UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Desarrollo y aplicación de metodologías genómicas para el mejoramiento molecular de *Eucalyptus dunnii* mediante Mapeo por Asociación y Selección Genómica

Tesis presentada para optar al título de Doctora de la Universidad de Buenos Aires en el área de CIENCIAS BIOLÓGICAS

Lic. Natalia Cristina Aguirre

Directora de Tesis: **Dra. Susana Noemí Marcucci Poltri.**

Codirector de Tesis: **Dr. Máximo Rivarola.**

Consejero de Estudios: **Horacio Esteban Hopp.**

Lugar de trabajo: Instituto de Agrobiotecnología y Biología Molecular (IABiMo, UEDD INTA-CONICET), CICVyA, CNIA (INTA Castelar).

Buenos Aires 2020

Desarrollo y aplicación de metodologías genómicas para el mejoramiento molecular de *Eucalyptus dunnii* mediante Mapeo por Asociación y Selección Genómica

RESUMEN

El género *Eucalyptus* es uno de los más importantes del mundo y, por ende, su investigación y mejoramiento genético concentra esfuerzos internacionales. En Argentina, el programa de mejoramiento de *Eucalypt*us se lleva a cabo principalmente por INTA. *Eucalyptus dunnii* es una de las especies de eucaliptos con alta demanda para producir pasta celulósica de fibra corta, especialmente en aquellas regiones donde otras especies de excelente aptitud papelera como el *E. globulus* y *E. grandis* son afectadas por condiciones edafoclimáticas limitantes. Además, presenta buenas características de crecimiento y productividad para su uso como madera sólida, aunque deben ser mejoradas. En ese sentido, las nuevas estrategias de mejoramiento molecular son útiles para dicho propósito.

El objetivo del presente trabajo es aplicar mejoramiento molecular en *E. dunnii* mediante las estrategias de Mapeo por Asociación y Selección Genómica a través del desarrollo y empleo de sistemas de genotipificación de alto rendimiento.

Para ello, se optimizó un protocolo de ddRADSeq (*Double-digest restriction site-associated DNA sequencing*) desarrollándolo en su totalidad en el país. Este protocolo es versátil y útil tanto para un número reducido de muestras, pudiendo ser escalado a cientos de muestras de manera eficiente y robusta, generando información relevante sobre marcadores moleculares SNPs y SSR. Es la primera vez que se aplica esta metodología en *E. dunnii*, analizándose una población de polinización abierta de 308 individuos (perteneciente a una red de ensayos de orígenes y procedencias del programa de mejoramiento de INTA, implantada en Ubajay, Entre Ríos). Los resultados obtenidos se compararon con la información generada por la aplicación del microarreglo comercial EUChip60K. Se obtuvieron un total de 8.170 (ddRADSeq) y 19.045 (EUChip60K) SNPs, distribuidos a lo largo de los 11 cromosomas y mostrando cobertura de regiones genómicas complementarias.

De manera original para esta especie se evaluó la estrategia de Mapeo por Asociación. Se utilizó un modelo lineal mixto y datos fenotípicos para catorce caracteres correspondientes a variables de crecimiento, calidad de madera (densidad básica y propiedades químicas de la madera estimadas mediante NIR (*Near Infrared Reflectance*)) y tensiones de crecimiento. La estructura poblacional mostró dos grupos genéticos para ambas plataformas de genotipado concordantes entre sí y se evidenció un bajo desequilibrio de ligamiento. Se encontraron 7, 13 y 19 SNPs (ddRADSeq, EUChip60K y matriz conjunta; $p < 0,0001$) asociados para 12 de las 14 características fenotípicas estudiadas. Para los marcadores asociados se realizó una búsqueda de genes empleando como referencia el genoma de *E. grandis* y explorando una ventana de 70 Kpb alrededor de los mismos. Se localizaron alrededor de 100 genes, varios de ellos con funciones relacionadas a las características analizadas. Cincuenta genes de distintas rutas metabólicas de la síntesis de celulosa, xilanos, fenilpropanoides, terpenos, lacasas, peroxidasas, entre otros se ubicaron a menos de 500 Kpb de los marcadores encontrados,

Por último, se aplicó una prueba de concepto de Selección Genómica con ambas metodologías de genotipificación y se compararon los predictores de mérito genético de acuerdo con su exactitud teórica (evaluación de la varianza del error en la predicción) y habilidad predictiva (correlación entre el fenotipo observado y estimado) respecto de los cálculos de predicciones convencionales. La exactitud teórica para los modelos de predicción con información genómica fue superior respecto del cálculo convencional particularmente para los caracteres de heredabilidad <0,43 (cel20, dap11, dap20, db20). La información genómica generada por ddRADSeq, tuvo mejor desempeño, siendo prácticamente indistinto para el resto de los caracteres. En cambio, la habilidad predictiva fue muy similar para todos los modelos, y el método convencional fue levemente superior para algunos caracteres con heredabilidad <0,43.

Las estrategias de mejoramiento molecular evaluadas durante el desarrollo de esta tesis podrán ser incorporadas como herramientas para la selección temprana de genotipos superiores de los programas de mejoramiento de *E. dunnii* acortando los ciclos de mejora. Asimismo, permitirá disponer de nuevos marcadores y genes candidatos de interés que podrán ser validados en otras poblaciones e incorporados como información para la generación de modelos predictivos de Selección genómica.

Palabras claves: ddRADSeq, GBS, EUChip60K, selección temprana, genes candidatos, calidad de madera.

Development and application of genomic methodologies for molecular improvement of *Eucalyptus dunnii* by Association Mapping and Genomic Selection

ABSTRACT

The genus *Eucalyptus* is one of the most important in the world and, therefore, its research and genetic improvement concentrates international efforts. In Argentina, the *Eucalyptus* breeding program is mainly carried out by INTA. *Eucalyptus dunnii* is one of the eucalypts species with high demand to produce short-fibre cellulose pulp, especially in those regions where other species of excellent papermaking aptitude such as *E. globulus* and *E. grandis* are affected by limiting soil and climate conditions. In addition, it has good growth and productivity traits to be used as solid wood, although they must be improved. In this context, the new strategies of molecular breeding are useful for this purpose.

The objective of this work is to apply molecular breeding in *E. dunnii* through Association Mapping and Genomic Selection and to develop a high-performance genotyping systems. This is the first time that a ddRADSeq (Double-digest restriction site-associated DNA sequencing) protocol is optimized for the species. This protocol is versatile and useful for both a small number of samples, as well as it can be scaled up to hundreds of samples efficiently and robustly, generating relevant information on SNP and SSR molecular markers. This methodology was successfully applied to an open-pollinated population of *E. dunnii* (308 individuals, belonging to an origin and provenance field trials of the INTA breeding program, located in Ubajay, Entre Ríos). In parallel, and in order to evaluate the potential of this developed methodology, the information generated was analyzed and compared with that of coming from EUChip60K commercial microarray. A total of 8,170 (ddRADSeq) and 19,045 (EUChip60K) SNP were obtained, distributed along the 11 chromosomes and covering complementary genomic regions.

For the Association Mapping analysis (pioneer in this species) a mixed linear model and fourteen phenotypic traits data were evaluated: growth, wood quality (basic density and chemical properties of wood estimated by NIR (Near Infrared Reflectance)) and growth stresses. Genetic structure showed two sub populations for both genotyping platforms consistent with each other and a low linkage disequilibrium was evident. Seven, 13 and 19 (ddRADSeq, EUChip60K and joint matrix; $p < 0.0001$) associated SNPs were found for all phenotypic evaluated traits. The associated markers were aligned with the published *E. grandis* genome sequences and gene mapping positions were carried out within a 70 Kbp window. About 100 genes were located, several of which are potentially related to the analyzed traits. In addition, 50 genes from different metabolic pathways belonging to cellulose and xylanes synthesis, phenylpropanoids, terpenes, laccases, peroxidases, among others, described in public genomic resources were located within 500 Kbp of SNPs mapped in *E. grandis* genome.

Finally, a Genomic Selection concept test was applied with both genotyping methodologies and the breeding values predictors were compared according to their theoretical accuracy (evaluation of the variance of the error in the prediction) and predictive ability (correlation between the phenotype observed and estimated) with

respect to conventional predictions. The theoretical accuracy for the prediction models including genomic information was higher with respect to the conventional results and particularly ddRADseq had better performance with those traits with heritability<0.43 (cel20, dap11, dap20, db20). The other traits had almost no differences. In contrast, the predictive ability was very similar for all models, and the conventional method was slightly higher for some traits with heritability <0.43.

The molecular breeding strategies applied during this thesis will allow early selection of superior genotypes in the *E. dunnii* breeding programs by shortening the improvement cycles, in addition providing new markers and candidate genes of interest that can be validated. in other populations and incorporated as information for predictive models.

Keywords: ddRADSeq, GBS, EUChip60K, early selection, candidate genes, wood quality.

AGRADECIMIENTOS

Al Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) por la beca otorgada para la realización de este trabajo de Tesis.

Al Instituto de Agrobiotecnología y Biología Molecular (IABiMo, UEDD INTA-CONICET, ex Instituto de Biotecnología, IB) del Instituto Nacional de Tecnología Agropecuaria (INTA), a sus directivos por permitirme desarrollar esta Tesis con total libertad y por poner a disposición los materiales e instalaciones de los laboratorios y a todos mis compañeros que hacen del Instituto un segundo hogar.

Quiero agradecerle a mi directora de Tesis la Dra. Susana Marcucci Poltri por confiar en mí, alentarme en numerables oportunidades a capacitarme en cursos, ir a congresos y estadías de intercambio, por hacer lo imposible para que pueda asistir, así como para que pueda cumplir con el sueño de hacer un doctorado. Por brindarme paciencia, contención y trabajar codo a codo cuando fue necesario.

A todo el grupo de forestales del IB Pamela, Cintia, Juan Gabriel, Carolina y Martín gracias por ser mi compañía de cada día y por todos los consejos brindados y palabras de aliento en todo momento que necesité. Es un placer trabajar con ustedes y gracias por haberme recibido tan bien en el grupo de trabajo.

A Norma por depositar su confianza en mí y darme la tarea de poner a punto la técnica de GBS en *E. dunnii*, la oportunidad de capacitarme en el exterior y, luego de poder transmitir los conocimientos adquiridos capacitando a otros. He aprendido muchísimo de todas las experiencias. Gracias, que junto con Caro me aceptaron y compartimos dos años la oficina.

A Carla, por trabajar juntas en la puesta a punto de GBS en varias especies y en el dictado de varios Talleres sobre el tema, y a Pame, que ambas me ayudaron un montón con el trabajo de tesis, transmitiéndome toda su experiencia previa en cómo trabajar en un proyecto de tesis doctoral, en particular con mapeo por asociación y selección genómica y el manejo de R. Gracias por toda la paciencia en las mil preguntas que les hice.

A Pablo Pathauer, Eduardo Cappa, Javier Oberschelp, Leonel Harrand, Juan López y Martín Marcó por los materiales de la población de *E. dunnii* de Argentina, y algunos comandos en R, que sin ellos esta Tesis no hubiera sido posible.

A las "Mate's Girls": Laura, Mónica, Noelia, Flavia, Pamela y Carla que compartieron cada día de esta Tesis, por las salidas culturales, baby showers, las numerosas charlas, consejos y palabras de aliento. Gracias por hacer cada día en el trabajo mucho más llevadero.

A toda la unidad de genómica (UGB), Andre, Pablo, Vero, Marianne, Vale, Ana y Nico, por el cariño de siempre, la ayuda y paciencia brindada, y por prestarme los equipos y espacio para poder llevar adelante mi trabajo. Siempre tendré mi corazón en UGB. A Andre en particular, por haberme empujado a comenzar a realizar el doctorado, gracias por la gran ayuda.

A todo el grupo del "locutorio" (los que están y los que se fueron buscando nuevos horizontes), muchísimas gracias por las charlas, los mates, las risas que me brindaron.

A mis amigos del INTA Bus, Edu, Diana, Mari y Meli, por darme su cariño y siempre alentarme a seguir adelante. Con ustedes comencé a tener una familia dentro del INTA.

A mis amigas de toda la vida Pame y Mari por estar siempre, en las buenas y en las malas. Amigas las quiero mucho.

A Guada y Dali, que aunque no nos veamos seguido, siempre están y son un ejemplo a seguir de trabajo y valentía. Las quiero.

A las Bombonas: Meli, Nati, Jesi, Lu, Fio y Ani por transitar este camino de crecer como personas y convertirnos en profesionales juntas.

A Marco por bancarme en los muchos viajes de trabajo, horas de laburo extra horario, alentarme a seguir adelante en todos mis proyectos y a ayudarme a bajar un cambio en los momentos de mucho estrés, ¡Te amo!. A Isabel por siempre recordarme que soy capaz de hacer todo lo que me proponga.

A mi familia, por acompañarme en cada momento de mi vida y ser un gran apoyo. A mi hermana Juli por ser mi amiga incondicional, compañera de aventuras de todo tipo, y por hacer los gráficos con diseños exclusivos para la tesis, pósters, y el paper. A mis Abuelos, dos de los cuales me transmitieron el gen de las ciencias, por ser un gran ejemplo a seguir. A mis padres por estar conmigo cada día, por hacer hasta lo imposible para que pudiera seguir adelante y bancarme hasta para hacer un posgrado. No hay palabras para agradecerles lo mucho que hicieron y hacen por mí. ¡Los quiero muchísimo!

A mi familia y amigos

Parte de los resultados presentados en este trabajo de Tesis fueron publicados:

Aguirre, N.C.; Filippi, C.V.; Zaina, G.; Rivas, J.G.; Acuña, C.V.; Villalba, P.V.; García, M.N.; González, S.; Rivarola, M.; Martínez, M.C.; Puebla, A.F.; Morgante, M.; Hopp, H.E.; Paniego, N.B.; Marcucci Poltri, S.N. *Optimizing ddRADseq in Non-Model Species: A Case Study in Eucalyptus dunnii Maiden. Agronomy* **2019**, *9*, 484. https://doi.org/10.3390/agronomy9090484

Aguirre, N.C.; Villalba, P.V.; Filippi, C.V.; Rivas, J.G.; García, M.N.; Martínez, M.C.; Acuña, C.V.; López, J.A.; López, J.A.; Pathauer, P.; Palazzini, D.; Harrand, L.; Oberschelp, J.; Marcó, M.; Hopp, H.E.; Rodrigues, J.; Grattapaglia, D.; Cappa, E.P.; Paniego, N.B.; Marcucci Poltri, S.N. *Comparación entre GBS y EUChip60K para la obtención de SNP informativos en una población de Eucalyptus dunnii Maiden. Comparison between GBS and EUChip60k for obtaining informative SNP data in a Eucalyptus dunnii Maiden population.* VIII Reunión GeMFo Bella Vista, Corrientes, Argentina 21 al 23 de agosto de 2019 Editores/Compiladores: Juan Adolfo López, Mariano Agustín Hernández, Augusto Javier López. Pp45. ISBN 978-987-86-1630-8.

Natalia C. Aguirre, Pamela V. Villalba, Carla V. Filippi, Juan G. Rivas, Martín N. García, María C. Martínez, Cintia V. Acuña, Javier A. López, Juan A. López, Pablo Pathauer, Dino Palazzini, Leonel Harrand, Javier Oberschelp, Martín Marcó, Esteban F. Cisneros, Rocío Carreras, Horacio Hopp, José Rodrigues, Dario Grattapaglia, Eduardo P. Cappa, Norma B. Paniego, Susana N. Marcucci Poltri. *"GBS and EUChip60k SNP Data Comparison for GWAS on a Eucalyptus dunnii Breeding Population in Argentina"*. IUFRO Tree Biotechnology 2019 Meeting. Forests, Technology & Society. 23 to 29 of June. North Carolina State University, Raleigh, USA, Póster.

Aguirre NC, Filippi CV, Villalba PV, García MN, Rivas JG, Martínez MC, Acuña CV, López AJ, López JA, Cappa EP, Pathauer PS, Palazzini DA, Harrand L, Oberschelp J, Marco MA, Cisneros EF, Carreras R, Hopp HE, Rodrígues JC, Grattapaglia D, Marcucci Poltri SN. *Use of EUChip60K for the genetic diversity characterization of an Eucalyptus dunnii breeding population of Argentina.* Congreso internacional Eucalyptus 2018, IUFRO, Improvement and culture of Eucalyptus, Managing Eucalyptus plantations under global changes. Le Corum Conference Center, Montpellier, France. Del 17 al 20 de Septiembre de 2018. Póster.

Aguirre, N.; Filippi, C.; Acuña, C.; Villalba, P.; Martinez, C.; Rivas, G.; García, M.; López, J.; López, A.; Oberschelp, J.; Harrand, L.; Puebla, A.; Paniego, N.; Hopp, E.; Rivarola, M; Marcucci Poltri, S. *"Evaluación de las potencialidades del método de Genotyping by Sequencing (GBS-ddRADseq) para el Mejoramiento Asistido de Eucalyptus dunnii".* REDBIO Argentina 2017, 11 a 13 de septiembre, Centro Científico Tecnológico- CONICET Bahía Blanca, Bs As, Argentina. Póster.

Aguirre, Natalia Cristina; Filippi, Carla Valeria; Zaina, Giusi; Rivas, Juan Gabriel; Acuña, Cintia; Villalba, Pamela Victoria; García, Martín Nahuel; Scaglione, Davide; Morgante, Michele; González, Sergio; López, Juan Adolfo; López, Augusto Javier; Oberschelp, Javier; Harrand, Leonel; Martínez, María Carolina; Puebla, Andrea Fabiana; Hopp, Horacio Esteban; Rivarola, Máximo; Paniego, Norma Beatriz; Marcucci Poltri, Susana Noemí. *Desarrollo de una estrategia de genotipificación por secuenciación para el mejoramiento asistido de Eucalyptus dunnii. Development of a genotyping by sequencing strategy for assisted breeding of Eucalyptus dunnii*. Pág. 65 a 68. VII Reunión GeMFo, Trabajos Técnicos, San Miguel de Tucumán, Tucumán, Argentina, 24 al 26 de agosto de 2016, Grupo de Genética y Mejoramiento Forestal. Juan Adolfo López; Luis Fernando Fornes; compilado por Juan Adolfo López; Luis Fernando Fornes. 1a ed. - Bella Vista: Juan Adolfo López, 2016. Archivo Digital, ISBN 978-987-42-1792-9. DOI: 10.13140/RG.2.2.34983.96167.

N. C. Aguirre, C.V. Filippi, G. Zaina; D. Scaglione, M. Morgante, S. González, M. Rivarola, J. A. López, A. J. López, J. Oberschelp, L. Harrand, M.C. Martinez, A. F. Puebla, H. E. Hopp, N. Paniego; S. N. Marcucci Poltri. *Puesta a punto de Metodologías Genómicas para el Mejoramiento Molecular de Eucalyptus dunnii.* 5º Simposio Internacional de Biotecnología e Ingeniería ambiental, 25 al 29 de julio de **2016**, Campus Miguelete Universidad Nacional de San Martín, San Martín, Bs. As., Argentina. Presentación Oral.

ÍNDICE GENERAL

ABREVIATURAS, ACRÓNIMOS Y SIGLAS

Locus o ***loci***: se utilizan en esta tesis en el sentido sensu lato. Es decir, una posición genómica ocupada por uno o unos pocos marcadores moleculares ubicados en una secuencia (entre 66 y 250 pb) que puede o no ser codificante.

