a su metodología e inciden en el número de lecturas obtenidas por sitio, un factor crucial para los algoritmos de llamado de variantes. En este capítulo, se crearon dos bibliotecas, cada una de 48 muestras, las cuales se secuenciaron por separado y a distintos tiempos en el mismo servicio de secuenciación. Para que sea económicamente viable a nivel poblacional, se

redujo la profundidad de secuenciación por individuo a 4 X de cobertura en comparación con la puesta a punto inicial (10 X, capítulo 1), que implicó un número reducido de cinco muestras. Los parentales se incluyeron en todas las corridas de secuenciación para aumentar la profundidad de lectura y cobertura del genoma, además de utilizarse como control de la calidad de la secuenciación por llevarse a cabo en tandas separadas. Un factor limitante a la hora del llamado de alelos de la población fue el genoma de referencia utilizado para el alineamiento de las lecturas, en el que los individuos evaluados compartían entre un 30 y un 50% de similitud, dependiendo del largo de los fragmentos, como se vio en el capítulo 1. No obstante en este trabajo, aún con porcentajes de alineamiento bajos, fue posible encontrar variantes SNPs con las distintas estrategias utilizadas. La variación en el número de lecturas obtenidas dentro de los individuos evaluados podría explicar que algunos genotipos posean profundidades de lecturas medias menores a la media de la población o que tengan varias regiones en el genoma sin información. Este fenómeno se observó previamente en los protocolos de ddRADseq (Bilton et al. 2018; Cerca et al. 2021), y puede atribuirse a la naturaleza de esta técnica de representación reducida del genoma. Cuando las muestras exhiben una alta divergencia genética, pueden surgir datos faltantes, ya que solo se registran lecturas solapadas en un porcentaje menor al 100% de los individuos genotipificados. Dado el tamaño y la complejidad del genoma, es evidente que se requieran mayores esfuerzos de secuenciación para aumentar el número de lecturas por individuo. Esto permitiría incrementar los sitios que comparten información entre las muestras, mejorando así la recuperación de marcadores (Cerca et al. 2021).

El programa Stacks permite realizar el llamado de variantes agrupando las muestras por metapoblaciones, facilitando la creación de paneles con diferentes perspectivas o aproximaciones (Rochette et al. 2019). Al abordar los datos como provenientes de una población biparental, se reduce el número de SNPs, lo que facilita la identificación de las regiones más relevantes para diferenciar la población en cuestión. Además, la construcción de mapas genéticos a partir de la identificación de los alelos provenientes de cada parental simplifica la selección de marcadores con segregación mendeliana. Por su parte, al tratar los datos como una población natural, es decir sin las relaciones de parentesco entre los individuos, los paneles resultantes contienen un mayor número de SNPs, lo que puede ser útil para descubrir variantes que posiblemente se encuentran en baja frecuencia en la población analizada, pero que son relevantes para el programa de mejoramiento (Rochette et al. 2017). En este capítulo, se evaluaron ambas aproximaciones, pero se optó por la primera para continuar explorando la construcción de mapas genéticos, ya que la selección de los marcadores satisface mejor los requisitos establecidos por los algoritmos empleados para el mapeo genético.

La definición de las metapoblaciones, al momento de realizar el llamado de variantes, influye en el número de SNPs final del panel. Las profundidades medias de lectura por variante en el análisis como población biparental (mediana de 10,50 y media de 26,77 lecturas por sitio) fueron menores comparadas con el análisis de la población natural en su conjunto (media de 60,5 y una mediana de 35,9 lecturas por sitio respectivamente). Esto se debe a la manera en que el software busca las variantes: en el caso de la población biparental compara a los dos parentales y luego, en estas regiones, registra lo encontrado en las progenies, o en el caso de la población natural evalúa los loci de todos los individuos sin una priorización (Cerca et al. 2021). Sin embargo, como puede apreciarse en la diferencia entre la media y la mediana de ambos análisis, se observó una gran variabilidad. Según las recomendaciones de Gerard et al. (2018) para el análisis de poliploides, se sugiere utilizar más de 25 lecturas para obtener una fuerte correlación de genotipos verdaderos, considerando también el número de individuos incluidos en el estudio. En el capítulo anterior se tuvieron en cuenta estas recomendaciones por tratarse únicamente de 5 variedades de caña. No obstante, en el análisis de la progenie se disminuyó el umbral a un mínimo de 10 lecturas por sitio, para evitar una condición excesivamente restrictiva que pudiera ocasionar la pérdida de información. Además, se consideró la naturaleza biparental de la población, en la que los genotipos parentales tuvieron un mayor número de lecturas.

En total se identificaron 2066 marcadores para el panel A y 3907 para el panel B. El número de marcadores y su distribución por cromosoma fue similar en los paneles A (biparental) y B (generado al analizar las muestras como una población natural) y, a su vez, estos se correspondieron con los resultados obtenidos en el capítulo 1. Estos resultados son coherentes con el tamaño de los cromosomas y podría ser un indicio de que el número de variantes encontradas se relaciona con las regiones que segregan juntas (bloques de ligamiento). Las regiones distales de los cromosomas presentan una mayor variabilidad genética, mientras que en el centro se encuentra una menor frecuencia de variantes o incluso áreas sin información (regiones en blanco). Este patrón de distribución podría confirmar que las zonas centroméricas son más metiladas que las zonas distales que alojan más regiones codificantes (Garsmeaur et al. 2018; Xue et al. 2022). En un estudio en el que se analizaron los perfiles de metilación de caña de azúcar, aplicando la digestión del ADN con endonucleasas McrBC para enriquecer regiones eucromáticas, observaron que, áreas del genoma con baja metilación coincidían con factores de transcripción y genes relacionados con la sacarosa, tales como WRKY (proteínas que se unen al ADN), bZIP (del inglés: Basic Leucine Zipper), WOX (del inglés: WUSCHEL-related Homeobox), sacarosa-fosfato sintasa y fructosa-1,6-bisfosfatasa (Grativol et al. 2014). Además, observaron que los genes que se superponen con regiones del ADN que presentan diferentes niveles de

metilación, tienden a expresarse de manera diferente entre distintos tejidos (Grativol et al. 2014). Estos resultados proporcionan información valiosa para comprender la estructura genómica de la caña de azúcar.

Después de analizar la segregación alélica, de los 2066 SNPs del panel A, 183 SNPs resultaron adecuados para el mapeo genético. Los marcadores bialélicos codominantes que siguieron una segregación en proporción 1:2:1, es decir, donde ambos progenitores resultaron ser heterocigotas, fueron 142. Además, se encontraron 185 SNPs heterocigotas solo en el parental NA 78-724 y 103 SNPs solo en el parental LCP 85-384. Esto indica que el parental femenino presentó una mayor heterocigosis que el masculino (Fig. 2.9). Esta condición de alta heterocigosidad en las líneas parentales implicó que se descartaran numerosos SNPs debido a la distorsión en la segregación, como se ilustra en la Figura 2.10. Asimismo, el uso del software para mapeo genético diseñado para diploides y aplicada a datos poliploides puede profundizar este filtrado (Mollinari et al. 2019). En comparación con bases de datos de diploides, la proporción de SNPs complejos es significativamente mayor en este contexto poliploide (Taniguti et al. 2022).

El mapeo genético con 183 marcadores permitió generar un mapa marco para todos los grupos de ligamiento de caña de azúcar (Fig. 2.11). Sin embargo, cada uno de estos grupos de ligamientos se presenta como un conjunto de fragmentos anclados al genoma de referencia (Fig. 2.12). Esta característica es común para diferentes mapas genéticos desarrollados por diferentes grupos de trabajo usando distintos tipo de marcadores moleculares (Yan et al. 2018). Considerando esta situación, para estudios posteriores se recomienda intensificar los esfuerzos de secuenciación para aumentar el número de lecturas asignadas a cada muestra de manera de lograr un mayor solapamiento entre regiones genómicas a lo largo de la población, disminuyendo el número de datos faltantes. De esta manera, se llegarán a mapas genéticos más densos, que permitirán unir estos LG pequeños y lograr un número más cercano al de los grupos homólogos (número de cromosoma de referencia correspondiente).

La recomendación de intensificar los esfuerzos de secuenciación cobra mayor relevancia a la hora de generar mapas con programas diseñados para las poblaciones poliploides. El primer paso de esta estrategia es estimar el dosaje alélico a partir de las lecturas asignadas a uno u otro alelo. Aunque se sabe que en la caña de azúcar se presentan aneuploidías, los modelos aplicados en esta tesis permiten obtener información sobre los genotipos de los que se trabaja en los programas de mejoramiento, que suelen ser de ploidías desconocidas. Además, estos modelos permiten identificar los heterocigotas ocultos en los paneles analizados como diploides, ya que de tratarse de SNPs AAAAAAAa o aaaaaaaA, probablemente sean considerados homocigotas

para una u otra variante (De Lara et al. 2019). Estos análisis se podrían complementar con estudios citogenéticos o conteos cromosómicos para ir ajustando los modelos (Gerard et al. 2018). Pereira et al. 2018, utilizó otro software similar (SuperMASSA) para identificar los dosajes alélicos en alfalfa, papa y pasto varilla (*Panicum virgatum*). En los estudios realizados sobre la última especie mencionada, encontraron que la mitad de los marcadores no se podían agrupar o que generaban LGs muy pequeños, lo que coincide con los resultados obtenidos en esta tesis (Pereira et al. 2018). García et al. (2013) realizó experimentos similares a los que se describen en este trabajo, y encontraron que el número de loci variaba dentro de cada clase de ploidía. Se conoce que el número de cromosomas en los grupos homeólogos de la caña de azúcar no es constante, lo que implica una variabilidad genética en la ploidía de esta especie (García et al. 2013).

A pesar de que las metodologías mencionadas dependen mucho de los esfuerzos de secuenciación, ya que se basan en el número de lecturas para cada locus para el cálculo de los dosajes alélicos, son capaces de generar información muy valiosa para comprender el comportamiento de la meiosis en especies con un alto nivel de complejidad (Soares et al. 2021). La posibilidad de la visualización del dosaje para cada marcador (Fig. 2.14), permite tener un análisis detallado de cada locus, y poder seleccionar aquellos que resulten más eficientes para los análisis posteriores, como lo puede ser el desequilibrio de ligamiento. Además, es importante destacar el comportamiento de las progenies respecto a los progenitores varía por locus, y en muchos loci, el dosaje heredado coincide con uno u otro parental, mientras que en otros se dan situaciones intermedias.

El software empleado para la construcción de mapas genéticos en poliploides aborda los pasos i. (estimación de las fracciones de recombinación por pares y pruebas estadísticas asociadas) y iv. (cálculo de las probabilidades de la fase parental y actualización de la fracción de recombinación), que se mencionan en la introducción de este capítulo, especialmente en el caso de autopoliploides con niveles elevados de ploidía (Mollinari et al. 2019). En este contexto, se emplean las fracciones de recombinación derivadas de la verosimilitud obtenida mediante Modelos Ocultos de Markov (HMM), teniendo en cuenta la variabilidad continua en las dosis de alelos posibles. En lugar de asignar únicamente dos dosis posibles, representa la dosis de alelos con una distribución de probabilidad, lo que permite estimar más precisamente las tasas de recombinación (Mollinari et al. 2019). La Fig. 2.15 muestra que los datos analizados como especie diploide asigna a la mayor cantidad de SNPs como heterocigotas, pero aun así ~ 300 SNPs los considera como homocigotas para la referencia en alguno de los parentales, mientras

que cuando se los analiza teniendo en cuenta las dosis alélicas el número de homocigotas baja considerablemente.

El MAPpoly también permite resumir la información de cada marcador, como se observa en el ejemplo de la Fig. 2.16, donde se pueden contemplar los posibles dosajes alélicos de todos los individuos evaluados con sus probabilidades de ocurrencia. Además, se informa la posición del marcador en el genoma y las proporciones de los individuos que contienen uno u otro dosaje alélico y las indeterminaciones. A partir de las frecuencias de recombinación se formaron grupos de marcadores muy distanciados entre sí, por lo que no se pudo graficar en un mismo mapa todos los LGs, pero sí LGs por separado, donde se visualizan los SNPs en los parentales y con colores los dosajes alélicos. Se puede apreciar que hay mucha variabilidad en los LGs evaluados, por lo que el agregado de nuevos marcadores a este mapa marco permitiría la reconstrucción de un mapa más denso. Sin embargo, esta primera aproximación permite tener una idea de la factibilidad de evaluar genomas con tan alta complejidad, utilizando genotipificación por secuenciación.

Para entender mejor la estructura poblacional, se realizaron análisis multivariados a partir de los paneles de marcadores SNPs. El método DAPC permitió dividir la población en grupos definidos, una opción atractiva en comparación con el software STRUCTURE. A diferencia de este último, el método DAPC no exige que las poblaciones estén en equilibrio HW y, además, puede manejar grandes conjuntos de datos sin usar software de procesamiento paralelo, ya que puede implementarse en R. El DAPC no pudo implementarse para el panel como poliploide, ya que no puede manejar loci con ploidías tan elevadas, por lo que se decidió analizar los datos con un análisis de componentes principales (PCA), igual que en el capítulo 1 de esta tesis. En ambos métodos, se agruparon 10 progenies con la madre NA 78-724, mientras que el comportamiento del resto de las líneas fue diferencial. Como se ha mencionado anteriormente, el método DAPC difiere ligeramente de un PCA, ya que obliga a los datos a agruparse en k grupos según el modelo con el menor BIC. La Función Discriminante 1 logró separar al clúster 1 de los clústeres 2 y 3 (Fig. 2.18 a y b). Esta separación sugiere que este primer eje captura las principales diferencias genéticas entre estos grupos. El clúster 1, que incluye 43 genotipos y, aunque un poco más alejado, al parental LCP 85-384, representa una agrupación genética específica que podría estar relacionada con las características heredadas de este parental (círculo azul en Fig. 2.18 a). Por otro lado, la Función Discriminante 2 separó los clústeres 2 y 3, destacando las diferencias genéticas adicionales entre estos grupos. El clúster 2, con 12 individuos incluyendo el parental NA 78-724 (círculo amarillo en Fig. 2.18 a), parece reflejar la influencia genética de este segundo parental. El clúster 3, que agrupa 37 líneas, sugiere la existencia de una subpoblación con

características genéticas diferenciadas que podrían ser de interés para el mejoramiento genético. El Análisis de Componentes Principales (PCA) realizado sobre los 568 SNPs, codificados según sus dosajes alélicos, proporcionó una visión adicional de la estructura genética en la población estudiada (Fig. 2.19). El PC1 explicó el 4,5% de la varianza total, mientras que el PC2 explicó el 4,2%. Estos porcentajes, aunque relativamente bajos, son típicos en estudios genéticos complejos donde la variabilidad está distribuida en muchos componentes. A pesar de esto, los primeros dos componentes permiten observar patrones importantes de agrupamiento y diferenciación genética. Como ocurrió en otros estudios con métodos multivariados, para analizar las relaciones entre individuos mediante marcadores con segregación mendeliana, a menudo no se pueden caracterizar adecuadamente los genotipos poliploides (Medeiros et al. 2020; Pimenta et al. 2021). Esta limitación no está relacionada con la calidad de los marcadores, sino con la complejidad del genoma poliploide. El análisis PCA mostró que el parental NA 78-724, destacado con un rectángulo amarillo en la Figura 2.19, se agrupó junto a 10 líneas. Este agrupamiento coincide con el observado en el DAPC (Fig. 2.18), excepto por la línea 155, lo que refuerza la consistencia de estos genotipos como un grupo genéticamente similar. Por otro lado, el parental LCP 85-384, marcado con un rectángulo azul, se ubicó más alejado de otras muestras (Fig. 2.19). Este hallazgo es consistente con la separación observada en el DAPC. Un grupo de seis progenies (94, 98, 140, 145, 153 y 164), identificadas con un rectángulo rojo, se diferenció claramente de los dos parentales y de las demás progenies (Fig. 2.19). Estos resultados son consistentes con los PCA realizados sobre variables fenotípicas evaluadas en la familia realizada en paralelo a esta investigación en la Tesis Doctoral del Ing. Agr. José María García (*comunicación personal*). En el DAPC, estos individuos se registraron en el grupo 3, entre otras 31 líneas, y separadas de los parentales (Fig. 2.18). Esta diferenciación podría indicar la presencia de combinaciones alélicas únicas o adaptaciones específicas que las distinguen dentro de la población. La identificación de estas progenies destaca la utilidad de los análisis multivariados para revelar subestructuras genéticas que podrían ser importantes para estudios de asociación genética y mejoramiento.

El protocolo de laboratorio utilizado en este estudio permitió escalar en el número de muestras analizadas y la generación de marcadores moleculares del tipo SNPs a nivel poblacional. El uso de enzimas de restricción sensibles a la metilación permitió filtrar áreas metiladas del genoma, favoreciendo así el genotipado de regiones codificantes. En varias etapas del análisis, se remarcó la importancia de duplicar los esfuerzos de secuenciación para aumentar el número de lecturas por región. Además, si las secuenciaciones ocurren en diferentes momentos y se generan bibliotecas genómicas en ensayos independientes cada vez, se

recomienda incluir a los parentales en cada ensayo como control de la secuenciación, para garantizar la consistencia en calidad y cantidad de lecturas, y para aumentar específicamente las lecturas asignadas a estos individuos. Por su parte, el protocolo bioinformático resultó útil para identificar SNPs robustos utilizando el software Stacks, con genoma de referencia e identificando primero a los parentales, para caracterizar las variantes polimórficas y, posteriormente, a las progenies. Si bien se trabajó con un mínimo de 10 lecturas por sitio, la evaluación de 90 individuos facilitó la identificación de SNPs robustos. No obstante, para lograr una mayor recuperación de SNPs se requieren mayores profundidades de lectura por locus, lo que también beneficia la estimación de los dosajes alélicos. Las rutinas utilizadas permitieron la generación de mapas genéticos en híbridos de caña de azúcar a través de dos enfoques distintos. El primero adoptó un enfoque reduccionista, analizando los datos con algoritmos desarrollados para especies diploides y desarrollando un mapa marco al que podrían agregarse más marcadores al grupo de 183 SNPs. El segundo enfoque reflejó la naturaleza poliploide basándose en los dosajes alélicos de los loci SNP. Resulta interesante la evaluación de estos datos con la nueva versión poliploide del genoma de referencia del cultivar R570, aunque no fue evaluada en esta tesis debido a su reciente publicación (Healey et al. 2024). Por último, se evaluó la estructura poblacional, lo que permitió la identificación de agrupamientos consistentes y la diferenciación de ciertos genotipos que podrían indicar la presencia de combinaciones alélicas únicas. Estos hallazgos subrayan la importancia de utilizar múltiples enfoques analíticos para obtener una visión integral de la diversidad genética. Este trabajo desarrollado en paralelo a la tesis doctoral de Ing. Agr. José M García representa una fase inicial del desarrollo de estudios de mapeo de QTL y/o fenotipo-genotipo de caracteres agronómicos de interés en el programa de mejoramiento del cultivo. En este sentido, la evaluación fenotípica de la población de 90 individuos para el mejoramiento de caña de azúcar se realizó en el contexto del trabajo titulado "Caracterización feno-genotípica de familias híbridas de *Saccharum spp*. y clones derivados de cruzamientos con *S. spontaneum* para aumento de producción de fibra, biomasa y azúcar" (Facultad de Agronomía y Zootecnia, Universidad Nacional de Tucumán, aún sin publicar). Asimismo, los métodos establecidos en este capítulo pueden aplicarse al banco de germoplasma para el manejo y preservación de las accesiones, y para la selección de parentales para generar nuevas variedades ya sea por distancia genética combinada o no con datos de asociación fenotipo-genotipo para caracteres de interés. La utilización de marcadores SNP-ddRADseq en el mejoramiento de caña de azúcar es una innovación prometedora para acortar los tiempos del proceso de mejoramiento y obtener nuevas variedades respondiendo a las necesidades del sistema productivo.

Capítulo 3. Análisis de la variabilidad anatómica de los entrenudos en variedades seleccionadas para la acumulación de azúcar o fibra e identificación de genes candidato para características agroindustriales.

Introducción

La caña de azúcar se ha cultivado principalmente por su alto contenido de sacarosa en los tallos, el sustrato principal para producir azúcar y etanol. Durante muchos años, no se diferenciaron las variedades cultivadas para la obtención de estos productos, ya que ambas industrias utilizaban los jugos de fermentación y las melazas de los tallos de caña para la producción de azúcar comestible y biocombustibles de primera generación (Ferreira et al. 2016). Sin embargo, con la aparición de tecnologías de segunda generación para la obtención de bioenergía, es posible aprovechar la biomasa, especialmente la celulosa, que antes se desperdiciaba (Ali et al. 2019). Esto ha llevado a incorporar en los programas de mejoramiento genético materiales con altos contenidos de celulosa en los tejidos del tallo y bajos porcentajes de ligninas y amilopectinas, componentes de las paredes celulares que dificultan el aprovechamiento energético de la biomasa (García et al. 2022; García et al. 2023b). Por esta razón, resulta fundamental estudiar el metabolismo de los polisacáridos y los factores clave en su regulación, así como los reguladores de la biosíntesis de la pared celular, ya que estas vías pueden alterarse según el objetivo del programa de mejoramiento (Souza et al. 2019).

La porción cosechable de la planta de caña de azúcar es su tallo, con una estructura formada por nudos y entrenudos cilíndricos interconectados. En los entrenudos se encuentran varios tejidos esenciales: la epidermis que protege el tallo, el parénquima ubicado en la médula, almacena nutrientes, el colénquima que aporta flexibilidad, y el esclerénquima que proporciona rigidez. Además, el tejido vascular, compuesto por xilema y floema, transporta agua, minerales y productos de la fotosíntesis (Moore et al. 2014). Estos tejidos trabajan juntos para asegurar la estructura, el soporte y el funcionamiento adecuado del tallo. Las variedades de caña de azúcar exhiben divergencias en el tamaño y la configuración de sus haces vasculares (HV), así como en el número y distribución dentro del tejido parenquimático del tallo (Costa et al. 2013). Cada haz vascular comprende el xilema, que consiste en células del protoxilema flanqueadas por dos elementos del metaxilema, así como el floema y las células del parénquima que conectan ambos tejidos. Los HV localizados en la médula del tallo en la caña de azúcar están rodeados por una

vaina de células fibrosas, cuya función es separar el agua del xilema del apoplasto de los tejidos de almacenamiento del tallo (Jacobsen et al. 1992; Walsh et al. 2005). Esta envoltura de fibras presenta paredes celulares lignificadas y/o suberizadas (Moore et al. 2014). En contraste, los HV ubicados en la corteza del tallo cumplen principalmente una función de sostén y no participan activamente en el transporte a larga distancia del floema (Walsh et al. 2005). Las células del esclerénquima no solo circundan los haces vasculares, sino que también establecen conexiones con las capas de la epidermis, desempeñando un papel de anclaje y asegurando la integridad de la estructura vascular (Santiago et al. 2010).

La necesidad de mejorar las características de los cultivos energéticos para eficientizar el uso de la biomasa estimuló el estudio de la biogénesis de la pared celular, así como la eficiencia de la actividad fotosintética y la translocación de los productos fotosintéticos (Chang et al. 2022). En este sentido, se realizaron estudios de genómica comparativa que permitieron la identificación de nuevos genes y regiones asociadas con la pared celular, y en la síntesis de polisacáridos como celulosa y hemicelulosa en especies vinculadas (de Carvalho et al. 2019). Además, para mejorar la calidad de la biomasa y facilitar su utilización como materia prima, se han llevado a cabo simulaciones bioinformáticas entre regiones genómicas de plantas de interés, como *Sorghum bicolor* y *Miscanthus sinensis*, para identificar posibles QTLs vinculados a la composición de la pared celular y características de eficiencia de bioconversión (de Carvalho et al. 2019).

La caracterización a gran escala de los SNPs ofrece una visión integral de los patrones de variaciones tanto dentro de un mismo individuo como de una población, así como sus posibles impactos en las funciones metabólicas (Song et al. 2016). La utilización de estas variaciones en las secuencias de ADN se traduce en herramientas potentes para el genotipado de especies de interés agrícola como la caña de azúcar, facilitando la exploración y explotación de los recursos genéticos para el mejoramiento del cultivo. Actualmente, la disponibilidad de genomas de referencia permite localizar las variantes genéticas identificadas en regiones funcionales específicas proporcionando datos acerca de cómo las mutaciones afectan al proteoma. Así, es habitual utilizar genomas secuenciados y caracterizados para asociar las variantes con funciones específicas, buscando su relación o proximidad a los genes candidato (de Carvalho et al. 2019). El empleo de marcadores SNPs basados en genes que regulan las rutas metabólicas de los polisacáridos, la síntesis de lignina o asociados a genes de resistencia a enfermedades podría ser útil para descubrir nuevas variantes alélicas (Ming et al. 2002). Esto es relevante ya que uno de los principales objetivos de los programas de mejoramiento es incrementar el contenido de

azúcar, mientras que reducir el contenido de lignina puede facilitar la sacarificación de la celulosa para la producción de energía (Perera et al. 2023).

La evaluación genotípica del germoplasma de caña de azúcar como información adicional al fenotipado resulta fundamental para que los programas de mejoramiento puedan utilizar progenitores genéticamente más diversos o con mejores aptitudes para rasgos de interés. Los paneles de marcadores se aplican ampliamente al análisis de la diversidad, mapeo genético, estudios de asociación, selección asistida y selección genómica. Además, estos paneles permiten estudiar al mismo tiempo la diversidad alélica y los haplotipos de los genes o alelos vinculados a ciertos rasgos (Nayak et al. 2014). En la literatura se observa que la presencia de variantes asociadas a genes candidato en los paneles de SNP son relevantes en los estudios de mapeo de asociación para distintos caracteres agroindustriales (Xiong et al. 2023). Los estudios de asociación se han empleado para caracterizar resistencia a enfermedades y rasgos anatómicos, proporcionando información valiosa para la selección asistida y la selección genómica en los programas de mejoramiento de caña de azúcar (Chen et al. 2022; Xiong et al. 2023). Utilizando estos marcadores genéticos vinculados a funciones biológicas, los mejoradores pueden seleccionar eficientemente variedades superiores de caña de azúcar con las características deseadas, mejorando en última instancia la productividad del cultivo y la resistencia a las enfermedades.