A	adenina
ABLUP	*Best linear unbiased prediction* para matriz A de pedigrí
ADN o DNA	Ácido desoxirribonucleico o *Deoxyribonucleic acid*
EMMA	*Efficient Mixed Model Association*
Ampure XP	*Marca comercial de perlas magnéticas para purificar ADN*
ANOVA	Análisis de la varianza
ApeKI	Enzima de restricción ApeKI
VCF	Formato de llamada variante o *Variant Call Format*
AT	Altura Total
at20	Altura total a los 20 años
at6	Altura total a los 6 años
BAM	archivo *Binary Alignment Map*
BCUN	*Boomi Creek y Unumungar State Forest*, Australia
BED	formato *Browser Extensible Data*
BIC	Criterio de Información Bayesiano de Bayesian Information Criterion
BLUP	*Best linear unbiased prediction*
C	Citosina
°C	Grados centígrados
CABA	Ciudad Autónoma de Buenos Aires
cel	celulosa
cel20	celulosa total a los 20 años
CESA	ruta de biosíntesis de celulosa y xilano
CICVyA	Centro de Investigación en Ciencias Veterinarias y Agronómicas
Chip	Chip o microarreglo
CMLM	Modelo lineal mixto comprimido o *Compressed Mixed Linear Model*
CNIA	Centro Nacional de Investigaciones Agropecuarias
CO_2	Dióxido de carbono
CONICET	Consejo Nacional de Investigaciones Científicas y Técnicas
CP	Componentes principales
CTAB	Bromuro de cetiltrimetilamonio o *Cetyl Trimethyl Ammonium Bromide*
CV	Coeficiente de variación
d.e.	Desvío estándar
DAP	Diámetro a 1,30 m de altura o Diámetro a la altura del pecho
dap11	diámetro a la altura del pecho a los 11 años

dap20	diámetro a la altura del pecho a los 20 años
dap6	diámetro a la altura del pecho a los 6 años
DAPC	Análisis Discriminante de Componentes Principales *o Discriminant Analysis of Principal Components*
DArT	Tecnología de matrices de diversidad o *Diversity arrays technology*
db o db20	densidad básica o densidad básica a los 20 años
ddRADseq	*Double-digest Restriction-site Associated DNA sequencing*
DHAC	Death Horse Track Region y Acacia Creek, Australia
DL	Desequilibrio de ligamiento
EEA	Estación Experimental Agropecuaria
EE.UU.	Estados Unidos
EMMA	*Efficient Mixed Model Association*
ET	Exactitud Teórica
Eucgr	prefijo de genes de *Eucalyptus*
EUChip60k	Microarreglo comercial de *Eucalyptus*
Extet- extet20	Extractivos etanólicos o Extractivos etanólicos a los 20 años
Exttot - exttot20	Extractivos totales o extractivos totales a los 20 años
F_{ST}	Índice de fijación o *fixation index*
FAO	Organización de las Naciones Unidas para la Alimentación y la Agricultura
Faostat	Estadísticas de la Organización de las Naciones Unidas para la Alimentación y la Agricultura
FDR	*False Discovery Rate*
for o for6	forma de fuste o forma de fuste a los 6 años
FASTA	formato FASTA
Fastq	formato fastq de calidad de secuencias
FSC	*Forest Stewardship Council*
G	Guayacilo
G	Matriz genómica
G	Guanina
GAPIT	*Genomic Association and Prediction Integrated Tool*
Gb	Giga pares de bases
GBLUP	modelo *Genomic Best linear unbiased predictor*
GBS	Genotipado por secuenciación o *Genotyping by Sequencing*
GC	Genes candidatos, del inglés *Candidate Gene*
gff3	*General Feature Format 3*
GL	Grupos de ligamiento
GO	*Gene Ontology*
GWAS	Estudio de Asociación de Genoma Amplio o *Genome wide association study*
H	Matriz combinada de relaciones
ha	hectárea

HBLUP	*Best linear unbiased prediction* para matriz combinada de relaciones H
He	Heterocigosis esperada
Ho	Heterocigosis Observada
HP	Habilidad Predictiva
HSC	Huerto semillero clonal
HSP	Huerto semillero de progenies
IABiMo	Instituto de Agrobiotecnología y Biología Molecular (UEDD INTA-CONICET)
I+D+i	Investigación, desarrollo e innovación
ie	"id est" o "es decir"
INDEC	Instituto nacional de estadísticas y censos
InDel	Inserciones/deleciones
INTA	Instituto Nacional de Tecnología Agropecuaria
IR o ir20	Índice de rajado en rollizos o índice de rajado en rollizo a los 20 años
ISA	Instituto Superior de Agronomía, Portugal
K	Matriz de parentesco o mil
Kg/m^3	kilogramos por metro cúbico
klas20	Lignina Klason a los 20 años
Km	Kilómetro
Kpb	Kilo pares de bases
LD-kNNi	algoritmo *Linkage Disequilibrium k-nearest neighbor genotype*
SE	lecturas o secuencias Single end
lig20	Lignina total a los 20 años
LVL	Maderas de chapas laminadas
M	metro
MA	Mapeo por Asociación
MAB	Mejoramiento asistido por marcadores *o Marker Assisted Breeding*
MAF	Frecuencia alélica mínima o *Minimum allele frequency*
MAS	Selección asistida por marcadores o *Marker Assited Selection*
matriz Q	Matriz de estructura poblacional
MboI	Enzima de restricción MboI
MDF	Tablero de fibras de densidad media
MiddRADseq	*Mi Double-digest Restriction-site Associated DNA sequencing*
Min.	Minutos
MISA	software MicroSAtellite
MiSeq	Equipo de secuenciación de la empresa Illumina
MLM	Modelo Lineal Mixto o *Mixed Linear Model*
mm	milímetro
MOE	Módulo de elasticidad
MOR	Módulo de rotura

Mpb o Mb	Mega pares de bases
N	Cualquiera de las Base nucleotídicas del ADN
n°	número
NEB	*New England Biolabs*
NextSeq	Equipo de secuenciación de la empresa Illumina
ng	nanogramos
NGS	*Next Generation Sequencing*
NIR	Reflectancia en el infrarrojo cercano o *Near Infrared Reflectance*
NovaSeq	Equipo de secuenciación de la empresa Illumina
NSW	New South Wales
OC	Oaky Creek, Australia
OP	población de polinización abierta o *Open Pollinated population*
org	organización
OSB	tableros de fibras orientadas
P	Frecuencia alélica media del alelo minoritario o p-valor
P1	Protocolo 1
P2	Protocolo 2
pb	pares de bases
PCA	Análisis de componentes principales
PCR	Reacción en cadena de la polimerasa o *Polymerase Chain Reaction*
PE	Población de entrenamiento
PE	lecturas o secuencias pareadas Paired End
PEFC	Programa para el Reconocimiento de Certificación Forestal
PEV	Varianza predicha del error o *Prediction error variance*
PF	*Pasing filter* o lecturas que pasaron el filtro de calidad
PIC	Información del contenido polimórfico o *Polymorphic information content*
Pmol	Pico mol
p	Frecuencia alélica media del alelo minoritario
SBP	proteínas de unión a ribonucleasa S o *S-ribonuclease binding protein*
PstI	Enzima de restricción PstI
PV	Población de validación
PyMES	Pequeñas y medianas empresas
q	Frecuencia alélica media del alelo mayoritario
QTLs	*loci* de caracteres cuantitativos o *Quantitative Trait Loci*
r^2	correlación
R^2	Varianza fenotípica explicada por un marcador
RADseq	Secuenciación de ADN asociada a los sitios de restricción o *Restriction-site Associated DNA sequencing*
REML	Inferencia de máxima verosimilitud restringida
RFU	Unidades de fluorescencia relativa o relative fluorescence units

RILs	Línea endogámica recombinante o *Recombinant Inbred Lines*
S	Siringilo
SAM	Archivo en formato *Sequence Alignment Map*
SD	Procedencia de plantación de Oliveros, Argentina
SDRLK	S-domain receptor like kinase o Receptor de la subfamilia del dominio S tipo quinasa
SE	Lecturas o secuencias *Single End*
seg	Segundo
SG	Selección Genómica
S/G o S:G o sg20	Relación de monómeros de lignina Siringilo-Guayacilo y S:G a los 20 años
SNPs	Polimorfismo de Nucleótido Sinple o *Single Nucleotide Polymorphism*
SphI	Enzima de restricción SphI
SphI-HF	Enzima SphI de alta fidelidad o *High Fidelity*
ssGBLUP	modelo *single step Genomic Best linear unbiased predictor*
SSR	Repetición de secuencia simple polimórfica o *Simple sequence Repeat* o microsatélites
SY	*South of Yabra State Forest*, Australia
T	timina
Tn	toneladas
UEDD	Unidad Ejecutora de Doble Dependencia
US$	Dólares
USD	Dólares estadounidenses
VBP	Valor bruto de la producción
VC	valores de cría
vs	versus

1 INTRODUCCIÓN

1. LOS RECURSOS FORESTALES MUNDIALES

Las poblaciones forestales son sistemas biológicos altamente complejos, con ciclos de vida largos y con una gran diversidad genética. Ésta es extremadamente importante ya que les permite a las especies forestales sobrevivir a los constantes cambios ambientales y proporciona una base fundamental para el mejoramiento genético.

A nivel global, estos sistemas forestales cubren alrededor del 31% del área total de la Tierra (aproximadamente 4 billones de hectáreas). Con una distribución heterogénea, los cinco países con mayor riqueza forestal (la Federación de Rusia, Brasil, Canadá, Estados Unidos de América y China) suman un 53% del área mencionada según el último análisis de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO, del inglés *Food and Agriculture Organization of the United Nations*). La mayoría de los bosques del mundo son bosques naturales, observando que ha habido un aumento progresivo de los bosques plantados (Figura 1.1).

Figura 1.1. *Proporción de superficie forestal natural y plantada a nivel mundial*. Tomado de: www.fao.org/forest-resources-assessment/es.

Actualmente, se destaca la importancia de los bosques como sumideros de dióxido de carbono ya que absorben y almacenan carbono en la biomasa por encima y por debajo del suelo; son hábitats para la conservación de la biodiversidad; son proveedores de servicios ambientales (aire y agua limpios, mitigación de los efectos del cambio climático, recreación, actividades culturales y espirituales) y son importantes como sostén de los medios de vida y de oportunidades económicas. Desempeñan una función fundamental en la lucha contra la pobreza rural, el logro de la seguridad alimentaria y medios de subsistencia decentes. En ese sentido, los bosques suministran a la población mundial madera y productos forestales no maderables (FAO, 2015) y en los países de bajos ingresos la leña sigue siendo el producto maderable más importante (Figura 1.2).

Figura 1.2. *Proporción de uso de madera como leña a nivel mundial*. Tomado de: www.fao.org/forest-resources-assessment/es.

Como se mencionó anteriormente (Figura 1.1), el área de bosques plantados aumentó en más de 105 millones de ha desde 1990. La tasa media anual de incremento entre 1990 y 2000 fue de 3,6 millones de ha. La tasa alcanzó su máximo nivel y llegó a 5,9 millones de ha por año en el período 2000-2005, para luego disminuir a 3,3 millones de ha (2010-2015) por año, conforme se reducía la plantación en América del Norte, Asia oriental, Asia meridional y sudoriental, y Europa (FAO, 2015).

Las analogías y diferencias entre los bosques naturales y los bosques plantados son asuntos muy debatidos entre las partes interesadas que estudian el cambio del bosque. Los bosques naturales contribuyen a la conservación de la diversidad de los genotipos y al mantenimiento de la composición natural de las especies arbóreas, a la estabilidad de su estructura y a la dinámica ecológica. Los bosques plantados se suelen establecer como bosques de producción o con el propósito de proteger el suelo y el agua. Adecuadamente gestionados, estos bosques pueden proporcionar varios productos y servicios y contribuir a reducir la presión sobre los bosques naturales (FAO, 2015).

Consecuentemente, ha habido un aumento en la utilización de especies cultivadas y un marcado declive en el empleo de especies nativas en la industria forestal; en particular, debido a que las especies cultivadas tienen rápido crecimiento y el principal objetivo de estas plantaciones es el incremento de producción por hectárea implantada (Buongiorno & Zhu, 2014). Entre los principales géneros de rápido crecimiento encontramos a *Populus*, *Acacia*, *Acer, Pinus, Eucalyptus,* entre otros, siendo este último preponderante por su mayor implantación en el mundo (Figura 1.3).

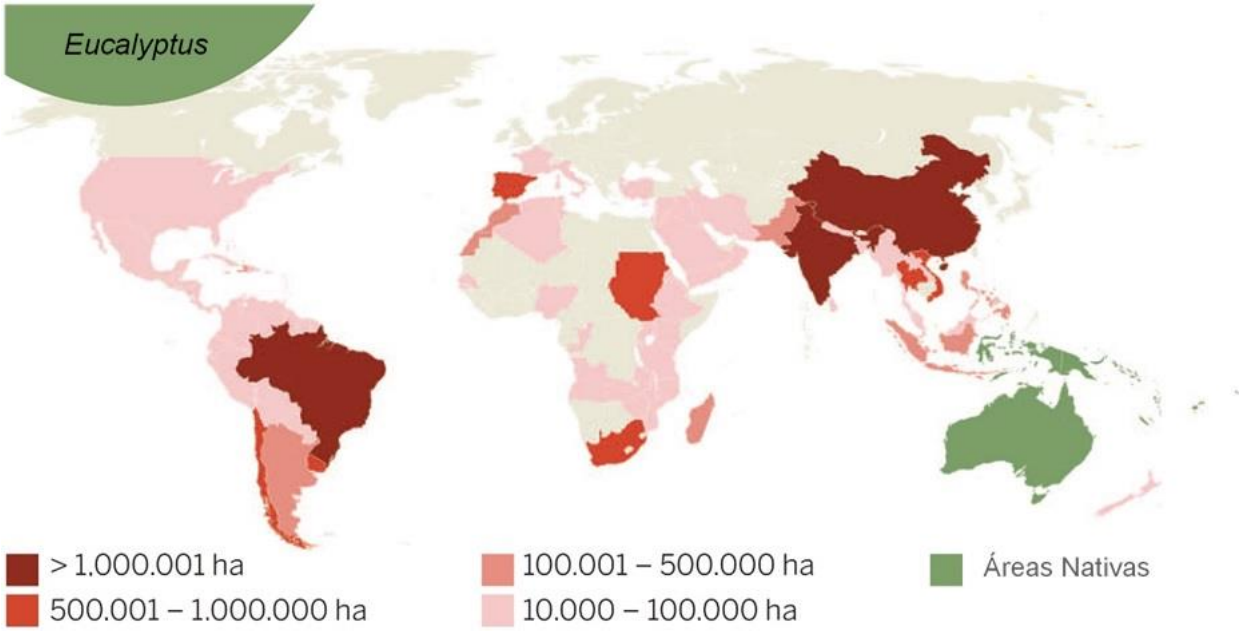


Figura 1.3: *Distribución mundial de Eucalyptus: Áreas Nativas y plantaciones.* Referencia: Cantidad de hectáreas (ha) implantadas por país. Fuente: Wingfield et al. (2015).

En 1955 se estimaba que la superficie plantada con eucaliptos era de alrededor de 700.000 ha en todo el mundo (FAO, 1981). Actualmente, las plantaciones comerciales de eucaliptos abarcan alrededor de 4 millones de ha en 58 países y regiones, incluyendo Australia; otros 50 tienen plantaciones experimentales u ornamentales (FAO, 2015). Algunas de estas regiones podrían emprender plantaciones comerciales en los próximos años (http://www.fao.org/3/AC459S/AC459S04.htm).

2. PRODUCCIÓN Y COMERCIO MUNDIAL DE PRODUCTOS FORESTALES

Los principales productos forestales maderables son: la madera en rollo industrial; la madera aserrada; los paneles de madera; la fibra utilizada para manufacturar papel y cartón; el papel y el cartón como productos finales y el combustible de leña, carbón vegetal y pellets (FAO, 2018a).

La madera en rollo industrial es toda la madera en rollo utilizada para cualquier propósito que no sea la energía. Se compone de: madera para pulpa; madera aserrada y madera de chapa; y otra madera en rollo industrial (por ejemplo, madera en rollo utilizada para postes de cercas y postes de teléfono o electricidad; FAO, 2018). La madera aserrada comprende los tablones, durmientes (travesaños), vigas, viguetas, tablas, listones, "madera" propiamente dicha, etc.

La categoría de productos de paneles de madera consiste en madera contrachapada (incluidos los tableros compensados y las maderas de chapas laminadas denominadas LVL), aglomerados, tableros de partículas,

tableros de fibras orientadas (OSB) y tableros de fibra como el MDF (tablero de fibras de densidad media), CLT (madera laminada cruzada); Parallam entre otros.

La fibra utilizada para fabricar papel y cartón incluye el papel recuperado (papel de desecho), otra pasta de fibra y la pulpa de madera utilizada para hacer papel. El grupo de productos de papel y cartón comprende los papeles gráficos (papel de prensa, de impresión y de escritura) y otros papeles y cartones. Este último se subdivide a su vez en papel y cartón de embalaje, papel doméstico e higiénico y otros papeles y cartones.

El combustible de madera es madera en rollo que se utiliza como combustible para cocinar, calentar o producir energía e incluye la madera usada para hacer carbón y pellets. Incluye madera cosechada de los tallos principales, ramas y otras partes de los árboles y astillas (chips) de madera que se hacen directamente (es decir, en el bosque) de madera en rollo para ser usadas como combustible.

En el año 2018 la producción y comercio mundiales de todos los principales productos madereros registraron sus valores más altos según el último informe de FAO (FAO, 2018b) (Figura 1.4). La producción, las importaciones y las exportaciones de madera en rollo, madera aserrada, paneles de madera, pulpa de madera, carbón de leña y pellets alcanzaron su máximo desde 1947, cuando la FAO comenzó a informar sobre las estadísticas de productos de los bosques del mundo. En 2018, el crecimiento de la producción de los principales grupos de productos basados en la madera osciló entre el 1% (tableros de madera) y el 5% (madera en rollo industrial). El más rápido crecimiento se produjo en Asia y el Pacífico, América del Norte y América del Sur y Regiones europeas, probablemente debido a un crecimiento económico positivo en estas áreas. El año 2018 no fue tan fuerte para la industria del papel. La producción mundial de papel y cartón fue del 1,5% debido principalmente a las interrupciones en el suministro de papel recuperado y la continua sustitución de los medios de comunicación impresos con productos digitales.

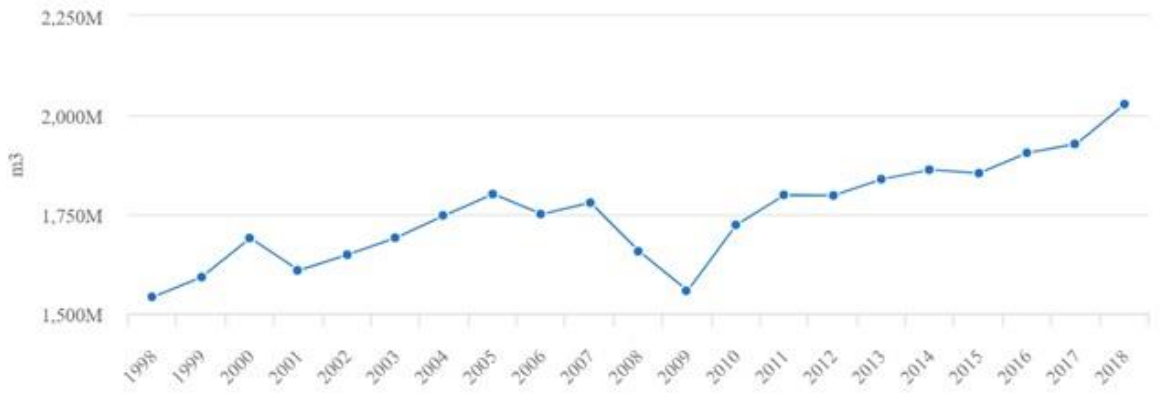


Figura 1.4: *Producción Mundial Anual de Madera en rollo industrial entre 1998 y 2018.* Fuente: http://www.fao.org/faostat/es/#data/FO.

La producción de pellets de madera ha aumentado drásticamente en los últimos años, debido principalmente a la demanda generada por la bioenergía y objetivos establecidos por la Comisión Europea. En 2018, la producción global creció un 11%, alcanzando los 37 millones de toneladas, de los cuales más de la mitad (24 millones de toneladas) se comercializaron internacionalmente.

Según estudios realizados por la FAO (2018), el aumento de la demanda de productos forestales y de sus derivados continuará en alza, junto con el crecimiento demográfico y con la modificación de nuevos estilos de vida, ya sea debido a una expansión de la nueva clase media, a la transición mundial hacia una vida predominantemente urbana u a otros factores. Es por esto por lo que será necesario adoptar técnicas de producción más eficientes para poder satisfacer estas demandas crecientes (FAO, 2018). Sin embargo, un factor determinante son las inversiones en el sector forestal que están directamente ligadas a los ingresos que se pueden obtener a partir de los bosques. Las nuevas políticas medioambientales y energéticas, así como las variaciones en las economías regionales afectarán la producción forestal. En ese sentido, también debe ser tenida en cuenta la producción de bioenergía a partir de residuos forestales como fuente de biomasa en un contexto mundial de agotamiento de los combustibles fósiles.

En la actualidad, los mayores productores de bienes madereros se ubican en tres regiones distintas: la región de Asia y el Pacífico, América del Norte y Europa (http://www.fao.org/faostat/es/#data/FO, FAOSTAT, 2018). Estas regiones, en conjunto, representaron alrededor del 90 por ciento del total del comercio mundial de productos forestales (FAO, 2018). La mayor parte del comercio internacional de estos bienes es entre países dentro de cada una de estas regiones o entre ellas (Figura 1.5). Como se puede observar en las siguientes figuras, los principales países productores varían dependiendo del artículo a fabricar. Por ejemplo, para la creación de tableros de madera el mayor productor es China con más de 100 millones de metros cúbicos de producción mundial seguido por Estados Unidos con un poco más de 38 millones de metros cúbicos (Figura 1.6). Sin embargo, estos valores se invierten para la producción de pulpa para papel en donde los mayores productores son Estados Unidos y Canadá (Figura 1.7) (FAO, 2018).

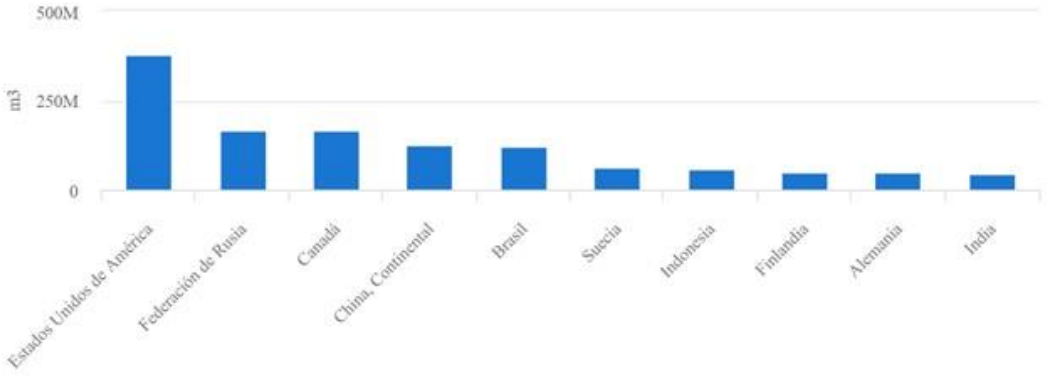


Figura 1.5: *Producción mundial de Madera en rollo industrial. Principales países productores. Promedio entre 1998 y 2018* (Referencia: M=millón. China, continental: no incluye los territorios de Hong Kong, Macao y Taiwán). Fuente: http://www.fao.org/faostat/es/#data/FO.

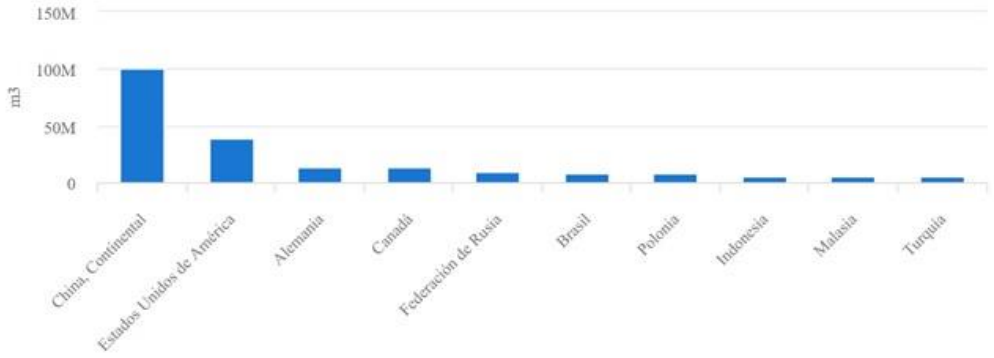


Figura 1.6: *Producción mundial de Tableros de Madera. Principales países productores. Promedio entre 1998 y 2018* (Referencia: M=millón. China, continental: no incluye los territorios de Hong Kong, Macao y Taiwán). Fuente: http://www.fao.org/faostat/es/#data/FO.

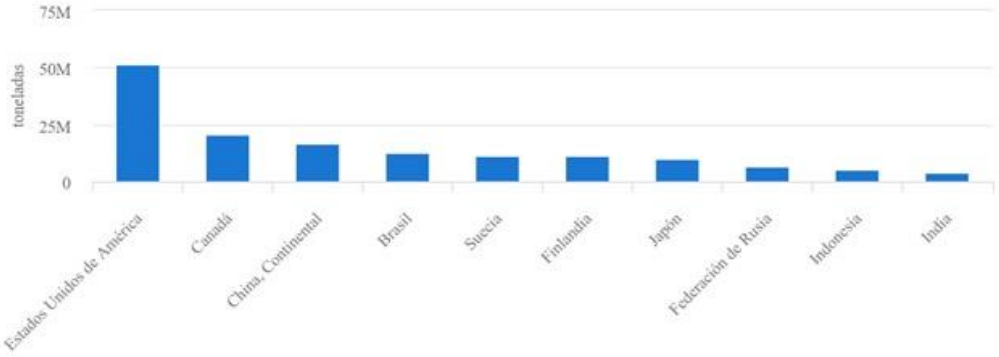


Figura 1.7: *Producción mundial de Pulpa para Papel. Principales países productores. Promedio entre 1998 y 2018* (Referencia: M= millón. China, continental: no incluye los territorios de Hong Kong, Macao y Taiwán). Fuente: http://www.fao.org/faostat/es/#data/FO.

A pesar de los altos niveles de producción que se pueden apreciar dentro de estas regiones, todavía son insuficientes para satisfacer la demanda en estos países. Por lo tanto, se ve incrementada la dependencia de las importaciones generando así las inversiones en producción maderera en otras regiones forestales, primordialmente en América del Sur. Esto a su vez ocasionó que en América Latina y el Caribe se haya incrementado la producción, el consumo y el comercio de la mayoría de los productos forestales, principalmente de aquellos que se elaboran a partir de maderas provenientes de plantaciones de mejora. Las principales especies comerciales que se producen en el continente son el eucalipto (principalmente *Eucalyptus grandis, E. globulus* e híbridos) y el pino (*Pinus taeda, P. elliottii*).

La demanda de productos forestales solamente se puede satisfacer mediante una gestión sostenible que incluya una producción más eficiente; en particular, porque se está trabajando a partir de un recurso estático o en disminución. Con esta problemática en mente, en la última década, ha habido un incremento en el campo de la investigación en estudios relacionados con el aumento de la producción de biomasa de los árboles (Ragauskas et al., 2006) y una mejora en la eficiencia de la utilización de los mismos sumado al interés de emplear la biomasa forestal para la producción de bioenergía. Por estas razones, se debe explorar el potencial para optimizar la composición genética de árboles para lograr una mayor productividad en sus ambientes de cultivo (Nasholm et al., 2014). La implementación de estrategias genómicas modernas como el Mapeo por Asociación (MA) y Selección Genómica (SG o GS del inglés *Genomic Selection*) no sólo permitirán lograr estos objetivos sino también disminuir el tiempo de obtención de los mismos.

3. SITUACIÓN FORESTAL EN ARGENTINA

Argentina posee una gran diversidad de ambientes debido a su variedad topográfica y climática que depende del rango de latitud y longitud. Dichos ambientes incluyen bosques subtropicales y templados, de los cuales 27.221.721 hectáreas son clasificadas como Tierras Forestales (más de 0,5 ha cubiertas en más de un 10% por árboles de una altura superior a 5 m) y 65.441.306 como Otras Tierras Boscosas (tierra no clasificada como "bosque" con más de 0,5 ha con árboles de una altura superior a 5 m y una cubierta de dosel de 5 a 10 por ciento, FAO, 2012).

Asimismo, Argentina tiene entre 8 y 20 millones de ha disponibles para la forestación, de las cuales 5 millones no compiten en el uso con otras alternativas agropecuarias. La amplia diversidad de climas, suelos y especies, el marco jurídico propicio para las inversiones forestales, los bajos costos de producción y las altas tasas de crecimiento y rotaciones cortas de ciertas especies forestales, ofrecen ventajas comparativas especiales para la implantación de bosques cultivados. Además, cuenta con un amplio espectro de empresas industriales, alta capacidad de adaptación y buen potencial para el empleo calificado (FAO, 2012).

De dicha área disponible para forestación, sólo cuenta con 1,3 millones de ha de bosques cultivados

(también denominados bosques implantados, cultivados o plantaciones forestales), representados principalmente por especies exóticas de rápido crecimiento de los géneros *Pinus* 54%, *Eucalyptus* 32%, *Populus* y *Salix* con el 11% y otras especies con el 3 %. Esta superficie de bosques implantados se concentra principalmente en las regiones mesopotámica, pampeana y patagónica andina (Figura 1.8; FAO, 2012). El 80% de las plantaciones se concentra en la región Mesopotámica, en las provincias de Misiones, Corrientes y Entre Ríos, quedando un 20 % que se reparte en diversas regiones del país (FAO, 2012). Las extracciones de madera a partir de bosques implantados han aumentado más de 3 veces en los últimos 30 años. En el año 2017, el total de madera extraída en el país fue de 12.589.588 tn, de las cuales 7.845.782 tn (62,3%) fueron de coníferas (siendo el 58% de *Pinus sp.*), 4.119.755 tn (32,7%) de *Eucalyptus sp.*, y 593.605 tn (4,7%) de la familia de *Salicaeas* (Figura 1.9; Secretaría de Agroindustria, 2018).

Figura 1.8: *Distribución de las plantaciones Forestales en la República Argentina.* Referencia: Plantaciones indicadas en color verde. Fuente: El estado de los recursos genéticos forestales en el mundo- Informe Nacional- Argentina- Dirección de Producción Forestal del Ministerio de Agricultura, Ganadería y Pesca de la Nación para FAO, 2012.

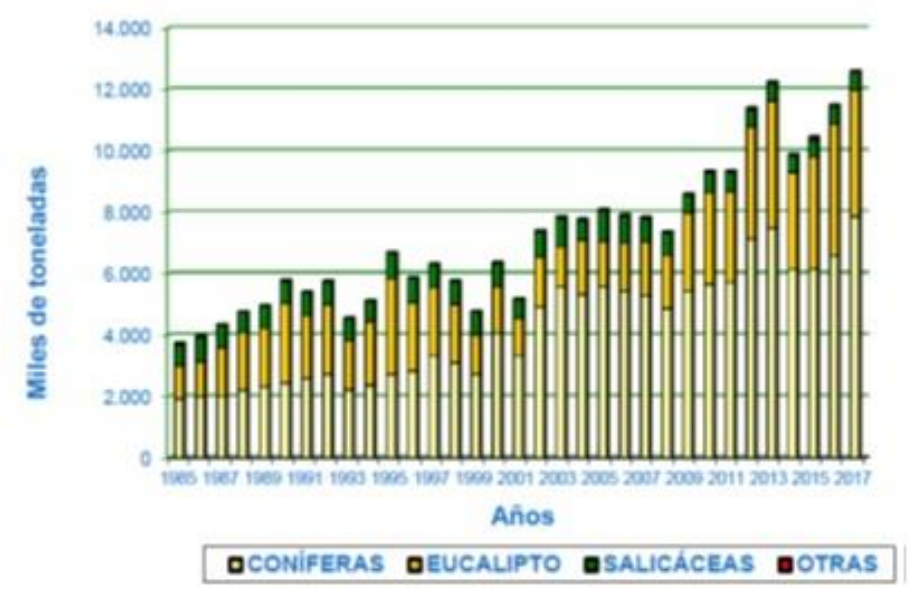


Figura 1.9: *Evolución de las extracciones de madera por especie en la República Argentina*. Fuente: Sector Forestal 2017, Dirección Nacional de Desarrollo Foresto Industrial, Secretaría de Agroindustria, Ministerio de Producción y Trabajo (2018).

Desde hace ya varios años que el Estado Nacional promueve, a través de un apoyo económico no reintegrable (A.E.N.R., Ley 25.080), el establecimiento de plantaciones forestales en el país como una alternativa económica e, incluso, existe un incentivo adicional para que la plantación se realice con material genético de calidad superior o seleccionado.

Un aspecto estratégico del desarrollo forestal nacional es garantizar la conservación y el uso sostenible de los recursos genéticos disponibles de las principales especies de cultivo. En el desarrollo forestal actual, la silvicultura intensiva proporciona una estrategia fundamental para la producción industrial de madera.

Generalmente las plantaciones forestales se basan en especies exóticas (eucalipto, pino, etc.) porque son especies de crecimiento extremadamente acelerado permitiendo planificar con rapidez y con una cierta seguridad futuros polos forestales (FAO, 2018). Una de las ventajas, conjuntamente con la producción de madera a gran escala, es que son los únicos cultivos aceptados hasta el momento para la comercialización de Bonos de carbono, que son los bonos que pagan los países emisores de dióxido de carbono (CO_2) a los productores por la captación de dicho compuesto. A su vez, es importante destacar su uso como reemplazo para la madera de especies nativas (*Araucaria angustifolia*, *Cedrela balansae*, *Nothofagus pumilio*, etc.) que era utilizada para fines "poco nobles" (cajones, tarimas, envases, etc.) y que no justificaban el empleo de madera de valor, pudiendo recurrir a especies como los eucaliptos o los pinos (Braier et al., 2004).

A raíz de los compromisos asumidos bajo el Acuerdo de París del 2015, en la Argentina en el 2019 se presentó un plan llamado ForestAr 2030 (Plan Estratégico Forestal y Foresto Industrial 2030,

https://www.argentina.gob.ar/forestar2030), que involucra a varios ministerios, y que busca reducir las emisiones de CO_2. ForestAr 2030 es una plataforma multisectorial que apunta a la conservación y ampliación del patrimonio forestal argentino y la activación de una economía forestal que impulse el desarrollo social, económico y ambiental.

Las proyecciones de demanda de madera (*World Wildlife Fund*), calculan que por el consumo de madera y para que no se pierdan más bosques nativos se deben plantar globalmente siete millones de hectáreas por año. Para ello, Argentina podría contribuir ya que cuenta con un alto potencial para crecer en forestaciones de alta productividad. Cerca del 40% de las plantaciones forestales en el país se encuentran certificadas por sellos de gestión sostenibles, como son la certificación de Gestión Forestal FSC (Forest Stewardship Council) y PEFC (Programa para el Reconocimiento de Certificación Forestal). Estas plantaciones proveen materia prima al 95% de las industrias de base forestal del país, que incluyen alrededor de 2.700 PyMES que emplean en forma directa cerca de 100.000 personas (Vilella, 2019).

En 2017, según el INDEC, el valor bruto de la producción (VBP) alcanzó los US$14.000 millones, equivalente al 7,3% del valor agregado industrial. Pero hay déficit en el campo de las exportaciones de origen forestal: en 2017 fueron de US$551 millones, el récord fue en 2011 con US$1.120 millones, y actualmente persiste el déficit comercial de US$558 millones. De las exportaciones, el papel representaba el 33% del total, seguido por la pasta de madera con el 21%; madera y sus manufacturas, con el 20%; productos gráficos, con el 7%; muebles, con el 1%, y otros.

En este contexto, brevemente el plan ForestAr 2030 prevee (https://www.argentina.gob.ar/forestar2030) para el 2030:

-Incrementar la superficie forestal plantada a 2 millones de hectáreas (+ 50%), industrializar distintas regiones del país y realizar inversiones en infraestructura crítica (trenes, puertos, energía, comunicación, entre otros) para escalar otras actividades en las provincias y crear 187 mil empleos de calidad.

-Revertir el déficit histórico en la balanza comercial, exportando 2.500 millones USD.

-Poner en valor a los bosques nativos ampliando la gestión sostenible y el reconocimiento de los servicios ecosistémicos que proveen a las comunidades y a toda la sociedad.

-Incrementar el agregado de valor en la cadena foresto industrial apoyándose en políticas permanentes de investigación, desarrollo e innovación (I+D+i).

-Aportar a la adaptación y mitigación del cambio climático. Contribuir significativamente al cumplimiento de la meta absoluta a través de la reducción de las emisiones y el aumento de las capturas de gases de efecto invernadero debido a la gestión sostenible de los bosques nativos, las plantaciones forestales y toda la cadena de valor asociada.

-Asegurar la sostenibilidad de los proyectos y prácticas asociados al presente Plan que resguardan sitios de

alto valor de conservación, biodiversidad y patrimonio cultural.

4. MEJORAMIENTO GENÉTICO FORESTAL: PRINCIPALES CARACTERES DE INTERÉS

La finalidad de un programa de mejora forestal es desarrollar individuos o poblaciones genéticamente superiores, en otras palabras, mejorar las características cuantitativas y cualitativas de rendimiento y calidad en comparación a las utilizadas comercialmente (Marcó, 2005). Las características forestales principales por mejorar son: crecimiento (Diámetro, Altura y Volumen), forma del fuste (Calidad del Fuste y de los Rollizos), y calidad de la madera (que puede estar dirigida a la producción de fibras, fundamentalmente papel, o bien para usos sólidos). El mejoramiento genético en los programas de INTA, se ha focalizado últimamente a considerar a la calidad de madera como primordial, además de la producción (Figura 1.10).

Para que un programa de mejoramiento sea exitoso se lo debe planear a corto y largo plazo. A corto plazo lo que se tiene en cuenta es la ganancia en la productividad, y a largo plazo, se debe considerar la diversidad genética de las poblaciones para que sean capaces de afrontar los cambios venideros (Savolainen & Kärkkäinen, 1992). Además, es necesario conocer cómo los caracteres a mejorar se relacionan entre sí y con otros caracteres que no estén siendo seleccionados en ese momento, tal que cuando se mejora una característica el efecto potencial sobre la otra pueda ser predicho (Poke et al., 2005). Las correlaciones fenotípicas indican la presencia de relaciones entre caracteres que podrían deberse a una respuesta similar a condiciones ambientales o a asociaciones genéticas. Las correlaciones genéticas son importantes para determinar la potencial selección simultánea o independiente de caracteres (Poke et al., 2005).

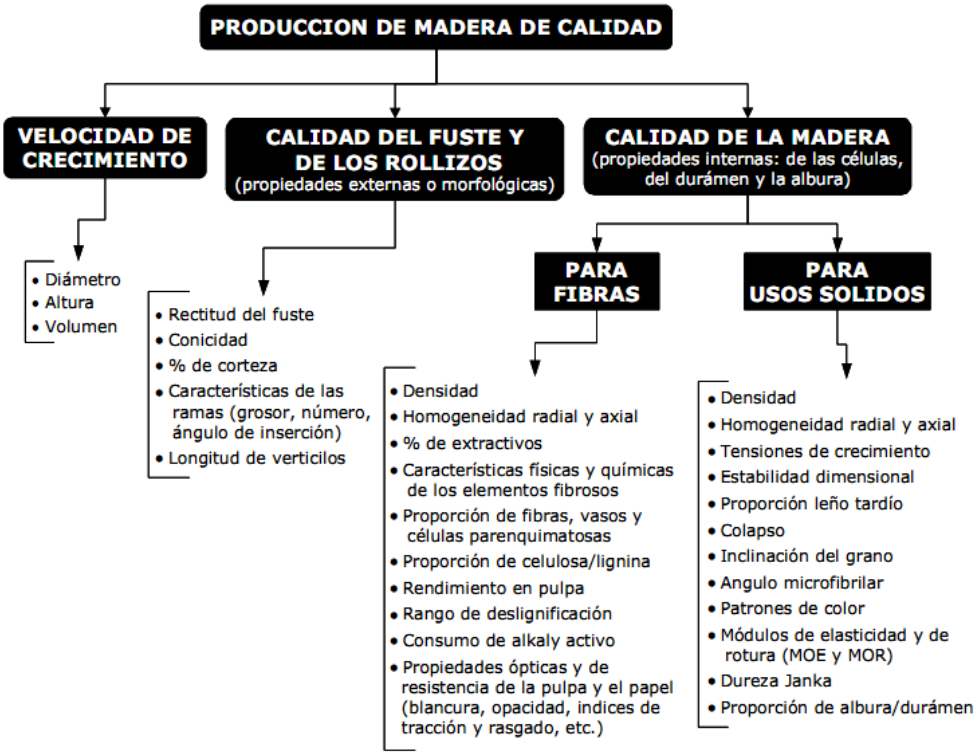


Figura 1.10: *Principales características de interés en los programas de mejora forestal para calidad de madera.* Fuente: López, 2005.