En este Capítulo, se caracterizó la variabilidad en la anatomía en entrenudos del cultivar comercial (LCP 85-384) y del biotipo de caña energía (INTA 05-3116), contrastantes fenotípicamente a campo, complementando los datos de variabilidad genotípica obtenidos en capítulos anteriores. Se utilizó un enfoque bidimensional que involucró etapas críticas del ciclo de cultivo (patrones temporales) y diferentes posiciones de entrenudos a lo largo del tallo (patrones espaciales). La descripción anatómica incluyó 1- análisis histoquímico de la deposición de lignina 2- descripción cuantitativa de los elementos de los haces vasculares, 3- relación entre el número de haces vasculares con el área del entrenudo. Paralelamente, se predijo el efecto de las variantes alélicas identificadas en los capítulos anteriores de este trabajo de Tesis para explorar regiones funcionales a partir de la exploración de genes candidato para el metabolismo de los polisacáridos, de la estructura de la pared celular o implicados en la respuesta al estrés abiótico.

Materiales y métodos

Material vegetal y diseño experimental

En este experimento se utilizaron dos híbridos de caña de azúcar con fenotipos contrastantes: LCP 85-384, un cultivar comercial con contenido alto en sacarosa y bajo en fibra, e INTA 05-3116, un biotipo de caña energía y con contenido bajo en sacarosa, analizados también en el capítulo 1. Tallos de caña de azúcar de soca 2 (rebrote después de la segunda cosecha) fueron recolectados al azar en el campo experimental de la Estación Experimental INTA Famaillá (27°03'S, 65°25'O, 363 msnm, provincia de Tucumán, Argentina) de plantas cultivadas en parcelas de una sola hilera de 6 m de largo y un espaciamiento entre hileras de 1,6 m. Se cortaron cinco tallos de cada genotipo a nivel del suelo en cada una de las siguientes etapas de crecimiento y desarrollo: macollaje (diciembre), gran crecimiento (febrero), maduración temprana (mayo) y maduración tardía (agosto). Las etapas se eligieron por ser críticas para determinar el rendimiento agronómico a lo largo del ciclo de cultivo de la caña de azúcar (Moore y Frederick, 2014). El número de entrenudos muestreados en las diferentes posiciones de los entrenudos (IP) aumentó con el desarrollo de la planta, por lo que la IP 1 se recogió en el macollaje, las IP 1 y 5 en el gran crecimiento, las IP 1, 5, 10 y 15 en la maduración temprana y las IP 1, 5, 10, 15 y 20 en la maduración tardía. El IP se definió como el número de entrenudos contados desde el nivel del suelo; por lo tanto, los entrenudos conservan su posición a lo largo del desarrollo de la planta, lo que permite una interpretación más fácil de los cambios en la composición de los entrenudos. Además, se contó el número de entrenudos por debajo de la primera hoja con cuello visible (TVD) en cada fase para tener en cuenta la edad fisiológica del entrenudo en la interpretación de los datos. Los entrenudos de dos tallos recogidos (réplicas independientes) se utilizaron para este análisis.

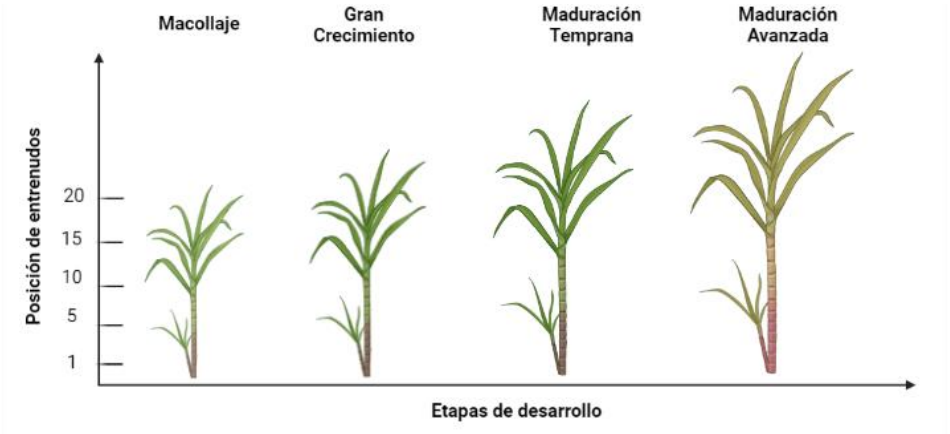

Figura 3.1. *Ilustración esquemática del diseño experimental utilizado para la caracterización de los entrenudos de la caña de azúcar.* La composición de los entrenudos se estudió en un perfil temporal que incluye cuatro etapas críticas de desarrollo en el ciclo de crecimiento de la planta. En cada etapa de desarrollo, se evaluó un perfil espacial que incluye varias posiciones de entrenudos (IP) desde el nivel del suelo. Figura creada con BioRender.com

Anatomía de entrenudos de caña de azúcar y biotipo de caña energía

Para la determinación histoquímica de la lignina, se cortaron manualmente secciones de tallo de aproximadamente un centímetro en el punto medio del entrenudo. Las secciones se fijaron en una solución FAA (formol-alcohol-ácido acético) durante 48 h y se conservaron en alcohol al 70%. Se realizaron cortes a mano con bisturí y se tiñeron con floroglucinol-HCl durante 5 a 10 min (Tao et al. 2009). Las secciones teñidas se enjuagaron en agua destilada, se montaron en un portaobjetos y se examinaron mediante campo brillante en un microscopio óptico de transmisión. Las muestras se observaron con un objetivo de 20X. Las figuras se capturaron con un tamaño de imagen de 1704 x 2272 píxeles. Se determinó la longitud y el ancho de los haces vasculares y el diámetro del metaxilema. Se seleccionó un mínimo de 20 haces vasculares para realizar las mediciones, para cada una de las dos réplicas seleccionadas por genotipo.

Análisis estadístico

El análisis de datos de la variabilidad anatómica en híbridos de caña con fenotipos contrastantes se realizó utilizando el software R v 2.5.2 (R Core team, 2018). Se utilizó el análisis de varianza (ANOVA) con un enfoque de Modelos Lineales Generalizados Mixtos (GLMM) (Garibaldi et al. 2016) para analizar los efectos del genotipo, el IP y la etapa de desarrollo en los componentes anatómicos de los entrenudos de la caña de azúcar. Debido a restricciones biológicas (no se disponía del conjunto completo de entrenudos en todos los estadios de desarrollo) fueron necesarios dos análisis estadísticos independientes para analizar los efectos del IP y del estadio de desarrollo. El primer análisis incluyó los efectos del genotipo, la IP y la interacción genotipo:IP y se dividió por estadios de desarrollo (perfil temporal). El segundo análisis incluyó los efectos del genotipo, el estadio de desarrollo y la interacción genotipo:estadio de desarrollo y se dividió por IP (perfil espacial). Se incorporó un efecto aleatorio de múltiples mediciones para los haces vasculares. Se utilizaron los índices del criterio de información de Akaike (AIC) y del criterio de información bayesiano (BIC) para seleccionar el modelo de mejor ajuste para cada análisis estadístico, y se utilizó la prueba Tuckey como prueba de comparación de medias post-hoc. Se utilizó la función VarIdent (paquete nlme) para modelizar la heteroscedasticidad (Garibaldi et al. 2016).

Predicción del efecto funcional de los SNPs y anotación génica

Se generó una lista de 261 posiciones genómicas correspondientes a la población en estudio (en formato VCF). Esta lista surgió de integrar los datos de los paneles de SNPs de los parentales obtenidos en el capítulo 1 (No-150) y de la progenie obtenidos en el capítulo 2. Luego, se utilizó el sistema de predicción de efectos de las variantes (VEP) de Ensembl Plants (versión 51) para predecir el efecto funcional y asignar una función molecular inferida a partir de la

notación del genoma de referencia (en formatos FASTA y gff3). Para identificar genes ubicados río arriba o río abajo de una variante determinada se utilizó el parámetro por defecto de 5000 bases (5 kb).

De las 261 posiciones analizadas, se seleccionaron los SNPs ubicados en las regiones 3' y 5' UTR, en sitios de splicing alternativo y en exones. A su vez, se descartaron aquellos SNPs con anotación funcional de "Proteína hipotética conservada", debido a la falta de información sobre su implicancia en el metabolismo, lo que resultó en la eliminación de 8 loci. Posteriormente, se seleccionaron cuatro SNPs por estar vinculados con posibles genes candidato para la selección de materiales de caña, afines al metabolismo de los polisacáridos, de la estructura de la pared celular o implicados en la respuesta al estrés abiótico. Las tablas y figuras se realizaron utilizando los paquetes readxl (Wickham et al. 2023), writexl (Ooms 2023), tidyverse (Wickham et al. 2019) y ggplot2 (Wickham 2016) del entorno R.

Validación de los SNPs

Para la validación, se realizó una búsqueda bibliográfica de estudios que identificaron variantes SNPs en caña de azúcar utilizando el mismo genoma de referencia empleado en esta tesis (Garsmeaur et al. 2018). Luego, se compararon las posiciones genómicas y se verificaron los alelos para confirmar que se trataba del mismo SNP.

Resultados

Anatomía de entrenudos de caña de azúcar y del biotipo de caña energía

Explorando técnicas complementarias a los paneles de SNPs, se realizó un análisis histológico de entrenudos de tallos, para evaluar la variabilidad en la anatomía de estos tejidos en cañas contrastantes fenotípicamente a campo, utilizando distintos patrones temporales y espaciales. En las secciones transversales de los entrenudos de caña de azúcar evaluadas, se reconocieron dos regiones: la médula central y la corteza, en la periferia del tallo, adyacente a la epidermis (Fig. 3.2). En la médula, los haces vasculares (HV) estaban en baja densidad y dispersos con una distribución organizada, separados por células de parénquima de almacenamiento (Fig. 3.2 c-h). Por el contrario, los haces vasculares se encontraban en mayor concentración y apretados cerca de la corteza, conformando parte del tejido de sostén del tallo (Fig. 3.2 a, b). La tinción de color rojo correspondiente a la detección de lignina con el reactivo floroglucinol ácido permitió visualizar su acumulación en el esclerénquima de los HV de ambos híbridos. La forma de la vaina del esclerénquima fue diferente en ambos genotipos, siendo romboidal en INTA 05-3116 y oval en LCP 85-384 (Fig. 3.2 e, h).

En términos generales, la longitud de los HV exhibió similitudes a lo largo del ciclo de crecimiento y desarrollo de la planta. No obstante, se destacaron HV notablemente más extensos en el caso de IP1 durante la fase de gran crecimiento para el cultivar LCP 85-384, y durante la etapa de maduración temprana y tardía para el INTA 05-3116 (Fig. 3.3 a, Apéndice 2 Tabla 3.1). En IP1 se observó una tendencia hacia mayores longitudes de HV en INTA 05-3116 en comparación con LCP 85-384 en tres etapas de desarrollo, a excepción de la fase de gran crecimiento (Fig. 3.3 a, Apéndice 3 Tabla 3.1). Se identificaron diferencias estadísticamente significativas durante la maduración temprana en la interacción genotipo:IP. En esta etapa, se observó que en el IP1 en INTA 05-3116 (536,8 µm), los HV presentaron una longitud de 117 µm mayor en comparación con los HV del mismo entrenudo en LCP 85-384 (419,3 µm). Estas discrepancias se intensificaron al comparar la longitud de los HV en el IP1 del INTA 05-3116 con los IPs superiores del genotipo LCP 85-384. Durante la maduración tardía, el IP1 de INTA 05-3116 (474,6 µm) mostró una longitud promedio de los HV de 93 µm mayor que el mismo entrenudo del LCP 85-384 (381,9 µm) (Fig. 3.3 a, Apéndice 3 Tabla 3.1).

Los HV resultaron más anchos en LCP 85-384 que los de INTA 05-3116, destacándose tendencias desde la fase de gran crecimiento, aunque la interacción genotipo:IP solo resultó estadísticamente significativa en la maduración tardía (Fig. 3.3b, Apéndice 1 Tabla 3.2). En esta última etapa, el IP10 (405,4 µm) en LCP 85-384 mostró HV de 58,1 y 55,7 µm más anchos que los respectivos anchos de los IPs 10 y 15 de INTA 05-3116 (347,3 y 349,7 µm, respectivamente) (Apéndice 3 Tabla 3.2). Analizando el IP15 (420,2 µm) en LCP 85-384 en maduración tardía mostró que los HV resultaron un 72,9, 70,5 y 77,7 µm más anchos que los IPs 10 (347,3 µm), 15 (349,7 µm) y 20 (342,5 µm) en INTA 05-3116 (Fig. 3.3 b, Apéndice 3 Tabla 3.1). En esta dimensión analizada, se identificaron variaciones propias para cada genotipo. La tendencia observada en LCP 85-384 fue en aumento progresivo desde la base hacia el ápice del tallo. En maduración temprana, por ejemplo, los HV pasaron de 374,4 µm (IP1) a 390,2 µm (IP15) (Fig. 3.3b, Apéndice 3 Tabla 3.1). En cambio, INTA 05-3116 mostró una tendencia opuesta, disminuyendo de abajo hacia arriba, como se evidenció en la maduración temprana, donde los HV descendieron de 362,9 µm (IP1) a 338,4 µm (IP15) (Fig. 3.3 b, Apéndice 3 Tabla 3.1). Esta diferencia en la tendencia se mantuvo consistentemente incluso en maduración tardía, donde los anchos de HV en LCP 85-384 (364,0 µm en IP1 a 402,9 µm en IP20) contrastaron con los de INTA 05-3116 (362,1 µm en IP1 a 342,5 µm en IP20).

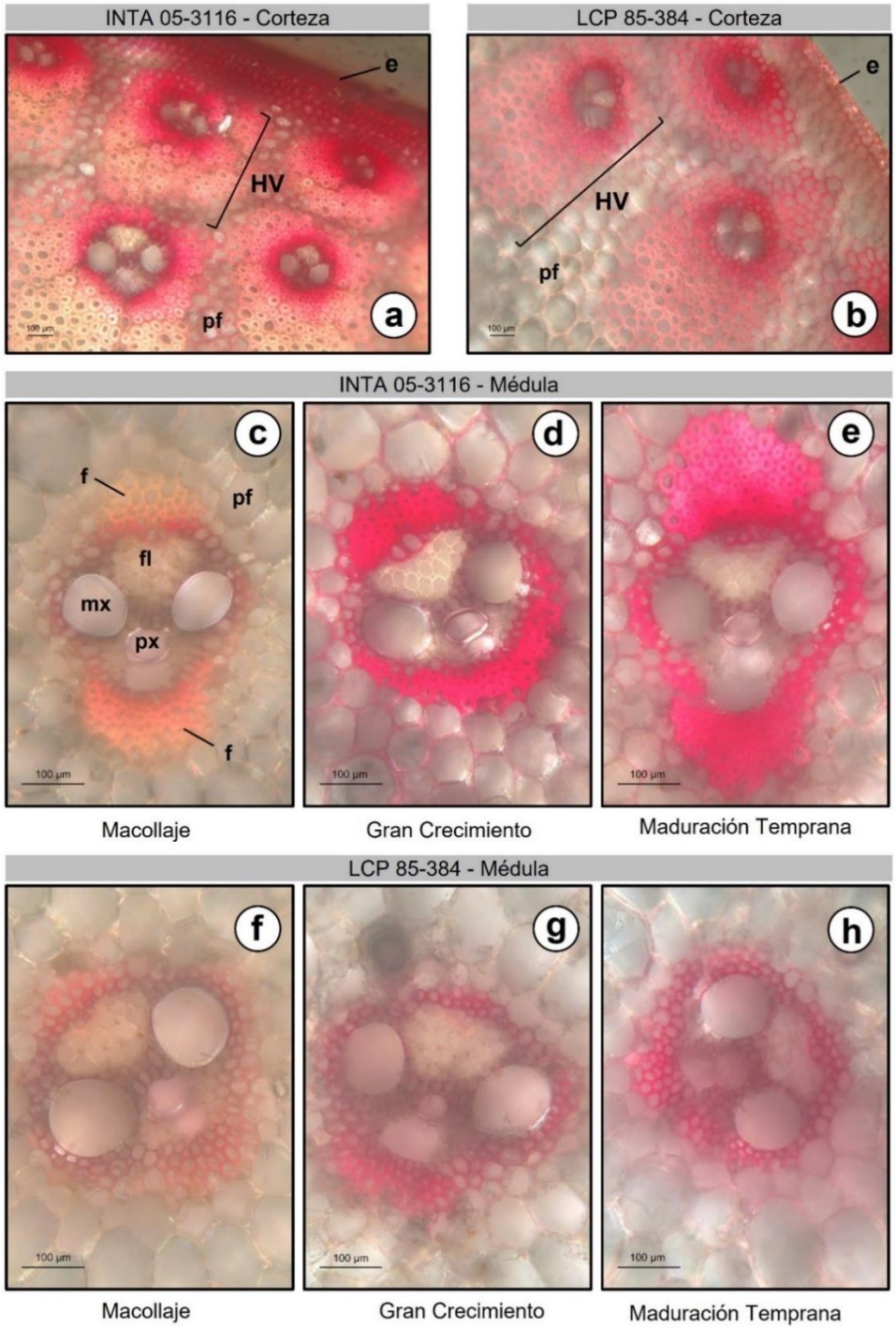

Figura 3.2. *Sección transversal de entrenudos de los cultivares LCP 85-384 (cultivar comercial) e INTA 05-3116 (biotipo caña energía) teñidas con floroglucinol ácido (rojo), a lo largo de tres estadios de desarrollo.* Imágenes representativas de la región de la corteza de entrenudos de INTA 05-3116 (a) y LCP 85-384 (b); región de la médula en tres estadios de desarrollo en IP1 en INTA 05-3116 (c, d, e) y en LCP 85-384 (f, g, h). HV= haz vascular, mx= metaxilema, px= protoxilema, fl= floema, pf= parénquima fundamental, e= epidermis, f= fibra. Barra de escala 100 µm.

Aunque los valores del diámetro del metaxilema de los IPs para cada genotipo fueron estables durante las etapas de macollaje y maduración temprana, LCP 85-384 presentó una tendencia a valores más altos que en el biotipo de caña energía en las cuatro etapas de

desarrollo analizadas. En maduración tardía, los IP superiores de LCP 85-384 mostraron una tendencia a valores más altos que en los entrenudos inferiores, pasando de 121,5 µm en IP20 a 109,4 µm en IP1; en contraste con los valores de los IP de INTA 05-3116, que fueron similares a lo largo del mismo perfil espacial (Apéndice 3 Tabla 3.1). El número de HV para cada genotipo fue estable a lo largo del perfil espacial dentro de cada fase de crecimiento. Excepto en macollaje, se observó un mayor número de HV en LCP 85-384 que en INTA 05-3116 en las restantes etapas de desarrollo de la planta (Fig. 3.3 c, Apéndice 3 Tabla 3.1).

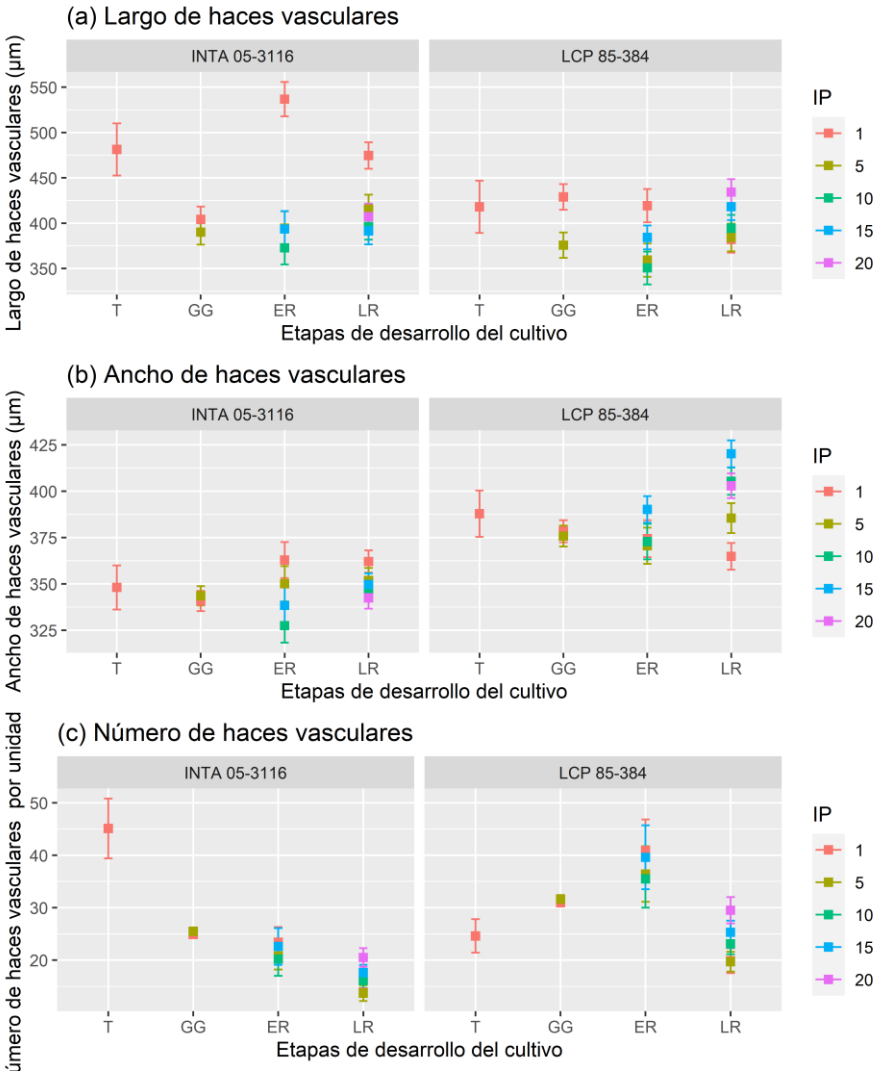

Figura 3.3. Medias de los efectos principales de las medidas anatómicas de los entrenudos de LCP 85-384 (cultivar comercial) e INTA 05-3116 (biotipo de caña energía) en cuatro estadios de desarrollo. Largo y ancho de los haces vasculares en micrómetro (µm) y número de haces por unidad de medición. T= Macollaje; GG= Gran crecimiento; ER=Maduración temprana; LR=Maduración tardía; IP= entrenudo

Predicción del efecto funcional de los SNPs

Para predecir las posibles implicancias funcionales de los SNPs, se recurrió al sistema de predicción de efectos de las variantes (VEP) de Ensembl Plants (Fig. 3.4). Este análisis se aplicó a

una lista de 261 posiciones genómicas correspondientes a SNPs informativos de la población en estudio. Se observó un patrón de localización similar al explorado en el capítulo 1 con respecto al porcentaje de variantes identificadas por cromosoma para el efecto de los SNPs. Las variantes genéticas ubicadas río arriba de los genes representaron el 23,4% del total (61 SNPs), mientras que las variantes río abajo de los genes constituyeron el 28,0% (73 SNPs), conformando más de la mitad de las variantes detectadas, con una suma del 51,3% (Fig. 3.4). Además, se encontraron 21 variantes intergénicas (8,1%) y 60 SNPs intrónicos (23,0%) (Fig. 3.4). Se identificaron 20 variantes con cambio de sentido y 16 sinónimas. Por su parte, 3 SNPs se localizaron en la región de *splicing* alternativo. Finalmente, se hallaron 6 variantes en la región 3' UTR y 1 en la 5' UTR (Fig. 3.4).

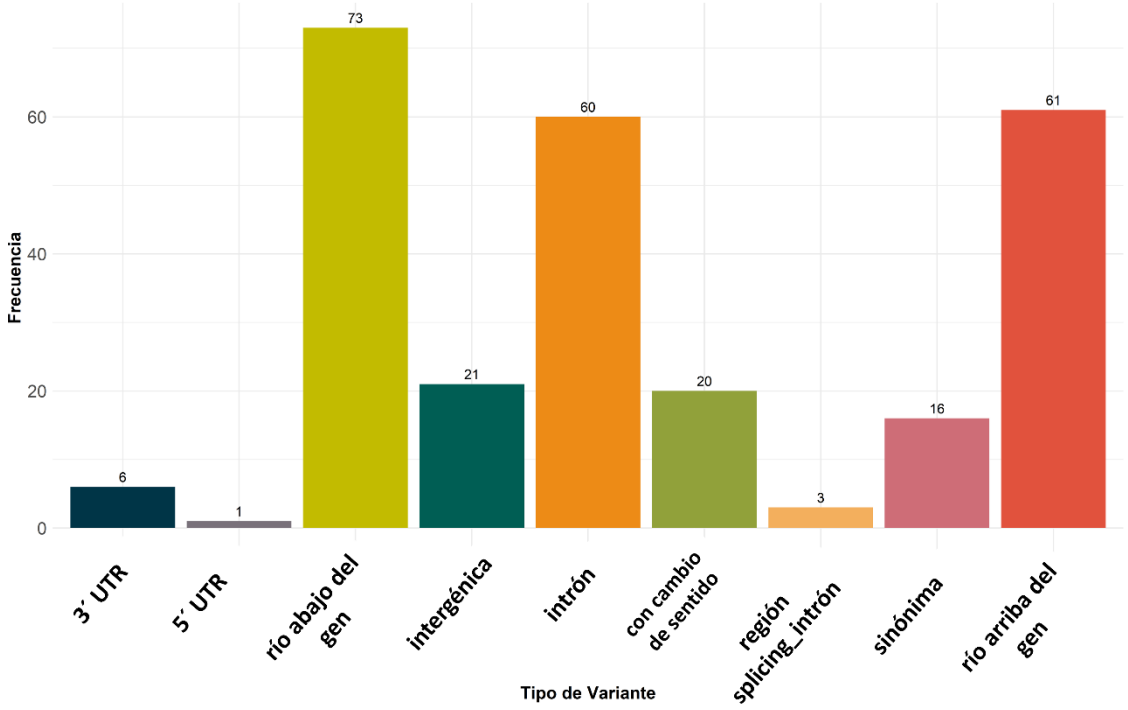

Figura 3.4. *Análisis funcional in silico de SNPs.* Tipos de variantes encontradas para el panel de 261 SNPs y su frecuencia de aparición.