5. COMPOSICIÓN DE LA MADERA Y MEDICIONES DE SU CALIDAD

La madera está compuesta por una mezcla de polímeros: celulosa (40%), lignina (30%), hemicelulosa (xilano) (20%) y extractivos (5%) (Walker, 2006). Las tres primeras son sustancias poliméricas que conforman la pared celular (Plomion et al., 2001; Turner & Somerville, 1997) y los extractivos son pequeñas moléculas no estructurales involucradas en la defensa y metabolismo de las células vivas del árbol (Núñez, 2004). La medición de la proporción de estos distintos compuestos es fundamental para analizar la calidad de la madera de acuerdo con su uso posterior (para fibras o usos sólidos, Figura 1.8). El análisis de la madera se realiza comúnmente a través de distintos ensayos químicos. Sin embargo, la cuantificación de los componentes químicos de la madera es notoriamente difícil y, en muchos casos, costosa. Es por ello que se han desarrollado métodos quimiométricos rápidos y más baratos a partir del uso de espectros de NIR (Reflectancia en el Infrarrojo cercano, del inglés *Near-InfraRed spectroscopy*) para la estimación de la composición química de la madera (Rodrigues et al., 1998; Da Silva Perez et al., 2007; Schimleck et al., 2000). De esta manera, utilizando esta metodología, la lignina puede cuantificarse considerando la fracción soluble e insoluble en ácido (lignina Klason; Schwanninger & Hinterstoisser, 2002). Por otro lado, además del contenido total, puede estimarse la composición de los monómeros que la forman, que en el caso de eucalipto está integrada por los monolignoles: siringilo (S) y guayacilo (G) y, particularmente la relación de estos monómeros (S:G) es de

importancia en la industria papelera porque, según ésta, se facilitan la delignificación química necesaria para la producción del papel (Clarke, 2009; Rencoret et al., 2007). También puede estimarse la cantidad de extractivos (totales, etanólicos y acuosos), que como indica su nombre, son aquellos compuestos que se extraen con distintos solventes de la madera, importantes para la industria papelera como para la defensa del árbol.

6. EL GÉNERO *Eucalyptus*

El género *Eucalyptus* comprende más de 700 especies de árboles nativos de Australia e islas al norte (Ladiges et al., 2003), algunas de las cuales se encuentran entre los árboles más plantados (Doughty, 2000; Figura 1.3). Muestran propiedades excelentes como el rápido crecimiento, la calidad de la madera, un nivel excepcional de variabilidad y una adecuada capacidad de propagación vegetativa, además de que sus plantaciones ayudan a mitigar las presiones humanas sobre los bosques nativos (Rezende et al., 2014). Las especies 'Big Nine' que se usan en la silvicultura de plantación son *E. grandis*, *E. saligna*, *E. pellita*, *E. urophylla*, *E. globulus*, *E. dunnii*, *E. nitens*, *E. tereticornis* y *E. camaldulensis* (Grattapaglia & Kirst, 2008). *E. grandis* subtropical es probablemente la especie más ampliamente plantada del género para la producción industrial de madera y el primer eucalipto con genoma secuenciado (Myburg et al., 2014).

Figura 1. 11. *Eucalyptus dunnii*

En Argentina, el programa de mejoramiento de *Eucalyptus* se lleva a cabo principalmente por el Instituto Nacional de Tecnología Agropecuaria (INTA). En los últimos años, como consecuencia del cambio climático, se han registrado heladas excepcionales, tanto tempranas como tardías, de mayor duración, intensidad y extensión territorial (De la Peña et al., 2012). Este panorama afecta al desarrollo temprano de plantaciones de *E. grandis*. Por otra parte, para la industria celulósica, *E. globulus* es la especie más requerida, pero su área de plantación se restringe al sudeste bonaerense, alejada de las industrias consumidoras, tanto de Argentina como de Uruguay.

Eucalyptus dunnii Maiden (Figura 1.11, goma blanca de Dunn), es una especie diploide (2n =22), que ocurre naturalmente en el noreste de Nueva Gales del Sur, Australia, principalmente en los márgenes de las selvas tropicales (Figura 1.12). Esta especie ha aumentado su aceptación como una alternativa a las plantaciones de *E. grandis*, creciendo mejor en sitios más secos y más susceptibles a las heladas (Darrow, 1994; Thomas et al., 2009), aunque aún sigue siendo un ejemplo de especie menos estudiada o no modelo del género.

El programa de mejoramiento genético forestal de INTA propone como especie alternativa a *E. dunnii* introduciéndola a fines de los años 70 en el país (Marcó & White, 2002), ya que muestra mayor tolerancia a

bajas temperaturas que *E. grandis*, y buen crecimiento en el norte de Buenos Aires, sur de Santa Fe y Entre Ríos, cerca de las industrias consumidoras. Sin embargo, aunque esta especie presenta mayor densidad de madera que *E. grandis*, tiene menor aptitud para usos de madera sólida, debido a su alto nivel de tensiones de crecimiento y menor estabilidad dimensional, evidenciadas por una mayor proporción de rajaduras después del apeo. A pesar de esto último, es posible su mejoramiento ya que estas propiedades se encuentran bajo un fuerte control genético, siendo su heredabilidad de 0,48 ±0.21 (López et al., 2012). Además, la selección de individuos de *E. dunnii* con características deseables para la industria celulósico-papelera permitiría una mejora sustantiva de la calidad industrial.

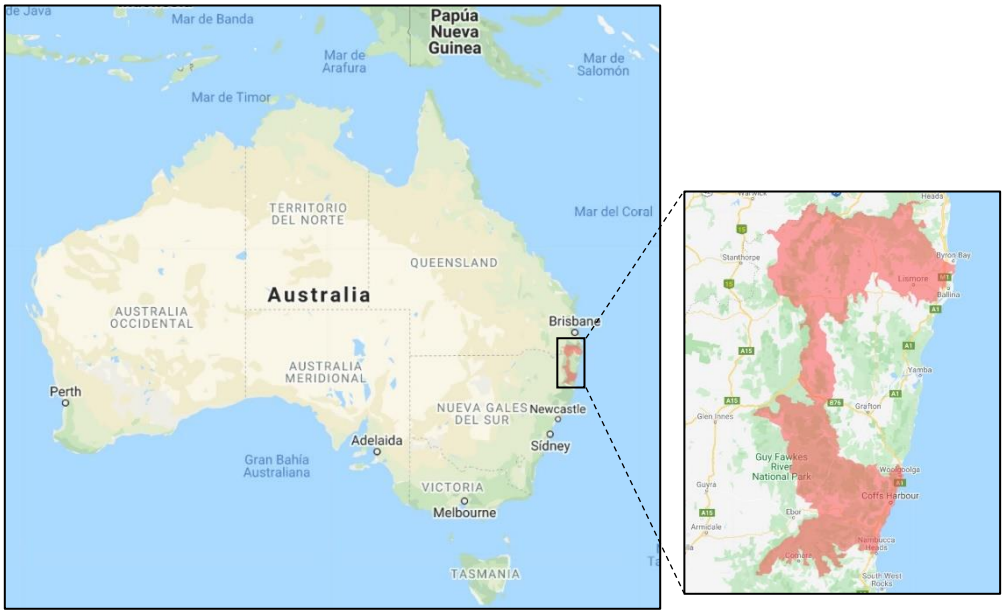


Figura 1.12. *Distribución Natural de E. dunnii en Australia. Referencia:* puntos rojos: poblaciones de *E. dunnii*. Fuente: https://www.environment.nsw.gov.au/threatenedspeciesapp/profile.aspx?id=20100.

Con el fin de comenzar el programa de mejoramiento genético de la especie, entre 1991-1992, el INTA instaló una serie de ensayos de orígenes/procedencias/progenies en 6 sitios de la Mesopotamia (Marcó & White, 2002). A una edad promedio de 5 años se realizaron las estimaciones de ganancia genética utilizando un índice combinado de selección a través del cual fue posible capturar el 87% y 70% de la máxima ganancia posible para volumen y forma, respectivamente (Marcó & White, 2002). Haciendo uso de dicha información, a los 11 años de edad, 2 ensayos ubicados en la Provincia de Entre Ríos fueron raleados y para ser transformados en Huertos Semilleros de Progenies (HSP, López et al., 2012). Al mismo tiempo, los genotipos de mayor ganancia genética fueron movilizados vía injerto e instalados en un Huerto Semillero Clonal (HSC) en el Instituto de Recursos Biológicos del INTA Castelar, donde las condiciones climáticas son muy favorables para la producción de semilla (34º 37´ S, 58º 40´ O).

7. LAS HERRAMIENTAS Y RECURSOS DISPONIBLES PARA EL MEJORAMIENTO MOLECULAR DE *Eucalyptus*

El mejoramiento asistido por marcadores moleculares es empleado en el ciclo de mejora de especies forestales y agronómicas, porque permite la selección temprana de los mejores individuos para caracteres fenotípicos aún no expresados. Su aplicación brinda un impacto positivo directo porque ofrece un aumento en la ganancia genética en un menor período de tiempo, lo que es extremadamente relevante debido al largo período generacional y al ciclo de las especies forestales (Neale & Kremer, 2011; Varshney et al., 2017).

El muestreo eficiente del genoma de la planta con marcadores genéticos suficientes e informativos, juega un papel clave en el mejoramiento, conservación y estudios evolutivos. En las últimas décadas, los investigadores han desarrollado diferentes tipos de marcadores moleculares útiles para estos fines. Hoy en día, los SNPs (*Single Nucleotide Polymorphism* o Polimorfismo de Nucleótido Simple) se han convertido en los marcadores de elección, debido a su alta abundancia en los genomas, estabilidad, codominancia y automatización del proceso de genotipado (Torkamaneh et al., 2018).

Los SNPs son actualmente los más utilizados en el mejoramiento, en los enfoques de mapeo de genes y QTL, principalmente debido al menor costo por dato puntual y a la relativa facilidad en el diseño del ensayo, así como en la interpretación de los resultados (Bajgain et al., 2016).

Como su nombre lo indica, los SNPs son variaciones en una secuencia que involucran la sustitución de un nucleótido cuando se comparan dos alelos (ya secuenciados) de cromosomas homólogos, generalmente provocados por errores durante la división celular (Thavamanikumar et al., 2011). Los SNPs son abundantes y están presentes uniformemente a lo largo de todo el genoma de un individuo, sin embargo, no todas las variaciones de una base en una secuencia son consideradas SNPs, para esto, la modificación de la base debe estar presente en al menos un 1% de la población, es decir, con una frecuencia alélica mínima o MAF (*Minimum Allele Frequency*) de 0,01 (Brookes, 1999). La restricción en cuanto a la frecuencia es lo que distingue un SNP de una mutación puntual. Basándose en la definición de Brookes, los marcadores SNPs no incluyen las inserciones/deleciones (InDel) (Khlestkina & Salina, 2006). La gran mayoría de los SNPs son bialélicos, es decir, tienen dos alelos los cuales están representados por una sustitución de una base por otra, incluyendo las transiciones purina-purina (A-G) o pirimidina-pirimidina (C-T) y las transversiones purina-pirimidina o pirimidina-purina (A-C, A-T, G-C o G-T). Por este motivo, se necesita una alta densidad de SNPs para obtener un nivel de información adecuada, como la obtenida por SSR (*Simple Sequence Repeat* o repetición de secuencia simple polimórfica o microsatélites), ya que estos últimos son multialélicos. Sin embargo, la gran ventaja de los SNPs por sobre los demás tipos de marcadores, es que tienen la capacidad de ser identificados de forma automatizada, de manera que pueden encontrarse millones de SNPs a la vez (Torkamaneh et al., 2018).

Como se mencionó anteriormente, los SNPs se distribuyen a lo largo de todo el genoma, incluyendo regiones codificantes, regiones intrónicas, regiones reguladoras y regiones intergénicas. Los SNPs se pueden clasificar en "no sinónimos", cuando se encuentran en regiones codificantes y alteran la secuencia aminoacídica de tal forma que generan un cambio en la función de la proteína o en la expresión de esta (Ramensky, 2002). De esta forma, existen variaciones funcionales que son capaces, por ejemplo, de generar la susceptibilidad a alguna patología, pudiendo estar localizados en la región promotora del gen, alterando su actividad transcripcional (modulando la unión de factores de transcripción), en intrones (modulando la estabilidad de la proteína) o en sitios de *splicing* o en regiones intragénicas. Otro tipo de SNPs son los llamados "sinónimos" o silenciosos, los cuales no alteran al aminoácido (Khlestkina & Salina, 2006). Sin embargo, los codones sinónimos no están presentes en frecuencias iguales en genes/genomas y varían entre organismos, lo que se conoce como sesgo de uso de codones o *codon usage bias*. Estas frecuencias están relacionadas con la abundancia de ARNt afines por lo cual, en un organismo dado, los codones más utilizados se traducirán más rápidamente y otorgarán una producción más eficiente de proteínas en las células (Komar, 2016).

En eucaliptos, se han descripto SNPs en genes candidatos, principalmente en *E. camaldulensis, E. globulus, E. nitens, E. loxophleba*, *E. urophylla* con un rango de 1/16 pb a 1/108 pb dependiendo del gen y posiciones en intrones y exones (Hendre et al., 2011; Külheim et al., 2009; Mandrou et al., 2012).

Para el género *Eucalyptus* se publicó un conjunto de 768 SNPs con una metodología de microarreglo GGGT (*Golden Gate genotyping technology*, Illumina; Grattapaglia et al., 2011), un microarreglo o chip de DArT (*Diversity Array Technology*; Sansaloni et al., 2010) con 7680 sondas y un chip de 60 mil SNPs (EUChip60K, Silva-Junior et al., 2015). Si bien el conjunto de SNPs de GGGT mostró elevado porcentaje de polimorfismo en *E. grandis*, esta tasa disminuyó significativamente para *E. globulus*, se estimó en sólo un 10% para las otras especies y además esta metodología de genotipificación ya fue descontinuada (Grattapaglia et al., 2012). Por otro lado, el microarreglo DArT, aunque brinda una gran cantidad de datos genotípicos, tiene la desventaja de estar compuesto por marcadores dominantes y de ser una tecnología prestada exclusivamente por una empresa australiana (Sansaloni et al., 2010).

En el año 2015 se desarrolló el sistema comercial EUChip60K (Silva-Junior et al, 2015), que es una herramienta que permite genotipificar miles de SNPs de acceso público dentro del género *Eucalyptus*, con gran velocidad de procesamiento y bajo costo. Los SNPs incluidos en este chip fueron descubiertos a partir de la resecuenciación de nueva generación o masiva (*Next Generation Sequencing*, NGS) de 240 individuos del género, y su comparación con el genoma de referencia de *E. grandis* (Myburg et al, 2014), disponible en Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html). La forma en la que fue desarrollado maximiza la cobertura del genoma y satisface los requisitos esenciales de alta precisión y reproducibilidad de las genotipificaciones, ya que se obtiene el mismo conjunto de SNP entre estudios independientes en la misma

especie (Silva-Junior et al, 2015). En el caso de *E. dunnii*, esta especie está representada sólo por un 5% de los individuos utilizados en el desarrollo de dicho chip, con 12 individuos de una única población natural (Población Natural Acacia Creek, NSW, Australia), reportando en la validación 17.014 SNPs con MAF >0,01 (Silva-Junior et al, 2015). Sin embargo, el mismo proceso de descubrimiento de los SNPs introduce un sesgo de los datos, porque sólo los polimorfismos de un pequeño número de individuos de determinadas poblaciones de cada especie se encuentran representados en dicha plataforma. Esto puede afectar a las frecuencias alélicas poblacionales, sesgándolas hacia alelos comunes (Albrechtsen et al., 2010; B. Li & Kimmel, 2013a), y a las relaciones genéticas entre individuos, no pudiendo ser corregidas fácilmente (Bajgain et al., 2016).

Como se mencionó anteriormente, *Eucalyptus dunnii* está poco representada en el chip y no existen hasta el momento desarrollos de marcadores o estudios del genoma para esta especie en particular. Desarrollar estrategias para la búsqueda de marcadores en esta especie que es importante para Argentina y para otros países con clima subtropical es de particular interés.

1.7.1 El Genoma de Eucalyptus

El tamaño de genoma promedio para el subgénero *Symphyomyrtus* es de 650 millones de pares de bases (Mpb) estimado por citometría de flujo (Grattapaglia & Bradshaw Jnr, 1994). En ese sentido, el genoma de eucalipto posee un tamaño superior si se lo compara con el genoma de la planta modelo *Arabidopsis thaliana* de 135 Mpb en cinco cromosomas (*Arabidopsis Genome Initiative*, 2000) o con el genoma de *Prunus persica* de 265 Mpb en ocho cromosomas (*The International Peach Genome Initiative* et al., 2013). Por el contrario, si se lo compara con el tamaño del genoma de las coníferas que van de 10 a 30 Gb (por ejemplo *Norway spruce* tiene un tamaño de 20 Gb, N=12; Nystedt et al., 2013) el genoma de eucalipto es mucho más pequeño. Hace algunos años, un grupo de la comunidad científica se propuso encarar la secuenciación del genoma completo de *E. grandis* mediante la estrategia de secuenciación por *shot-gun.* Este reto llevó a que en enero del 2011 se liberara, para el conocimiento público, un borrador del genoma con una profundidad de 8X (Myburg et al., 2014; *Eucalyptus Genome Database* http://phytozome.jgi.doe.gov/) mediante la utilización de tres librerías distintas y utilizando secuenciadores Sanger (ABI 3730XL). Se empleó como material "BRASUZ1" que es un individuo de 17 años derivado de una autofecundación.

A la fecha (Myburg et al., 2014; *Eucalyptus Genome Database* http://phytozome.jgi.doe.gov/), se obtuvo un 94% del genoma secuenciado organizado en 11 cromosomas (605 Mpb) y 4941 andamios o *scaffolds* pequeños (alrededor de 85 Mpb) que aún no pudieron ser anclados dado que corresponden a regiones altamente repetitivas; sin embargo, el 98,98% de las regiones codificantes se encuentran en los 11 cromosomas ensamblados y un muy bajo porcentaje en los *scaffolds* pequeños.

Otro de los genomas del género *Eucalyptus* secuenciado es el de *E. camaldulensis*, para lo cual se usó una combinación del método convencional de Sanger y los métodos de NGS, seguido de un análisis bioinformático (Hirakawa et al., 2011). La longitud total de las secuencias genómicas no redundantes así obtenidas fue de 655 Mb, pero las mismas no se encuentran ensambladas en cromosomas discretos, si no que consta de 81.246 *scaffolds* y 121.194 *contigs* de menor longitud. Hasta la fecha, no se dispone de la secuencia del genoma de *E. dunnii* siendo el genoma de referencia más pulido y avanzado del género el de *E. grandis.*

1.7.2 Nuevas metodologías genómicas basadas en secuenciación de nueva generación

Una alternativa a la detección de SNPs mediante sistemas fijos (*arrays* o microarreglos, también llamados chips) y que evita el sesgo introducido por este tipo de sistemas (Poland et al., 2012) son los métodos de reducción de la complejidad del genoma (con enzimas de restricción) combinados con la secuenciación NGS. Estas herramientas permiten realizar diferentes estudios a nivel poblacional sin la necesidad de contar con información genómica previa de la especie, pudiendo ser aplicada en organismos no modelo, con un costo menor que empleando microarreglos ya que no necesita del desarrollo de un chip (Andrews et al., 2016; Davey et al., 2011). Las metodologías de *Restriction site-associated DNA sequencing* (RADseq, o secuenciación de ADN asociada a los sitios de restricción, Baird et al., 2008), *Genotyping by sequencing* (GBS, o genotipado por secuenciación, Elshire et al., 2011) y sus protocolos derivados, han surgido recientemente como enfoques genómicos prometedores para el descubrimiento SNPs a gran escala y a lo largo de todo el genoma. Se basan en obtener secuencias NGS de una porción del genoma (representación reducida) de varias muestras al mismo tiempo (multiplexadas), no requieren de un genoma de referencia o conocimientos previos de polimorfismos, y combinan el descubrimiento de marcadores y el genotipado en un solo proceso o protocolo. Por lo tanto, proporcionan una estrategia rápida, de alto rendimiento y rentable para llevar a cabo un análisis de todo el genoma, y que puede ser aplicada a especies no modelo y conjuntos de germoplasma únicos, para obtener información de variantes o polimorfismos exclusivos (Andrews et al., 2016; Davey et al., 2011).

Estos enfoques implican la digestión del ADN con enzimas de restricción y luego la secuenciación de un subconjunto de los fragmentos generados, seleccionándolo por un rango de tamaño específico. Con el objetivo de mejorar algunas de las debilidades del protocolo de RADseq original, como por ejemplo para evitar un paso de corte de fragmentos al azar por sonicación (Revisado en Andrews et al., 2016; Timm et al., 2018), los investigadores han desarrollado muchas metodologías derivadas, incluyendo 2b-RADseq (Wang et al., 2012), ezRADseq (Toonen et al., 2013) y ddRADseq (Peterson et al., 2012). El protocolo de *Double digest Restriction site-associated DNA sequencing*, o ddRADseq, utiliza dos enzimas de restricción diferentes para cortar el ADN: una de corte raro (es decir, una enzima con un sitio de reconocimiento de más de 4 pares de bases) y una de corte frecuente. A su vez, sólo se secuencian los fragmentos que caen entre ambos sitios de restricción y dentro de un rango de tamaño específico (Timm et al., 2018). Esto reduce la cantidad de secuencias necesarias para

alcanzar una cobertura o profundidad óptima de cada *locus*, así como el porcentaje de datos faltantes, en comparación con RADseq.