Anotación génica de los 261 loci

Se realizó la anotación funcional por inferencia usando la base de datos Ensembl Plants para comprender mejor el impacto biológico de las variantes genéticas identificadas. A partir de los 261 loci analizados, se identificaron 28 SNPs ubicados en regiones codificantes, tales como regiones 3' y 5' UTR, *splicing* alternativo y exones como posibles candidato para estudios de asociación (Tabla 3.1; Apéndice 3 Tabla 3.3). La Tabla 3.1 detalla la información correspondiente a cada SNP, incluyendo su posición en relación con el genoma de referencia (cv R570), alelos, tipo de variante y los genes a los que están asociados. Además, se buscó vincular estos genes a las distintas vías o ciclos del metabolismo. De estos SNPs, 8 están asociados a proteínas de membrana (de los cuales 5 pertenecen al mismo gen), 3 están relacionados con el funcionamiento celular, 5 están vinculados al crecimiento y desarrollo, 2 están asociados al estrés, 1 combina las dos categorías anteriores. Por otro lado, se registró 1 SNP relacionado a cada una de las siguientes rutas o ciclos metabólicos: biosíntesis de lignina, biosíntesis de purinas, flujo citoplasmático, regulaciones transcripcionales, ácidos grasos, estructuras de las membranas y metabolismo secundario. Asimismo, se identificaron 2 variantes asociadas a glicosiltransferasas.

De los 60 SNPs localizados dentro de intrones, se identificaron 7 variantes como posibles candidato para genes implicados en la síntesis de polisacáridos (SNP 8, ID: 22325_137; SNP 21, ID: 2070_95; SNP 22, ID: 2070_93; SNP 23, ID: 2070_43; vinculados con hexosiltransferasas) y la modificación de las paredes celulares (SNP 234; ID: 122049_17; vinculado a la proteína de reducción de la acetilación de la pared celular). Además, se encontraron dos intrones en factores de transcripción de genes relacionados con la síntesis de sacarosa, como el factor de transcripción para BZIP (SNP 42, ID: 15715_29) y para WRKY (SNP 177, ID: 96556_143). Los detalles de estos sitios polimórficos se encuentran en el Apéndice 3, Tabla 3.3.

Tabla 3.1. *Análisis funcional in silico de SNPs.* Descripción de 28 variantes SNPs seleccionadas a modo de ejemplo y su asociación al metabolismo vegetal. SNP_ID: nombre del SNP; Gen ID: nombre del gen asociado al SNP; SNP POS: posición del SNP en pares de bases, según el genoma de referencia de la caña de azúcar del cv R570; SNP: variante de referencia/variante alternativa; Tipo de variante: implicancia de la variante; Anotación funcional - Gen: nombre de la proteína, enzima o factor de transcripción asociado al SNP; Metabolismo: ruta o ciclo al que pertenece el gen del SNP asociado.

N°	SNP_ID	Cromosoma	Gen_ID	SNP POS	SNP	Tipo de variante	Anotación funcional - Gen	Metabolismo
40	14275_129	Sh01	Sh01_g028810	48784756	A /G	sinónima	Proteína TRANSMISIÓN DE POLEN ABERRANTE 1	Crecimiento y desarrollo
51	17545_60	Sh01	Sh01_g034620	57483546	C /T	con cambio de sentido	Proteína de unión a la miosina 1	Flujo citoplasmático
53	20736_42	Sh01	Sh01_g039470	64985402	C /T	con cambio de sentido	Factor de transcripción sensible al etileno ERF073	Respuesta a estrés
56	21119_126	Sh01	Sh01_g040050	65865714	C /T	5_prima_UTR	Inhibidor de la proteinasa de cisteína	Crecimiento y desarrollo
59	22095_107	Sh01	Sh01_g041730	68144573	A /G	con cambio de sentido	Proteína Dirigente	Biosíntesis de lignina
66	31133_133	Sh02	Sh02_g015170	28854584	C /T	sinónima	Fosforibosil aminoimidazol-succino carboxamida sintasa	Biosíntesis de purinas
71	31721_90	Sh02	Sh02_g016110	30678593	G /C	con cambio de sentido	Proteína vegetal isoprenilada 41 asociada a metales pesados	Crecimiento, desarrollo, respuesta a estrés
72	34111_44	Sh02	Sh02_g020250	36503815	C /T	3_prima_UTR	Proteínas de la familia SR proteína-cinasa específica II (SRPK4)	Crecimiento y desarrollo
73	34111_135	Sh02	Sh02_g020250	36503906	A /G	3_prima_UTR	Proteínas de la familia II de la proteína cinasa específica del SR	Crecimiento y desarrollo
77	38037_118	Sh02	Sh02_g026950	45863362	T /C	con cambio de sentido	Proteína de unión a ARN	Regulación postranscripcional
99	49187_47	Sh03	Sh03_g016700	29709876	T /G	sinónima	Acil-CoA sintetasa de cadena larga	Metabolismo Ácidos grasos
105	52535_52	Sh03	Sh03_g022050	38061066	C /T	sinónima	Proteína con dominio BSD	Crecimiento y desarrollo
108	54307_136	Sh03	Sh03_g025100	42405274	T /G	sinónima	Proteína 2 similar a TOM1	Proteinas de membranas
131	69171_28	Sh04	Sh04_g019850	34011701	G /A	sinónima	Plástido (P)ppGpp sintasa	Respuesta a estrés
148	77788_76	Sh05	Sh05_g008220	16430690	G /A	sinónima	Proteína VIP3 que contiene repeticiones WD	Funcionamiento celular
149	77979_119	Sh05	Sh05_g008530	16967967	T /G	con cambio de sentido	similar a la proteína Os11g0587100 (preniltransferasa)	Estructura de la membrana celular
153	84418_84	Sh06	Sh06_g007200	13860027	G /A	sinónima	similar a la proteína H0305E08.1 (RRM domain-containing protein)	Regulación postranscripcional
154	85290_8	Sh06	Sh06_g008680	16056817	A /G	sinónima	Alfa-bisabolol sintasa	Metabolismo secundario
193	101832_105	Sh07	Sh07_g017160	28468845	T /A	sinónima	Beta C2-xilosiltransferasa	Glicosiltransferasas
213	123364_8	Sh09	Sh09_g003130	5126406	G /A	sinónima	Transportador ABC de la familia B, miembro 21	Proteinas de membranas
244	127495_112	Sh10	Sh10_g009950	16910628	C /A	región splicing_intrón	Arabinosiltransferasa RRA1	Glicosiltransferasas
249	128509_49	Sh10	Sh10_g011180	20298999	C /T	sinónima	Proteasas de membrana dependientes de metales	Proteinas de membranas
250	128509_117	Sh10	Sh10_g011180	20299067	C /T	3_prima_UTR	Proteasas de membrana dependientes de metales	Proteinas de membranas
251	128509_130	Sh10	Sh10_g011180	20299080	C /T	3_prima_UTR	Proteasas de membrana dependientes de metales	Proteinas de membranas
252	128511_113	Sh10	Sh10_g011180	20299159	G /A	3_prima_UTR	Proteasas de membrana dependientes de metales	Proteinas de membranas
253	128511_49	Sh10	Sh10_g011180	20299223	T /A	3_prima_UTR	Proteasas de membrana dependientes de metales	Proteinas de membranas
257	133685_54	Sh10	Sh10_g018550	33613438	C /T	con cambio de sentido	Quinasa de serina/treonina ATR	Funcionamiento celular
260	133685_56	Sh10	Sh10_g018550	33613440	A /G	con cambio de sentido	Quinasa de serina/treonina ATR	Funcionamiento celular

La Tabla 3.2 proporciona una visión detallada de cómo varían los genotipos en diferentes muestras para cada uno de los cuatro SNPs seleccionados. Estos SNPs se distribuyen en distintos cromosomas y están asociados con genes específicos, lo que permite identificar variaciones genéticas potencialmente útiles para la identificación de variedades y la selección asistida en programas de mejoramiento genético. La información sobre los genotipos, incluyendo los dosajes alélicos en muestras analizadas, proporciona datos valiosos para entender la variabilidad genética y su aplicación práctica.

- El **SNP 59, ID: 22095_107** (cromosoma 1; A/G) presenta una mutación con cambio de sentido en el gen de las proteínas dirigentes, que son un grupo de proteínas que juegan un papel crucial en la biosíntesis de lignina y en la formación de compuestos fenólicos estructurales (Li et al. 2022). En la versión simplificada, simulando la diploidía, los genotipos INTA 05-3116, NA 78-724, LCP 85-384 y NA 56-30, evaluados en el capítulo 1, se registraron como heterocigotas, mientras que el HOCP 92-665 resultó homocigota AA. En los modelos en los que se evaluó el dosaje alélico, el parental LCP 85-384 también se registró como heterocigota, pero con distintas dosis alélicas para cada variante, con 5 copias del alelo de referencia y 3 del alelo alternativo. En contraste, el parental NA 78-724 resultó ser homocigota GG, es decir, para el alelo alternativo (Tabla 3.2).

- El **SNP 77, ID: 38037_118** (cromosoma 2; T/C) presenta una mutación con cambio de sentido en el gen de proteínas de unión al ARN en plantas, clave en la regulación postranscripcional de la expresión génica. Estas proteínas se unen al ARN mensajero (ARNm) y otros tipos de ARN para influir en su procesamiento, transporte, estabilidad, y traducción (Gururani 2023). Este caso es interesante porque en el biotipo INTA 05-3116 se encontró la variante alternativa en homocigosis, mientras que en el LCP 85-384 y NA 78-724 resultaron homocigotas para la referencia en la versión simplificada. Sin embargo, al evaluar a los parentales con sus dosajes alélicos el LCP 85-384 resultó heterocigota (con un genotipo de 6/2). Además, en las progenies 135, 153 y 172 se vio que sus genotipos son iguales al del INTA 05-3116, es decir, 0/8 (variante alternativa en homocigosis), combinación que no se encontró en los parentales (Tabla 3.2).

- El **SNP 193, ID: 101832_105** (cromosoma 7; T/A) presenta una mutación sinónima en el gen de la Beta C2-xilosiltransferasa, esencial para la síntesis y modificación de polisacáridos en la pared celular de plantas, proporcionando resistencia y flexibilidad, lo cual permite su

rápido crecimiento y adaptación a condiciones ambientales adversas (Takano et al. 2015; Hrmova et al. 2022). Para este SNP, se registró un genotipo homocigota TT, tanto en el INTA 05-3116 como NA 56-30, mientras que para las demás accesiones fueron heterocigotas (Tabla 3.2).

- El **SNP 244, ID: 127495_112** (cromosoma 10; C/A) presenta una mutación la región de *splicing* en el gen de la arabinosiltransferasa RRA1, que es una enzima glicosiltransferasa, es crucial en la incorporación de arabinosa a los polisacáridos de la pared celular, contribuyendo a su robustez y flexibilidad para soportar el crecimiento y el estrés mecánico, así como a la adaptación a condiciones ambientales adversas (Zhang et al. 2023). Para este SNP, se registró un genotipo heterocigota en INTA 05-3116, NA 56-30, LCP 85-384, NA 78-724, tanto en la versión simulada como diploide como al evaluar los dosajes alélicos. En contraste, el genotipo de HOCP 92-665 fue homocigota. Al evaluar los parentales con dosajes alélicos, NA 78-724 mostró 5 copias del alelo C y 3 del alelo A, mientras que en LCP 85-384 se observaron 4 copias de cada alelo (Tabla 3.2).

Tabla 3.2. *Asignación de genotipos para cada SNP.* Se seleccionaron cuatro SNPs: 22095_107, 38037_118, 101832_105 y 127495_112 para ejemplificar la información que se puede extraer de cada variante. A su vez, se informa el genotipo de las muestras analizadas. SNP_ID: nombre del SNP; Gen ID: nombre del gen asociado al SNP; NA 78-724 diplo: panel No-150 del capítulo 1, analizado como diploide; LCP 85-384 diplo: panel No-150 del capítulo 1, analizado como diploide; NA 78-724 ploi8: panel de SNPs con los dosajes alélicos asignados del capítulo 2; LCP 85-384 78-724 ploi8: panel de SNPs con los dosajes alélicos asignados del capítulo 2.

N°	SNP_ID	Cromosoma	Gen ID	SNP	INTA 05-3116	NA 78-724 diplo	LCP 85-384 diplo	HOCP 92-665	NA 56-30	NA 78-724 ploi8	LCP 85-38 ploi8
59	22095_107	Sh01	Sh01_g041730	A/G	0/1	0/1	0/1	0/0	0/1	0/8	5/3
77	38037_118	Sh02	Sh02_g026950	T/C	1/1	0/0	0/0	0/1	0/1	8/0	6/2
193	101832_105	Sh07	Sh07_g017160	T/A	0/0	0/1	0/1	0/1	0/0	5/3	5/3
244	127495_112	Sh10	Sh10_g009950	C/A	0/1	0/1	0/1	0/0	0/1	5/3	4/4

Evaluación de la heterocigosis

La caña de azúcar es un cultivo altamente heterocigótico. En comparación con las gramíneas diploides, donde cada gen tiene una sola copia, en la caña de azúcar estos genes suelen tener de 2 a 6 copias, pudiendo llegar hasta 15 (Souza et al. 2019). Por esta razón, se evaluó la proporción de alelos diferentes en los SNPs identificados. A partir de la información de las 261 posiciones analizadas, se realizó el recuento de los SNPs en los que se obtuvieron individuos heterocigotas y homocigotas, pero para facilitar la visualización, solamente se muestran los datos de los 28 SNPs seleccionados en la Tabla 3.1 (Tabla 3.3 y Tabla 3.4). Del total de las variantes, 72 SNPs resultaron heterocigotas en todos los individuos analizados, y 216

variantes fueron heterocigotas en 90 o más individuos. Estos extremos podrían ser de interés particular para estudios de asociación genética. Uno de los pocos casos en los que se encontraron menos genotipos heterocigotas fue en el SNP 73, ID: 34111_135. En este caso particular, se observaron 43 en heterocigosis y 53 en homocigosis.

Tabla 3.3. *Heterocigosis en caña de azúcar*. Número de individuos homocigotas y heterocigotas para cada SNP. SNP_ID: nombre del SNP.

N°	SNP_ID	Cromosoma	SNP	Heterocigotas	Homocigotas
40	14275_129	Sh01	A /G	97	0
51	17545_60	Sh01	C /T	84	13
53	20736_42	Sh01	C /T	97	0
56	21119_126	Sh01	C /T	87	10
59	22095_107	Sh01	A /G	96	1
66	31133_133	Sh02	G /C	97	0
71	31721_90	Sh02	C /T	84	13
72	34111_44	Sh02	A /G	89	8
73	34111_135	Sh02	T /C	43	53
77	38037_118	Sh02	T /G	95	2
99	49187_47	Sh03	C /T	68	29
105	52535_52	Sh03	T /G	97	0
108	54307_136	Sh03	T /G	97	0
131	69171_28	Sh04	G /A	86	11
148	77788_76	Sh05	T /G	85	12
149	77979_119	Sh05	G /A	79	18
153	84418_84	Sh06	A /G	97	0
154	85290_8	Sh06	T /A	95	2
193	101832_105	Sh07	A /G	91	6
213	123364_8	Sh09	C /A	96	1
244	127495_112	Sh10	C /T	97	0
249	128509_49	Sh10	C /T	97	0
250	128509_117	Sh10	C /T	97	0
251	128509_130	Sh10	G /A	97	0
252	128511_113	Sh10	T /A	97	0
253	128511_49	Sh10	C /T	82	15
257	133685_54	Sh10	C /T	82	15
260	133685_56	Sh10	A /G	82	15

Al evaluar el comportamiento de cada individuo, se destacó el elevado número de loci heterocigotas, con un promedio de 245 loci heterocigotas por individuo (Tabla 3.4). Al comparar a los parentales NA 78-724 y LCP 85-384 en sus versiones diploide y poliploide, observamos que, al evaluar los dosajes alélicos, se registró un mayor número de SNPs como heterocigotas. Esto sugiere que el segundo análisis refleja mejor la naturaleza heterocigota en poliploides.

Tabla 3.4. *Heterocigosis en caña de azúcar*. Número de loci homocigotas y heterocigotas para cada individuo evaluando las 261 posiciones genómicas. NA 78-724 diplo: panel No-150 del capítulo 1 (resaltado en amarillo), analizado como diploide; LCP 85-384 diplo: panel No-150 del capítulo 1 (resaltado en violeta), analizado como diploide; NA 78-724 ploi8: panel de SNPs con los dosajes alélicos asignados del capítulo 2 (resaltado en amarillo); LCP 85-384 78-724 ploi8: panel de SNPs con los dosajes alélicos asignados del capítulo 2 (resaltado en violeta).

Individuos	Loci Heterocigotas	Loci Homocigotas
INTA 053116	152	109
NA7872 diplo	213	48
LCP85384 diplo	240	21
HOCP 92665	205	56
NA 5630	195	63
NA7872 ploi8	250	11
LCP8538 ploi8	260	1
Pro83	239	22
Pro84	245	16
Pro85	250	11
Pro86	255	6
Pro87	249	12
Pro88	240	21
Pro89	232	29
Pro90	241	20
Pro91	256	5
Pro92	243	18
Pro93	242	19
Pro94	253	8
Pro95	246	15
Pro96	255	6
Pro97	248	13
Pro98	245	16
Pro99	255	6
Pro100	243	18
Pro101	254	7
Pro102	241	20
Pro103	249	12
Pro104	233	28
Pro105	240	21
Pro106	252	9

Individuos	Loci Heterocigotas	Loci Homocigotas
Pro107	246	15
Pro108	247	14
Pro109	247	14
Pro110	248	13
Pro111	251	10
Pro112	249	12
Pro113	226	35
Pro114	250	11
Pro115	242	19
Pro116	250	11
Pro117	256	5
Pro118	241	20
Pro119	233	28
Pro120	254	7
Pro121	241	20
Pro122	243	18
Pro123	248	13
Pro124	256	5
Pro125	248	13
Pro126	256	5
Pro127	256	5
Pro128	252	9
Pro129	243	18
Pro130	241	20
Pro131	255	6
Pro132	257	4
Pro133	248	13
Pro134	247	14
Pro135	256	5
Pro136	246	15
Pro137	247	14
Pro138	246	15
Pro139	247	14
Pro140	240	21

Individuos	Loci Heterocigotas	Loci Homocigotas
Pro141	238	23
Pro142	245	16
Pro143	248	13
Pro144	240	21
Pro145	236	25
Pro146	252	9
Pro147	255	6
Pro148	244	17
Pro149	253	8
Pro150	243	18
Pro151	250	11
Pro152	237	24
Pro153	255	6
Pro154	247	14
Pro155	244	17
Pro156	245	16
Pro157	242	19
Pro158	253	8
Pro159	247	14
Pro160	241	20
Pro161	248	13
Pro162	246	15
Pro163	260	1
Pro164	249	12
Pro165	246	15
Pro167	240	21
Pro168	257	4
Pro169	245	16
Pro170	241	20
Pro171	245	16
Pro172	245	16
Pro174	254	7

Validación de los SNPs

Se realizó una búsqueda en la bibliografía para ver si los SNPs hallados en este trabajo fueron reportados previamente. En este sentido, se tomó el trabajo de Souza et al. (2019) y se encontró que 20 de los 261 loci con función asignada coinciden con los reportados para la variedad SP80-3280 (Tabla 3.5). Cabe destacar que, en ese trabajo, los autores utilizaron otra metodología para la generación de SNPs, desde la preparación de la biblioteca genómica, tipo de secuenciación y análisis bioinformático (Souza et al. 2019). La Tabla 3.5 informa la denominación numérica asignada a cada variante SNP identificada en esta tesis, el nombre del SNP asignado, el cromosoma y la posición en pares de bases según la referencia del cv R570, la variante de referencia y la alternativa, así como el nombre de la proteína, enzima o factor de transcripción asociado al SNP. Además, se especifica la posición referenciada al genoma de la variedad SP80-3280 (Tabla 3.5).

Tabla 3.5. *Comparación de SNPs encontrados en este trabajo y en Souza et al. (2019).* 20 variantes SNPs compartidas entre la publicación Souza et al. (2019) y este trabajo de Tesis. N°: número de SNP de la Tabla 3.1; SNP_ID: nombre del SNP; R570 Cromosoma: cromosoma de la caña de azúcar del cv R570; Gen_ID: nombre del gen asociado al SNP; SNP_POS: posición del SNP en pares de bases, según el genoma de referencia de la caña de azúcar del cv R570; SNP: variante de referencia/variante alternativa; gene_sp80= nombre del gen de la variedad SP80-3280; SP80_contig= contig de la variedad SP80-3280. Anotación funcional - Gen: nombre de la proteína, enzima o factor de transcripción asociado al SNP.

N°	SNP_ID	r0 Cromoso	Gen_ID	SNP_POS	SNP	gene_sp80	SP80_contig	Anotación funcional - Gen
40	14275_129	Sh01	Sh01_g028810	48784756	A /G	uti_cns_0123139.4	uti_cns_0123139	Proteína TRANSMISIÓN ABERRANTE DE POLEN 1
51	17545_60	Sh01	Sh01_g034620	57483546	C /T	uti_cns_0076942.3	uti_cns_0076942	Proteína de unión a la miosina 1
53	20736_42	Sh01	Sh01_g039470	64985402	C /T	uti_cns_0046871.9	uti_cns_0046871	Factor de transcripción ERF073 que responde al etileno
56	21119_126	Sh01	Sh01_g040050	65865714	C /T	uti_cns_0180774.4	uti_cns_0180774	Inhibidor de la proteinasa de cisteína
59	22095_107	Sh01	Sh01_g041730	68144573	A /G	uti_cns_0041399.8	uti_cns_0041399	Proteína Dirigente
66	31133_133	Sh02	Sh02_g015170	28854584	C /T	uti_cns_0022996.1	uti_cns_0022996	Fosforibosil aminoimidazol-succino carboxamida sintasa
71	31721_90	Sh02	Sh02_g016110	30678593	G /C	uti_cns_0213914.3	uti_cns_0213914	Proteína vegetal isoprenilada 41 asociada a metales pesados
72	34111_44	Sh02	Sh02_g020250	36503815	C /T	uti_cns_0140879.1	uti_cns_0140879	Proteínas de la familia SR proteína quinasa específica II (SRPK4)
77	38037_118	Sh02	Sh02_g026950	45863362	T /C	uti_cns_0229455.1	uti_cns_0229455	Proteínas de unión al ARN
99	49187_47	Sh03	Sh03_g016700	29709876	T /G	uti_cns_0047315.1	uti_cns_0047315	Acil-CoA sintetasa de cadena larga 4
105	52535_52	Sh03	Sh03_g022050	38061066	C /T	uti_cns_0209082.3	uti_cns_0209082	Proteína con dominio BSD
108	54307_136	Sh03	Sh03_g025100	42405274	T /G	uti_cns_0271902.2	uti_cns_0271902	Proteína 2 similar a TOM1
131	69171_28	Sh04	Sh04_g019850	34011701	G /A	uti_cns_0320494.1	uti_cns_0320494	Sintasa de plastidio (P)ppGpp
148	77788_76	Sh05	Sh05_g008220	16430690	G /A	uti_cns_0291462.1	uti_cns_0291462	Proteína VIP3 que contiene repeticiones WD
153	84418_84	Sh06	Sh06_g007200	13860027	G /A	uti_cns_0006206.6	uti_cns_0006206	Similar a la proteína H0305E08.1 (proteína que conti
154	85290_8	Sh06	Sh06_g008680	16056817	A /G	uti_cns_0339519.1	uti_cns_0339519	Sintasa de alfa-bisabolol
193	101832_105	Sh07	Sh07_g017160	28468845	T /A	uti_cns_0235753.1	uti_cns_0235753	Beta C2-xilosiltransferasa
213	123364_8	Sh09	Sh09_g003130	5126406	G /A	uti_cns_0126667.2	uti_cns_0126667	Transportador ABC de la familia B, miembro 21
249	128509_49	Sh10	Sh10_g011180	20298999	C /T	uti_cns_0130101.2	uti_cns_0130101	Proteasa de membrana dependiente de metal similar
257	133685_54	Sh10	Sh10_g018550	33613438	C /T	uti_cns_0183331.5	uti_cns_0183331	Quinasa de serina/treonina ATR

Discusión

Para mejorar la producción de bioenergía a partir de fuentes derivadas de la caña de azúcar, resulta fundamental comprender mejor la biosíntesis de su pared celular. Esto permitiría el desarrollo de cultivares con mayor biomasa y paredes celulares más susceptibles a la hidrólisis. En este capítulo se ha caracterizado la variabilidad anatómica de los entrenudos en el cultivar comercial LCP 85-384 y el biotipo de caña energía INTA 05-3116, los cuales presentan diferencias fenotípicas marcadas a campo. Además, se exploró el impacto biológico de variantes alélicas identificadas en el capítulo 2 sobre genes candidatos relacionados con el metabolismo de la sacarosa, y el metabolismo de la lignina.

Las regiones reconocidas en la caracterización de las dos variedades de caña fueron similares a lo observado por distintos autores en la misma especie y en otras gramíneas (Da Costa et al. 2019; Zhong et al. 2019). En el centro del tallo, en la región conocida como médula, las células de parénquima son más grandes que los tipos de células circundantes. En estudios previos, se hipotetizó que los haces vasculares se forman en el punto de elongación o antes del proceso de diferenciación celular, ya que su número no varía significativamente, al igual que los resultados observados en este capítulo (Matos et al 2013). De forma análoga, el tamaño de los haces vasculares (HV) de cada genotipo no tuvo variaciones significativas a lo largo del cultivo. Sin embargo, al evaluar los genotipos, se observó que INTA 05-3116 tendió a tener HV más largos que LCP 85-384 en varias etapas de desarrollo, mientras que LCP 85-384 presentó HV más anchos, especialmente durante la última etapa evaluada (Figura 3.3; Apéndice Tabla 3.1). En cuanto al impacto de las etapas de desarrollo, los resultados durante la maduración temprana, con diferencias significativas en la longitud de HV entre los genotipos, sugieren que INTA 05-3116 puede estar mejor adaptado para un crecimiento rápido en esta fase (Figura 3.3; Apéndice Tabla 3.2). Durante la maduración tardía, las diferencias observadas resaltan la estabilidad de ciertas características morfológicas en cada genotipo. LCP 85-384 mantiene un mayor ancho de HV (Figura 3.3; Apéndice Tabla 3.2), lo que podría implicar una mayor estabilidad en condiciones adversas y una menor probabilidad de vuelco debido a su altura (Valarmathi 2021). Por otro lado, INTA 05-3116 sigue mostrando una mayor longitud de HV, lo que es observable en el campo en relación con la altura de las plantas y sus mayores posibilidades de vuelco (Figura 3.3; Apéndice Tabla 3.2). En otro trabajo similar, determinaron que el área entre los haces vasculares aumentó significativamente en 5300 µm2 entre la elongación y la emergencia de la inflorescencia (Matos et al 2013). Por lo tanto, el aumento del área del entrenudo del tallo antes de la floración fue explicado por un incremento en el tamaño de las células parenquimáticas y no un cambio en el tamaño de los haces vasculares (Matos et al. 2013).

Como la mayoría de los cultivares de caña energía (Sandhu et al. 2015, 2016; Fanelli et al.2020), el INTA 05-3116 está estrechamente relacionado con el ancestro silvestre de alto contenido en fibra *S. spontaneum* (García et al. 2022). La influencia de esta especie ancestral fue evidente a nivel histológico, en donde el INTA 05-3116 mostró un menor número de HV y diámetro de los vasos del metaxilema que el cultivar de caña de azúcar, lo que probablemente sea heredado de *S. spontaneum* (Poelking et al. 2015). Además, los resultados histoquímicos revelaron que la deposición de lignina en INTA 05-3116 se extendía a las paredes celulares del parénquima, patrón que fue descripto anteriormente en *S. spontaneum* y *S. Robustum* (Poelking et al. 2015; Llerena et al. 2019). La mayor intensidad de la tinción de floroglucinol en los tejidos

evaluados del INTA 05-3116, en comparación con el LCP 85-384, se debió probablemente al mayor contenido total de fibra del INTA 05-3116, que fue 121% superior al LCP 85-384, en la etapa de maduración tardía y no a una diferencia en el contenido de lignina, que fue similar en ambos genotipos evaluados (García et al 2023).

De forma complementaria a los estudios sobre la anatomía de la pared celular, investigamos la anotación funcional de los SNP identificados en el capítulo 2 para vincularlos a genes candidato para características agroindustriales. Las anotaciones funcionales de SNP contribuyen al descubrimiento de variantes implicadas en regulación génica. Esta información es relevante para proponer estrategias más focalizadas y eficaces en el estudio de la base genética de características complejas. Los paneles de marcadores moleculares caracterizados en esta tesis (capítulos 1 y 2) permitieron identificar 261 posiciones genómicas presentes la población en estudio sobre los cuales se realizó un proceso de anotación funcional. Para ello se predijo el efecto de las variantes (VEP) usando la base de datos Ensembl Plants. Este proceso eficientiza la interpretación de los SNPs, al clasificarlos utilizando términos establecidos en la base de datos Gene Ontology para las categorías función molecular, compartimentalización y proceso biológico (McLaren et al. 2016). Esta información facilita el análisis y priorización de variantes genéticas. Además, estas pueden visualizarse en el contexto de la estructura génica y los dominios proteicos. Como ya se ha mencionado a lo largo de este trabajo de tesis, estas herramientas son fundamentales para el avance en la caracterización de la caña de azúcar, aunque el desarrollo de las mismas sigue siendo limitado por la baja disponibilidad de genomas completos de alta calidad para ser usados como referencias. En este sentido, los proyectos de abordaje genómico se ven acotados en el porcentaje de lecturas alineadas con las referencias disponibles. En coincidencia con este trabajo, Chaves et al. (2022) reportó alineamientos del 50% lecturas provenientes de GBS contra la referencia del cv R570 (Garsmeur et al. 2018), no obstante, la proporción de los SNPs con datos funcionales fue baja. Estos autores describieron patrones de distribución genómica similares a los encontrados en esta tesis. Respecto a las implicancias funcionales de los SNPs, también reportan un menor número de loci en las regiones codificantes en comparación con las regiones no. Otros autores reportaron una abundancia de variantes génicas especialmente en las regiones reguladoras aguas arriba de los genes, lo que puede afectar el nivel de expresión entre las distintas copias de los genes (Souza et al. 2019). En este trabajo, se destacó que la frecuencia de SNPs en los extremos 3'UTR (6 SNPs) fue mayor que en las 5'UTR (1 SNP), un resultado similar fue descripto por Parida et al. (2016). Estos autores sugieren que las secuencias no codificantes 5'UTR están funcionalmente más conservadas que las 3'UTR, ya que contienen secuencias que regulan la expresión.

Mediante la utilización de paneles de marcadores moleculares asociados a genes candidato, es posible concentrarse en las regiones funcionales del genoma. Esta aproximación facilita la comparación de características contrastantes en las variedades a mejorar, como el contenido de azúcar o la estructura de la pared celular de los distintos genotipos (Khanbo et al. 2021). Sin embargo, es necesario realizar múltiples pruebas para validar los resultados obtenidos en genomas complejos como la caña de azúcar, con altos niveles de ploidía y de heterocigosis. No obstante, la presencia de SNPs asociados a una amplia gama de funciones biológicas sugiere que estos marcadores genéticos pueden aportar información sobre múltiples aspectos del crecimiento, el desarrollo y la respuesta a diferentes condiciones ambientales. La tabla Apéndice 3 Tabla 3.3 describe la anotación de las 261 variantes SNPs, lo cual permite explorar variantes con impacto potencial en el metabolismo de una ruta, ciclo o vía particular. En la Tabla 3.1 se recopila la información sobre 28 SNPs, en lo que se examinó con gran precisión cómo un cambio de base puede afectar diversos aspectos celulares y moleculares que ocurren en la población de estudio.

Por ejemplo, la identificación de SNPs asociados a las paredes celulares provee información respecto a roles críticos en el transporte de nutrientes, la comunicación celular y la respuesta a factores ambientales, sugiriendo que los SNPs identificados pueden tener un impacto significativo en la adaptación y la supervivencia de la planta. En este trabajo se detectaron varios SNPs vinculados con proteínas y enzimas claves para las paredes celulares. El SNP 59, ID: 22095_107 presenta una mutación con cambio de sentido en el gen de una proteína dirigente. Estas proteínas no tienen actividad enzimática propia, pero dirigen la formación estereoespecífica de compuestos fenólicos, particularmente lignanos, al guiar la acción de otras enzimas. En la biosíntesis de ligninas, estas proteínas guían la polimerización de monolignoles. Los lignanos y ligninas refuerzan las paredes celulares, actuando como barreras físicas en la respuesta de las plantas a estreses bióticos y abióticos, al ayudar a mantener la integridad de la pared celular y regular la transpiración (Li et al. 2022). Por su parte el SNP 234, ID:122049_17 ubicado en un intrón está vinculado con el gen involucrado en la regulación de la acetilación de los componentes de la pared celular (RWA, por sus siglas en inglés, de "Reduction of Wall Acetylation"), como los polisacáridos (pectinas, hemicelulosas). La acetilación afecta las propiedades físicas y químicas de la pared celular, incluyendo su porosidad, rigidez y capacidad de interacción con otras moléculas.

La estructura de la pared celular influye en la capacidad de la planta para retener agua y regular el transporte de solutos, lo cual es esencial para la supervivencia en condiciones adversas. Las glicosiltransferasas son un grupo de enzimas que participan en la biosíntesis de

polisacáridos como la celulosa, hemicelulosa y pectina en las paredes celulares de las plantas, entre otras rutas metabólicas. La presencia de SNPs asociados a genes que codifican para glicosiltransferasas, como la beta c2-xilosiltransferasa (SNP 193, ID: 101832_105) y la arabinosiltransferasa (SNP 244, ID: 127495_112) podrían ser utilizados para evaluar el comportamiento de estas enzimas en el metabolismo de la caña de azúcar. La actividad de la beta C2-xilosiltransferasa en la síntesis de hemicelulosa contribuye a estas propiedades estructurales, permitiendo a la planta mantener su forma y resistir estrés mecánico (Bencúr et al. 2005). La modificación de los componentes de la pared celular, incluida la incorporación de xilosa por la beta C2-xilosiltransferasa, permitiría a las plantas a adaptarse a diversas condiciones ambientales, como la sequía y la salinidad, al fortalecer la estructura celular y regular el transporte de agua y solutos (Takano et al. 2015; Hrmova et al. 2022). Por su parte, la actividad de la arabinosiltransferasa en la incorporación de arabinosa en los polisacáridos contribuye a las propiedades estructurales, permitiendo a la planta mantener su forma y resistir el estrés mecánico. La identificación de variantes específicas asociadas a estas enzimas podría revelar patrones de regulación o eventos de selección que han afectado la evolución de estas vías metabólicas. En este trabajo, se registraron variante localizadas en 6 intrones vinculados con genes implicados en la síntesis de polisacáridos, específicamente con hexosiltransferasas (SNP 8, ID: 22325_137; SNP 21, ID: 2070_95; SNP 22, ID: 2070_93; SNP 23, ID: 2070_43). Las hexosiltransferasas juegan un papel crucial en la síntesis de polisacáridos, como la celulosa y la hemicelulosa, que son componentes esenciales de la pared celular. Además, estas enzimas pueden estar implicadas indirectamente en el metabolismo de la sacarosa y otros oligosacáridos, facilitando la formación de moléculas complejas a partir de azúcares simples. Se identificó un SNP en cada uno de los factores de transcripción WRKY y bZIP, los cuales regulan genes involucrados en la fotosíntesis y en la producción de compuestos intermedios necesarios para la síntesis de sacarosa, como la sacarosa-fosfato sintasa. Además, estos factores de transcripción están implicados en la respuesta de las plantas a diversas formas de estrés, tanto biótico como abiótico (Xue et al. 2022).

La identificación de varios SNPs asociados al mismo gen, como es el caso del gen de la proteasa de membrana dependiente de metales (SNPs de 249 a 253 vinculados con el Gen ID: Sh10_g011180, Tabla 3.1), puede indicar la presencia de regiones genómicas sujetas a una fuerte presión de selección, lo que a su vez implicaría una mayor variabilidad fenotípica en la población para determinados caracteres (Bowles 2002; Ernst y Klaffke 2003). Al igual que lo observado en el capítulo 2, la Tabla 3.2 presenta las discrepancias con relación al estado de las variantes SNP cuando se caracterizan como de segregación mendeliana respecto a la poliploide.

Un ejemplo es el SNP 59: ID: 22095_107 que resultó en un genotipo heterocigota (A/G) en el análisis simplificado (simulando que se trata de una especie diploide), mientras que para el parental NA 78-724 analizando los dosajes alélicos resultó en homocigota para el alelo alternativo (0/8). Un caso inverso se dio en el SNP 77; ID: 38037_118 donde el LCP 85-384 resultó ser homocigota para la referencia (T/T) con el análisis diploide y heterocigota con el poliploide (6/2). Basándonos en estos resultados, recomendamos que, para genomas tan complejos como este, se priorice la asignación de recursos durante la secuenciación para obtener profundidades por locus que faciliten la determinación precisa de los dosajes alélicos.

Varios autores han destacado altos niveles de heterocigosidad en la caña de azúcar (Gutierrez et al. 2018; Souza et al. 2019). A pesar de la identificación de múltiples variantes alélicas en un mismo loci, las mutaciones que implican una sola variante genética son más comunes que aquellas con tres o cuatro SNPs. Por su parte, resultados de análisis de citología en cultivares modernos han demostrado que la mayoría de los cromosomas en caña usualmente forman bivalentes, lo que podría explicar que la mayoría de los SNPs sigan un comportamiento bialélico (Vieira et al. 2018; Souza et al. 2019). Sin embargo, la presencia de varias copias génicas reportadas en la bibliografía (Souza et al. 2019), nos llevó a evaluar la heterocigosis de la población biparental bajo estudio (Tablas 3.3 y 3.4). El alto número de loci heterocigotas en la mayoría de los individuos, incluso la presencia de individuos completamente heterocigotas (72 genotipos) resalta la diversidad genética dentro de la población, sugiriendo una alta capacidad de adaptación a diferentes entornos o condiciones ambientales. El SNP 73 ID: 34111_135, uno de los pocos SNPs en los cuales se identificaron más variantes homocigotas que heterocigotas, podría indicar una selección o una presión evolutiva específica en ese locus. La exploración de este sitio genético en otras variedades de caña de azúcar o, incluso de otras especies del género *Saccharum,* resulta interesante para conocer la implicancia en términos de la historia evolutiva o la adaptación de la población. Otros autores han reportado diferentes niveles de heterocigosis en las especies fundadoras de caña de azúcar, a partir de calcular los índices de Shannon, a través de la evaluación de 36 marcadores microsatélites (Nayak et al. 2014). Los resultados mostraron índices de Shannon para *S. spontaneum* (0,492)*, S. officinarum* (0,456)*, S. Barberi* (0,423)*, S. robustum* (0,427) *y S. sinense* (0,383). Estos valores sugieren una mayor heterocigosidad en las dos primeras especies fundadoras, lo que contribuye a la variabilidad genética de sus progenies.

La validación de las variantes SNP mediante la comparación con los SNP reportados por Souza et al. (2019) otorga valor adicional a nuestros hallazgos (Tabla 3.5). La consistencia en la identificación de estas variantes en los mismos loci, utilizando una metodología diferente en cuanto a la preparación de la biblioteca genómica (kit de Illumina TruSeq), secuenciación con

lecturas largas (~5 kb) y análisis bioinformático, respalda la robustez de nuestros resultados. Además, se empleó una variedad de caña de azúcar diferente (SP80-3280). En el estudio de Souza et al. (2019), los autores se propusieron ensamblar las regiones genéticas de la variedad SP80-3280, ampliamente cultivada en Brasil, mediante comparaciones con otras especies de gramíneas. Una vez obtenido el ensamblado, realizaron un mapeo contra el genoma de referencia del sorgo y el del cultivar R570 de caña de azúcar (el mismo utilizado en este trabajo de tesis) para identificar los SNPs. Los resultados obtenidos son de gran utilidad los programas de mejoramiento locales tanto para para llevar a cabo estudios de asociación de genoma completo (GWAS), como de selección genómica, en conjunto con análisis fenotípicos complementarios.

En este capítulo se realizó una caracterización de la variabilidad anatómica de los entrenudos en el cultivar comercial LCP 85-384 y el biotipo de caña energía INTA 05-3116, aportando información útil de las estructuras físicas y morfológicas de los segmentos del tallo. Cuando se evaluaron los genotipos, se observó que INTA 05-3116 tendió a tener HV más largos que LCP 85-384 en varias etapas de desarrollo, mientras que LCP 85-384 presentó HV más anchos, especialmente durante la etapa previa a una cosecha tardía. En cuanto al impacto de las etapas de desarrollo, la longitud de HV entre los genotipos fue significativamente mayor durante la maduración temprana, lo que sugiere que el biotipo de alta fibra puede estar mejor adaptado para un crecimiento rápido en esta fase. A su vez, se logró la exploración del impacto biológico de 261 SNPs de alta calidad y representatividad. Se registraron un mayor número de variantes polimórficas en regiones reguladoras (río arriba o debajo de genes), respecto a regiones codificantes. Los datos generados en este capítulo permitieron detectar varios polimorfismos vinculados con proteínas y enzimas claves para las paredes celulares, que intervienen en la síntesis de polisacáridos y lignina que agregan valor a los paneles de SNP desarrollados. En concordancia con los resultados obtenidos por otros grupos de trabajo, se encontraron altos niveles de heterocigosis en la población analizada, incluso la presencia de 72 individuos completamente heterocigotas para los 261 loci evaluados.

Síntesis final

Poseer un conocimiento detallado de la diversidad genética del germoplasma utilizado en el mejoramiento de la caña de azúcar es crucial para determinar la estructura poblacional dentro de las colecciones de germoplasma y comprender los patrones de diversidad. Esto permite seleccionar progenitores adecuados para cruzamientos dirigidos, con el objetivo de lograr la introgresión efectiva de alelos favorables para los rasgos de interés. En los programas de mejoramiento de instituciones públicas, la adopción de herramientas moleculares generalmente se ve restringida por la falta de recursos, financiamiento e infraestructura. Esta condición limita la complementación de las estrategias de mejoramiento clásico con las innovaciones emergentes de las tecnologías genómicas. Sin embargo, estos avances recientes en este campo ofrecen alternativas eficientes para la caracterización molecular de poblaciones de mejoramiento, alentando a los programas más pequeños a diversificar sus estrategias mediante la incorporación de datos moleculares. Esta tesis se enfocó en desarrollar marcadores moleculares de tipo SNP a nivel poblacional en caña de azúcar, empleando técnicas de genotipificación por secuenciación. Se exploró su potencial en estudios de diversidad, mapeo genético y el efecto funcional de los loci SNP. Este trabajo se enmarca en las prioridades estratégicas del área de biotecnología para el mejoramiento de la caña de azúcar en INTA.

En este trabajo, se puso a punto un protocolo de ddRADSeq *(digestión con dos enzimas de restricción*) en un número reducido de genotipos de caña de azúcar del programa de mejoramiento de INTA. Luego, el protocolo fue escalado a un mayor número de muestras, proporcionando información relevante sobre marcadores moleculares SNPs a nivel poblacional. Este protocolo fue evaluado desde el laboratorio hasta el análisis bioinformático, comparando distintas plataformas de secuenciación y parámetros para obtener SNPs confiables. Como resultado se generó un panel de 47.964 SNPs para cinco líneas del programa de mejoramiento, y otro panel de 2.066 SNPs para la población de mejoramiento NA 78-724 x LCP 85-384.

Con estos datos se generaron mapas genéticos para la población de caña de azúcar a través de dos enfoques. El primero se basó en algoritmos desarrollados para especies diploides, generando un mapa marco preliminar representativo de los 10 grupos de ligamiento de la caña de azúcar. El segundo enfoque se basó en estimar los dosajes alélicos para reflejar la naturaleza poliploide, generando un mapa de haplotipos que permita estimar la contribución de cada parental sobre los loci en estudio. Disponer de herramientas y métodos para desarrollar estos mapas es crucial para entender la estructura del genoma sustentar el mejoramiento de variedades para características de interés agroindustrial.

Se evaluó la estructura poblacional, lo que permitió la identificación de agrupamientos consistentes y la diferenciación de conjuntos de genotipos, reflejando la presencia de combinaciones alélicas únicas.

Se realizó una caracterización de la variabilidad anatómica de los entrenudos en el cultivar comercial LCP 85-384 y el biotipo de caña energía INTA 05-3116. El INTA 05-3116 mostró menor número de haces vasculares y diámetro del metaxilema y su deposición de lignina se extendió más hacia el tejido parenquimático, en comparación con la caña comercial a lo largo del ciclo de crecimiento. Se observaron haces vasculares más largos en INTA 05-3116 que en LCP 85-384 en varias etapas de desarrollo, mientras que este último presentó haces vasculares más anchos.

Finalmente, se evaluó el panel generado en relación con genes de interés agroindustrial, especialmente aquellos asociados al metabolismo de polisacáridos y otros componentes de las paredes celulares. Se exploró el impacto biológico de 261 SNPs de alta calidad y representatividad, destacándose la identificación de 28 SNPs en regiones codificantes con potencial impacto biológico. Entre estos, se detectaron 15 polimorfismos vinculados a proteínas y enzimas clave en la síntesis de polisacáridos y otros componentes de la pared celular.

Conclusiones

La tecnología ddRADseq demostró ser eficiente para el desarrollo de SNPs en caña de azúcar. La selección de enzimas de restricción sensibles a la metilación resultó adecuada para muestrear el genoma de manera homogénea a lo largo de todos los cromosomas. Se obtuvieron paneles de SNPs en cantidad y representación global del genoma, conforme a lo esperado por el método. Esta metodología fue reproducible tanto intra-muestras (en distintas bibliotecas y momentos, así como en diferentes plataformas de secuenciación) como entre muestras (en una población de 90 individuos).

La metodología es adecuada para programas de mejoramiento genético con bajo financiamiento, ya que puede aplicarse en distintos momentos según la necesidad y el presupuesto. La intensidad de secuenciación en caña de azúcar es crucial. Es posible obtener marcadores SNPs robustos utilizando lecturas pareadas de un tamaño mínimo de 150 pb, una profundidad de lectura mínima de 10, un máximo de 4 millones de lecturas por individuo y un número mínimo de 90 individuos. Sin embargo, como recomienda la literatura, se comprobó que, para la asignación correcta de dosajes alélicos, la profundidad mínima de lecturas debe ser de 56.

El genoma de referencia utilizado permitió la identificación de numerosas variantes SNPs, potencialmente útiles y aplicables en el programa de mejoramiento del INTA; sin embargo, las tasas de alineamiento con el genoma monoploide del R570 sugieren que se trata de un cultivar genéticamente distante de las variedades argentinas y americanas. Utilizar una referencia más cercana o, incluso, la de los mismos parentales permitiría un mejor aprovechamiento de los datos de secuenciación. Además, con estos mismos datos, se podría realizar un nuevo análisis utilizando la referencia poliploide, que posee 5,04 Gb de datos organizados en 10 grupos homólogos, lo que resultaría en una mayor cobertura del genoma en comparación con el ensamblado de 530 Mb utilizado aquí como la referencia.

El panel de SNPs resultó eficiente para captar la variabilidad entre las accesiones de los programas de mejoramiento genético de INTA y USDA evaluados, demostrando su potencial para la selección de parentales de futuras poblaciones de mejoramiento.

La estrategia de secuenciación de menor intensidad aplicada a la progenie de la población de mejoramiento, habiendo definido previamente los paneles de SNPs en las líneas parentales con una profundidad de 56x, resultó eficiente para la generación de un panel de 2066 marcadores.

La representación genómica de los SNPs permitió reconstruir 38 grupos de ligamiento parciales correspondientes a los 10 cromosomas básicos de caña de azúcar. Dada la complejidad del genoma, el mapeo genético basado en marcadores de segregación mendeliana requiere paneles de marcadores más abundantes debido a la baja proporción de marcadores con esta segregación.

La estimación de los dosajes alélicos y el mapeo basado en algoritmos poliploides permitió la reconstrucción de los haplotipos, reflejando la contribución de cada parental en las progenies de la población de mapeo. Considerando que la caña de azúcar es una especie alopoliploide que presenta aneuploidías, es fundamental contar con mayor profundidad de lecturas por loci para lograr una asignación de dosajes alélicos más representativa y de mayor cobertura genómica.

Para realizar una selección de parentales para mejoramiento de materiales multipropósito, las mediciones a nivel de haces vasculares y distribución de lignina necesitan ser complementadas por mediciones químicas de los componentes básicos de las paredes celulares para ser informativos.

La predicción de la función de los SNPs corroboró que la metodología implementada fue exitosa para explorar regiones codificantes del genoma, incluyendo aquellas con posible implicancia en el metabolismo de los polisacáridos y otros componentes de la pared celular, posibilitando la identificación de genes candidato que pueden ser relevantes para futuros estudios.

Lic. Catalina Molina	Dr. Alberto Acevedo	Dra. Norma Paniego
Doctorando	Director	Co-Directora

Bibliografía

[1] Abreu L, Grassi M, Carvalho L, Silva J, Oliveira J, Bressiani J, Pereira G (2020) Energy cane vs sugarcane: Watching the race in plant development. Industrial Crops and Products, 156, 112868. https://doi.org/10.1016/j.indcrop.2020.112868.

[2] Acevedo A, Molina C, García JM, Federico ML; Erazzú LE (2021) Current Status of Sugarcane (Saccharum spp.) Breeding under Sustainable Conditions in Argentina; Universitat Politècnica de València. 129-147

[3] Acevedo A, Tejedor MT, Erazzú LE, Cabada S, Sopena R (2017) Pedigree comparison highlights genetic similarities and potential industrial values of sugarcane cultivars. Euphytica 213. https://doi.org/10.1007/s10681-017-1908-2

[4] Aguirre NC, Filippi CV, Zaina G, Rivas JG, Acuña CV, Villalba PV, García MN, González S, Rivarola M, Martínez MC, Puebla AF, Morgante M, Hopp HE, Paniego NB, Poltri SNM (2019) Optimizing DDRADseq in non-model species: A Case Study in Eucalyptus dunnii Maiden. Agronomy 9. https://doi.org/10.3390/agronomy9090484

[5] Aitken KS (2021) History and Development of Molecular Markers for Sugarcane Breeding. Sugar Tech 24:341–353. https://doi.org/10.1007/S12355-021-01000-7

[6] Ali A, Khan M, Sharif R, Mujtaba M, Gao SJ (2019) Sugarcane omics: An update on the current status of research and crop improvement. Plants 8:1–24. https://doi.org/10.3390/plants8090344

[7] Andrews S (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[8] Andru S, Pan YB, Thongthawee S, Burner DM, Kimbeng CA (2011) Genetic analysis of the sugarcane (Saccharum spp.) cultivar "LCP 85-384". I. Linkage mapping using AFLP, SSR, and TRAP markers. Theor Appl Genet 123:77–93. https://doi.org/10.1007/S00122-011-1568-X

[9] Arro J, Park J-W, Wai CM, VanBuren R, Pan Y-B, Nagai C, Silva J da, Ming R (2016) Balancing selection contributed to domestication of autopolyploid sugarcane (Saccharum officinarum L.). Euphytica 209:477–493. https://doi.org/10.1007/s10681-016-1672-8

[10] Asnaghi C, Roques D, Ruffel S, Kaye C, Hoarau JY, Télismart H, Girard JC, Raboin LM, Risterucci AM, Grivet L, D'Hont A (2004) Targeted mapping of a sugarcane rust resistance gene (Bru1) using bulked segregant analysis and AFLP markers. Theor Appl Genet 108:759–764. https://doi.org/10.1007/S00122-003-1487-6

[11] Audano PA, Beck CR (2023) Small allelic variants are a source of ancestral bias in structural variant breakpoint placement. bioRxiv, doi: 10.1101/2023.06.25.546295

[12] Balsalobre TWA, da Silva Pereira G, Margarido GRA, Gazaffi R, Barreto FZ, Anoni CO, Cardoso-Silva CB, Costa EA, Mancini MC, Hoffmann HP, de Souza AP, Garcia AAF, Carneiro MS (2017) GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. BMC Genomics 18:1–19. https://doi.org/10.1186/s12864-016-3383-x

[13] Bencúr P, Steinkellner H, Svoboda B, Mucha J, Strasser R, Kolarich D, Hann S, Köllensperger G, Glössl J, Altmann F, Mach L (2005) Arabidopsis thaliana β1,2-xylosyltransferase: an unusual glycosyltransferase with the potential to act at multiple stages of the plant N-glycosylation pathway. Biochemical Journal, 388(2):515-525. doi: 10.1042/BJ20042091

[14] Benedetti P (2018) Primer relevamiento del cultivo de caña de azúcar de la República Argentina a partir de imágenes satelitales para la campaña 2018. Inta 2–6

[15] Bilton TP, Schofield MR, Black MA, Chagné D, Wilcox PL, Dodds KG (2018) Accounting for Errors in Low Coverage High-Throughput Sequencing Data When Constructing Genetic Maps Using Biparental Outcrossed Populations, Genetics, Volume 209, Issue 1, 1 May 2018, Pages 65–76, https://doi.org/10.1534/genetics.117.300627

[16] Blischak PD, Kubatko LS, Wolfe AD (2017) SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. https://doi.org/10.1101/120261

[17] Borcard D, François G, Legendre P (2018) Numerical Ecology with R, 2 nd. Springer International Publishing, Cham, Switzerland.