El protocolo original ddRADseq fue desarrollado en base a datos de animales, y se ha aplicado ampliamente en el descubrimiento de marcadores SNP y genotipado para varias especies de dicho reino. Asimismo, ha sido aplicado en distintas plantas (Nazareno et al., 2017; Pyne et al., 2017; Roy et al., 2017; Vargas et al., 2017; Zhou et al., 2014), como en forestales y frutales (revisado en Parchman et al., 2018), y se desarrollaron algunos protocolos para plantas en general (Peterson et al., 2014; Yang et al., 2016). Sin embargo, debido a la diversidad y complejidad de los genomas de las plantas, los diferentes pasos del protocolo ddRADseq requieren de un acondicionamiento u optimización para lograr mejores resultados específicos. Dichos pasos incluyen la selección del par de enzimas de restricción, la determinación del rango de tamaños de fragmentos óptimo, la idoneidad y el rendimiento de la plataforma de secuenciación, la profundidad de secuenciación y la estrategia de búsqueda de marcadores o SNPs. Además, debido a que esto implica pasos de prueba, el desarrollo de un protocolo optimizado para establecer la metodología en un pequeño número de muestras de plantas es ineludible, principalmente para laboratorios con presupuestos limitados.

No obstante, un posible problema del uso de GBS es la generación de datos perdidos cuando se analiza conjuntamente un gran número de individuos durante la secuenciación, debido al polimorfismo en los sitios de corte de las enzimas y a la variación en el número de secuencias entre individuos y *loci*. Esto puede causar un sesgo en las estadísticas genéticas de la población, diferente al obtenido con los chips de SNPs. Sin embargo, para resolver este inconveniente, es posible sólo utilizar los *loci* con un porcentaje mínimo de individuos con datos faltantes o bien, si esto deteriora demasiado el número de datos, existen dos maneras de resolver este problema: (i) aumentar la profundidad de secuenciación y/o (ii) imputar los datos perdidos utilizando, por ejemplo, la información de relación genética entre genotipos de la población (Bajgain et al 2015; Andrews et al., 2016). El aumento de la profundidad se logra aumentando el número de secuencias NGS obtenidas por cada muestra a analizar, pero este aumento viene de la mano de un aumento en los costos de la obtención de los datos, debido al uso de más reactivos de secuenciación. Por otro lado, la imputación, que es la sustitución de datos faltantes por algún valor, en otras palabras, "completar" datos faltantes con valores plausibles a través de diversas estrategias, debe realizarse de todos modos, ya que muchas herramientas utilizadas en el análisis genómico requieren conjuntos de datos completos (Figura 1.13). Además, la imputación bien realizada mejora el poder de los análisis posteriores (Money et al., 2015; Torkamaneh et al., 2018). Existen varios algoritmos de imputación, y los más comunes se basan en la utilización de paneles de referencia (información de genoma y genotipos de referencia) que ayudan a la precisión de la imputación (Torkamaneh et al., 2018). Sin embargo, los organismos que no son modelo no poseen paneles de genotipos de referencia disponibles. El método de imputación del genotipo vecino más cercano, LD-kNNi (*Linkage Disequilibrium-k-nearest neighbor genotype*), es un algoritmo que no requiere mapas físicos o genéticos, y está diseñado para trabajar en datos de

genotipo no ordenados (sin fase) de especies con una gran proporción de heterocigosis, como las forestales. Explota el hecho de que los marcadores útiles para la imputación a menudo no están físicamente cerca del genotipo faltante, sino que se distribuyen por todo el genoma. LinkImpute es un programa basado en LD-kNNi, y mostró mayor velocidad, precisión comparable, y menor sesgo en las estimaciones de frecuencia de alelos al ser comparado con varios métodos de imputación de genotipos, entre ellos los más utilizados (Money et al., 2015).

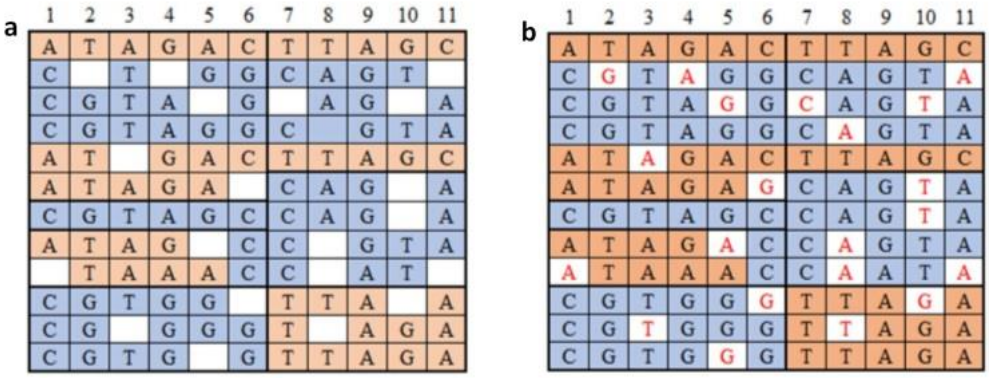


Figura 1.13: *Esquema de imputación de datos*. Datos de GBS (a) antes y (b) después de la imputación. Matrices con individuos en filas y SNPs en columnas. Imputación basada en fases. Los marcadores ubicados dentro del mismo bloque de DL están sombreados en el mismo color. Fuente: Torkamaneh et al., 2018.

Una aplicación colateral de esta técnica NGS en plantas es el descubrimiento rápido y rentable de *loci* de SSR (Barchi et al., 2011; Torales et al., 2018). Los SSR tienen numerosos usos, incluido el desarrollo de mapas de ligamiento, el mapeo de *loci* de rasgos cuantitativos (QTL), la selección asistida por marcadores, la identificación de cultivares o clones, los estudios de estructura y diversidad genética de las poblaciones, además de otros, que aún desempeñan un papel importante en esta "era genómica" (Hodel et al., 2016).

En lo que respecta a especies forestales, un reciente aumento en el número de estudios que utilizan GBS sugiere que es probable que se convierta en una de las estrategias más comunes para generar datos genómicos para una variedad de aplicaciones (Parchman et al., 2018). Parchman et al. (2012), Chen, et al., (2013) y Pan et al., (2015), aplicaron este tipo de bibliotecas exitosamente en coníferas (*Pinus y Picea*), siendo especies no modelo, sin genoma de referencia y con gran cantidad de ADN repetitivo.

En cuanto a la aplicación de GBS en *Eucalyptus*, tanto Grattapaglia et al. (2011) como Faria et al. (2012) evaluaron en *E. grandis* (18 y 24 individuos) *y E. globulus* (18 y 24 individuos) la viabilidad y el rendimiento de este tipo de métodos (RADseq y GBS, con una única enzima de restricción PstI y ApeKI, respectivamente), y utilizando el genoma de referencia de *E. grandis* permitió obtener una gran cantidad de SNPs (200 mil totales y 42.300 entre ambas especies; 10.861 en *E. grandis* y 2.134 en *E globulus*; respectivamente).

En el presente trabajo se describe el desarrollo de un novedoso protocolo ddRADseq optimizado y de menor costo en *E. dunnii*, utilizando un pequeño número de muestras y una estrategia de escalada a un número elevado de muestras y versátil para ser aplicado fácilmente a cualquier especie vegetal.

8. APLICACIÓN DE LAS HERRAMIENTAS GENÓMICAS PARA EL MEJORAMIENTO MOLECULAR DE *EUCALYPTUS*

1.8.1 Detección de QTL

Desde siempre se reconoce el gran interés en encontrar conexiones entre el genotipo y el fenotipo (Weir, 2008), siendo esto el objetivo fundamental de la genética (Botstein & Risch, 2003). Comprender la base molecular que explica la variación fenotípica es la meta principal de los estudios genéticos, pero también es importante entender cómo estos conocimientos pueden ser aplicados en forma directa a los programas de mejora. Por este motivo, en los últimos años, ha habido una gran cantidad de trabajos científicos en donde se realizó la búsqueda de QTL (del inglés *Quantitative Trait Loci*) para identificar genes responsables de la variación cuantitativa de distintas características complejas de importancia económica y/o evolutiva. Desde trabajos en humanos (Corder et al., 1994; Kerem et al., 1989) hasta en plantas (Thornsberry et al., 2001; Zhu et al., 2008) y aún en eucaliptos (Freeman et al, 2009; Gion et al., 2000; Kullan et al., 2012; Poke et al., 2005; Thumma et al., 2010) entre otros trabajos.

El objetivo de la detección de QTL es encontrar marcadores moleculares que estén ligados a los genes que gobiernan los caracteres de interés, dado por su localización (dentro o cerca de los marcadores).

Hay dos aproximaciones distintas que se pueden utilizar para identificar las variantes genotípicas (*loci*, alelos) ligadas a las diferencias fenotípicas de un carácter: Análisis de Ligamiento (detecta posición y orden de los marcadores y QTL) y Mapeo por Asociación (MA) o Asociación genética (Mackay & Powell, 2007; Mackay, 2001).

1.8.2 Mapeo de ligamiento genético versus Mapeo por Asociación Genética

El análisis de detección de QTL mediante ligamiento genético corresponde a la aproximación clásica en donde se utilizan familias o poblaciones biparentales que fueron creadas a partir del cruzamiento de dos parentales con fenotipos contrastantes para el carácter en estudio y que garantizan la suficiente segregación genética del mismo. Esta estrategia permite organizar los Grupos de Ligamiento (GL) y posicionar los marcadores y caracteres dentro de ellos. Por otro lado, en el MA se usan poblaciones naturales y/o una colección de individuos, muchas veces de ascendencia desconocida, para detectar la variación genotípica responsable de las diferencias del carácter estudiado. Una diferencia importante entre las dos estrategias es la

tasa de recombinación en cada una de las poblaciones, que da como resultado resoluciones de mapeo distintas. En ese sentido, en los análisis de ligamiento tradicionales, son pocos los eventos de recombinación que llegan a producirse resultando en un mapeo de baja resolución a diferencia del MA en donde la recombinación y la diversidad genética natural se analizan para lograr un mapeo de alta resolución (Zhu et al., 2008) (Figura 1.14).

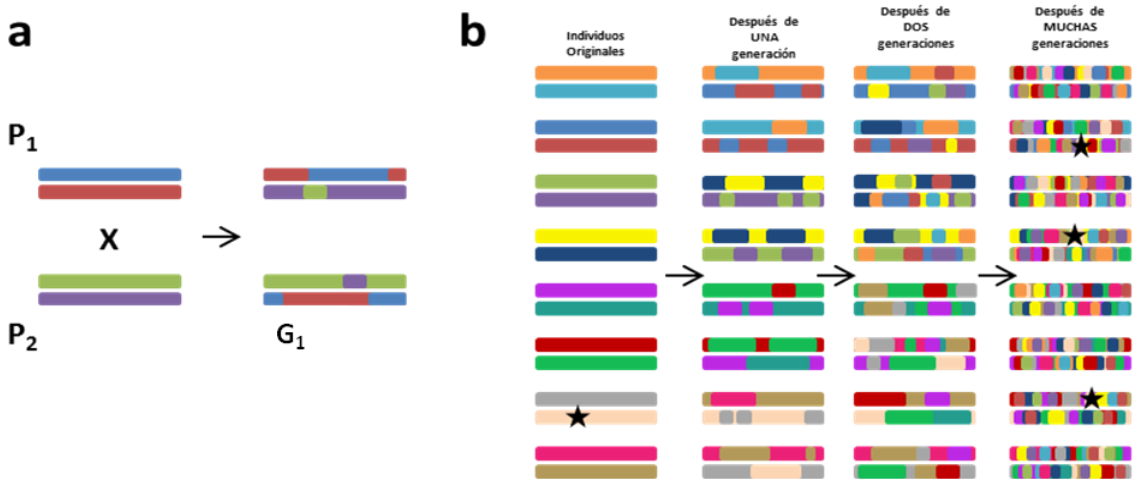


Figura 1.14. *Mapeo por ligamiento y mapeo por asociación.* **a.** Población de mapeo biparental utilizada en el análisis de mapeo por ligamiento. P1 y P2 son los parentales iniciales que poseen fenotipo contrastante para el carácter de estudio y los que generan la población de mapeo (G1). En esta estrategia hay sólo algunas oportunidades de recombinación (segmentos rojos y azules y, verdes y violetas de la G1) dando así un mapeo de baja resolución. **b.** Sobre una población natural (individuos originales) se quiere detectar una variante alélica (estrella negra) responsable del fenotipo de interés. Mediante el mapeo por asociación se logra un mapeo de alta resolución gracias a los múltiples eventos de recombinación (representados por los segmentos de diferentes colores) dado a lo largo de muchas generaciones y a la gran cobertura de marcadores que se utilice. Fuente: Modificado de Zhu et at. (2008).

Cabe aclarar que a pesar de que los estudios de QTL en poblaciones biparentales son exitosos para el descubrimiento de QTL, sin embargo, cuentan con la limitante importante de su utilidad solamente para el mismo pedigrí de la población de análisis y requieren de una validación cuando se involucran distintos fondos genéticos. Además, sólo son útiles los alelos que difieren entre los padres y por lo tanto segregan. Dado que la resolución del mapeo se basa únicamente en la cantidad de eventos de recombinación que ocurren durante el desarrollo de la población (Mitchell-Olds, 2010), es necesario contar con numerosas progenies para analizar. Para evitar estas limitaciones, se puede aumentar el número de entrecruzamientos, utilizando poblaciones multiparentales (Alqudah et al., 2019), siendo las RILs en los cultivos agrícolas, las poblaciones de preferencia. Sin embargo, estos desarrollos son prácticamente imposibles de lograr en especies forestales (por el tiempo y por la gran heterocigosis que poseen). Frente a estas cuestiones, el MA posee varias ventajas sobre el mapeo por ligamiento tradicional como son la rapidez en establecer asociaciones porque no se crean poblaciones de

mapeo, la diversidad genética es mayor, ofrecen mayor resolución debido al número de recombinaciones ancestrales y por lo tanto poseen un mayor número de alelos diversos (Buckler & Thornsberry, 2002; Flint-Garcia et al., 2003; Yu et al., 2006).

1.8.3 Mapeo por Asociación Genética

El fundamento del MA es la búsqueda de asociaciones significativas entre un marcador y un rasgo fenotípico, asumiendo que el marcador está en desequilibrio de ligamiento (DL) con un *locus* causal que afecta al rasgo (Tan, 2018). El MA puede realizarse ya sea en base a genes candidatos o por el análisis global del genoma (*Genome wide association study*, GWAS). La estrategia de GWAS tiene el potencial de identificar una arquitectura genética más completa y de permitir la asignación de función en los casos donde no se dispone de Genes Candidatos determinados, especialmente para caracteres cuantitativos (Zhu et al., 2008). Entre los recursos estadísticos para MA, el modelo mixto propuesto en Yu et al. (2006) para múltiples niveles de relación, es el más utilizado, aunque también se pueden tratar los efectos de marcadores como aleatorios (Goddard et al., 2009).

Una desventaja potencialmente grave del mapeo de asociación es que así como las asociaciones entre marcador y rasgos fenotípico se pueden deber a una vinculación con polimorfismos causales, también pueden surgir de la presencia de estructura de poblacional (Zhao et al., 2007). Por ejemplo, los estudios de asociación genética en líneas cultivadas endogámicas se enfrentan al problema de las tasas infladas de falsos positivos debido a la existencia de estructura de la población y a la relación genética entre los organismos, causadas por la historia genealógica a menudo compleja en la mayoría de ellos. La aplicación de pruebas estadísticas convencionales de independencia entre un marcador genético y un fenotipo es propensa a asociaciones espurias porque es probable que el marcador y el fenotipo estén correlacionados a través de la estructura de la población, lo que viola el supuesto de independencia bajo la hipótesis nula (Kang et al., 2008).

El MA es una estrategia novedosa para la detección de *loci* que contribuyen a la variación de un carácter. Esta metodología ha sido favorecida en los últimos años por el avance de las tecnologías de genotipificación de alta procesividad (del inglés *high throughput genotyping*) mencionadas anteriormente, por el desarrollo de mejores métodos estadísticos (Zhu et al., 2008), y por la disponibilidad informática y de programas estadísticos libres para el procesamiento de un gran volumen de datos.

Sin embargo, el poder de los estudios de asociación está basado en el nivel de desequilibrio de ligamiento (DL) de la población que se evalúa. Si se trabaja con cultivos agrícolas autógamos (por ej. Girasol, trigo) que tienen un alto grado de mejora, los cromosomas poseen un DL alto, en otras palabras, tienen fragmentos grandes que no recombinan entre sí. Por eso, tienen la ventaja que, aunque el marcador se encuentre alejado

del gen es posible detectarlo como marcador ligado al carácter porque el DL decae lentamente con la distancia al gen responsable del fenotipo. Por el contrario, por la historia evolutiva y del mejoramiento reciente de los árboles, los fragmentos que no recombinan son cortos (el DL se extiende unas pocas pares de bases del gen de interés) y la detección del gen sólo es posible cuando el marcador se encuentra muy cercano al gen (Rafalski, 2002)(Figura 1.15).

En especies forestales cultivadas con poco grado de domesticación o especies nativas, GWAS es especialmente útil debido a la rápida caída del DL que se manifiesta, gracias al tipo de reproducción predominantemente cruzada y a las recombinaciones genéticas múltiples que ocurren. Por lo tanto, si se cuenta con un barrido genómico de amplia cobertura, resulta altamente probable que un marcador que se haya determinado como asociado, esté ubicado a una distancia física cercana a la variante funcional o incluso sea dicha variante (Neale & Kremer, 2011; Tan, 2018). Por ejemplo, Porth et al. (2013) y Mckown et al. (2014) realizaron GWAS para caracteres de madera, biomasa, ecofisiológicos y de fenología en *Populus trichocarpa* utilizando un chip de 34K SNP. Del mismo modo, un estudio en *Pinus contorta* identificó 11 *loci* que explican el 50% de la variación fenotípica en serotina, mediante el uso de GBS (Parchman *et al*, 2012). El primer GWAS en *Eucalyptus* fue aplicado en *E. globulus* (Cappa *et al*, 2013), mediante el microarreglo de DArT, identificando 16 marcadores asociados a crecimiento y a dos rasgos de lignina. Más recientemente, se aplicó con éxito en rasgos de productividad y resistencia a enfermedades (Resende et al., 2017b) en híbridos de *Eucalyptus* mediante la aplicación del EUChip60K.

La figura 1.15 ilustra un ensayo de asociación genética para la identificación de la variación alélica que influye sobre un carácter determinado, en este caso altura **(a)**, en dos escenarios distintos: bajo desequilibrio de ligamiento **(b)** y alto desequilibrio de ligamiento **(c)**. Se determinan los SNPs de cada genotipo generalmente tomando una muestra pequeña y luego extendiendo al total de la población estudiada para los SNPs seleccionados. En el primer escenario, bajo DL **(b)** y para el SNP superior ("A" vs. "C"), los árboles que poseen el SNP "A" tienen una altura promedio de 92. Mientras que, los que tienen "C" su altura promedio es sólo de 56, coincidiendo con una asociación significativa entre el genotipo SNP y el fenotipo altura. Para el segundo SNP ("G" vs. "T"), los árboles que poseen el alelo "G" tienen un promedio de altura de 74, y los árboles que tienen "T" también tienen un promedio de altura de 74, indicando que este SNP no tiene asociación con el fenotipo altura. Por lo tanto, el bajo desequilibrio de ligamiento resultaría típicamente en una pequeña región cromosómica (pocos cientos de bases, señalado en la figura por cajas coloreadas). En el escenario de alto desequilibrio de ligamiento **(c)**, la región cromosómica que rodea al SNP es mucho más grande, algunos cientos de kilobases, y es común entre los individuos (también señalada con caja coloreada). Por lo tanto, el resultado es que un SNP con significativa asociación a un carácter no esté muy cerca del gen y al seleccionar por este marcador, en otra generación, puede romperse ese ligamiento por recombinación genética y perder la

conexión con el carácter que es lo que se busca para el mejoramiento asistido por marcadores (MAB, del inglés *Marker Assisted Breeding*).