[18] Bourke PM, Voorrips RE, Visser RGF, Maliepaard C (2018) Tools for genetic studies in experimental populations of polyploids. Front Plant Sci 9:1–17. https://doi.org/10.3389/fpls.2018.00513

[19] Bowles D (2002) A multigene family of glycosyltransferases in a model plant, Arabidopsis thaliana. Biochem Soc Trans. 30 (2): 301–306. doi: https://doi.org/10.1042/bst0300301

[20] Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. https://doi.org/10.1093/BIOINFORMATICS/BTM308

[21] Budeguer F, Enrique R, Perera MF, Racedo J, Castagnaro AP, Noguera AS, Welin B (2021) Genetic Transformation of Sugarcane, Current Status and Future Prospects. Front Plant Sci 12:2467. https://doi.org/10.3389/FPLS.2021.768609/BIBTEX

[22] Calviño M., Bruggmann R., & Messing J. (2008). Screen of genes linked to high-sugar content in stems by comparative genomics. Rice, 1(2), 166–176. https://doi.org/10.1007/s12284-008-9012-9

[23] Carlquist S, Schneidery EL (2011) Origins and nature of vessels in monocotyledons. 13. scanning electron microscopy studies of xylem in large grasses. International Journal of Plant Sciences, doi: 10.1086/658155

[24] Carvalho-Netto O, Bressiani J, Soriano H, Fiori C, Santos J, Barbosa G, Xavier M, Landell M, Pereira G (2014) The potential of the energy cane as the main biomass crop for the cellulosic industry. Chem Biol Technol Agric 1:1–8. https://doi.org/10.1186/s40538-014-0020-2

[25] Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013), Stacks: an analysis tool set for population genomics. Mol Ecol, 22: 3124-3140. https://doi.org/10.1111/mec.12354

[26] Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. G3 Genes|Genomes|Genetics 1:171–182. https://doi.org/10.1534/G3.111.000240

[27] Cerca J, Maurstad MF, Rochette NC, Rivera-Colón AG, Rayamajhi N, Catchen JM, Struck TH (2021) Removing the bad apples: A simple bioinformatic method to improve loci-recovery in de novo RADseq data for non-model organisms. Methods Ecol Evol 12:805–817. https://doi.org/10.1111/2041-210X.13562

[28] Chang W, Chen H, Jiao G, Dou Y, Liu L, Qu C, Li J, Lu K (2022) Biomolecular Strategies for Vascular Bundle Development to Improve Crop Yield. Biomolecules 12:1–14. https://doi.org/10.3390/biom12121772

[29] Chaves S, Ostengo S, Bertani RP, Peña Malavera AN, Cuenya MI, Filippone MP, Castagnaro AP, Balzarini MG, Racedo J (2022) Novel alleles linked to brown rust resistance in sugarcane. Plant Pathol 71:1688–1699. https://doi.org/10.1111/PPA.13605

[30] Chen MH, L Kuo, PO Lewis (2014) Bayesian phylogenetics: methods, algorithms, and applications. Chapman and Hall/CRC. Boca Raton.

[31] Chen X, Huang Z, Fu D, Fang J, Zhang X, Feng X, Xie J, Wu B, Luo Y, Zhu M, Qi Y (2022) Identification of Genetic Loci for Sugarcane Leaf Angle at Different Developmental Stages by Genome-Wide Association Study. Front Plant Sci 13:841693. https://doi.org/10.3389/FPLS.2022.841693/BIBTEX

[32] Clevenger J, Chavarro C, Pearl SA, Ozias-Akins P, Jackson SA (2015) Single nucleotide polymorphism identification in polyploids: A review, example, and recommendations. Mol Plant 8:831–846. https://doi.org/10.1016/j.molp.2015.02.002

[33] Clevenger JP, Ozias-Akins P (2015) SWEEP: A tool for filtering high-quality SNPs in polyploid crops. G3 Genes, Genomes, Genet 5:1797–1803. https://doi.org/10.1534/g3.115.019703

[34] Collucci D, Bueno RCA, Milagres AMF, Ferraz A (2019) Sucrose content, lignocellulose accumulation and in vitro digestibility of sugarcane internodes depicted in relation to internode maturation stage and Saccharum genotypes. Ind Crops Prod 139:111543. https://doi.org/10.1016/j.indcrop.2019.111543

[35] Costa THF, Masarin F, Bonifácio TO, Milagres AMF, Ferraz A (2013) The enzymatic recalcitrance of internodes of sugar cane hybrids with contrasting lignin contents. Ind Crops Prod 51:202–211. https://doi.org/10.1016/j.indcrop.2013.08.078

[36] Costet L, Le Cunff L, Royaert S, Raboin LM, Hervouet C, Toubi L, Telismart H, Garsmeur O, Rousselle Y, Pauquet J, Nibouche S, Glaszmann JC, Hoarau JY, D'Hont A (2012) Haplotype structure around Bru1 reveals a narrow genetic basis for brown rust resistance in modern sugarcane cultivars. Theor Appl Genet 125:825–836. https://doi.org/10.1007/S00122-012-1875-X

[37] Da Costa RMF, Pattathil S, Avci U, Winters A, Hahn MG, Bosch M (2019) Desirable plant cell wall traits for higher-quality miscanthus lignocellulosic biomass. Biotechnol Biofuels 12:1–18. https://doi.org/10.1186/s13068-019-1426-7

[38] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330

[39] de Carvalho LM, Borelli G, Camargo AP, de Assis MA, de Ferraz SMF, Fiamenghi MB, José J, Mofatto LS, Nagamatsu ST, Persinoti GF, Silva N V., Vasconcelos AA, Pereira GAG, Carazzolle MF (2019) Bioinformatics applied to biotechnology: A review towards bioenergy research. Biomass and Bioenergy 123:195–224. https://doi.org/10.1016/j.biombioe.2019.02.016

[40] De Lara LAC, Santos MF, Jank L, Chiari L, De Vilela MM, Amadeu RR, Dos Santos JPR, Da Pereira GS, Zeng ZB, Garcia AAF (2019) Genomic selection with allele dosage in Panicum maximum Jacq. G3 Genes, Genomes, Genet 9:2463–2475. https://doi.org/10.1534/g3.118.200986

[41] Di Pauli V, Fontana P, Pérez Gómez S, Felipe A, Vargas Corbalán M, et al. (2017) Análisis comparativo entre la respuesta a roya marrón y la presencia del gen Bru1 en el germoplasma de caña de azúcar de INTA. 4º Congreso Argentino de Fitopatología.

[42] Di Pauli V, Fontana P, Pérez Gómez S, Lizárraga J, Sopena R (2015) Optimización de un método de detección múltiple para royas de caña de azúcar . Ciencia y Tecnología de los Cultivos Industriales: Caña de azúcar. Año 5. N° 7. ISSN 1853-7677. Pág. 34-38.

[43] Di Pauli V, Fontana PD, Lewi DM, Felipe A, Erazzú LE (2021) Optimized somatic embryogenesis and plant regeneration in elite Argentinian sugarcane (Saccharum spp.) cultivars. J Genet Eng Biotechnol 19:171.

[44] Ernst C, Klaffke W (2003) Chemical Synthesis of Uridine Diphospho-d-xylose and UDP-l-arabinose. The Journal of Organic Chemistry. 68 (14), 5780-5783. DOI: 10.1021/jo034379u

[45] Evans DL, Hughes B, Joshi SV (2022) Comparative whole plastome and low copy number phylogenetics of the core Saccharinae and Sorghinae. bioRxiv, doi: 10.1101/2022.01.04.474895

[46] Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048. https://doi.org/10.1093/BIOINFORMATICS/BTW354

[47] Ewing B, Green P (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res. 8(3): 186–194.

[48] Fanelli A, Reinhardt L, Matsuoka S, Ferraz A, da Franca Silva T, Hatfield RD, Romanel E (2020) Biomass composition of two new energy cane cultivars compared with their ancestral Saccharum spontaneum during internode development. Biomass Bioenergy 141, 105696. https://doi.org/10.1016/j.biombioe.2020.105696.

[49] Felsenstein J. 2004. Inferring phylogenies. Sinauer associates. Sunderland.

[50] Ferreira SS, Hotta CT, Poelking VG de C, Leite DCC, Buckeridge MS, Loureiro ME, Barbosa MHP, Carneiro MS, Souza GM (2016) Co-expression network analysis reveals transcription factors associated to cell wall biosynthesis in sugarcane. Plant Mol Biol 91:15–35. https://doi.org/10.1007/s11103-016-0434-2Fickett N, Gutierrez A, Verma M, Pontif M, Hale A, Kimbeng C, Baisakh N (2018) Genome-wide association mapping identifies markers associated with cane yield components and sucrose traits in the Louisiana sugarcane core collection. Genomics 0–1. https://doi.org/10.1016/j.ygeno.2018.12.002

[51] Fickett ND, Ebrahimi L, Parco AP, Gutierrez A V., Hale AL, Pontif MJ, Todd J, Kimbeng CA, Hoy JW, Ayala-Silva T, Gravois KA, Baisakh N (2020) An enriched sugarcane diversity panel for utilization in genetic improvement of sugarcane. Sci Reports 2020 101 10:1–12. https://doi.org/10.1038/s41598-020-70292-8

[52] Fontana PD, Orru' L, Fontana CA, Cocconcelli PS, VIgnolo GM, Salazar SM (2019) Soil microbial community responses to differential sugarcane management strategies as revealed by 16S metagenomis. Proc Int Soc Sugar Cane Technol 30, 502–510

[53] Fontana PD, Rago AM, Fontana CA, Vignolo GM, Cocconcelli PS, Mariotti JA (2013) Isolation and genetic characterization of Acidovorax avenae from red stripe infected sugarcane in Northwestern Argentina. Eur J Plant Pathol 137: 525.

[54] Fontana PD, Tomasini N, Fontana CA, Di Pauli V, Cocconcelli PS, Vignolo GM, Salazar SM (2018) MLST reveals a separate and novel clonal group for Acidovorax avenae strains causing red stripe in sugarcane from Argentina. Phytopathology 109, 358-365.

[55] Fu J, McKinley B, James B, Chrisler W, Meng L, Gaffrey MJ, Mitchell HD, Orr G, Swaminathan K, Marshall-Colon A (2023) The stem cell-type transcriptome of bioenergy sorghum reveals the spatial regulation 3 of secondary cell wall networks 4 5 Author list. https://doi.org/10.46936/lser.proj.2020.51398/60000192

[56] Fu YB (2014) Genetic diversity analysis of highly incomplete snp genotype data with imputations: An empirical assessment. G3 Genes, Genomes, Genet 4:891–900. https://doi.org/10.1534/g3.114.010942

[57] Fukuda H, Ohashi-Ito K (2018) Vascular tissue development in plants. Current Topics in Developmental Biology, doi: 10.1016/BS.CTDB.2018.10.005

[58] Garcia AAF, Mollinari M, Marconi TG, Serang OR, Silva RR, Vieira MLC, Vicentini R, Costa EA, Mancini MC, Garcia MOS, Pastina MM, Gazaffi R, Martins ERF, Dahmer N, Sforça DA, Silva CBC, Bundock P, Henry RJ, Souza GM, Van Sluys MA, Landell MGA, Carneiro MS, Vincentz MAG, Pinto LR, Vencovsky R, Souza AP (2013) SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Sci Rep 3:1–10. https://doi.org/10.1038/srep03399

[59] Garcia JM, Merino MF, Felipe A (2023a) Caracterización morfológica de dos nuevas variedades de caña de azúcar: INTA NA 03-663 e INTA NA 03-617. Revista Agronómica del Noroeste Argentino 42(2), 79-83.

[60] García JM, Molina C, Acevedo A, Silva M, Gómez LD, Erazzú LE (2020) Caña de azúcar en Argentina: situación actual y uso para fines bioenergéticos. Optimización de los procesos de extracción de biomasa sólida para uso energético. España: UP Valencia, pp. 145-158

[61] García JM, Molina C, Simister R, Taibo CB, Setten L, Erazzú LE, Gómez LD, Acevedo A (2023b) Chemical and histological characterization of internodes of sugarcane and energy-cane hybrids throughout plant development. Ind Crops Prod 199:116739. https://doi.org/10.1016/J.INDCROP.2023.116739

[62] García JM, Silva MP, Simister R, McQueen-Mason SJ, ERAZZÚ LE, GÓMEZ LD, Acevedo A (2022) "High variability for cell-wall and yield components in commercial sugarcane (Saccharum spp.) progeny contrasts with parental lines and energy cane" Journal of Crop Improvement https://doi.org/10.1080/15427528.2021.2011521

[63] Garibaldi LA, Oddi F, Aristimuño FJ, Behnisch AN (2016) Modelos generalizados aplicados a la Economía en el lenguaje R. 1a ed. - Bariloche: el autor. pp. 197. ISBN: 978-987-42-3009-6

[64] Garsmeur O, Droc G, Antonise R, Grimwood J, Potier B, Aitken K, Jenkins J, Martin G, Charron C, Hervouet C, Costet L, Yahiaoui N, Healey A, Sims D, Cherukuri Y, Sreedasyam A, Kilian A, Chan A, Van Sluys MA, Swaminathan K, Town C, Bergès H, Simmons B, Glaszmann JC, Van Der Vossen E, Henry R, Schmutz J, D'Hont A (2018) A mosaic monoploid reference sequence for the highly complex genome of sugarcane. Nat Commun 9. https://doi.org/10.1038/s41467-018-05051-5

[65] Gerard D (2021) Scalable bias-corrected linkage disequilibrium estimation under genotype uncertainty. Hered ,The Genet Soc 127:357–362. https://doi.org/10.1038/s41437-021-00462-5

[66] Gerard D, Ferrão LFV, Garcia AAF, Stephens M (2018) Genotyping polyploids from messy sequencing data. Genetics 210:789–807. https://doi.org/10.1534/genetics.118.301468

[67] Glyn J, Rod E, Merry B, Yates D, Yates W, Yates B, Digges P, Forber G, Todd M, Berding N, Cox M, Hogarth M, Bailey R, Leslie G, Irvin J (2004) Sugarcane, Second edi. Blackwel Publishing Company

[68] Grativol C, Regulski M, Bertalan M, McCombie WR, Da Silva FR, Zerlotini Neto A, Vicentini R, Farinelli L, Hemerly AS, Martienssen RA, Ferreira PCG (2014) Sugarcane genome sequencing by methylation filtration provides tools for genomic research in the genus Saccharum. Plant J 79:162–172. https://doi.org/10.1111/tpj.12539

[69] Grivet L, Arruda P (2002) Sugarcane genomics: Depicting the complex genome of an important tropical crop. Curr Opin Plant Biol 5:122–127. https://doi.org/10.1016/S1369-5266(02)00234-0

[70] Gururani MA (2023) Plant RNA-binding proteins as key players in abiotic stress physiology. J Exp Biol Agric Sci 11:41–53. https://doi.org/10.18006/2023.11(1).41.53

[71] Gutierrez AF, Hoy JW, Kimbeng CA, Baisakh N (2018) Identification of genomic regions controlling leaf scald resistance in sugarcane using a Bi-parental mapping population and selective genotyping by sequencing. Front Plant Sci 9:1–10. https://doi.org/10.3389/fpls.2018.00877

[72] Healey AL, Garsmeur O, Lovell JT, Shengquiang S, Sreedasyam A, Jenkins J, Plott CB, Piperidis N, Pompidor N, Llaca V, Metcalfe CJ, Doležel J, Cápal P, Carlson JW, Hoarau JY, Hervouet C, Zini C, Dievart A, Lipzen A, Williams M, Boston LB, Webber J, Keymanesh K, Tejomurthula S, Rajasekar S, Suchecki R, Furtado A, May G, Parakkal P, Simmons BA, Barry K, Henry RJ, Grimwood J, Aitken KS, Schmutz J (2024) The complex polyploid genome architecture of sugarcane. Nature. https://doi.org/10.1038/s41586-024-07231-4

[73] Hrmova M, Stratilová B, Stratilová E (2022) Broad Specific Xyloglucan:Xyloglucosyl Transferases Are Formidable Players in the Re-Modelling of Plant Cell Wall Structures. International Journal of Molecular Sciences. 23(3):1656. https://doi.org/10.3390/ijms23031656

[74] Hundia N, Kabir N, Mehta S, Pokhriyal A, Chua ZE, Rajaram A, Lutz M, Kumar A (2022) Genotype Imputation Using K-Nearest Neighbors and Levenshtein Distance Metric. Int Conf ICT Converg 2022-October:272–277. https://doi.org/10.1109/ICTC55196.2022.9952611

[75] INDEC: Instituto Nacional de Estadísticas y Censos- Resultados definitivos - Censo Nacional Agropecuario (2018) Buenos Aires. Retrieved December 21, 2021, from https://www.indec.gob.ar/ftp/cuadros/economia/cna2018_resultados_definitivos.pdf

[76] Jacobsen KR, Fisher DG, Maretzki A, Moore PH (1992) Developmental Changes in the Anatomy of the Sugarcane Stem in Relation to Phloem Unloading and Sucrose Storage. Bot Acta 105:70–80. https://doi.org/10.1111/j.1438-8677.1992.tb00269.x

[77] Jombart, T. (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24: 1403-1405. doi: 10.1093/bioinformatics/btn129

[78] Jombart T. and Ahmed I. (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. doi: 10.1093/bioinformatics/btr521

[79] Kane AO, Pellergini VOA, Espirito Santo MC, Ngom BD, García JM, Acevedo A, Erazzú LE, Polikarpov I (2021) Evaluating the Potential of Culms from Sugarcane and Energy Cane Varieties Grown in Argentina for Second-Generation Ethanol Production. Waste Biomass Valorization 2021 1–15. https://doi.org/10.1007/S12649-021-01528-5

[80] Khanbo S, Tangphatsornruang S, Piriyapongsa J, Wirojsirasak W, Punpee P, Klomsa-ard P, Ukoskit K (2021) Candidate gene association of gene expression data in sugarcane contrasting for sucrose content. Genomics 113:229–237

[81] Knaus BJ, Grünwald NJ (2017). "VCFR: a package to manipulate and visualize variant call format data in R." Molecular Ecology Resources, 17(1), 44–53. ISSN 757, http://dx.doi.org/10.1111/1755-0998.12549.

[82] Koseva BS, Macdonald SJ, Blumenstiel JP, Walters JR, Holder MT (2017) Application of RAD-seq in Evolutionary Genomics of Non-Model Organisms. University of Kansas

[83] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923

[84] Leal MV, Walter A, Seabra JA (2013) Sugarcane as an energy source. Biomass Convers Biorefin 3:17–26

[85] Li X, Liu Z, Zhao H, Deng X, Su Y, Li R, Chen B (2022) Overexpression of Sugarcane ScDIR Genes Enhances Drought Tolerance in Nicotiana benthamiana. Int J Mol Sci 23. https://doi.org/10.3390/IJMS23105340/S1

[86] Llerena JPP, Figueiredo R, dos Santos Brito M, Kiyota E, Mayer JLS, Araujo P, Schimpl FC, Dama M, Pauly M, Mazzafer P (2019) Deposition of lignin in four species of Saccharum. Sci. Rep. 9 (1), 5877. https://doi.org/10.1038/s41598-019- 42350-3.

[87] López de Heredia U (2016) Las técnicas de secuenciación masiva en el estudio de la diversidad biológica. Munibe Ciencias Nat 64. https://doi.org/10.21630/mcn.2016.64.07

[88] Mahadevaiah C, Appunu C, Aitken K, Suresha GS, Vignesh P, Mahadeva Swamy HK, Valarmathi R, Hemaprabha G, Alagarasan G, Ram B (2021) Genomic Selection in Sugarcane: Current Status and Future Prospects. Front Plant Sci 12:2019. https://doi.org/10.3389/FPLS.2021.708233/BIBTEX

[89] Manimekalai R, Suresh G, Govinda Kurup H, Athiappan S, Kandalam M (2020) Role of NGS and SNP genotyping methods in sugarcane improvement programs. Crit Rev Biotechnol 40:865–880. https://doi.org/10.1080/07388551.2020.1765730

[90] Margarido GRA, Souza AP, Garcia AAF (2007) OneMap: software for genetic mapping in outcrossing species. Hereditas 144:78–79. https://doi.org/10.1111/J.2007.0018-0661.02000.X

[91] Matos DA, Whitney IP, Harrington MJ, Hazen SP (2013) Cell Walls and the Developmental Anatomy of the Brachypodium distachyon Stem Internode. PLoS One 8:e80640. https://doi.org/10.1371/JOURNAL.PONE.0080640

[92] Matsuoka S, Kennedy AJ, Gustavo E, Santos D, Tomazela AL, Rubio LCS (2014) Energy Cane : Its Concept, Development, Characteristic and Prospects.

[93] McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, Kennedy M, Amirebrahimi M, Weers BD, McKinley B, Mattison A, Morishige DT, Grimwood J, Schmutz J, Mullet JE (2018) The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. Plant J 93:338–354. https://doi.org/10.1111/tpj.13781

[94] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F (2016) The Ensembl Variant Effect Predictor. Genome Biol 17:1–14.

[95] Medeiros C, Almeida Balsalobre TW, Carneiro MS (2020) Molecular diversity and genetic structure of Saccharum complex accessions. PLoS One 15:e0233211. https://doi.org/10.1371/JOURNAL.PONE.0233211

[96] Ming R, Wang YW, Draye X, Moore PH, Irvine JE, Paterson AH (2002) Molecular dissection of complex traits in autopolyploids: Mapping QTLs affecting sugar yield and related traits in sugarcane. Theor Appl Genet 105:332–345. https://doi.org/10.1007/S00122-001-0861-5/METRICS

[97] Ministerio de Hacienda. Secretaría de Política Económica. Informes de cadenas de valor - Azucarera. (2018) – Año 3 - N° 3. Disponible en: https://www.economia.gob.ar/peconomica/docs/SSPMicro_Cadenas_de_valor_Azucar.pdf

[98] Molina C, Aguirre NC, Vera PA, Filippi CV, Puebla AF, Marcucci Poltri SN, Paniego NB, Acevedo A (2022) ddRADseq-mediated detection of genetic variants in sugarcane. Plant Mol Biol 2022 1:1–15. https://doi.org/10.1007/S11103-022-01322-4

[99] Mollinari M, Garcia AAF (2019) Linkage Analysis and Haplotype Phasing in Experimental Autopolyploid Populations with High Ploidy Level Using Hidden Markov Models. G3 Genes|Genomes|Genetics 9:3297–3314. https://doi.org/10.1534/G3.119.400378

[100] Mollinari M, Olukolu BA, Pereira GS, Khan A, Gemenet D, Yencho GC, Zeng ZB. (2020) Unraveling the Hexaploid Sweetpotato Inheritance Using Ultra-Dense Multilocus Mapping. G3 Genes|Genomes|Genetics 10:1:281–292. https://doi.org/10.1534/g3.119.400620

[101] Moore PH, Botha FC, Sataloff RT, Johns MM, Kost KM (2014) Sugarcane: Physiology, Biochemistry, and Functional Biology. Wiley Blackwell

[102] Nayak SN, Song J, Villa A, Pathak B, Ayala-Silva T, Yang X, Todd J, Glynn NC, Kuhn DN, Glaz B, Gilbert RA, Comstock JC, Wang J (2014) Promoting Utilization of Saccharum spp. Genetic Resources through Genetic Diversity Analysis and Core Collection Construction. PLoS One 9:e110856. https://doi.org/10.1371/JOURNAL.PONE.0110856

[103] O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. Molecular Ecology, 27(16), 3193–3206. https://doi.org/10.1111/mec.14792

[104] Ooms J (2023). writexl: Export Data Frames to Excel 'xlsx' Format_. R package version 1.4.2. https://CRAN.R-project.org/package=writexl

[105] Ostengo S, Serino G, Perera MF. et al. (2021) Sugarcane Breeding, Germplasm Development and Supporting Genetic Research in Argentina. Sugar Tech. https://doi.org/10.1007/s12355-021-00999-z

[106] Ostengo S, Serino G, Perera MF. et al. (2021) Sugarcane Breeding, Germplasm Development and Supporting Genetic Research in Argentina. Sugar Tech. https://doi.org/10.1007/s12355-021-00999-z

[107] Palacio FX, Apodaca MJ, Crisci J V. (2020) Análisis multivariado para datos biológicos

[108] Paradis E (2010). "pegas: an R package for population genetics with an integrated–modular approach." Bioinformatics, 26, 419-420.

[109] Parida SK, Kalia S, Pandit A, Nayak P, Singh RK, Gaikwad K, Srivastava PS, Singh NK, Mohapatra T (2016) Single nucleotide polymorphism in sugar pathway and disease resistance genes in sugarcane. Plant Cell Rep 35:1629–1653. https://doi.org/10.1007/S00299-016-1978-Y/FIGURES/3

[110] Paterson PH, Moore PH, Tew TL (2012) The Gene Pool of Saccharum Species and Their Improvement. doi: 10.1007/978-1-4419-5947-8_3

[111] Pereira GS, Garcia AAF, Margarido GRA (2018) A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids. BMC Bioinformatics 19:1–10. https://doi.org/10.1186/s12859-018-2433-6

[112] Perera MF, Ostengo S, Malavera ANP, Balsalobre TWA, Onorato GD, Noguera AS, Hoffmann HP, Carneiro MS (2023) Genetic diversity and population structure of Saccharum hybrids. PLoS One 18:e0289504. https://doi.org/10.1371/journal.pone.028950

[113] Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model

[114] Pimenta RJG, Aono AH, Burbano RCV, Coutinho AE, da Silva CC, dos Anjos IA, Perecin D, Landell MG de A, Gonçalves MC, Pinto LR, de Souza AP, Gonzaga Pimenta RJ, Aono AH, Carlos R, Burbano V, Coutinho AE, Cristina Da Silva C, Antônio Dos Anjos I, Perecin D, Guimarães De Andrade Landell M, Gonçalves MC, Pinto LR, Pereira De Souza A (2021) Genome-wide approaches for the identification of markers and genes associated with sugarcane yellow leaf virus resistance. Sci Reports | 11:15730. https://doi.org/10.1038/s41598-021-95116-1

[115] Poelking VGC, Giordano A, Ricci-Silva ME, Williams TCR, Peçanha DA, Ventrella MC, Rencoret J, Ralph J, Barbosa MHP, Loureiro M (2015) Analysis of a modern hybrid and an ancient sugarcane implicates a complex interplay of factors in affecting recalcitrance to cellulosic ethanol production. PLOS One 10 (8). https://doi.org/10.1371/journal.pone.0134964.

[116] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[117] R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

[118] Raboin LM, Pauquet J, Butterfield M, D'Hont A, Glaszmann JC (2008) Analysis of genome-wide linkage disequilibrium in the highly polyploid sugarcane. Theor Appl Genet 116:701–714. https://doi.org/10.1007/S00122-007-0703-1/FIGURES/7

[119] Rago AM, Pérez Gómez SG, Fontana PD (2012) Caña de Azúcar. Identificación y manejo de las enfermedades en Argentina. Ediciones INTA. 52 pp. ISBN 978-987-679-118.

[120] Ravinet M, Meier J. Speciation & Population Genomics: a how-to-guide. Physalia Courses. February 2020. Available: https://speciationgenomics.github.io/filtering_vcfs/species. PLoS One 7. https://doi.org/10.1371/journal.pone.0037135

[121] Rochette NC, Catchen JM (2017) Deriving genotypes from RAD-seq short-read data using Stacks. Nat Protoc 12:2640–2659. https://doi.org/10.1038/nprot.2017.123

[122] Rochette NC, Rivera-Colón AG, Catchen JM (2019) Stacks 2: Analytical methods for paired end sequencing improve RADseq-based population genomics. Mol Ecol 28:4737–4754. https://doi.org/10.1111/mec.15253

[123] Sandhu HS, Comstock JC, Gilbert RA, Gordon VS, Korndörfer P, El-Hout N, Arundale RA (2015) Registration of 'UFCP 78–1013' sugarcane cultivar. J. Plant Regist. 9 (3), 318–324. https://doi.org/10.3198/jpr2014.08.0050crc.

[124] Sandhu HS, Gilbert RA, Comstock JC, Gordon VS, Korndörfer P, Arundale RA, El-Hout N (2016) Registration of 'UFCP 82–1655' sugarcane. J. Plant Regist. 10 (1), 22–27. https://doi.org/10.3198/jpr2014.10.0074crc.

[125] Sandhu HS, Gilbert RA, Comstock JC, Gordon VS, Korndörfer P, Arundale RA, El-Hout N (2016) Registration of 'UFCP 82-1655' Sugarcane. Journal of Plant Registrations 10 (1): 22–27. doi:10.3198/jpr2014.10.0074crc.

[126] Sandhu KS, Shiv A, Kaur G, Meena MR, Raja AK, Vengavasi K, Mall AK, Kumar S, Singh PK, Singh J, Hemaprabha G, Pathak AD, Krishnappa G, Kumar S (2022) Integrated Approach in Genomic Selection to Accelerate Genetic Gain in Sugarcane. Plants 2022, Vol 11, Page 2139 11:2139. https://doi.org/10.3390/PLANTS11162139

[127] Santchurn D, Badaloo M, Zho M, Labuschagne MT (2019) Genetic studies on sugar and fibre accumulation patterns among different types of energy canes in Mauritius. Sugar Tech 21, 879-890.

[128] Santiago R, Quintana J, Rodríguez S, Díaz EM, Legaz ME, Vicente C (2010) On the nature and distribution of sclereids in sugarcane leaves. Pakistan J Bot 42:2867–2881

[129] Serang O, Mollinari M, Garcia AAF (2012) Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. PLoS One 7:1–13. https://doi.org/10.1371/journal.pone.0030906

[130] Shearman JR, Pootakham W, Sonthirod C, Naktang C, Yoocha T, Sangsrakru D, Jomchai N, Tongsima S, Piriyapongsa J, Ngamphiw C, Wanasen N, Ukoskit K, Punpee P, Klomsa-ard P, Sriroth K, Zhang J, Zhang X, Ming R, Tragoonrung S, Tangphatsornruang S (2022) A draft chromosome-scale genome assembly of a commercial sugarcane. Sci Reports 2022 121 12:1–8. https://doi.org/10.1038/s41598-022-24823-0

[131] Smith MR (2019) Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets. Biology Letters, 15(2): 20180632.

[132] Soares NR, Mollinari M, Oliveira GK, Pereira GS, Lucia M, Vieira C (2021) Meiosis in Polyploids and Implications for Genetic Mapping: A Review. https://doi.org/10.3390/genes12101517

[133] Song J, Yang X, Resende MFR Jr, Neves LG, Todd J, Zhang J, Comstock JC and Wang J (2016) Natural Allelic Variations in Highly Polyploidy Saccharum Complex. Front. Plant Sci. 7:804. doi: 10.3389/fpls.2016.00804

[134] Sopena R, Felipe A, Rago A, Erazzú LE, Mariotti, et al. (2015) Nuevas Variedades de Caña de Azúcar Desarrolladas por el INTA Famaillá. URL: https://repositorio.inta.gob.ar/handle/20.500.12123/13652

[135] Souza GM, Van Sluys M-AA, Lembke CG, Lee H, Margarido GRA, Hotta CT, Gaiarsa JW, Diniz AL, Oliveira MDM, Ferreira SS de S, Nishiyama MY, Ten-Caten F, Ragagnin GT, Andrade PDM, De Souza RF, Nicastro GG, Pandya R, Kim C, Guo H, Durham AM, Carneiro MS, Zhang J, Zhang X, Zhang Q, Ming R, Schatz MC, Davidson B, Paterson AH, Heckerman D (2019) Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. Gigascience 8:1–18. https://doi.org/10.1093/gigascience/giz129Spurlock D, Phelan J (2019) Accuracy and completeness are required: A response to Dreher et al (2019). Nursing Forum, doi: 10.1111/NUF.12398

[136] Stoler N, Nekrutenko A (2021) Sequencing error profiles of Illumina sequencing instruments. NAR Genomics Bioinforma 3:1–9. https://doi.org/10.1093/nargab/lqab019

[137] sugarcane. Nature. https://doi.org/10.1038/s41586-024-07231-4

[138] Takano S, Matsuda S, Funabiki A, Furukawa J, Yamauchi T, Tokuji Y, Nakazono M, Shinohara Y, Takamure I, Kato K (2015) The rice RCN11 gene encodes β1,2-xylosyltransferase and is required for plant responses to abiotic stresses and phytohormones. Plant Sci. 236:75–88.

[139] Taniguti CH, Taniguti LM, Amadeu RR, Lau J, De Siqueira Gesteira G, De T, Oliveira P, Ferreira GC, Da G, Pereira S, Byrne D, Mollinari M, Riera-Lizarazu O, Franco Garcia AA (2022) Developing best practices for genotyping-by-sequencing analysis using linkage maps as benchmarks. bioRxiv 2022.11.24.517847. https://doi.org/10.1101/2022.11.24.517847

[140] Tao S, Khanizadeh S, Zhang H, Zhang S (2009) Anatomy, ultrastructure and lignin distribution of stone cells in two Pyrus species. Plant Sci 176:413–419. https://doi.org/10.1016/j.plantsci.2008.12.011

[141] Thiappan Selvi A, Nair NV (2009). Molecular Breeding in Sugarcane. International Journal of Agriculture, Enviornment and Biotechnology 3(1):115-127. Print ISSN : 0974-1712. Online ISSN : 2230-732X.

[142] Thirugnanasambandam PP, Hoang N V., Henry RJ (2018) The challenge of analyzing the sugarcane genome. Front Plant Sci 9:1–18. https://doi.org/10.3389/fpls.2018.00616

[143] Trujillo-Montenegro JH, Rodríguez Cubillos MJ, Loaiza CD, Quintero M, Espitia-Navarro HF, Salazar Villareal FA, Viveros Valens CA, González Barrios AF, De Vega J, Duitama J, Riascos JJ (2021) Unraveling the Genome of a High Yielding Colombian Sugarcane Hybrid. Front Plant Sci 12. https://doi.org/10.3389/FPLS.2021.694859

[144] Valarmathi R (2021) Anatomy of Tolerance Mechanisms in Sugarcane Crop to Abiotic Stresses. 107-121. doi: 10.1007/978-981-19-3955-6_6

[145] Vermerris, W. (2011). Survey of Genomics Approaches to Improve Bioenergy Traits in Maize, Sorghum and Sugarcane. Journal of Integrative Plant Biology, 53(2), 105–119. https://doi.org/10.1111/j.1744-7909.2010.01020.x

[146] Vieira MLC, Almeida CB, Oliveira CA, Tacuatiá LO, Munhoz CF, Cauz-Santos LA, Pinto LR, Monteiro-Vitorello CB, Xavier MA, Forni-Martins ER (2018) Revisiting meiosis in sugarcane: Chromosomal irregularities and the prevalence of bivalent configurations. Front Genet 9:1–12. https://doi.org/10.3389/fgene.2018.00213

[147] Vining KJ, Salinas N, Tennessen JA, Zurn JD, Sargent DJ, Hancock J, Bassil N V. (2017) Genotyping-by-sequencing enables linkage mapping in three octoploid cultivated strawberry families. PeerJ 2017:e3731. https://doi.org/10.7717/peerj.3731

[148] Walsh KB, Sky RC, Brown SM (2005) The anatomy of the pathway of sucrose unloading within the sugarcane stalk. Funct Plant Biol 32:367–374. https://doi.org/10.1071/FP04102

[149] Walton J (2020) The 5 Countries That Produce the Most Sugar. In: 5 Ctries. that Prod. Most sugar. https://www.investopedia.com/articles/investing/101615/5-countries-produce-most-sugar.asp. Accessed 22 Jan 2021

[150] Wei T, Simko V (2021). R package 'corrplot': Visualization of a Correlation Matrix. (Version 0.92), https://github.com/taiyun/corrplot.

[151] Wei X, Jackson P, Stringer J, Cox M (2008) Relative economic genetic value (rEGV)—an improved selection index to replace net merit grade (NMG) in the Australian sugarcane variety improvement program. Proc Aust Soc Sugar Cane Technol 30: 174–181

[152] Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." _Journal of Open Source Software_, *4*(43), 1686. doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686

[153] Wickham H, Bryan J (2023). readxl: Read Excel Files_. R package version 1.4.3 https://CRAN.R-project.org/package=readxl

[154] Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. https://ggplot2.tidyverse.org

[155] Wu R, Ma CX, Painter I, Zeng ZB (2002) Simultaneous Maximum Likelihood Estimation of Linkage and Linkage Phases in Outcrossing Species. Theor Popul Biol 61:349–363. https://doi.org/10.1006/tpbi.2002.1577

[156] Xiong H, Chen Y, Pan YB, Shi A (2023) A Genome-Wide Association Study and Genomic Prediction for Fiber and Sucrose Contents in a Mapping Population of LCP 85-384 Sugarcane. Plants 12:1041. https://doi.org/10.3390/PLANTS12051041/S1

[157] Xue Y, Zou C, Zhang C, Yu H, Chen B, Wang H (2022) Dynamic DNA methylation changes reveal tissue-specific gene expression in sugarcane. Front Plant Sci 13:1–15. https://doi.org/10.3389/fpls.2022.1036764

[158] Yadav S, Jackson P, Wei X, Ross EM, Aitken K, Deomano E, Atkin F, Hayes BJ, Voss-Fels KP (2020) Accelerating Genetic Gain in Sugarcane Breeding Using Genomic Selection. Agron 2020, Vol 10, Page 585 10:585. https://doi.org/10.3390/AGRONOMY10040585

[159] Yadav S, Wei X, Joyce P, Atkin F, Deomano E, Sun Y, Nguyen LT, Ross EM, Cavallaro T, Aitken KS, Hayes BJ, Voss-Fels KP (2021) Improved genomic prediction of clonal performance in sugarcane by exploiting non-additive genetic effects. Theor Appl Genet 134:2235–2252. https://doi.org/10.1007/S00122-021-03822-1/TABLES/4

[160] Yang X, Kandel R, Song J, You Q, Wang M, Wang J (2018) Sugarcane genome sequencing and genetic mapping. 3–34. https://doi.org/10.19103/as.2017.0035.02

[161] Yang X, Pengcheng L, Zefeng Y, Chenwu X (2017a) Genetic mapping of quantitative trait loci in crops. Crop J 5:175–184. https://doi.org/10.1016/j.cj.2016.06.003

[162] Yang X, Song J, You Q, Paudel DR, Zhang J, Wang J (2017b) Mining sequence variations in representative polyploid sugarcane germplasm accessions. BMC Genomics 18:1–16. https://doi.org/10.1186/s12864-017-3980-3

[163] Yang X, Sood S, Glynn N, Islam MS, Comstock J, Wang J (2017c) Constructing high-density genetic maps for polyploid sugarcane (Saccharum spp.) and identifying quantitative trait loci controlling brown rust resistance. Mol Breed 37:1–12. https://doi.org/10.1007/S11032-017-0716-7/TABLES/3

[164] Yates AD, Allen J, Amode RM, Azov AG, Barba M, Becerra A, Bhai J, Campbell LI, Carbajo Martinez M, Chakiachvili M, Chougule K, Christensen M, Contreras-Moreira B, Cuzick A, Da Rin Fioretto L, Davis P, De Silva NH, Diamantakis S, Dyer S, Elser J, Filippi C V., Gall A, Grigoriadis D, Guijarro-Clarke C, Gupta P, Hammond-Kosack KE, Howe KL, Jaiswal P, Kaikala V, Kumar V, Kumari S, Langridge N, Le T, Luypaert M, Maslen GL, Maurel T, Moore B, Muffato M, Mushtaq A, Naamati G, Naithani S, Olson A, Parker A, Paulini M, Pedro H, Perry E, Preece J, Quinton-Tulloch M, Rodgers F, Rosello M, Ruffier M, Seager J, Sitnik V, Szpak M, Tate J, Tello-Ruiz MK, Trevanion SJ, Urban M, Ware D, Wei S, Williams G, Winterbottom A, Zarowiecki M, Finn RD, Flicek P (2022) Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. Nucleic Acids Res 50:D996–D1003. https://doi.org/10.1093/NAR/GKAB1007

[165] Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, Yuan X, Zhu M, Zhao S, Li X, Liu X. (2021) rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated tool for Genome-wide Association Study. Genomics Proteomics Bioinformatics. Mar 2:S1672-0229(21)00050-4. doi: 10.1016/j.gpb.2020.10.007. Epub ahead of print. PMID: 33662620

[166] You Q, Yang X, Peng Z, Islam MS, Sood S, Luo Z, Comstock J, Xu L, Wang J (2019) Development of an Axiom Sugarcane100K SNP array for genetic map construction and QTL identification. Theor Appl Genet 132:2829–2845. https://doi.org/10.1007/S00122-019-03391-4/FIGURES/4

[167] Zhang J, Zhang Q, Li L, Tang H, Zhang Q, Chen Y, Arrow J, Zhang X, Wang A, Miao C, Ming R (2019) Recent polyploidization events in three Saccharum founding species. Plant Biotechnol J 17:264–274. https://doi.org/10.1111/pbi.12962

[168] Zhang L, Prabhakar PK, Bharadwaj VS, Bomble YJ, Peña MJ, Urbanowicz BR (2023) Glycosyltransferase family 47 (GT47) proteins in plants and animals. Essays Biochem 67:639–652. https://doi.org/10.1042/EBC20220152

[169] Zhang Q, Qi Y, Pan H, Tang H, Wang G, Hua X, Wang Y, Lin L, Li Z, Li Y, Yu F, Yu Z, Huang Y, Wang T, Ma P, Dou M, Sun Z, Wang Y, Wang H, Zhang X, Yao W, Wang Y, Liu X, Wang M, Wang J, Deng Z, Xu J, Yang Q, Liu Z, Chen B, Zhang M, Ming R, Zhang J (2022) Genomic insights into the recent

chromosome reduction of autopolyploid sugarcane Saccharum spontaneum. Nat Genet 2022 546 54:885–896. https://doi.org/10.1038/s41588-022-01084-1

[170] Zhong R, Cui D, Ye ZH (2019) Secondary cell wall biosynthesis. New Phytol 221:1703–1723. https://doi.org/10.1111/nph.15537

Apéndice Capítulo 1

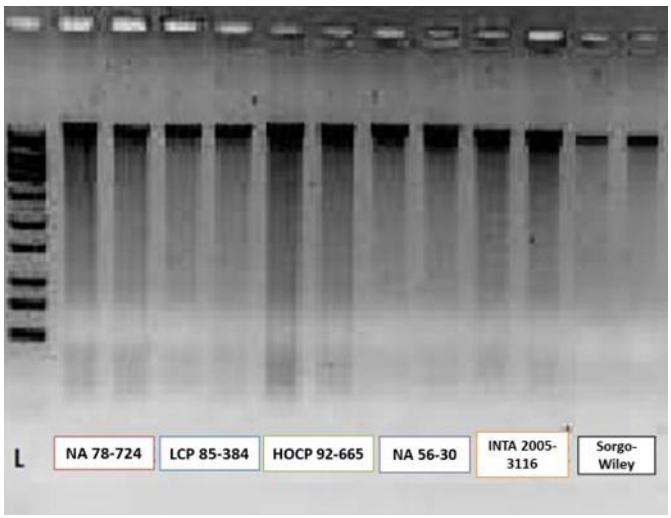

Figura 1.1. *Gel de agarosa con ADN genómico de los genotipos analizados*. L: marcador de peso molecular (1kb). 5 µl de muestra fueron sembrados en cada calle. Distintas intensidades se deben a diferentes concentraciones (Tabla 1).

Tabla 1.1. *Cuantificación de la concentración* e *índices de pureza de ADN genómico.* Contenido de ácido nucleicos medidos con Nanodrop ((ng/ul)). Relaciones de los índices de absorbancia de las muestras analizadas (A260/A280 y A260/A230) y contenido de ADN medido con Qubit (ng/ul).

Genotipo	Nanodrop (ng/ul)	Absorbancia 260/280	Absorbancia 260/230	QUIBIT (ng/ul)
NA 78-724	181,78	1,91	2,39	20
LCP 85-384	169,68	1,94	2,39	11
HOCP 92-665	181,42	1,9	2,46	29,3
NA 56-30	136,41	1,99	2,49	10,4
INTA 05-3116	163,93	1,91	2,31	18,7

Tabla 1.2. *Número de lecturas por muestra para dos plataformas de secuenciación Illumina.* Número de lecturas por muestra para la secuenciación NextSeq y NovaSeq y la relación de entre dichas plataformas por genotipo.

Genotipo	*NextSeq*	*NovaSeq*	*NovaSeq: NextSeq*
NA 78-724	3.944.615	29057716	7,4
LCP 85-384	3.437.305	24093598	7,0
HOCP 92-665	2.943.710	20986808	7,1
NA 56-30	2.658.076	18674372	7,0
INTA 05-3116	3.168.549	23580998	7,4

Tabla 1.3. *Análisis funcional in silico.* Consecuencias de los SNPs para a. Ne-70 FS 1 y b. No-150 FS 5

a. Ne-70 FS1		**Variantes procesadas**	9,094	**Overlapped genes**	6,618
Más severas		*Todas las consecuencias*		*Región Codificante*	
Tipo de consecuencias	cantidad	**Tipo de consecuencias**	cantidad	**Tipo de consecuencias**	cantidad
splice_donor_variant	3	splice_donor_variant	3	stop_gained	5
splice_acceptor_variant	4	splice_acceptor_variant	4	stop_lost	4
stop_gained	5	stop_gained	5	start_lost	5
stop_lost	4	stop_lost	4	missense_variant	894
start_lost	5	start_lost	5	synonymous_variant	868
missense_variant	894	missense_variant	894		
splice_region_variant	67	splice_region_variant	76		
synonymous_variant	861	synonymous_variant	868		
5_prime_UTR_variant	69	5_prime_UTR_variant	69		
3_prime_UTR_variant	147	3_prime_UTR_variant	148		
intron_variant	1,255	intron_variant	1,314		
upstream_gene_variant	2,179	upstream_gene_variant	3,889		
downstream_gene_variant	1,531	downstream_gene_variant	4,526		
intergenic_variant	2,070	intergenic_variant	2,070		

b. No-150 FS5		**Variants processed**	47,964	**Overlapped genes**	11,764
Más severas		*Todas las consecuencias*		*Región Codificante*	
Tipo de consecuencias	cantidad	**Tipo de consecuencias**	cantidad	**Tipo de consecuencias**	cantidad
splice_acceptor_variant	9	splice_acceptor_variant	9	stop_gained	55
splice_donor_variant	16	splice_donor_variant	16	stop_lost	14
stop_gained	55	stop_gained	55	start_lost	19
stop_lost	14	stop_lost	14	missense_variant	4,071
start_lost	19	start_lost	19	stop_retained_variant	11
missense_variant	4,071	missense_variant	4,071	synonymous_variant	3,586
splice_region_variant	446	splice_region_variant	504		
stop_retained_variant	7	stop_retained_variant	11		
synonymous_variant	3,541	synonymous_variant	3,586		
5_prime_UTR_variant	481	5_prime_UTR_variant	486		
3_prime_UTR_variant	1,014	3_prime_UTR_variant	1,018		
intron_variant	9,270	intron_variant	9,662		
upstream_gene_variant	11,428	upstream_gene_variant	21,216		
downstream_gene_variant	7,759	downstream_gene_variant	25,306		
intergenic_variant	9,834	intergenic_variant	9,834		

Tabla 1.4. Largo por cromosoma de los genomas de referencia publicados en la base de datos del NCBI.

Largo de cromosoma (bp)	Saccharum hybrid cv R570	*S. hybrid* cv CC_01_1940	*S. officinarum cv LA-purple* *	*S. officinarum* cv LA-purple **	*S. spontaneum cv Np-X* *	*S. spontaneum cv Np-X* ***	*Sorghum bicolor* cv BTx623
1	72,286,729	88,339,610	1,009,619,034	126,202,379	386,138,825	96,534,706	80,884,392
2	52,638,177	65,155,311	813,567,220	101,695,903	314,321,578	78,580,395	77,742,459
3	54,887,512	70,187,887	704,817,492	88,102,187	319,972,482	79,993,121	74,386,277
4	45,576,573	55,866,964	709,685,400	88,710,675	273,307,304	68,326,826	68,658,214
5	23,282,433	34,980,033	632,133,836	79,016,730	288,376,667	72,094,167	71,854,669
6	35,411,752	39,115,154	505,128,628	63,141,079	243,136,612	60,784,153	61,277,060
7	30,096,520	35,448,535	551,872,265	68,984,033	220,037,474	55,009,369	65,505,356
8	24,320,556	32,638,014	481,975,548	60,246,944	200,069,948	50,017,487	62,686,529
9	36,579,524	36,164,970	586,230,261	73,278,783	237,609,866	59,402,467	59,416,394
10	34,651,920	41,117,929	501,742,417	62,717,802	243,887,692	60,971,923	61,233,695

*Suma de las 8 copias de cada cromosoma (x=8)

**promedio de las 8 copias de cada cromosoma (x=8)

***promedio de las 4 copias de cada cromosoma (x=4)

Apéndice Capítulo 2

Tabla 2.1. *Cuantificación de la concentración* e *índices de pureza de ADN genómico.* Contenido de ácido nucleicos medidos con Nanodrop (ng/ul). Relaciones de los índices de absorbancia de las muestras analizadas (A260/A280 y A260/A230) y contenido de ADN medido con Qubit (ng/ul).

Muestra	Nanodrop (ng/ul)	Absorbancia 260/280	Absorbancia 260/230	Qubit (ng/ul)
83	183,4	1,84	1,73	43,3
84	258,50	1,93	2,35	66,8
85	176,80	1,85	1,79	54,4
86	142,80	1,72	1,26	74,3
87	129,7	1,81	1,42	33,8
88				20,7
89	169,8	1,88	1,67	22
90	56,5	1,87	1,38	16,3
91	186,80	1,81	1,59	74,7
92	81,9	1,89	1,89	18,1
93	152,70	1,86	2,32	88,9
94	89,30	1,84	1,84	23,1
95	103,40	1,86	1,78	23,2
96				30,9
97	101,70	1,94	2,02	25,3
98	58,10	1,96	1,87	26,7
99	85,60	1,75	1,59	55,5
100	76,00	1,90	1,80	19,7
101	131,2	1,74	1,25	47,6
102	76,40	1,82	1,42	32,5
103	65,0	1,67	1,04	36,8
104	109,90	1,85	1,41	24,0
105				34,1
106	76,60	1,86	1,94	36,9
107	88,30	1,80	1,31	17,0
108	102,30	1,76	1,09	45,5
109	146,60	0,63	0,64	25,1
110	66,50	1,82	1,68	23,5
111	67,5	1,76	1,50	19,8
112	151,0	1,93	1,98	20,6
113	95,80	1,70	1,14	57,2
114	85,1	1,67	0,97	40,3
115	102,70	1,82	1,40	49,8
116	111,6	1,82	1,21	10
117	65,7	1,70	1,29	43,1

Tabla 2.1. *Continuación*

Muestra	Nanodrop (ng/ul)	Absorbancia 260/280	Absorbancia 260/230	Qubit (ng/ul)
118	**217,5**	**1,88**	**1,96**	**58,4**
119	107,10	1,83	1,50	70,3
120	117,2	1,93	2,90	14,4
121	99,1	1,76	1,12	14,2
122				25,8
123	72,90	1,84	1,59	18,1
124	72,90	1,77	1,30	30,2
125	178,40	1,90	2,18	51,1
126	125,2	1,85	1,95	25,7
127	106,5	1,75	1,34	50,1
128	71,3	1,66	0,92	43,8
129	54,20	1,84	1,26	26,5
130	60,40	1,89	1,48	30,7
131	78,7	1,69	0,83	26,7
132	74,10	1,84	1,32	24,6
133	74,8	1,63	0,89	60,7
134	160,4	1,81	1,87	74,2
135	68,90	1,82	1,42	25,1
136	114,5	1,81	1,51	17,9
137				40,7
138	56,5	1,89	1,85	9,82
139	106,50	1,87	2,04	48,6
140	85,80	1,92	1,93	27,0
141	57,90	1,92	1,88	11,0
142	83,80	1,98	1,94	14,7
143	25,2	1,13	0,27	18,2
144	152,3	1,73	1,34	69,2
145	104,80	1,93	1,94	17,1
146	159,3	1,88	1,65	10,4
147	65,8	1,94	1,64	16,2
148	125,9	1,73	1,40	74,6
149	48,4	1,83	1,39	11,9
150	42,6	1,70	1,19	25
151	53,30	1,93	1,78	11,0
152	88,30	1,70	1,08	54,2

Tabla 2.1. *Continuación*

Muestra	Nanodrop (ng/ul)	Absorbancia 260/280	Absorbancia 260/230	Qubit (ng/ul)
153	92,70	1,86	1,59	13,8
154	135,9	1,80	1,46	69,7
155	133,60	1,88	1,90	40,9
156	80,70	1,89	2,14	29,5
157	68,7	1,84	1,60	8,04
158	64,90	1,89	1,77	21,0
159	62,40	1,92	1,73	27,0
160	**67,5**	**1,81**	**1,23**	**10,4**
161	56,70	1,93	1,85	15,6
162	108,40	1,65	1,03	50,2
163	130,3	1,82	1,37	40,6
164	57,20	1,75	1,21	25,7
165	176,2	1,81	1,68	117
167	102,3	1,86	1,64	66
168	105,60	1,84	1,72	64,2
169	109,9	1,89	1,83	12,9
170				21,4
171	40,00	1,44	0,47	19,6
172	84,00	1,79	1,59	53,3
174	72,4	1,75	1,58	43,8
LCP 85-384				36,4
NA 78-724				29,4

Tabla 2.2. *Datos analíticos del mapa de ligamiento poliploide.* a) número de marcadores por cromosoma; b) número de marcadores por grupo homólogo (HG).

a)

Cromosoma	N° marcadores
1	123
2	59
3	64
4	69
5	27
6	52
7	40
8	36
9	44
10	54
	568

b)

HG	N° marcadores
1	6
2	4
3	4
4	7
5	4
6	4
7	4
8	6
9	6
10	6
	51

Apéndice Capítulo 3

Tabla 3.1. Medias, desvío estándar e intervalos de confianza (IC) inferior y superior de LCP 85-384 (cultivar comercial) e INTA 05-3116 (biotipo caña energía) en cuatro estadios de desarrollo para las variables largo y ancho de haces vasculares en micrómetro (µm), diámetro de metaxilema en micrómetro (µm) y número de haces por unidad de medición. HV= haces vasculares; SD= desvío estándar; IC= intervalos de confianza.