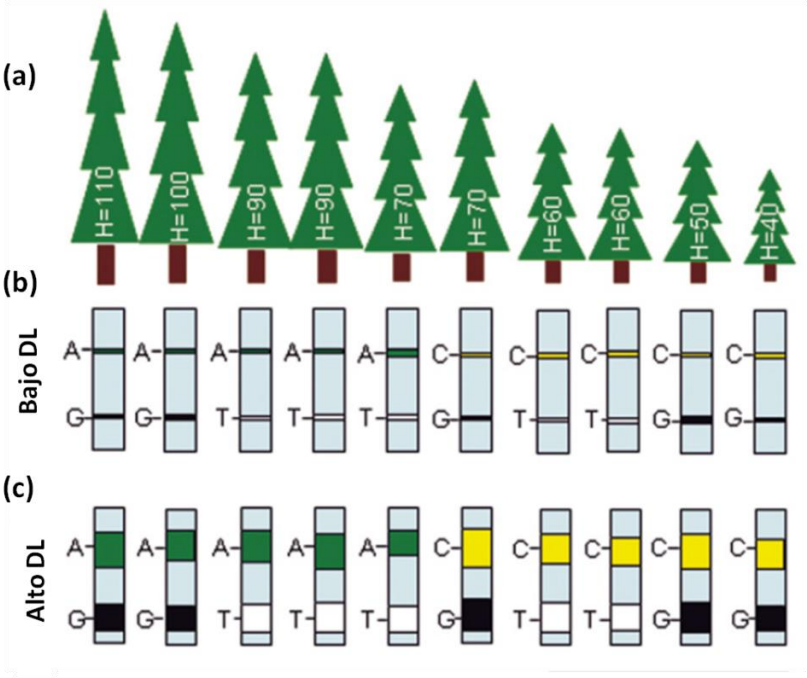


Figura 1.15: *Ensayo de asociación genética para la identificación de la variación alélica que influye sobre un carácter fenotípico "altura", en un escenario de bajo o alto desequilibrio de ligamiento (DL).* **a.** El carácter de interés se mide en árboles no relacionados tomados de una población. Como ejemplo se midió la altura (H, del inglés *height*) de cada árbol a una edad definida. **b.** Genotipo de SNPs en un escenario de bajo desequilibrio de ligamiento y, **c.** en alto desequilibrio de ligamiento. Se señala con cajas coloreadas la amplitud del desequilibrio de ligamiento. Fuente: Modificado de Groover (2007).

Hay varios factores genéticos y no genéticos que pueden afectar la estructura del DL (Flint-Garcia et al., 2003) como son, la recombinación, el efecto de selección, la deriva génica, el sistema de apareamiento, la presencia de estructuración críptica en la población, entre otros. Por ende, el nivel de DL varía entre especies e incluso entre poblaciones de la misma especie. Por lo tanto, también depende del criterio utilizado para la elección de las poblaciones que se utilicen en el estudio. Conocer el nivel de DL es importante para decidir la densidad de marcadores moleculares requeridos para poder capturar la mayor cantidad de QTL presentes en las poblaciones de MA.

El MA involucra varias etapas o pasos que permiten la determinación de asociaciones marcador-carácter; los cuales son: (1) la *fenotipificación* de los caracteres de interés en las poblaciones de MA, (2) la *genotipificación* mediante marcadores moleculares (dentro de Genes Candidatos o distribuidos al largo del genoma), (3) la determinación de la existencia de *estructura poblacional* y/o *estructura familiar* entre los individuos dentro de las poblaciones y, (4) el *mapeo por asociación* propiamente dicho. Todos estos pasos permiten una correcta identificación de secuencias polimórficas asociadas a los distintos caracteres de interés

mediante el empleo de modelos estadísticos que incluyen la información de estructura poblacional y familiar (Yu et al., 2006).

1.8.4 Mapeo por Asociación de Genoma Amplio (Genome Wide Association Studies)

El MA de genoma amplio permite independizarse de la necesidad de tener información o conocimiento previo de los genes candidatos responsables del fenotipo de interés, ya que el abordaje involucra un escaneado amplio del genoma de la especie para la búsqueda de asociaciones. Además, explora otras regiones que pueden ser portadoras del sector del genoma involucrado en la expresión del carácter, y que de otra manera no se incluirían en los análisis. Para estos análisis, se requiere de una adecuada densidad de marcadores moleculares que estén ampliamente distribuidos para cubrir todo el genoma. Hace algunos años, los altos costos impedían la utilización de este tipo de técnicas fuera de consorcios de investigación. La disminución de los costos de genotipificación, por medio de tecnologías de alta procesividad, permitió que el GWAS se difundiera, y posibilitó el desarrollo de varios estudios empleando este abordaje. De esta manera, el GWAS fue utilizado en trabajos en humanos (Chang et al., 2018; DeWan et al., 2006; Hirschhorn & Daly, 2005; Klein et al., 2005), animales (Sharmaa et al., 2015; Yin & König, 2019), plantas (Burghardt et al., 2017; Parchman et al., 2012; Xiao et al., 2017) y hasta su implementación en eucalipto (Kainer et al., 2019; Müller et al., 2017; Resende et al., 2017b), en particular por nuestro grupo de trabajo (Cappa et al., 2013).

1.8.5 Selección Genómica

Se denomina selección genómica (SG) al empleo de las asociaciones entre un fenotipo y muchos marcadores con una cobertura amplia y densa en el genoma para la elección de los mejores individuos en el mejoramiento genético (Meuwissen et al., 2001). A diferencia de la selección asistida por marcadores (MAS) que utiliza marcadores para buscar un número pequeño de *loci* con grandes efectos, la SG se basa en evaluar todos los *loci* que causan la variación fenotípica entre los individuos con una cobertura de densa de marcadores (Hayes & Goddard, 2010; Isik, 2014).

La SG estima simultáneamente el efecto de todos los marcadores en individuos genotipados y fenotipados a través de la aplicación de métodos estadísticos novedosos (Cappa et al., 2019). Al construir dichos modelos en base al análisis de una población con ambos tipos de datos (Población de entrenamiento o PE), la suma de los efectos de los marcadores puede emplearse para predecir los valores de cría (VC) o *breeding value* de individuos que sólo fueron genotipados (Población de aplicación o selección), sin la necesidad de fenotiparlos (Meuwissen et al., 2001). Consecuentemente, la población de aplicación debe tener algún grado de relación genética con la PE (ser descendientes o estar relacionados de algún modo) y debe existir DL entre marcadores y *loci* causales de los rasgos (Tan, 2018; Desta & Ortiz, 2014).

Desde que la SG ha transformado el mejoramiento genético del ganado, a partir de su aplicación exitosa en varios programas de cría en todo el mundo (Goddard et al., 2010), los mejoradores de cultivos agronómicos y forestales la han adoptado con gran expectativa (Isik, 2014). Sin embargo, comparando con los programas de mejora genética de cultivos agrícolas y animales, la aplicación de SG en forestales está en los comienzos (Isik, 2014). No obstante, el impacto de la SG en especies forestales sería mucho mayor, porque presentan gran variación genética para casi cualquier rasgo, proporcionando una oportunidad para aumentar considerablemente la ganancia genética (Grattapaglia, 2014; Isik, 2014).

Así, desde el principio de la década de 2010, ha habido un incremento en trabajos que evaluaron la aplicación de la SG en especies forestales (Grattapaglia & Resende, 2011; Resende et al., 2012; Resende et al., 2012; Zapata-Valenzuela et al., 2012; Zapata-Valenzuela et al., 2013). Trabajos recientes en *Picea* (El-Dien et al., 2015; Ratcliffe et al., 2015; Ratcliffe et al., 2017b), *Pinus pinaster* (Bartholomé et al., 2016; Isik et al., 2015) y especialmente en *Eucalyptus* (Cappa et al., 2019, 2017; Resende et al., 2017), muestran resultados prometedores en la habilidad de predicción de caracteres cuantitativos. Por ejemplo, en dos trabajos de *Eucalyptus* donde se aplicó SG, en uno de ellos se estimó una reducción del 50% del tiempo de una generación de mejoramiento, por ejemplo de 12 a 6 años (Grattapaglia et al., 2011a), y en el otro se obtuvieron precisiones coincidentes a las logradas por la selección fenotípica convencional (Resende et al., 2012a), sin embargo, permitiría acortar los tiempos de selección, al no esperar al crecimiento del árbol para que se expresen los caracteres deseados y poder realizar su análisis.

Un enfoque de SG alternativo, y equivalente al descripto por Meuwissen et al. (2001), es el BLUP genómico o GBLUP (Strandén & Garrick, 2009; VanRaden, 2008). Este método predice los valores de cría mediante el mejor predictor lineal insesgado (BLUP o *best linear unbiased prediction*) utilizando una matriz de parentesco entre individuos (matriz G) calculada a partir de datos genotípicos obtenidos mediante marcadores moleculares. Al utilizar dicha matriz G en lugar de la matriz convencional de relaciones esperadas provenientes del pedigrí (matriz A), GBLUP predice los VCs con mayor exactitud, y así incrementa la ganancia genética y la precisión de la selección para la próxima generación del ciclo de mejora (Nejati-Javaremi et al., 1997; Villanueva et al., 2005).

De este modo, GBLUP es un enfoque muy promisorio a ser aplicado en el mejoramiento genético de especies forestales, como lo han demostrado muchos estudios empíricos (Bartholomé et al., 2016), en particular en *Eucalyptus* (Durán et al., 2017; Grattapaglia et al., 2011). Sin embargo, GBLUP solo brinda valores de cría para los árboles genotipados, y esto limita su implementación en los programas de mejoramiento debido a que, por un tema de costos, no todos los individuos de los ensayos pueden ser genotipados (Cappa et al., 2019). Como ejemplo, Cappa et al. (2019) genotiparon 999 árboles de un total de 2706, y Cappa et al. (2017)187 árboles de un total de 2026.

Una metodología superadora, denominada GBLUP en un solo paso (ssGBLUP o *single step* GBLUP o HBLUP) fue propuesta por Misztal et al. (2009), Christensen & Lund (2010) y Legarra et al. (2011). En la misma, se toman en consideración, dentro del modelo de predicción, tanto los árboles genotipados como los no genotipados, y requiere que la totalidad de ellos cuente con datos fenotípicos evaluados. Para ello, combina la matriz A de relación de parentesco entre todos los individuos a partir del pedigrí, genotipados y no genotipados, con la matriz G de los individuos genotipados en una matriz denominada H (Cappa et al., 2019).

Así, ssGBLUP puede predecir también el valor de cría de los árboles no genotipados. Comúnmente, las relaciones de parentesco que se asumen como teóricas reconocen a las madres como no vinculadas entre sí, es decir relación genética "0". Sin embargo, esto en la mayoría de los casos, no es cierto. Por lo tanto, estos valores de relaciones entre madres se acercan más a la realidad cuando se consideran las relaciones evaluadas mediante marcadores moleculares en gran escala, y se pueden realizar estimaciones más precisas que cuando se utilizan los análisis convencionales de BLUP. Además, la información adicional generada al incluir datos genómicos en el ssGBLUP actúa como un puente de relación genética que conecta la información entre individuos y padres, facilitando una mejor utilización de la información durante el análisis BLUP (Cappa et al., 2017). Por lo tanto, se pueden obtener valores de cría más confiables y precisos de los árboles y así aumentar la probabilidad de un correcto orden de prioridad a la hora de seleccionar individuos para la próxima generación de mejora (Cappa et al., 2019). Varios estudios en animales (Christensen et al., 2012; Croué & Ducrocq, 2017; Guo et al., 2015) indicaron que ssGBLUP es más preciso que utilizar BLUP con pedigrí tradicional (ABLUP; Henderson, 1984) y que aplicar GBLUP. Sin embargo, el uso de ssGBLUP en especies vegetales es reciente y escaso (Cappa et al., 2019). Esta metodología fue aplicada en trigo, utilizando tanto GBS (Pérez-Rodríguez et al., 2017), como ddGBS (Ashraf et al., 2016), y en especies forestales, como en *Picea glauca* (Blaise Ratcliffe et al., 2017a) usando microarreglos tipo Illumina. En el género *Eucalyptus* se aplicó en *E. grandis* con datos provenientes de DArTs (Cappa et al., 2017) y en híbridos interespecíficos de *E. grandis* × *E. urophylla* y de *E. grandis* × *E. camaldulensis* (Cappa et al., 2019), a partir de SNPs del microarreglo EUChip60k.

HIPÓTESIS

- Es posible realizar mejoramiento molecular (asociación genética y selección genómica) en poblaciones diversas, poco domesticadas y que presentan bajo desequilibrio de ligamiento si se cuenta con un alto número de marcadores moleculares polimórficos.
- El empleo de paneles de marcadores de alta cobertura genómica y que segreguen en la población, permite identificar polimorfismos moleculares en las regiones genómicas responsables de aquellas características de interés que varíen en la población
- El análisis con marcadores SNP obtenidos a partir de la estrategia de GBS y mediante el empleo del panel comercial EuChip60K permite relevar diferentes regiones del genoma.
- Los genomas de *E. grandis* y *E. dunnii* están filogenéticamente muy relacionados, existiendo una elevada sintenia entre estas especies.
- En las regiones genómicas asociadas a características de interés en *E. dunnii* existen genes importantes que pueden ser identificados explorando el genoma de *E. grandis*.
- Las relaciones de parentesco entre genotipos "realizadas" (calculadas mediante el uso de marcadores moleculares) mejoran el cálculo de las predicciones de valores genéticos respecto de las relaciones genéticas teóricas obtenidas mediante el pedigrí.

OBJETIVOS

Objetivo general

Desarrollar y aplicar sistemas de genotipificación de alto rendimiento para su utilización en el mejoramiento molecular de *Eucalyptus dunnii* mediante dos estrategias: Mapeo por Asociación y Selección Genómica.

Objetivos específicos

1. Desarrollar herramientas para estudios genómicos de amplia cobertura en *E. dunnii*.

a) Poner a punto la metodología de GBS incluyendo la selección de enzimas de restricción óptimas para el genoma de *E. dunnii*, armado de las genotecas de GBS, secuenciación por NGS y generación de marcadores SNPs sobre un número reducido de genotipos de la población en estudio.
b) Evaluación de la totalidad de la población de mejoramiento de *E. dunnii* por la metodología de GBS ajustada.

2. Caracterizar fenotípica y genotípicamente la población objetivo.

a) Analizar los datos fenotípicos de crecimiento evaluados a distintas edades, índice de rajado de madera y datos de calidad de madera estimados mediante NIR para distintas propiedades físicas y químicas de la madera.
b) Analizar y comparar los datos genotípicos de SNP obtenidos por GBS y mediante el panel comercial EuChip60k. Aplicar los criterios de filtrado, imputación y curación de los datos de acuerdo con parámetros de calidad y genéticos (LD, MAF, etc) requeridos por cada sistema de genotipado. Calcular los estadísticos de genética de poblaciones que permitan describir la población en estudio.

3. Aplicar las metodologías de Mapeo por Asociación y Selección Genómica y la detección de genes candidatos potencialmente útiles para el mejoramiento molecular:

a) Realizar análisis de mapeo por asociación utilizando modelos lineales mixtos. Implementar la búsqueda y caracterización funcional "*in silico*" de los marcadores asociados y genes vecinos utilizando el genoma de *E. grandis* como referencia.
b) Realizar una prueba de concepto de Selección Genómica en la población de *E. dunnii* que permita predecir los valores de cría de los individuos mediante corrección de pedigrí.

2 MATERIALES Y MÉTODOS

1. Caracterización Fenotípica de la Población de Mejoramiento de *E. dunnii*

2.1.1 Población de mejoramiento de E. dunnii y caracterización fenotípica

La población de *E. dunnii* evaluada forma parte de una red de ensayos de la especie en la Mesopotamia y Región pampeana, desarrollada por profesionales de la EEA de Concordia (Entre Ríos) y Bella Vista (Corrientes) de INTA. Dicha red de ensayos comenzó a gestarse en 1989 con la recolección de semillas de polinización abierta (OP: Open pollination) de 72 familias provenientes de cinco fuentes de semillas (Tabla 2.1.1). Sesenta familias fueron colectadas de árboles madre de rodales nativos de *E. dunnii*, a partir de cuatro orígenes del estado de New South Wales (NSW) en Australia. Las familias restantes provinieron de plantaciones de *E. dunnii* ubicadas cerca de Oliveros, Santa Fe, Argentina (fuente de semillas local , de procedencia australiana conocida, Moleton, NSW; Tabla 2.1.1; Marcó & White, 2002), donde fueron seleccionadas de 12 árboles madre, superiores en forma y volumen de fuste.

Tabla 2.1.1. *Fuente de semillas de E. dunnii de la población de mejoramiento de Ubajay*. Fuente de semillas; origen/procedencia: poblaciones nativas de New South Wales (NSW) en Australia, y origen local en Argentina; Coordenadas geográficas (latitud y longitud); Altitud; N° Familias: cantidad de árboles cosechados por cada origen/procedencia, que corresponde a una familia de medios hermanos en OP. Fuente: López et al. (2012).

Fuente de Semilla	Origen/Procedencia	Latitud Sur-Longitud	Altitud (m)	N° Familias
BCUN	Boomi Creek, NSW	28° 25'/152° 41' Este	300	10
	Unumungar State Forest, 10 km al Este de Woodenbong, NSW	28° 25'/152° 42' Este	300	2
DHAC	Death Horse Track Region, Este de Legume, NSW	28° 25'/152° 20' Este	600–700	26
	Acacia Creek, a 25 km al noroeste de Urbenville, NSW	28° 23'/152° 20' Este	600–750	4
OC	Oaky Creek, NSW	28° 36'/152° 31' Este	520	9
SY	South of Yabra State Forest, NSW	28° 36'/152° 29' Este	540	9
SD	Individuos seleccionados de plantaciones comerciales en Oliveros, Santa Fe, Argentina. Origen nativo: Moleton, NSW	30° 10'/152° 10' Este (origen nativo)	420	12
		32° 33' 60° 51' Oeste (plantación)	27	

Figura 2.1.1: *Distribución de la red de ensayos de los programas de mejoramiento de INTA para E. grandis y E. dunnii*. Porción de América del Sur donde se ve Uruguay, el sur de Brasil y Paraguay y el noreste de Argentina. Ubicación de 12 ensayos en la región de la Mesopotamia (provincias de Misiones, Corrientes y Entre Ríos) y la región Pampeana Argentina. *E. grandis*: Ensayos del 1 al 6; *E. dunnii*: ensayos del 7 al 12. Sitio 9: Población de mejoramiento bajo estudio en la presente tesis, municipio de Ubajay, Entre Ríos. Fuente: Marcó & White (2002).

La población de mejoramiento a evaluada en la presente tesis corresponde al sitio número nueve de la Figura 2.1.1. Dicha población fue implantada en el año 1.991 en el municipio de Ubajay (31°45’ de latitud Sur y 58°15’ de longitud Oeste, altitud de 40 m) en la provincia de Entre Ríos, Argentina. La cantidad total de individuos fue de 1520 y se empleó un diseño experimental de 20 bloques completos con parcelas de árboles individuales con una distancia de 3 m × 3 m (Marcó & White, 2002).

A la edad de 6 años, se evaluó la supervivencia de la población (1.458 ejemplares de 1.500 totales), y se midió la altura total (at6 o AT a los 6 años), el diámetro a la altura del pecho (dap6 o DAP a los 6 años, diámetro del tronco a una altura de 1,3 m) y forma del tronco o fuste (for6, Tabla 2.1.2; Marcó & White, 2002). Para las mediciones de diámetro a la altura del pecho se utilizó una cinta diamétrica Forestry Suppliers, Inc. Modelo 347D (EE.UU.) y para altura total, un hipsómetro laser Haglöf® modelo L400 (Suecia). La forma de fuste fue evaluada en una escala arbitraria del uno al cuatro, donde uno indicó la mejor y cuatro la peor forma (M. Marcó & White, 2002).

Tabla 2.1.2. *Caracteres fenotípicos medidos en la población de Ubajay*. Nombre del carácter; edad a la cual fue medido el mismo; código o abreviatura; n: número de individuos con datos; Unidad de medida; Estadísticas poblacionales: Media, desvío estándar (d.e.), mediana, mínimo y máximo. Fuente: Marcó & White (2002), López et al. (2012).

	Caracter	Edad de medición (años)	Código	n	Unidad de medida	Media	d.e.	Mediana	Mínimo	Máximo
1	**Diámetro a la altura del pecho**	6	**dap6**	1458	cm	16,33	3,07	16,30	6,00	27,00
2	**Altura total**		**at6**	1458	m	17,96	2,30	18,00	7,00	24,00
3	**Forma de fuste**		**for6**	1458	1 al 4	2,54	1,06	3,00	1,00	4,00
4	**Diámetro a la altura del pecho**	11	**dap11**	316	cm	24,77	3,48	24,50	14,20	35,30
5	**Diámetro a la altura del pecho**	20	**dap20**	318	cm	35,31	5,59	35,00	21,70	56,00
6	**Altura total**		**at20**	318	m	33,96	3,88	33,80	23,30	44,00
7	**Índice de rajado de rollizo**		**ir20**	318		1,01	0,59	0,87	0,21	3,51
8	**Extractivos etanólicos**		**extet20**	269		3,29	1,20	3,10	1,00	8,80
9	**Extractivos totales**		**exttot20**	269		5,09	1,41	4,90	2,10	11,40
10	**Lignina Klason**		**klas20**	269		24,99	0,66	24,90	23,00	27,00
11	**Lignina Total**		**lig20**	269		28,44	1,04	28,40	24,80	31,80
12	**Relación Siringilo/Guayacilo**		**sg20**	269		1,83	0,07	1,83	1,62	2,01
13	**Celulosa Total**		**cel20**	269		46,86	1,20	46,90	42,70	50,10
14	**Densidad básica**		**db20**	269	Kg/m^3	516,52	67,43	523,80	313,00	666,90

A partir de dichas mediciones, Marcó & White (2002) seleccionaron los mejores individuos en base a su forma y crecimiento (cálculo del volumen a partir del diámetro y la altura), a través de las estimaciones de ganancia genética utilizando un índice combinado de selección, donde se capturó el 87% y 70% de la máxima ganancia posible para volumen y forma, respectivamente.