Genotipo	*Estadio*	*IP*	*Largo HV*	*Largo HV SD*	*IC inferior*	*IC superior*	*Ancho HV*	*Ancho HV SD*	*IC inferior*	*IC superior*	*Diámetro de Metaxilema*	*Diámetro de Metaxilema SD*	*IC inferior*	*IC superior*	*Número de HV por unidad*	*Número de HV por unidad SD*
LCP 85-384	*Macollaje*	1	418	28.7	292.6	543.4	387.9	12.5	334.2	441.6	125.3	1.7	118.1	132.4	24.64	3.2
	Gran crecimiento	1	428.9	14.2	390.0	467.7	378.4	5.9	362.0	394.8	116	2.5	109.9	122.4	30.97	0.79
	Gran crecimiento	5	375.7	14	336.4	415.0	375.8	5.6	360.1	391.4	119.5	2.4	114.2	126.3	31.62	0.81
	Maduración temprana	1	419.3	18.4	378.0	460.6	374.4	10	352.2	396.6	120	3.3	112.7	127.2	40.91	5.89
	Maduración temprana	5	359.2	18.3	318.0	400.4	370.6	9.8	348.6	392.5	124.4	3.2	117.3	131.6	36.42	5.25
	Maduración temprana	10	350.6	18.1	309.6	391.6	372.8	9.6	351.3	394.2	126.8	3.1	119.8	133.7	35.46	5.52
	Maduración temprana	15	384.3	13.2	355.0	413.7	390.2	7.2	374.2	406.3	125.6	2.4	120.3	130.9	39.63	6.14
	Maduración tardía	1	381.9	14.6	349.2	414.6	364.9	7.2	348.9	380.8	109.8	2.9	102.7	116.2	19.92	2.4
	Maduración tardía	5	384.2	15.3	350.8	417.6	385.5	8.1	367.4	403.6	119.7	3.1	113.9	128.4	19.68	1.88
	Maduración tardía	10	394.5	14.6	361.8	427.3	405.4	7.3	389.1	421.8	129.5	2.9	122.6	136.3	23.05	1.99
	Maduración tardía	15	418.1	14.6	385.4	450.8	420.2	7.3	403.9	436.6	129.2	2.9	122.8	136.5	25.27	2.16
	Maduración tardía	20	434.3	14.2	401.9	466.7	402.9	6.7	387.9	417.8	124	2.8	114.8	128.1	29.48	2.51
INTA 05-3116	*Macollaje*	1	481.3	28.9	358.9	603.8	348.1	11.9	310.4	385.9	290.8	15.3	95.3	105.9	45.05	5.73
	Gran crecimiento	1	404.1	14.2	365.3	442.9	340.8	5.5	327.8	353.9	106.5	2.2	101.7	111.2	24.92	0.73
	Gran crecimiento	5	390.1	13.8	350.3	429.9	343.6	5.2	329.3	358.0	111.1	2.2	105.2	116.0	25.45	0.74
	Maduración temprana	1	536.8	19	494.9	578.7	362.9	9.7	342.5	383.3	107.2	3.1	100.6	113.7	23.35	2.87
	Maduración temprana	5	394.5	18.6	353.0	435.9	350.1	9.5	329.0	371.3	110.7	3.1	103.9	117.5	20.78	2.57
	Maduración temprana	10	372.7	18.1	331.7	413.7	327.5	9.2	307.0	348.1	108.8	3	102.3	115.5	20.23	3.15
	Maduración temprana	15	393.6	19.7	350.8	436.5	338.4	10.1	315.9	360.9	118.4	3.3	111.0	125.7	22.61	3.51
	Maduración tardía	1	474.6	14.7	441.7	507.4	362.1	6	349.5	374.6	109.3	2.8	102.6	115.0	13.88	1.69
	Maduración tardía	5	415.6	15.8	381.6	449.6	351.9	6.7	336.8	366.9	112	3	103.9	118.0	13.71	1.45
	Maduración tardía	10	396.3	14.4	363.8	428.9	347.3	5.8	334.4	360.2	116.9	2.7	109.7	122.7	16.06	1.39
	Maduración tardía	15	391.5	14.8	358.6	424.3	349.7	6	336.2	363.1	113.5	2.8	105.0	118.2	17.61	1.52
	Maduración tardía	20	407.1	14.4	374.5	439.6	342.5	5.8	329.6	355.5	102.1	2.7	94.6	107.5	20.54	1.76

Tabla 3.2. Medias, desvío estándar e intervalos de confianza (IC) inferior y superior de LCP 85-384 (cultivar comercial) e INTA 05-3116 (biotipo de caña energía) para los entrenudos 1,5, 10, 15 y 20 para las variables de largo y ancho de haces vasculares en micrómetro (µm), diámetro de metaxilema en micrómetro (µm) y número de haces por unidad de medición. HV= haces vasculares; SD= desvío estándar; IC= intervalos de confianza.

Genotipo	*Estadio*	*IP*	*Largo HV*	*Largo HV SD*	*IC inferior*	*IC superior*	*Ancho HV*	*Ancho HV SD*	*IC inferior*	*IC superior*	*Diámetro de Metaxilema*	*Diámetro de Metaxilema SD*	*IC inferior*	*IC superior*	*Número de HV por unidad*	*Número de HV por unidad SD*
LCP 85-384	*Macollaje*	1	418.1	17.4	377.9	458.2	388.4	8.4	369.1	407.7	123.5	2.8	117.1	129.9	23.1	3.1
	Gran crecimiento	1	428.9	18.0	387.3	470.5	378.4	8.6	358.6	398.2	116.0	2.9	109.4	122.7	30.8	0.75
	Maduración temprana	1	419.3	17.4	379.1	459.6	374.5	8.2	355.6	393.4	119.6	2.8	113.2	126.0	37.3	0.83
	Maduración tardía	1	394.8	17.7	353.9	435.6	365.8	8.3	346.5	385.0	110.3	2.8	103.8	116.7	18.2	3.89
	Gran crecimiento	5	375.8	11.7	347.1	404.4	375.8	7.0	358.7	392.8	119.4	2.9	112.3	126.5	31.8	5.3
	Maduración temprana	5	362.6	11.4	334.7	390.5	373.8	6.7	357.2	390.3	123.9	2.8	116.9	130.8	38.6	4.9
	Maduración tardía	5	383.9	12.6	353.0	414.8	385.8	7.6	367.3	404.3	119.5	3.1	112.0	127.0	19.4	6.1
	Maduración temprana	10	353.7	9.2	328.1	379.2	375.6	8.6	351.6	399.6	125.2	2.9	117.3	133.2	38.4	2.6
	Maduración tardía	10	394.6	9.9	367.0	422.1	405.4	9.3	379.6	431.1	128.1	3.0	119.7	136.6	23.0	1.9
	Maduración temprana	15	384.3	16.0	345.1	423.5	390.2	6.9	373.3	407.2	123.8	1.2	120.9	126.8	38.3	2.1
	Maduración tardía	15	418.0	22.6	362.7	473.4	420.2	9.8	396.2	444.1	129.0	1.7	124.9	133.1	26.9	2.2
	Maduración tardía	20	434.2	13.5	376.0	492.5	407.5	4.8	386.8	428.2	124.2	1.3	118.5	129.8	28.8	2.3
INTA 05-3116	*Macollaje*	1	466.4	18.0	428.0	504.7	351.1	7.8	334.5	367.8	102.2	2.6	96.7	107.6	44.6	5.2
	Gran crecimiento	1	412.5	18.1	370.7	454.3	342.1	8.0	323.7	360.5	106.4	2.6	100.5	112.4	25.2	0.9
	Maduración temprana	1	525.5	19.0	481.7	569.4	363.1	8.2	344.2	382.0	107.0	2.7	100.9	113.1	24.5	0.8
	Maduración tardía	1	474.5	18.2	432.5	516.5	359.6	8.0	341.2	378.0	109.4	2.6	103.4	115.3	15.3	2.3
	Gran crecimiento	5	390.1	11.2	365.4	414.7	343.9	6.6	329.5	358.4	110.7	2.7	104.7	116.7	25.2	2.7
	Maduración temprana	5	394.4	11.4	366.5	422.3	350.1	6.7	333.7	366.4	111.0	2.8	104.3	117.8	20.1	3.5
	Maduración tardía	5	415.6	12.5	384.9	446.3	351.9	7.4	333.7	370.1	113.0	2.9	105.9	120.1	14.0	3.3
	Maduración temprana	10	372.7	9.0	351.4	394.0	328.0	8.0	309.0	347.0	108.6	2.8	102.0	115.2	18.6	1.5
	Maduración tardía	10	396.3	9.5	370.1	422.6	347.2	8.2	324.4	370.0	116.8	2.9	108.9	124.8	16.1	1.6
	Maduración temprana	15	393.7	23.2	341.1	446.3	338.4	9.8	316.3	360.5	115.9	1.6	112.3	119.5	23.4	1.4
	Maduración tardía	15	391.5	22.6	336.3	446.7	349.7	9.3	327.0	372.3	113.9	1.4	110.5	117.3	16.5	1.5
	Maduración tardía	20	409.9	13.3	367.5	452.3	342.5	3.8	330.4	354.5	102.1	1.2	98.3	105.8	21.2	1.6

Tabla 3.3. *Análisis funcional in silico de SNPs.* Descripción de 261 variantes SNPs y su asociación al metabolismo vegetal. SNP_ID: nombre del SNP; Gen ID: nombre del gen asociado al SNP; SNP POS: posición del SNP en pares de bases, según el genoma de referencia de la caña de azúcar del cv R570; SNP: variante de referencia/variante alternativa; Tipo de variante: implicancia de la variante; Gen Biológico de anotación funcional: nombre de la proteína, enzima o factor de transcripción asociado al SNP.

N°	SNP_ID	Cromosoma	Inicio Gen	Fin Gen	Gen ID	SNP POS	SNP	Tipo de variante	Anotacion funcional - Gen
1	9282_6	Sh01	3607193	3607909	Sh01_g002680	3594861	C /T	intergénica	Débilmente similar a la proteína OSJNBa0009P12.16
2	9282_6	Sh01	3608206	3609559	Sh01_g002690	3594861	C /T	intergénica	Reductasa de alfa-pirona tetracétida 1
3	9282_6	Sh01	3611008	3612683	Sh01_g002700	3594861	C /T	intergénica	Dihidroflavonol-4-reductasa
4	9282_6	Sh01	3612685	3615709	Sh01_g002710	3594861	C /T	intergénica	Dihidroflavonol 4-reductasa
5	9282_6	Sh01	3616931	3621271	Sh01_g002720	3594861	C /T	intergénica	Fosfatasa ácida
6	9282_6	Sh01	3674561	3675040	Sh01_g002730	3594861	C /T	intergénica	Proteína hipotética conservada
7	10696_14	Sh01	4018838	4022360	Sh01_g002880	4019271	G /T	río arriba del gen	Proteína hipotética conservada
8	22325_137	Sh01	6860376	6861811	Sh01_g004730	6861623	C /T	intrón	Hexosiltransferasa
9	24531_122	Sh01	8940352	8941299	Sh01_g006160	8986204	T /C	río abajo del gen	Similar a proteína expresada
10	24531_122	Sh01	8950733	8953669	Sh01_g006170	8986204	T /C	río abajo del gen	Miembro 25 de la familia de transportadores ABC B
11	24531_122	Sh01	8987073	8987493	Sh01_g006180	8986204	T /C	río abajo del gen	Factor de transcripción bHLH84
12	24547_65	Sh01	8999610	9002696	Sh01_g006220	9002098	A /C	río arriba del gen	Subunidad 50 del factor de estimulación de la escisión
13	370_51	Sh01	11012217	11013869	Sh01_g007780	11012295	C /T	río arriba del gen	Transportador de folato C cloroplástico
14	371_143	Sh01	11012217	11013869	Sh01_g007780	11012435	T /C	río arriba del gen	Transportador de folato C cloroplástico
15	371_96	Sh01	11012217	11013869	Sh01_g007780	11012482	G /A	río arriba del gen	Transportador de folato C cloroplástico
16	483_79	Sh01	11326042	11326349	Sh01_g007960	11326218	T /C	intrón	Reductasa de difosfato de 4-hidroxi-3-metilbut-2-enilo 2C cloroplástico
17	483_79	Sh01	11350959	11354762	Sh01_g007970	11326218	T /C	intrón	p10Sh086H20
18	528_118	Sh01	11402384	11403755	Sh01_g008030	11405804	G /A	río arriba del gen	Proteína hipotética conservada
19	609_146	Sh01	11637691	11639826	Sh01_g008100	11638378	A /C	intrón	Similar a proteína expresada
20	1281_58	Sh01	13338899	13348697	Sh01_g009130	13340417	C /T	intrón	Quinasa de proteína serina/treonina putativa IREH1
21	2070_95	Sh01	15362440	15367158	Sh01_g010480	15364933	G /A	intrón	Hexosiltransferasa
22	2070_93	Sh01	15362440	15367158	Sh01_g010480	15364935	T /A	intrón	Hexosiltransferasa
23	2070_43	Sh01	15362440	15367158	Sh01_g010480	15364985	G /A	intrón	Hexosiltransferasa
24	2131_124	Sh01	15526636	15533807	Sh01_g010540	15530796	A /G	intrón	ATPasa 1C de tipo de membrana plasmática
25	2131_124	Sh01	15540945	15544979	Sh01_g010550	15530796	A /G	intrón	Quinasa VIP2 de hexakisfosfato de inositol y difosfato de pentakisfosfato de inositol
26	2131_124	Sh01	15544981	15547997	Sh01_g010560	15530796	A /G	intrón	Quinasa VIP1 de hexakisfosfato de inositol y difosfato de pentakisfosfato de inositol
27	3165_56	Sh01	18198427	18202722	Sh01_g011970	18202939	G /A	río abajo del gen	Reductasa NOL 2C cloroplástico de clorofil(ide) b
28	3165_20	Sh01	18204026	18204622	Sh01_g011980	18202975	T /A	río abajo del gen	Factor de transcripción de la familia Myb PHL12
29	3385_138	Sh01	18789961	18793513	Sh01_g012340	18791970	C /T	río abajo del gen	Homólogo de fosfatasa de proteína serina/treonina BSL2
30	6737_106	Sh01	28715966	28728597	Sh01_g017730	28716594	C /T	río abajo del gen	Sintetasa de glutamil-ARNt
31	7109_59	Sh01	29579188	29581380	Sh01_g018240	29578159	G /C	río abajo del gen	Similar a proteína expresada
32	7109_59	Sh01	29583343	29584110	Sh01_g018250	29578159	G /C	río abajo del gen	Reductasa de versicolorina
33	7109_59	Sh01	29585924	29586691	Sh01_g018260	29578159	G /C	río abajo del gen	Reductasa de versicolorina
34	7109_59	Sh01	29587003	29592050	Sh01_g018270	29578159	G /C	río abajo del gen	Proteína 8 LRR del grupo Ras intracelular de plantas
35	7109_59	Sh01	29592138	29592326	Sh01_g018280	29578159	G /C	río abajo del gen	Proteína 8 LRR del grupo Ras intracelular de plantas
36	7109_59	Sh01	29595855	29596640	Sh01_g018290	29578159	G /C	río abajo del gen	Reductasa de versicolorina

Tabla 3.1. *continuación.*

N°	SNP_ID	Cromosoma	Inicio Gen	Fin Gen	Gen ID	SNP POS	SNP	Tipo de variante	Anotacion funcional - Gen
37	7109_59	Sh01	29615693	29616421	Sh01_g018300	29578159	G /C	río abajo del gen	Proteína similar a reductasa de aldehído dependiente de NADPH 2C cloroplástico
38	7109_59	Sh01	29620358	29621143	Sh01_g018310	29578159	G /C	río abajo del gen	Reductasa de versicolorina
39	7109_59	Sh01	29627563	29627976	Sh01_g018320	29578159	G /C	río abajo del gen	Proteína similar a reductasa de aldehído dependiente de NADPH 2C cloroplástico
40	14275_129	Sh01	48780528	48786411	Sh01_g028810	48784756	A /G	sinónima	Proteína ABERRANT POLLEN TRANSMISSION 1
41	15668_18	Sh01	52881210	52886748	Sh01_g031520	52883359	G /A	intrón	Proteína de fluorescencia alta de clorofila 107
42	15715_29	Sh01	52965703	52967105	Sh01_g031600	52978947	A /G	intrón	Factor de transcripción BZIP (Fragmento)
43	15715_29	Sh01	52974266	52980077	Sh01_g031610	52978947	A /G	intrón	Homólogo de proteasa Lon 2C mitocondrial
44	15716_130	Sh01	52974266	52980077	Sh01_g031610	52979090	G /A	intrón	Homólogo de proteasa Lon 2C mitocondrial
45	15716_12	Sh01	52974266	52980077	Sh01_g031610	52979208	T /C	intrón	Homólogo de proteasa Lon 2C mitocondrial
46	15716_7	Sh01	52974266	52980077	Sh01_g031610	52979213	G /A	intrón	Homólogo de proteasa Lon 2C mitocondrial
47	15716_7	Sh01	52980798	52981501	Sh01_g031620	52979213	G /A	intrón	Homólogo de proteasa Lon 2C mitocondrial
48	15716_7	Sh01	52995493	52995755	Sh01_g031630	52979213	G /A	intrón	Proteína hipotética conservada
49	16802_24	Sh01	55609755	55611382	Sh01_g033320	55610672	T /C	intrón	Homólogo de proteína interactuante con sintasa de óxido nítrico
50	16802_21	Sh01	55609755	55611382	Sh01_g033320	55610675	C /A	intrón	Homólogo de proteína interactuante con sintasa de óxido nítrico
51	17545_60	Sh01	57481757	57484587	Sh01_g034620	57483546	C /T	con cambio de sentido	Proteína 1 de unión a miosina
52	20269_42	Sh01	63851400	63855439	Sh01_g038840	63853892	A /T	intrón	Similar a posible dominio de unión a metal en proteína de la familia inhibidora de RNasa L 2C RLI
53	20736_42	Sh01	64984272	64985502	Sh01_g039470	64985402	C /T	con cambio de sentido	Factor de transcripción sensible al etileno ERF073
54	20736_42	Sh01	64987113	64987382	Sh01_g039480	64985402	C /T	con cambio de sentido	Proteína hipotética conservada
55	20736_42	Sh01	64998190	64999664	Sh01_g039490	64985402	C /T	con cambio de sentido	Factor de transcripción sensible al etileno 1
56	21119_126	Sh01	65865707	65866388	Sh01_g040050	65865714	C /T	5_prima_UTR	Inhibidor de cisteína proteínica
57	22093_29	Sh01	68137797	68145908	Sh01_g041730	68144366	C /G	con cambio de sentido	Proteína hipotética conservada
58	22093_48	Sh01	68137797	68145908	Sh01_g041730	68144385	G /T	con cambio de sentido	Proteína hipotética conservada
59	22095_107	Sh01	68147544	68148380	Sh01_g041740	68144573	A /G	con cambio de sentido	Proteína dirigent
60	22095_107	Sh01	68137797	68145908	Sh01_g041730	68144573	A /G	con cambio de sentido	Proteína hipotética conservada
61	22095_107	Sh01	68149601	68150607	Sh01_g041750	68144573	A /G	con cambio de sentido	Proteína PIR
62	23633_124	Sh01	71904198	71907130	Sh01_g044530	71887627	G /C	intergénica	Familia NRT1/PTR 6.3 de proteínas
63	23633_124	Sh01	71919865	71920610	Sh01_g044540	71887627	G /C	intergénica	Similar a deshidrogenasa de malato [NADP] 2C precursor de cloroplasto
64	27353_125	Sh02	1747634	1749025	Sh02_g001140	1748576	A /G	con cambio de sentido	Proteína hipotética conservada
65	28913_135	Sh02	2210778	2220103	Sh02_g001550	2215265	T /A	intrón	Similar a secuencia genómica completa del andamio 8 del cromosoma chr3
66	31133_133	Sh02	28852622	28856523	Sh02_g015170	28854584	C /T	sinónima	Similar a fosforibosilaminoimidazol-succinocarboxamida sintasa
67	31133_90	Sh02	28852622	28856523	Sh02_g015170	28854627	T /C	intrón	Similar a fosforibosilaminoimidazol-succinocarboxamida sintasa
68	31133_89	Sh02	28852622	28856523	Sh02_g015170	28854628	A /G	intrón	Similar a fosforibosilaminoimidazol-succinocarboxamida sintasa
69	31133_70	Sh02	28852622	28856523	Sh02_g015170	28854647	T /A	intrón	Similar a fosforibosilaminoimidazol-succinocarboxamida sintasa
70	31133_55	Sh02	28852622	28856523	Sh02_g015170	28854662	G /A	intrón	Similar a fosforibosilaminoimidazol-succinocarboxamida sintasa
71	31721_90	Sh02	30674121	30679473	Sh02_g016110	30678593	G /C	con cambio de sentido	Proteína de planta asociada a metal pesado 41
72	34111_44	Sh02	36503678	36506910	Sh02_g020250	36503815	C /T	3_prima_UTR	Similar a SRPK4
73	34111_135	Sh02	36503678	36506910	Sh02_g020250	36503906	A /G	3_prima_UTR	Similar a SRPK4
74	36952_148	Sh02	43193853	43197582	Sh02_g025020	43194508	A /G	intrón	Transportador orgánico de cationes/carnitina 7
75	37622_48	Sh02	45035060	45036886	Sh02_g026400	45056945	G /A	intrón	ATPasa transportadora de fosfolípidos 2

Tabla 3.1. *continuación.*

N°	SNP_ID	Cromosoma	Inicio Gen	Fin Gen	Gen ID	SNP POS	SNP	Tipo de variante	Anotacion funcional - Gen
76	37622_48	Sh02	45056148	45060247	Sh02_g026410	45056945	G /A	intrón	Proteína UPF0553
77	38037_118	Sh02	45860831	45866790	Sh02_g026950	45863362	T /C	con cambio de sentido	Proteína similar a unión de ARN
78	59190_24	Sh03	624191	625563	Sh03_g000400	624971	C /T	intrón	Proteína que contiene dominio B3 Os01g0234100
79	59190_24	Sh03	625608	635933	Sh03_g000410	624971	C /T	intrón	Proteína que contiene dominio B3 At3g19184
80	59820_24	Sh03	763340	769912	Sh03_g000530	766930	T /A	río arriba del gen	TRZ4 de tRNAse Z 2C mitocondrial
81	47124_82	Sh03	2368252	2368902	Sh03_g001570	2381688	T /A	intergénica	Dioxygenasa de ácido giberélico 2-beta 1
82	47124_82	Sh03	2370195	2371230	Sh03_g001580	2381688	T /A	intergénica	Dioxygenasa de ácido giberélico 2-beta 1
83	47124_82	Sh03	2372371	2372861	Sh03_g001590	2381688	T /A	intergénica	Proteína B2
84	47124_82	Sh03	2394447	2396069	Sh03_g001600	2381688	T /A	intergénica	Amidasa A putativa del asparagina N-acetil-beta-glucosaminilo de péptido-N4
85	47124_82	Sh03	2398300	2400549	Sh03_g001610	2381688	T /A	intergénica	Amidasa A del asparagina N-acetil-beta-glucosaminilo de péptido-N4
86	51868_119	Sh03	3663934	3675816	Sh03_g002150	3673559	G /A	intrón	Polimerasa de ARN dependiente de ARN
87	59030_92	Sh03	5778489	5785354	Sh03_g003170	5841806	G /A	río abajo del gen	Similar a secuencia genómica completa del andamio 6 del cromosoma chr4
88	59030_92	Sh03	5787807	5794952	Sh03_g003180	5841806	G /A	río abajo del gen	Quinasa 2 relacionada con CDPK
89	59030_92	Sh03	5833603	5834710	Sh03_g003190	5841806	G /A	río abajo del gen	Subunidad M del factor de inicio de traducción eucariota 3
90	59030_92	Sh03	5837484	5838554	Sh03_g003200	5841806	G /A	río abajo del gen	Subunidad M del factor de inicio de traducción eucariota 3
91	42718_131	Sh03	10666798	10668608	Sh03_g006370	10672413	A /G	río arriba del gen	DExH-box helicasa de ARN dependiente de ATP DExH3
92	42718_131	Sh03	10671155	10674410	Sh03_g006380	10672413	A /G	río arriba del gen	Aldolasa de fructosa-bisfosfato
93	43189_124	Sh03	12118555	12121659	Sh03_g007070	12119407	G /A	intrón	3-beta-hidroxiesteroide-delta-isomerasa
94	43189_124	Sh03	12122742	12124844	Sh03_g007080	12119407	G /A	intrón	Similar a proteína Os08g0190200
95	46170_58	Sh03	20695376	20696348	Sh03_g011980	20699150	G /A	río arriba del gen	Proteína hipotética conservada
96	46170_58	Sh03	20704696	20710926	Sh03_g011990	20699150	G /A	río arriba del gen	Similar a transportador ABC tipo MDR
97	46686_59	Sh03	22383985	22385358	Sh03_g012800	22382742	C /T	río abajo del gen	Factor de transcripción bHLH87
98	48405_112	Sh03	27391719	27396942	Sh03_g015400	27396082	A /G	intrón	Similar a proteína Os01g0624000
99	49187_47	Sh03	29703315	29710846	Sh03_g016700	29709876	T /G	sinónima	Sintetasa de acil-CoA de cadena larga 4 (Fragmento)
100	49187_112	Sh03	29703315	29710846	Sh03_g016700	29709941	G /A	intrón	Sintetasa de acil-CoA de cadena larga 4 (Fragmento)
101	49894_12	Sh03	31592347	31594058	Sh03_g017620	31593543	A /G	intrón	Proteína que contiene dominio de homología de receptor 2C transmembrana y dominio RING 1
102	51301_51	Sh03	35129621	35132854	Sh03_g020020	35130488	C /T	río abajo del gen	Proteína hipotética conservada
103	52014_73	Sh03	37040789	37042716	Sh03_g021220	37044305	C /T	río arriba del gen	Similar a proteína Os01g0242300
104	52014_73	Sh03	37043671	37044971	Sh03_g021230	37044305	C /T	río arriba del gen	Similar a proteína Os02g0788500
105	52535_52	Sh03	38058175	38061455	Sh03_g022050	38061066	C /T	sinónima	Proteína que contiene dominio BSD
106	52656_64	Sh03	38284938	38286313	Sh03_g022370	38284738	G /C	río abajo del gen	Hidrolasa de ubiquitina 13
107	54306_124	Sh03	42403545	42408352	Sh03_g025100	42405218	A /G	intrón	Proteína tipo TOM1 2
108	54307_136	Sh03	42403545	42408352	Sh03_g025100	42405274	T /G	sinónima	Proteína tipo TOM1 2
109	54307_136	Sh03	42417598	42419132	Sh03_g025110	42405274	T /G	sinónima	Subunidad 5 de succinato deshidrogenasa 2C mitocondrial
110	56120_42	Sh03	47168114	47169504	Sh03_g028340	47170727	T /C	río abajo del gen	Similar a proteína P0696G06

Tabla 3.1. *continuación.*

N°	SNP_ID	Cromosoma	Inicio Gen	Fin Gen	Gen ID	SNP POS	SNP	Tipo de variante	Anotacion funcional - Gen
111	56556_142	Sh03	48693945	48694130	Sh03_g029330	48704317	G /C	río arriba del gen	Transferasa 1 de hidroxicinamoil espermidina
112	56557_96	Sh03	48707041	48713520	Sh03_g029340	48704430	A /T	río arriba del gen	Desmetilasa de lisina específica JMJ705
113	56557_96	Sh03	48715358	48716467	Sh03_g029350	48704430	A /T	río arriba del gen	Similar a endopeptidasa de cisteína
114	56557_96	Sh03	48734450	48735666	Sh03_g029360	48704430	A /T	río arriba del gen	Proteína que contiene repetición de tetratricopeptido 1
115	57148_35	Sh03	50159721	50161464	Sh03_g030490	50160414	A /T	río abajo del gen	Subunidad TOM20 del receptor de importación mitocondrial
116	73257_132	Sh04	4413607	4415777	Sh04_g003050	4414423	A /G	con cambio de sentido	Proteína hipotética conservada
117	74875_95	Sh04	7722040	7725551	Sh04_g005130	7724241	T /C	intrón	Proteína TIC 2C cloroplástico
118	63659_87	Sh04	19200671	19201400	Sh04_g010660	19202026	G /A	río arriba del gen	Fosfatidilinositol/fosfatidilcolina transferasa SFH10
119	63913_142	Sh04	20076949	20079076	Sh04_g011160	20078548	A /C	intrón	CRL de liasa de cromóforo 2C cloroplástico
120	65009_37	Sh04	23447797	23450293	Sh04_g013080	23450380	A /G	río arriba del gen	Proteína hipotética conservada
121	67025_35	Sh04	28868047	28870150	Sh04_g016340	28863197	G /T	río arriba del gen	Peroxidasa putativa (Fragmentos)
122	67025_35	Sh04	28870784	28871044	Sh04_g016350	28863197	G /T	río arriba del gen	Proteína PBP1 de unión a calcio
123	67907_85	Sh04	31147548	31148409	Sh04_g017660	31148359	C /G	río arriba del gen	Proteína secretora 12 rica en repeticiones de cisteína
124	67951_7	Sh04	31197797	31198339	Sh04_g017710	31212095	A /C	río arriba del gen	Jun o 2 de polcalcina
125	67951_7	Sh04	31200080	31200740	Sh04_g017720	31212095	A /C	río arriba del gen	Proteína hipotética conservada
126	67951_7	Sh04	31201507	31204774	Sh04_g017730	31212095	A /C	río arriba del gen	Endoglucanasa 6
127	68281_44	Sh04	31863202	31863773	Sh04_g018390	31872627	C /T	río arriba del gen	Receptor 3.3 de glutamato
128	68281_44	Sh04	31865121	31867754	Sh04_g018400	31872627	C /T	río arriba del gen	Transportador bidireccional de azúcar SWEET
129	68281_107	Sh04	31872799	31873804	Sh04_g018410	31872690	C /T	río arriba del gen	Componente SEC3A del complejo exocisto
130	68405_136	Sh04	32176362	32177627	Sh04_g018660	32179452	A /C	río abajo del gen	Desaturasa de ácidos grasos DES2
131	69171_28	Sh04	34005538	34012193	Sh04_g019850	34011701	G /A	sinónima	Similar a sintetasa de (P)ppGpp en plástidos
132	69423_89	Sh04	34554248	34556658	Sh04_g020250	34555757	A /C	río abajo del gen	Factor de elongación Tu 2C cloroplástico
133	70295_126	Sh04	37003544	37003933	Sh04_g021550	37003495	C /T	río arriba del gen	Proteína que delimita la célula cortical
134	70295_126	Sh04	37015730	37016179	Sh04_g021560	37003495	C /T	río arriba del gen	Proteína rica en prolina de 14 kDa DC2
135	70295_126	Sh04	37027503	37027856	Sh04_g021570	37003495	C /T	río arriba del gen	Similar a proteína RCc3
136	71294_66	Sh04	39764471	39767360	Sh04_g023440	39766922	G /C	intrón	Proteína relacionada con WAT1
137	71294_67	Sh04	39764471	39767360	Sh04_g023440	39766923	A /C	intrón	Proteína relacionada con WAT1
138	72305_77	Sh04	42193842	42194154	Sh04_g025130	42203566	C /T	río abajo del gen	Factor de transcripción bHLH82
139	72305_77	Sh04	42195025	42195282	Sh04_g025140	42203566	C /T	río abajo del gen	Factor de transcripción bHLH69
140	72305_77	Sh04	42204979	42205749	Sh04_g025150	42203566	C /T	río abajo del gen	Helicasa de ARN dependiente de ATP de caja DEAD 1
141	72305_77	Sh04	42209744	42210014	Sh04_g025160	42203566	C /T	río abajo del gen	Helicasa de ARN dependiente de ATP de caja DEAD 1
142	72305_77	Sh04	42216597	42217148	Sh04_g025170	42203566	C /T	río abajo del gen	Fosfatasa del dominio C-terminal de la ARN polimerasa II similar a 4
143	72305_77	Sh04	42218132	42218317	Sh04_g025180	42203566	C /T	río abajo del gen	Proteína putativa similar a la sintasa de estrictosidina 3
144	73161_112	Sh04	43890277	43893127	Sh04_g026460	43890777	A /G	río arriba del gen	3-cetoacil-CoA tiolasa C peroxisomal
145	73161_122	Sh04	43890277	43893127	Sh04_g026460	43890787	A /G	río arriba del gen	3-cetoacil-CoA tiolasa C peroxisomal

Tabla 3.1. *continuación.*

N°	SNP_ID	Cromosoma	Inicio Gen	Fin Gen	Gen ID	SNP POS	SNP	Tipo de variante	Anotacion funcional - Gen
146	73161_124	Sh04	43890277	43893127	Sh04_g026460	43890789	C /T	río arriba del gen	3-cetoacil-CoA tiolasa C peroxisomal
147	73591_90	Sh04	44856995	44860721	Sh04_g027370	44859374	G /A	río arriba del gen	Proteína de dedo de zinc tipo NF-X1 NFXL1
148	77788_76	Sh05	16429742	16432894	Sh05_g008220	16430690	G /A	sinónima	Proteína que contiene repeticiones WD VIP3
149	77979_119	Sh05	16966691	16968739	Sh05_g008530	16967967	T /G	con cambio de sentido	Similar a la proteína Os11g0587100
150	93647_124	Sh06	5608118	5608527	Sh06_g002400	5608321	G /A	sinónima	Proteína hipotética conservada
151	83045_114	Sh06	10444073	10447661	Sh06_g004750	10447404	A /G	río abajo del gen	Similar a la proteína que contiene el dominio DDT
152	84156_79	Sh06	13260375	13262987	Sh06_g006780	13263137	T /A	río abajo del gen	Débilmente similar a la proteína OSIGBa0115M15.9
153	84418_84	Sh06	13856615	13861095	Sh06_g007200	13860027	G /A	sinónima	Similar a la proteína H0305E08.1
154	85290_8	Sh06	16056079	16057571	Sh06_g008680	16056817	A /G	sinónima	Alfa-bisabolol sintasa
155	85835_5	Sh06	17433056	17435418	Sh06_g009590	17474221	A /G	río abajo del gen	L-alotreonina aldolasa
156	88715_73	Sh06	25008113	25009622	Sh06_g014050	25015855	C /T	río arriba del gen	Proteína nuclear localizada con motivo AT-hook 13
157	88715_73	Sh06	25009649	25010666	Sh06_g014060	25015855	C /T	río arriba del gen	Proteína nuclear localizada con motivo AT-hook 1
158	88715_73	Sh06	25016378	25017713	Sh06_g014070	25015855	C /T	río arriba del gen	Similar a la proteína OSJNBa0086O06.15
159	88715_73	Sh06	25020676	25021608	Sh06_g014080	25015855	C /T	río arriba del gen	Proteína nuclear localizada con motivo AT-hook 21
160	88811_87	Sh06	25259857	25260357	Sh06_g014150	25318355	T /C	río abajo del gen	Proteína ribosomal 60S L12
161	88811_87	Sh06	25300847	25303630	Sh06_g014160	25318355	T /C	río abajo del gen	Proteína similar a FloricaulA / tipo leafy
162	88811_87	Sh06	25310467	25311543	Sh06_g014170	25318355	T /C	río abajo del gen	Quinasa asociada a la pared
163	88811_87	Sh06	25315957	25316436	Sh06_g014180	25318355	T /C	río abajo del gen	At1g06060
164	88811_87	Sh06	25351802	25353430	Sh06_g014190	25318355	T /C	río abajo del gen	Proteína ribosomal 60S L12-1
165	90656_113	Sh06	29608788	29613931	Sh06_g017000	29611274	G /T	río abajo del gen	Similar a la proteína OSIGBa0097P08.1
166	101227_12	Sh07	286078	286350	Sh07_g000150	268639	G /A	intergénica	Quinasa de serina/treonina STY8
167	101227_12	Sh07	287026	287401	Sh07_g000160	268639	G /A	intergénica	Proteína con repeticiones pentatricopéptido At1g62930 2C cloroplástica
168	101227_12	Sh07	322383	326976	Sh07_g000170	268639	G /A	intergénica	Proteína tipo Tubby-like F-box 12
169	101227_12	Sh07	327453	330080	Sh07_g000180	268639	G /A	intergénica	S-aciltransferasa
170	95538_85	Sh07	1078892	1083348	Sh07_g000680	1079822	T /G	río abajo del gen	Proteína NCA1
171	102695_94	Sh07	3669990	3670157	Sh07_g002730	3674263	G /C	río abajo del gen	Proteína hipotética conservada
172	102695_94	Sh07	3671066	3674835	Sh07_g002740	3674263	G /C	río abajo del gen	Proteína que contiene el dominio de dedo de zinc CCCH 55
173	102695_94	Sh07	3676520	3680665	Sh07_g002750	3674263	G /C	río abajo del gen	Transportador de cationes orgánicos/carnitina 7
174	104767_66	Sh07	9544132	9545855	Sh07_g006570	9544695	C /T	río arriba del gen	5-pentadecatrienil resorcinol O-metiltransferasa
175	96406_85	Sh07	13444883	13448106	Sh07_g008310	13446888	A /T	río arriba del gen	Proteína tipo anquirina
176	96406_132	Sh07	13444883	13448106	Sh07_g008310	13446935	C /A	río arriba del gen	Proteína tipo anquirina
177	96556_143	Sh07	13937392	13942398	Sh07_g008530	13941816	C /T	intrón	Factor de transcripción WRKY WRKY24
178	96568_56	Sh07	13953298	13953807	Sh07_g008560	13955115	A /T	río abajo del gen	Miembro de la superfamilia de transmembrana 9
179	96568_56	Sh07	13953809	13956613	Sh07_g008570	13955115	A /T	río abajo del gen	Miembro de la superfamilia de transmembrana 9

Tabla 3.1. *continuación.*

N°	SNP_ID	Cromosoma	Inicio Gen	Fin Gen	Gen ID	SNP POS	SNP	Tipo de variante	Anotacion funcional - Gen
180	96649_146	Sh07	14292508	14293996	Sh07_g008660	14351243	T /C	río abajo del gen	ATPasa transportadora de fosfolípidos 10
181	96649_146	Sh07	14347985	14349916	Sh07_g008670	14351243	T /C	río abajo del gen	BURP5
182	99734_87	Sh07	23070018	23073710	Sh07_g013280	23071332	A /G	río arriba del gen	Proteína hemo citocromo c
183	99734_150	Sh07	23070018	23073710	Sh07_g013280	23071395	A /G	río arriba del gen	Proteína hemo citocromo c
184	99734_150	Sh07	23084888	23085883	Sh07_g013290	23071395	A /G	río arriba del gen	Similar a la familia de fosfoesterasas tipo calcineurina
185	101130_78	Sh07	26641616	26642570	Sh07_g015960	26650785	G /C	intrón	Proteína que contiene dominios BTB/POZ y MATH 1
186	101130_78	Sh07	26648932	26651432	Sh07_g015970	26650785	G /C	intrón	Similar a la proteína H0112G12
187	101130_78	Sh07	26653766	26654740	Sh07_g015980	26650785	G /C	intrón	Proteína que contiene dominios BTB/POZ y MATH 1
188	101130_78	Sh07	26656112	26657104	Sh07_g015990	26650785	G /C	intrón	Proteína que contiene dominios BTB/POZ y MATH 2
189	101130_78	Sh07	26679926	26680916	Sh07_g016000	26650785	G /C	intrón	Proteína que contiene dominios BTB/POZ y MATH 2
190	101130_78	Sh07	26688364	26690038	Sh07_g016010	26650785	G /C	intrón	Proteína que contiene dominios BTB/POZ y MATH 2
191	101726_10	Sh07	28144261	28147732	Sh07_g016960	28146289	G /A	río arriba del gen	Similar a la familia de proteínas tipo quinasa de carbohidratos PfkB
192	101727_34	Sh07	28144261	28147732	Sh07_g016960	28146564	A /G	río arriba del gen	Similar a la familia de proteínas tipo quinasa de carbohidratos PfkB
193	101832_105	Sh07	28468220	28471442	Sh07_g017160	28468845	T /A	sinónima	Beta C2-xilosiltransferasa
194	102104_66	Sh07	29192883	29198424	Sh07_g017670	29198132	G /C	intrón	Proteína de ensamblaje de la citocromo c oxidasa COX15
195	110558_70	Sh08	2529055	2531127	Sh08_g001730	2531248	G /A	río abajo del gen	Hexosiltransferasa
196	110558_70	Sh08	2531150	2533506	Sh08_g001740	2531248	G /A	río abajo del gen	Hexosiltransferasa
197	110558_70	Sh08	2536444	2538295	Sh08_g001750	2531248	G /A	río abajo del gen	Transportador de calcio uniporter 6 2C mitocondrial
198	110558_70	Sh08	2541304	2542071	Sh08_g001760	2531248	G /A	río abajo del gen	Reductasa-dihidrofolato sintasa de timidilato bifuncional
199	110558_70	Sh08	2560669	2563394	Sh08_g001770	2531248	G /A	río abajo del gen	Proteína tipo kinesina KIN-7D 2C mitocondrial
200	111333_133	Sh08	4733022	4733445	Sh08_g003130	4733618	T /C	río arriba del gen	Proteína que contiene el dominio ACT ACR12
201	111334_56	Sh08	4734417	4736047	Sh08_g003140	4733781	T /A	río arriba del gen	Proteína que contiene el dominio ACT ACR12
202	111334_56	Sh08	4740084	4740251	Sh08_g003150	4733781	T /A	río arriba del gen	Proteína tipo HUA2-LIKE 3
203	112190_103	Sh08	7449643	7452150	Sh08_g004520	7451527	G /A	intrón	Proteína hipotética conservada
204	106091_15	Sh08	13232482	13237736	Sh08_g006950	13245722	A /G	río arriba del gen	Proteína putativa ribosoma-inactivadora cucurmosina
205	106091_15	Sh08	13240152	13240531	Sh08_g006960	13245722	A /G	río arriba del gen	Proteína que contiene el dominio de dedo de zinc CCCH 16
206	106945_142	Sh08	15467976	15470086	Sh08_g007990	15470400	C /T	río abajo del gen	Inositolfosfotransferasa de fosfatidilinositol
207	106945_142	Sh08	15470096	15470429	Sh08_g008000	15470400	C /T	río abajo del gen	Inositolfosfotransferasa de fosfatidilinositol
208	106945_142	Sh08	15473663	15475342	Sh08_g008010	15470400	C /T	río abajo del gen	Similar a la proteína expresada
209	106945_142	Sh08	15476354	15480690	Sh08_g008020	15470400	C /T	río abajo del gen	Factor de licencia de replicación del ADN MCM7
210	108926_97	Sh08	20483441	20483719	Sh08_g011080	20495574	A /G	sinónima	Proteína hipotética conservada
211	108926_97	Sh08	20491683	20497534	Sh08_g011090	20495574	A /G	sinónima	Similar a la proteína expresada
212	108927_133	Sh08	20491683	20497534	Sh08_g011090	20495690	T /C	con cambio de sentido	Similar a la proteína expresada
213	123364_8	Sh09	5121155	5127162	Sh09_g003130	5126406	G /A	sinónima	Transportador ABC familia B miembro 21
214	114111_147	Sh09	13463026	13482394	Sh09_g007120	13463862	G /A	intrón	Débilmente similar a la proteína Os05g0481700

Tabla 3.1. *continuación.*

N°	SNP_ID	Cromosoma	Inicio Gen	Fin Gen	Gen ID	SNP POS	SNP	Tipo de variante	Anotacion funcional - Gen
215	115772_115	Sh09	18758795	18761279	Sh09_g009320	18778481	A /T	intergénica	Proteína hipotética conservada
216	116564_21	Sh09	21045071	21051332	Sh09_g010510	21050955	A /T	intrón	Similar a la proteína Os05g0406200
217	116564_21	Sh09	21053730	21054965	Sh09_g010520	21050955	A /T	intrón	Proteína putativa con repeticiones pentatricopéptido At4g02750
218	116687_91	Sh09	21455456	21457013	Sh09_g010700	21456810	C /T	río arriba del gen	Proteína NRT1/ PTR FAMILY 3
219	118241_90	Sh09	25960726	25966878	Sh09_g013360	25963663	A /G	intrón	Proteína tipo SelT
220	119219_79	Sh09	27943821	27945461	Sh09_g014640	28030977	G /T	río abajo del gen	Proteína INDUCIDA POR DEFICIENCIA DE AZUFRE 1
221	119219_79	Sh09	27946254	27947566	Sh09_g014650	28030977	G /T	río abajo del gen	Similar a la proteína Os05g0506100
222	119219_79	Sh09	27960677	27964524	Sh09_g014660	28030977	G /T	río abajo del gen	Proteína putativa metilación del ADN dirigida por ARN 3
223	119219_79	Sh09	27977797	27979349	Sh09_g014670	28030977	G /T	río abajo del gen	Proteína putativa esterase/lipasa GDSL At3g48460
224	119219_79	Sh09	27990502	27994226	Sh09_g014680	28030977	G /T	río abajo del gen	Factor de producción de ribosomas 2 homólogo
225	119219_79	Sh09	27994764	27995582	Sh09_g014690	28030977	G /T	río abajo del gen	Similar a la proteína P0696G06.26
226	119219_79	Sh09	28000724	28001309	Sh09_g014700	28030977	G /T	río abajo del gen	Proteína hipotética conservada
227	119219_79	Sh09	28004520	28009730	Sh09_g014710	28030977	G /T	río abajo del gen	Calreticulina-3 (Fragmento)
228	119219_79	Sh09	28026822	28029908	Sh09_g014720	28030977	G /T	río abajo del gen	Cisteína proteinasa RD21A
229	119809_87	Sh09	29659519	29663343	Sh09_g015820	29659430	A /G	río abajo del gen	Proteína de transferencia de histidina pseudo 5
230	120307_111	Sh09	30709421	30710848	Sh09_g016720	30732620	C /T	río arriba del gen	Glucosiltransferasa
231	120307_111	Sh09	30734513	30735076	Sh09_g016730	30732620	C /T	río arriba del gen	Factor de iniciación de traducción eucariótico 3 subunidad F
232	120307_111	Sh09	30735078	30736028	Sh09_g016740	30732620	C /T	río arriba del gen	Factor de iniciación de traducción eucariótico 3 subunidad F
233	121440_99	Sh09	33248943	33253396	Sh09_g018530	33250897	C /A	intrón	Transportador facilitado de glucosa de la familia de transportadores de solutos 2C miembro 8
234	122049_17	Sh09	34733711	34738344	Sh09_g019640	34735285	C /T	intrón	Proteína REDUCCIÓN DE LA ACETILACIÓN DE LA PARED 3
235	134234_45	Sh10	391922	396027	Sh10_g000200	374735	G /A	río abajo del gen	Proteína tipo oxidasa monocobre SKU5
236	134234_45	Sh10	405103	407301	Sh10_g000210	374735	G /A	río abajo del gen	Proteína putativa
237	131573_35	Sh10	2874571	2876106	Sh10_g001800	2875014	C /T	río abajo del gen	Factor de transcripción TRAF

Tabla 3.1. *continuación.*

N°	SNP_ID	Cromosoma	Inicio Gen	Fin Gen	Gen ID	SNP POS	SNP	Tipo de variante	Anotacion funcional - Gen
238	135172_134	Sh10	6552345	6561675	Sh10_g004220	6555530	C /A	intrón	Proteína asociada a la resistencia a multidrogas 3
239	135625_44	Sh10	7807274	7811522	Sh10_g005010	7808455	A /T	río arriba del gen	Proteína que contiene el dominio de dedo de zinc CCCH 49
240	135625_48	Sh10	7807274	7811522	Sh10_g005010	7808459	A /C	río arriba del gen	Proteína que contiene el dominio de dedo de zinc CCCH 49
241	135626_64	Sh10	7807274	7811522	Sh10_g005010	7808679	C /T	río arriba del gen	Proteína que contiene el dominio de dedo de zinc CCCH 49
242	135626_64	Sh10	7811747	7812470	Sh10_g005020	7808679	C /T	río arriba del gen	Adenililtransferasa y sulfurtransferasa MOCS3-2
243	125527_54	Sh10	11097899	11098228	Sh10_g007020	11100316	C /T	río arriba del gen	Proteína hipotética conservada
244	127495_112	Sh10	16897294	16911056	Sh10_g009930	16910628	C /A	región splicing_intrón	Proteína hipotética conservada
245	127495_112	Sh10	16911058	16911511	Sh10_g009940	16910628	C /A	región splicing_intrón	Proteína hipotética conservada
246	127495_112	Sh10	16911540	16912047	Sh10_g009950	16910628	C /A	región splicing_intrón	Arabinosiltransferasa RRA1
247	128130_74	Sh10	19140524	19145805	Sh10_g010830	19145041	T /C	intrón	Transportador de fosfato PHO1-3
248	128496_124	Sh10	20259310	20272775	Sh10_g011170	20259301	A /T	río abajo del gen	Proteína hipotética conservada
249	128509_49	Sh10	20295594	20299491	Sh10_g011180	20298999	C /T	sinónima	Similar a la proteasa de membrana dependiente de metal
250	128509_117	Sh10	20295594	20299491	Sh10_g011180	20299067	C /T	3_prima_UTR	Similar a la proteasa de membrana dependiente de metal
251	128509_130	Sh10	20295594	20299491	Sh10_g011180	20299080	C /T	3_prima_UTR	Similar a la proteasa de membrana dependiente de metal
252	128511_113	Sh10	20295594	20299491	Sh10_g011180	20299159	G /A	3_prima_UTR	Similar a la proteasa de membrana dependiente de metal
253	128511_49	Sh10	20295594	20299491	Sh10_g011180	20299223	T /A	3_prima_UTR	Similar a la proteasa de membrana dependiente de metal
254	129133_54	Sh10	22263744	22265187	Sh10_g011850	22366511	A /T	intergénica	Proteína putativa ligasa de ubiquitina E3 RMA1H1
255	129133_54	Sh10	22283816	22295049	Sh10_g011860	22366511	A /T	intergénica	Subunidad epsilon del complejo de proteínas 1
256	129133_54	Sh10	22347505	22347966	Sh10_g011870	22366511	A /T	intergénica	Proteína hipotética conservada
257	133685_54	Sh10	33603089	33603577	Sh10_g018530	33613438	C /T	con cambio de sentido	Similar a la familia de proteínas tipo dedo de zinc
258	133685_54	Sh10	33604719	33606239	Sh10_g018540	33613438	C /T	con cambio de sentido	Factor de transcripción bHLH93
259	133685_54	Sh10	33609588	33616085	Sh10_g018550	33613438	C /T	con cambio de sentido	Quinasa de serina/treonina ATR
260	133685_56	Sh10	33609588	33616085	Sh10_g018550	33613440	A /G	con cambio de sentido	Quinasa de serina/treonina ATR
261	133685_56	Sh10	33629927	33630169	Sh10_g018560	33613440	A /G	con cambio de sentido	Proteína TSS