A los 11 años de edad (año 2002), la población de Ubajay fue raleada y transformada en Huerto Semillero de Progenies, seleccionando los mejores individuos en base a su forma y crecimiento (cálculo del volumen a partir de DAP y AT), a través de las estimaciones de ganancia genética utilizando un índice combinado de selección, donde se capturó el 87% y 70% de la máxima ganancia posible para volumen y forma, respectivamente (M. Marcó & White, 2002).

Sin embargo, para mantener la diversidad genética y procurando el mayor equilibrio posible en cuanto a la distribución espacial, se conservaron entre 6-8 individuos de las familias con mejor comportamiento (mejor forma y mayor crecimiento o volumen), 3-5 individuos de las familias de comportamiento intermedio y 1-2 ejemplares de las de peor comportamiento (López et al., 2012), totalizando 329 árboles. En ese mismo

momento, se volvieron a tomar medidas del diámetro a la altura del pecho (dap11), como fue explicado anteriormente (Tabla 2.1.2).

Al mismo tiempo (año 2002), los genotipos de mayor ganancia genética fueron movilizados vía injerto e instalados en un Huerto Semillero Clonal en el Instituto de Recursos Biológicos del INTA Castelar, donde las condiciones climáticas son muy favorables para la producción de semilla (34º 37´ S, 58º 40´ O, López et al., 2012).

Finalmente, en el año 2011, cuando la población de Ubajay presentaba 20 años de edad, se midió el diámetro a 1,30 m de altura (dap20) y la altura total (at20, Tabla 2.1.2). Luego, se realizó un muestreo destructivo con el objetivo de identificar individuos con bajo nivel de tensiones de crecimiento (o bajo índice de rajado), mayor estabilidad dimensional y una densidad de la madera similar al promedio de la especie (López et al., 2012) para su utilización como madera sólida.

La estimaron de la intensidad de las tensiones de crecimiento fue evaluada a través de la medición del índice de rajado en rollizos (ir20, López et al., 2012). Para ello, los ejemplares fueron cortados manualmente con motosierra a partir de 1,30 m de altura, y en cada uno de ellos se obtuvieron dos rollizos consecutivos de 2,50 m de longitud. Inmediatamente después del trozado, los extremos de ambos rollizos fueron cubiertos con bolsas plásticas a efectos de mantener la humedad natural y retardar el proceso de secado. De esta manera, las rajaduras observadas, casi exclusivamente, pueden atribuirse a la liberación de las tensiones de crecimiento. Luego de transcurridas 72 horas, en los extremos de ambos rollizos se tomaron imágenes (Figura 2.1.2), y a partir de las mismas fueron calculados los Índice de Rajado en rollizos (ir20) para cada ejemplar. Este índice relaciona el área ocupada por las rajaduras y el área total de la sección transversal que contiene a esas rajaduras (cara del rollizo, López et al., 2012). El ir20 a través del cual fue caracterizado cada individuo corresponde al promedio de cuatro caras evaluadas de los dos rollizos evaluados (Tabla 2.1.2).

Figura 2.1.2. *Variación del índice de rajado en las caras de rollizos de E. dunnii.* Imágenes de las caras de los rollizos 72 horas posteriores al apeo. Los individuos se muestran ordenados de izquierda a derecha, de menor a mayor índice de rajado. Huerto Semillero de Progenies de Ubajay, Entre Ríos, año 2011. Fuente: Juan A. López (EEA Bella Vista INTA).

A partir de listones obtenidos de los rollizos, se realizaron estimaciones de propiedades químicas de la madera mediante NIR (*Near Infrared spectroscopy* o Reflectancia en el infrarrojo cercano) en el servicio de

espectrometría del Instituto Superior de Agronomia (ISA) en Portugal. Se estimaron seis características: contenido de celulosa (cel20), extractivos totales (exttot20), extractivos etanólicos (extet20), contenido de lignina Klason (klas20), contenido total de lignina (lig20), relación de los monómeros siringilo y guayacilo de la lignina (sg20, Tabla 2.1.2). Además, fue estimada la densidad básica de la madera (db20, Kg/m^3, Tabla 2.1.2), determinando el peso anhidro con una balanza electrónica de precisión (0,001 g) y el volumen por el método de desplazamiento de fluidos (método de inmersión en agua).

Brevemente, para las estimaciones mediante espectros NIR, las muestras fueron secadas en una estufa y los listones se molieron hasta convertirlos en aserrín con un molino Retsch (Alemania), hasta obtener fracciones entre 1 y 2 mm que fueron microdigeridas para ser químicamente caracterizadas. Los espectros NIR fueron obtenidos por reflectancia difusa en un equipamiento Bruker modelo Vector 22 N/I/F (EE.UU.) con esfera de integración y sonda de fibras ópticas. Para hacer las determinaciones de las distintas propiedades, se construyeron curvas para la generación de modelos predictivos a partir de ensayos químicos adecuados, en donde se analizaron muestras que cubrieron la distribución de los espectros, para que sea lo más confiable y amplia posible. Dependiendo la característica química se analizaron entre 20-50 muestras químicamente y se validaron los modelos predictivos generados con otras 15-22 muestras (Rodrigues et al., 1998).

2.1.2 Ajuste de datos de características fenotípicas

Se ajustaron los datos de las catorce características fenotípicas para la totalidad de la población de Ubajay, la cual presentó entre 1.468 y 269 individuos con datos según el carácter a considerar (Tabla 2.1.2). Se verificó la normalidad de las distribuciones mediante la prueba de *Shapiro-Wilks* (Shapiro & Wilk, 1965) implementada en la función *shapiro.test* del paquete *stats* de R versión 3.6.1 (R Core Team, 2019).

A partir de dichos análisis, se ajustaron de acuerdo con si las distribuciones presentaban asimetría positiva (moda menor a la media o distribución sesgada hacia la izquierda) o negativa (moda mayor a la media o distribución sesgada hacia la derecha) y se transformaron mediante el logaritmo en base 10. Si las mismas presentaban leptocurtosis (distribución más en punta que la normal) se corrigieron los valores crudos por la inversa y si presentaban platicurtosis (distribución más en plana que la normal) los valores crudos fueron elevados al cuadrado.

La variable categórica de forma de fuste a los 6 años (for6) fue transformada a variable continua mediante *Normal Score* con el paquete *stats* en R (R Core Team, 2019; Jansson y Danell, 1993; Salvador Figueras y Gargallo, 2003).

Luego de transformar para normalidad, excepto for6, se estandarizaron todos los datos, restándole a todos los valores la media y dividiéndolos por el desvío estándar del carácter correspondiente.

Se utilizó la función *remlf90* mediante el programa *breedR* (Muñoz & Sánchez, 2014), con un modelo lineal mixto de árbol individual usando la inferencia de máxima verosimilitud restringida (REML, Patterson & Thompson, 1971), ajustando los caracteres para remover los efectos ambientales a gran escala (efecto de bloque).

Por último, se estimaron las correlaciones entre cada par de caracteres fenotípicos y su significancia mediante la función *pairs.panels* del paquete *psych* de R (Revelle, 2019). Esta información es útil para plantear estrategias de mejoramiento conjunto para los caracteres que correlacionen positivamente, o estrategias de mejoramiento paralelas si estas fueran negativas, con el fin de cubrir diferentes necesidades de la demanda del sector forestal.

2. Desarrollo de una Metodología de Genotipificación Masiva para *E. dunnii*

2.2.1 Material vegetal, extracción y cuantificación de ADN

Se optimizó y se aplicó un protocolo derivado de ddRADseq en dos muestras de *E. dunnii* (A y B), y posteriormente se amplió a otras 24 muestras (1 a 24). Las muestras pertenecen al programa de mejoramiento de *E. dunnii* del INTA (Anexo 7.1.1, Tabla 7.1.1). Se recogieron hojas jóvenes frescas, se secaron en un liofilizador (Labconco Corporation, Kansas City, MO, EE. UU.) y se conservaron en sílica gel hasta la extracción de ADN (ácido desoxirribonucleico). El ADN genómico se extrajo de las hojas liofilizadas siguiendo el método CTAB (Anexo 7.1.2) descrito por Hoisington et al. (1994) con modificaciones para *E. dunnii* como se describe en Marcucci Poltri et al. (2003). Su calidad fue verificada por Nanodrop (Thermo Fisher Scientific, Waltham, MA, EE. UU.) y electroforesis en gel de agarosa al 1% teñidos con bromuro de etidio. El ADN se cuantificó usando un fluorómetro Qubit 2.0 (Thermo Fisher Scientific).

2.2.2 Evaluación de enzimas óptimas para la digestión y rango de selección de tamaño de los fragmentos de ADN in silico

Se realizaron varias digestiones *in silico* del genoma de referencia de *E. grandis v2.0* (disponible en Phytozome https://phytozome.jgi.doe.gov/pz/portal.html, Myburg et al., 2014) para evaluar tanto el conjunto óptimo de enzimas de restricción para el genoma de *E. dunnii* y el número de fragmentos de ADN que se recuperarán mediante diferentes selecciones de tamaño (Parchman et al., 2018; Peterson et al., 2012). Las simulaciones se llevaron a cabo utilizando el paquete SimRAD (Lepais & Weir, 2014). Se probaron las combinaciones de enzimas de restricción PstI-MspI y SphI-MboI, según los trabajos de Peterson et al. (2014) y Scaglione et al. (2015), respectivamente. Además, se evaluaron diferentes selecciones de tamaño para lograr entre 1e4 a 5e4 fragmentos en una ventana óptima entre 50 a 100 pares de bases (pb) según lo sugerido por Peterson et al. (2012) o hasta 150 pb. Finalmente, se buscó un tamaño medio de inserto entre 295 a 420 pb,

obteniendo un tamaño de biblioteca final entre 350 y 600 pb. Este rango de tamaño es adecuado para la amplificación en puentes de las plataformas Illumina y para poder obtener un solapamiento mínimo de lecturas de secuencias pareadas o lecturas Paired End (PE) de 150 pb o más. De esta manera, se aprovecha una mayor cantidad de datos de bp.

Para simular la amplificación de los fragmentos con ambos extremos de sitios de corte de enzimas se aplicó la rutina *insilico.digest* para ambas combinaciones de enzimas y la rutina *adapt.select*. Finalmente, para seleccionar diferentes subpoblaciones de fragmentos para cada digestión se utilizó la opción *size.select*. Las medias de tamaños consideradas fueron de 320, 370, 420 pb, con dos anchos de ventana para simular selección manual (electroforesis en gel de agarosa, 100 o 150 pb) y automática (por SAGE ELF, 70 o 140 pb, para uno o dos pocillos de elución).

2.2.3 Evaluación de digestiones in vitro

Las digestiones *in vitro* del ADN de *E. dunnii* se realizaron utilizando las condiciones de reacción descritas por Scaglione et al. (2015). El perfil de los fragmentos obtenidos se visualizó en un gel de agarosa (Anexo 7.1.6, Figura 3) y electroforesis capilar en el sistema Fragment Analyzer 5200 (Advanced Analytical Technologies, Inc., Santa Clara, CA, EE. UU.) utilizando el kit de alta sensibilidad de ADN (Agilent Technologies, Santa Clara, CA, EE. UU.).

2.2.4 ddRADseq optimizado: Protocolo 1

Digestión: 150 ng de cada ADN de las muestras A y B se digirieron completamente usando SphI-HF y MboI (2,4 U por enzima, New England Biolabs (NEB), Ipswich, MA, EE. UU.), y se incubaron a 37 °C durante 90 min. La reacción se inactivó a 65 °C durante 20 min y se purificó con 1,5 volúmenes (×) de microesferas o perlas magnéticas Ampure XP (Beckman Coulter, Brea, CA, EE. UU.; Scaglione et al., 2015). En este punto, la digestión completa del ADN se evaluó mediante electroforesis en Fragment Analyzer (Advanced Analytical Technologies, Inc.), esperando una población de fragmentos distribuidos homogéneamente y menores a 3 Kpb aproximadamente para una digestión óptima.

Ligación: se utilizaron los adaptadores universales (oligonucleótidos bicatenarios) publicados por Peterson et al. (2014; Anexo 7.1.5). Específicamente, el Adaptador 2 (A2) fue diseñado con forma de "Y" para la amplificación específica de fragmentos con extremos con diferentes sitios de corte de enzimas. Los extremos adhesivos del adaptador 1 (A1) y A2 se cambiaron a los correspondientes a SphI y MboI, respectivamente. Se evaluaron distintas concentraciones de adaptadores A1 y A2: 2 y 5 pmol similares al protocolo de Scaglione et al. (2015), 2 pmol de ambos adaptadores, según Elshire et al. (2011), y 0,1 y 15 pmol, según Peterson et al. (2014) y se eligió la concentración que mostró menor cantidad de dímeros de adaptadores (2 pmol y 5 pmol

de A1 y A2, respectivamente).. La reacción utilizó 2,4 unidades Weiss de ADN ligasa T4 (Invitrogen, Carlsbad, CA, EE. UU) y se incubó durante 1 hora a 23 ° C, seguido de una incubación adicional durante 1 hora a 20 °C e inactivación durante 20 minutos a 65 °C (Scaglione et al., 2015). Luego se realizó una purificación con 1 × de Ampure XP por muestra antes de realizar la PCR (*Polymerase Chain Reaction* o reacción en cadena de la polimerasa).

PCR: para las reacciones se utilizaron los cebadores o *primers* de doble índice diseñados por Lange et al. (2014; Tabla suplementaria S2). Dichos oligonucleótidos presentan en un extremo una región necesaria para secuenciar en las plataformas Illumina, luego un índice (secuencia específica de 8 pb) que permite la identificación de cada genoteca, y por último una secuencia complementaria a los adaptadores. Se utilizó la ADN polimerasa de alta fidelidad (NEB Phusion) con el siguiente protocolo de ciclado (Scaglione et al., 2015): 3 minutos de desnaturalización inicial (95 °C), 10 ciclos de amplificación (30 seg a 95 °C, 30 seg a 60 °C, 45 seg a 72 °C), y 2 minutos de una extensión final (72 °C). Posteriormente se realizó una purificación de perlas Ampure XP 1,2 × por PCR.

Mezcla o agrupamiento (*pooling*): después de agregar los cebadores indexados por PCR, las genotecas obtenidas se mezclaron en forma equimolar según su concentración medida en fluorómetro (Qubit 2.0) y se concentraron por vacío y calor (30 a 45 °C por 45 o 60 min) en un equipo SpeedVac (Eppendorf, Hamburgo, Alemania).

Selección de tamaño: se aplicó una selección de tamaño manual (rango entre 450 y 550 pb, correspondiente a fragmentos de interés entre 310 y 410 pb, sin considerar adaptadores) mediante electroforesis en gel de agarosa de bajo punto de fusión al 1,5% (Bio-Rad Laboratories, Hercules, CA, EE. UU.) y corte de la agarosa mediante el uso de un bisturí. Finalmente, los fragmentos de ADN seleccionados se purificaron a partir del gel con un kit QIAquick Gel Extraction (Qiagen N.V., Hilden, Alemania; Scaglione et al., 2015).

Secuenciación: Se realizó una secuenciación PE (2 × 151 pb) en MiSeq (Illumina Inc., San Diego, CA, EE. UU.) para ambas muestras luego que las genotecas finales se cuantificaron cuantificaran mediante fluorómetro Qubit 2.0 (kit dsDNA de High Sensitivity, Thermo Fisher Scientific) y se verificó su calidad en el sistema Fragment Analyzer (kit de ADN de alta sensibilidad, Agilent). Se realizó una secuenciación PE (2 × 151 pb) en MiSeq (Illumina Inc., San Diego, CA, EE. UU.) para ambas muestras (Figura 2.1.1). Los datos de secuenciación fueron almacenados en el servidor del instituto IABiMo. Finalmente, el secuenciador otorga 2 archivos de formato fastq (archivos que contienen datos de las secuencias con la calidad de cada base especificada mediante el índice de calidad de Phred) por muestra.

Para más detalles, en la figura 2.2.1 se esquematiza un resumen del protocolo 1 y en el Anexo 7.1.6 se muestra una versión extendida.

2.2.5 ddRADseq optimizado: prueba piloto de Protocolo 2

El P1 ddRADseq optimizado se escaló posteriormente para obtener el P2 24-plex (muestras 1 a 24, anexo 7.1.7). En primer lugar, el P2 se puso a punto en 24 muestras y consistió en P1 con algunas modificaciones de la siguiente manera:

Ligación: se agregaron 24 códigos de barras o *barcodes* de longitud variable (4 a 9 pb) diseñados por Poland et al. (2012), que permiten agrupar las muestras en un paso anterior a la PCR. Además, dicha longitud variable evita una baja calidad de secuencia de las primeras bases de las lecturas dada por la homogeneidad de bases de los sitios de restricción (Elshire et al., 2011; Yang et al., 2016; Anexo 7.1.5). Además, cada reacción se realizó con 160 U de ligasa (NEB Cohesive End Ligation).

Agrupamiento de muestras: las ligaciones se agruparon de a 24, en función de la concentración de cada digestión cuantificada con Picogreen (Sigma-Aldrich) en un fluorómetro FluorStar Optima (BMG Labtech, Ortenberg, Alemania).

Selección de tamaño: se realizó una selección automática de tamaño en un casete de agarosa al 2% en el equipo SAGE ELF (Sage Science, Inc., Beverly, MA, EE. UU.) y se recogieron los fragmentos de 450 pb en promedio (entre 415 y 485 pb) de uno de los 12 pocillos disponibles. Posteriormente, se realizó un paso adicional de purificación 0,8 × de perlas Ampure XP, para sólo recuperar los fragmentos de ADN mayores a 300 pb.

PCR: se realizó una PCR para el grupo de genotecas (un par de *primers* con índices que identifican al grupo).

Secuenciación: el grupo de genotecas (*pool*) se secuenció primero en modo PE (2 × 250 pb) y a baja profundidad en un equipo MiSeq (Illumina Inc.), y finalmente se secuenció en modo PE (2 × 75 pb) y profundidad óptima en un secuenciador NextSeq 500 (Illumina Inc.).

Para más detalles, en la figura 2.2.1 se esquematiza un resumen del protocolo 2 y en el Anexo 7.1.7 se muestra una versión extendida. En la figura 2.2.1 se muestra un diagrama de los pasos de ambos protocolos (P1 y P2) y en la figura 2.2.2 se ofrece un esquema del armado de genotecas con el P2. Se esquematizan los sitios donde el ADN será digerido por ambas enzimas. Se muestra en detalle un fragmento de ADN con extremos cohesivos o *sticky* para ambas enzimas. Se pueden visualizar los adaptadores con códigos de barras

o *barcodes* y la porción final de las secuencias que incorporan los oligonucleótidos o *primers* en el paso de PCR. Estos *primers* incluyen secuencias índices y secuencias complementarias a la celda de flujo o *flowcell* de los secuenciadores Illumina. Con respecto a dichas porciones de los *primers*, las primeras permiten identificar los pooles de muestras en el P2 y son leídas por primers de secuenciación especiales. Las segundas permiten la fijación de los fragmentos a secuenciar a la celda de flujo, y su posterior amplificación clonal para que puedan ser detectadas en momento de la secuenciación. Por último, se esquematizan las porciones de un fragmento de las genotecas que efectivamente son secuenciadas y incluidas dentro del par de lecturas PE. Nótese que sólo se esquematiza un par de lecturas PE, pero por muestra se estima obtener entre 500 mil a 1 millón de lecturas PE. Se puede observar que cada muestra presenta un *barcode* que la identifica y se detallan los posibles SNPs a encontrar. En este caso el secuenciador NextSeq brindará dos archivos en formato fastq por cada pool de muestras (uno con lecturas forward o R1 y otra con las reverse o R2).

Figura 2.2.1. *Protocolos optimizados de ddRADseq para E. dunnii*. Se enumeran cada uno de los pasos involucrados. Referencia: PE: lecturas *Paired End*. PCR: *Polymerase Chain Reaction*. Fuente: modificado de Aguirre et al. (2019).

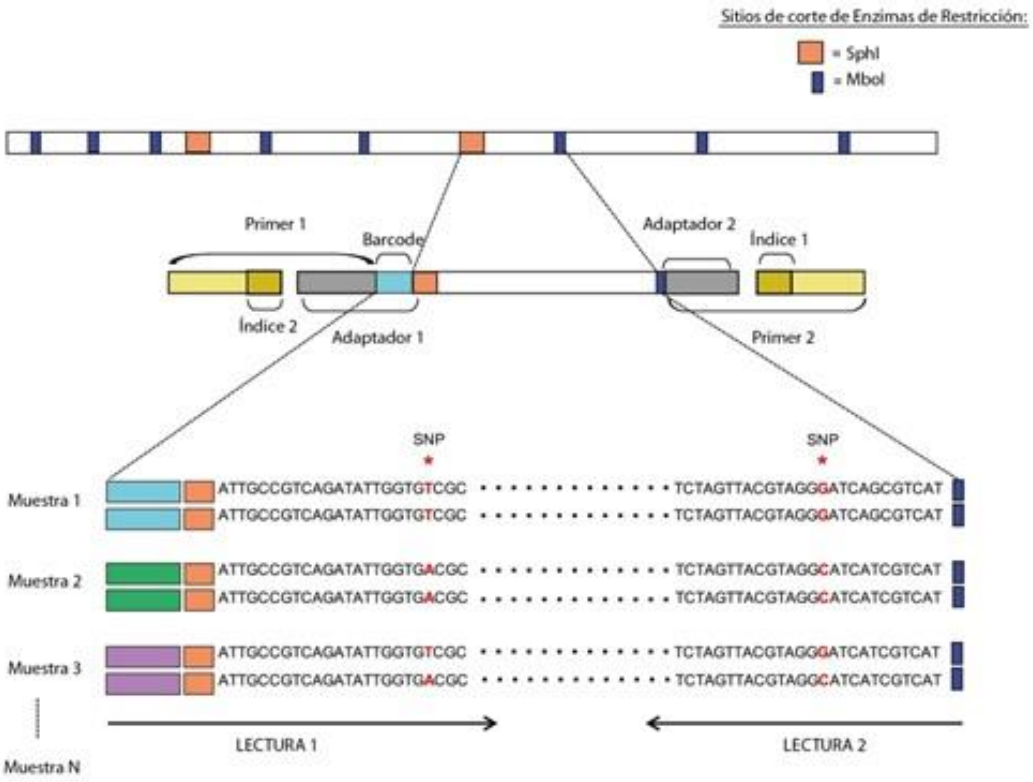


Figura 2.2.2. *Esquema general de genoteca obtenida con protocolo optimizado de ddRADseq 2.0.* ADN digerido con el par enzimático SphI-MboI, ligación de adaptadores con *barcodes*, amplificación por PCR con oligonucleótidos con índices. Se esquematizan fragmentos de tres muestras secuenciados por *PE* (Pair End: lecturas en ambos extremos de los fragmentos) y se localizan SNPs en ambos *loci*. Fuente: modificado de Truong et al. (2012).

2.2.6 Análisis bioinformático de secuencias de ddRADseq

El análisis bioinformático descripto a continuación es uno de los pasos fundamentales en el proceso de obtención de marcadores informativos a partir de las metodologías de GBS o ddRADseq. El mismo permite obtener una matriz de marcadores para todas muestras partiendo de archivos de texto en formato fastq que contienen millones de secuencias para cada muestra (Torkamaneh et al., 2016). Antes de comenzar dicho análisis, se verificó la calidad de secuenciación analizando los archivos fastq mediante el programa FastqC (Andrews, 2010). Este programa permite ver estadísticas de calidad generales de la secuenciación NGS, como número de lecturas o reads totales, porcentaje de contenido GC total y proporción de bases a lo largo de todas las lecturas, calidad de secuenciación por base a lo largo de toda la lectura, calidad promedio por lecturas, presencia de bases no asignadas y presencia de adaptadores, entre otros parámetros de calidad.

Aunque muchos programas de bioinformática y paquetes de R (R Core Team, 2019) pueden manejar este tipo de datos de secuenciación de representación reducida del genoma, *Stacks* (Catchen et al., 2013; Catchen et al., 2011; v1.48, Universidad de Oregón, Eugene, OR, EE. UU.) es uno los *softwares* o *pipelines* (programas)

que funciona igualmente bien cuando se trabaja con o sin un genoma de referencia. Este software fue desarrollado principalmente para organismos sin genomas de referencia y secuenciación RAD de alta profundidad. Además, *Stacks* se encuentra entre los programas con alta precisión para el llamado de SNP en este tipo de datos de secuenciación (Torkamaneh et al., 2016; Wickland et al., 2017).

Una vez obtenidos los datos utilizando ambos protocolos, éstos se analizaron con diferentes componentes de Stacks v1.48. Las muestras A y B (analizadas en MiSeq, y mediante dos archivos *fastq* por muestra) se usaron para comparar la eficiencia entre la definición de *loci de novo* o con el genoma de referencia de *E. grandis v2.0*, así como para evaluar la utilidad del P1. En el caso del P2, por otro lado, se utilizó genoma de referencia en las muestras 1 a 24 (analizadas en NextSeq 500, mediante dos archivos *fastq* conteniendo las lecturas del pool de las 24 genotecas). Además, para evaluar el rendimiento de las plataformas Illumina se analizaron las genotecas de las muestras que se secuenciaron dos veces (MiSeq y NextSeq, ambas repeticiones juntas) para evaluar el rendimiento de las plataformas Illumina.

Depurado de secuencias por calidad: El paquete de análisis informático de Stacks, en primer lugar, permite una limpieza por calidad de todas las secuencias del conjunto de muestras, a partir de los archivos formato fastq. Dicha limpieza por calidad es específica para genotecas de representación reducida, en particular para aquellas que utilizan enzimas de restricción, y es implementada mediante el componente *process_radtag.pl*. En esta limpieza se eliminaron las lecturas que presentaban: bases que no fueron definidas (N), bajo valor medio de índice de calidad de Phred (inferior a 10, o sea un error de asignación de bases mayor al 10%), ausencia de sitios de reconocimiento de enzimas y presencia de secuencias de adaptadores. Además, las muestras A y B se recortaron hasta 145 pb de largo, debido a que la calidad de las lecturas cayó en las últimas bases, según lo observado con FastqC (Andrews, 2010).

Por otro lado, también fueron filtradas de acuerdo con su calidad las secuencias de las muestras 1 a 24 a las que se les aplicó el P2. Además, debido a que estas muestras se encontraban agrupadas en dos archivos formato fastq, las mismas se demultiplexaron (se separaron las lecturas correspondientes a cada muestra, teniendo en cuenta el código de barras presente al inicio de cada lectura, que permite asignar cada una a una muestra en particular) generando dos archivos fastq por muestra. Finalmente, ya que las lecturas presentaban diferentes longitudes después de eliminar entre 4 a 9 pb correspondientes a los códigos de barras, y con el fin de obtener lecturas de la misma longitud, las mismas se truncaron en 66 pb (75 menos 9 pb). Para los análisis posteriores, se consideraron tanto las lecturas PE como SE que pasaron los filtros (SE que sólo pasó el filtro una lectura del par).

Búsqueda de loci y SNPs de novo y con genoma de referencia: Para definir los *loci* y finalmente los SNPs, el programa *Stacks*, en primer lugar, apila las secuencias que son 100% idénticas dentro de cada muestra. Estas

pilas o alelos (*tags* o *stacks*) pueden generarse sin genoma de referencia o con la guía de uno. En el paso siguiente, dentro de cada muestra, dichas pilas son contrastadas y comparadas en búsqueda de pequeñas variaciones entre las mismas, definiendo de esta manera *loci* dentro de individuos, los cuales pueden ser heterocigotas para uno o varios SNPs. Una vez analizado cada individuo, Stacks genera un catálogo con todos los *loci* encontrados para todos los individuos, donde también se contrastan todos los *loci* en busca de un pequeño número de bases cambiantes que definan SNPs nuevos entre individuos. Finalmente, todos los *loci* de cada individuo o muestra son comparados con el catálogo para la genotipificación final, donde se define que *loci* de la población está presente en cada individuo y que SNPs y variantes tiene con respecto a la población (Catchen et al., 2013).

Por lo tanto, se utilizó la pipeline *denovo_map.pl* (componente de *Stacks*) para buscar *loci* y SNPs *de novo* (solo para P1), mientras que *ref_map.pl.pl* se aplicó tanto en P1 y P2 para llamar SNPs después de mapear las lecturas contra el genoma de referencia de *E. grandis* con *Bowtie2* (parámetros predeterminados; Langmead & Salzberg, 2012) con parámetros default. En todos los análisis, se utilizó un mínimo de tres lecturas para definir un alelo (o pila) dentro de un individuo (-m 3). Particularmente para el análisis *de novo*, se permitieron tres bases diferentes entre alelos (pila, *tags* o *stacks*) para construir un *locus* dentro de un individuo (-M 3) y se permitieron tres bases diferentes entre *loci* para construir el catálogo (-n 3).

Después de ejecutar los tres análisis (análisis *de novo* para P1 y con referencia tanto para P1 como P2), se aplicó el componente *rxstacks* para filtrar los posibles errores de definición de genotipo. Por lo tanto, se aplicó el modelo de SNP acotado (bounded SNP model) y se retuvieron *loci* con logaritmo de las verosimilitudes mayores que -10 (menos diez). Además, fueron admitidos hasta una proporción de 0,05 de individuos con *loci* que presentaron más de dos alelos, y los haplotipos en exceso de *loci* individuales fueron podados de acuerdo con su prevalencia en la población.

Obtención de matrices finales (archivos VCF): Finalmente, para poder evaluar los resultados de ambos protocolos se exportaron seis matrices diferentes, cada una comprendiendo distintas muestras y filtros. El programa *Stacks v1.48* necesita del componente *populations* para exportar los datos de los genotipos en un formato VCF (*Variant Call Format* o formato de llamada variante) que permite la ejecución de análisis posteriores. Así, el componente *populations* se ejecutó con diferentes combinaciones de filtros, lo que resultó en seis archivos VCF para los SNP y *loci* ddRADseq polimórficos de diferentes grupos de muestras detallados a continuación.

Para las muestras A y B, se obtuvieron cuatro matrices básicas, una con la totalidad de los datos (marcadores totales) y otra sin datos faltantes (marcadores compartidos entre ambas muestras), tanto para el análisis *de novo* como para el análisis con genoma de referencia. Así, se contabilizaron los *loci* y SNPs que fueron encontrados

en cada muestra, con o sin genoma de referencia, y cuántos de ellos fueron compartidos entre ambas en cada análisis.

Para las muestras 1 a 24 derivadas de P2, se obtuvieron dos matrices, una para las genotecas secuenciadas en NextSeq y otra para dichas genotecas más sus repeticiones obtenidas en la secuenciación de MiSeq. Ambas matrices se filtraron por una frecuencia mínima de alelos (MAF) de 0,05, y *loci* con hasta un 20% de datos perdidos. Así, al igual que para las muestras A y B se contabilizaron los *loci* y SNPs que fueron encontrados en cada muestra y se evaluó la distribución de estos a lo largo del genoma de *E. grandis*.

2.2.7 Identificación de SSRs

Para las muestras A y B, se identificaron los SSR polimórficos presentes utilizando el software MicroSAtellite (Institute of Plant Genetics and Crop Plant Research, Gatersleben, Alemania), también conocido como MISA (Thiel et al., 2003), de acuerdo a lo realizado en el trabajo de Qin et al. (2017). La opción *fasta_sample* del módulo de *populations* (Stacks v1.48) se utilizó para obtener las secuencias de los dos haplotipos de cada muestra para cada *locus* en formato FASTA. Luego, de acuerdo con el mismo criterio utilizado por Torales et al. (2018), se buscaron SSR con un mínimo de cinco repeticiones para dinucleótidos, cuatro repeticiones para trinucleótidos y tres repeticiones para motivos de tetra, penta y hexanucleótidos. También se buscaron y analizaron los SSR polimórficos.

2.2.8 Evaluación de la robustez de los SNP – Comparación de las plataformas de secuenciación

La robustez y reproducibilidad para encontrar SNPs del P2 se evaluó comparando los conjuntos de datos de NextSeq y MiSeq de las 46 genotecas (23 NextSeq y sus 23 repeticiones a baja profundidad de MiSeq). Para ello se partió del archivo VCF filtrado por datos faltantes y MAF inferior al 20% y 0,05 respectivamente (obtenido en el apartado 2.2.6). Se calculó la matriz de disimilitud, mediante el método de idéntico por descendencia, entre las 46 genotecas directamente desde el VCF filtrado usando el paquete *SNPRelate* en R (R Core Team, 2019; Zheng et al., 2012). El dendrograma se construyó utilizando el análisis de agrupamiento jerárquico implementado en la opción *snpgdsHCluster* (*SNPRelate*) y se graficó con el paquete *ggplot2* de R (Wickham, 2016).

3. Caracterización Genotípica de la Población de Mejoramiento de *E. dunnii*

En primer lugar, se extrajo el ADN de la totalidad de la población (308 individuos) con el método de CTAB (Anexo 7.1.1), de la misma manera que en las 26 muestras iniciales para la optimización de los protocolos ddRADseq para la especie. Los mismos se cuantificaron mediante Qubit 2.0, y se verificó su calidad mediante Nanodrop y geles de agarosa al 1% teñidos con bromuro de etidio. Finalmente se realizaron diluciones del ADN a 15 ng/µl en agua bidestilada estéril.

2.3.1 Genotipificación con datos de secuencias ddRADSeq

Luego de la puesta a punto de ambos protocolos de genotipificación ddRADseq optimizados para *E. dunnii*, se aplicó el P2 a toda la población de mejoramiento, mediante el armado de 13 grupos de 24 muestras cada uno (incluyendo 4 muestras extra por requerimiento del protocolo) y fueron secuenciadas en un NextSeq 500 (Hospital Gutiérrez, CABA, Argentina). Dicha secuenciación se llevó a cabo con un kit comercial que brinda 150 pb de largo de lectura (150-cycle high output kit NextSeq, Illumina Inc.), cuyo rendimiento esperado es de 400 millones de lecturas PE. Se llevó a cabo una única corrida programada para obtener lecturas PE de 75 pb cada una (800 millones de lecturas totales), y se estimó un promedio de 61,5 millones de lecturas PE por grupo de muestras (800 millones en los 13 grupos).

Luego de la obtención de las secuencias de ddRADseq (de aquí en adelante se nombrará como GBS a todo lo referido a los datos de ddRADseq) se procedió con el análisis bioinformático de todas las muestras de la población de Ubajay.

Para la búsqueda de *loci* y marcadores SNPs en los datos de secuencias de GBS se utilizó nuevamente el software *Stacks v1.48* (Catchen et al., 2013), y se aplicaron los mismos pasos que a las 24 muestras de la prueba piloto del P2. En resumen, se depuraron las secuencias por calidad (truncando las lecturas en 66 pb) y se realizó una búsqueda de *loci* y SNPs *de novo* y con genoma de referencia. Particularmente, se permitieron dos bases de diferencia entre alelos (pila, tags o stacks) para construir un *locus* dentro de un individuo (-M 2) y se permitieron dos bases diferentes entre *loci* para construir el catálogo (-n 2). Se removieron los *stacks* o alelos con gran profundidad de secuencias, ya que es muy probable que provinieran de regiones repetitivas del genoma (-t). Al ser una especie diploide, sólo se consideraron *loci* constituidos por dos *stacks* (-X "ustacks: --max_locus_stacks 2"). Adicionalmente, debido a la gran cantidad de *loci* descubiertos por muestra, y siendo gran parte de ellos exclusivos de unos pocos individuos, fue necesario agregar filtros adicionales para que sea factible el análisis en el servidor del Instituto de Agrobiotecnología y biología molecular del INTA Castelar. Así, se agregaron filtros para obtener SNPs presentes como mínimo en un 50% de los individuos incluidos en el análisis y con una frecuencia alélica mínima (MAF) de 0,01. Para ambos tipos de análisis se implementó un

filtrado final de *loci* por calidad (posibles errores de definición de genotipo) con el módulo *rx_stacks.pl*. En este paso, lo que se modificó con respecto al análisis en las 24 muestras fue el logaritmo de las verosimilitudes de los *loci*, en este caso filtrando valores menores que -20 (menos veinte).

Finalmente, se ejecutó el componente *populations*, como paso final de ambos análisis para exportar los datos de los genotipos en un formato VCF. Luego de evaluar distintos filtros, se decidió utilizar un filtro de *loci* definidos con un mínimo de seis lecturas para considerarlos en la matriz final (-m 6), ya que el llamado de *locus* heterocigotas es más robusto al aumentar la profundidad de lecturas de ddRADseq utilizadas para definir un *locus* (de 3 a 6; Rochette & Catchen, 2017).

Para análisis posteriores se consideró la matriz obtenida (formato VCF) mediante el análisis con la referencia de *E. grandis*, a la cual se la denominará a partir de aquí en adelante como matriz de GBS. Esta decisión se tomó en base a que el análisis con genoma de referencia permite evaluar la distribución de los marcadores en el genoma y referenciar genes cercanos a marcadores asociados. Según los requerimientos de los programas a utilizar posteriormente dicha matriz fue convertida a formato *plink* (archivos .ped y .map), mediante el programa *plink v1.9* (C. C. Chang et al., 2015), según fue necesario.

2.3.2 Genotipificación con datos de EUChip60K

Para la obtención de datos genotípicos con el EUChip60k, se dispusieron 250 ng de ADN de los 308 individuos en 4 placas plásticas de 96 pocillos y se liofilizaron en un equipo Labconco (EE.UU.) durante 24 horas a -50ºC y 5-10 atmósferas. Luego se cubrieron con una tapa adhesiva, se envolvieron con film, y se enviaron al servicio brindado por NEOGEN (EE.UU., © NEOGEN CORPORATION https://genomics.neogen.com/en/plant#eucalyptus), para genotipificar mediante el microarreglo EUChip60K (Silva-Junior et al., 2015).

Una vez recibidos los archivos correspondientes a la genotipificación con el Chip se utilizó el módulo de genotipificación del programa GenomeStudio 2.0 para la designación alélica de los individuos ensayados. El programa necesita de tres archivos para realizar el análisis: (a) el archivo SNP Manifest, en el cual se encuentran listados los SNPs que se encuentran en el chip del ensayo (el conjunto de oligonucleótidos utilizados en el ensayo); (b) los archivos Sample Sheet, correspondiente a la ubicación de las muestras analizadas en cada placa enviada a genotipificar y (c) los archivos ".idat" resultantes de la corrida en el equipo IScan, que son los archivos en donde se encuentran las intensidades de los fluoróforos.

Estos tres archivos fueron cargados en el programa dentro del módulo de genotipificación y se creó un proyecto. Debido a que el Chip fue desarrollado para varias especies del género *Eucalyptus*, se utilizó el archivo de clusterización (.egt) correspondiente a la sección Maidenaria (ie.: *E. globulus*, *E. nitens* y *E. dunnii*)

del género *Eucalyptus*. Este archivo provee una mejor especificidad, sensibilidad y concordancia en la asignación de genotipos heterocigotas.

Luego, el programa recodificó las clases genotípicas (AA, AB y BB) de cada marcador SNP de acuerdo con la suma normalizada de las intensidades de los dos fluoróforos (Cy3 y Cy5).

La calidad de las recodificacioness genotípicas resultantes fue determinada mediante el análisis de dos parámetros que permiten determinar el nivel de calidad de cada genotipo asignado: el *GenTrain Score* y el *GenCall Score*. El primero establece el grado de confiabilidad del SNP (el cual va de 0 a 1 indicando la bondad de los agrupamientos de las clases genotípicas) y, el segundo, indica el grado de confiabilidad del individuo (también va de 0 a 1 indicando como el individuo se ajusta a la agrupación por clase en el conjunto de todos los SNPs analizados). Estos valores de confianza definen el *Call Rate* que identifica la eficacia con la que fue asignado un genotipo (los valores van de 0 a 1 mostrando el porcentaje de SNPs designados a una muestra determinada).

De esta manera, se aplicó un umbral de *Call Rate* de muestra del 90%, siendo consideradas genotipadas con éxito aquellas que superen dicho valor.

Asimismo, se aplicó un filtro técnico o *TECH* (de calidad de SNPs) sugerido por los autores que desarrollaron el Chip (Silva-Junior et al., 2015). Cuando se aplica dicho filtro *TECH*, solo se muestra con datos de genotipificación aquellos SNP que pasaron el filtro. Los SNP de baja calidad se reportan como dato perdido.

La matriz obtenida en este apartado fue elegida para continuar con los análisis subsecuentes, a la cual se la denominará a partir de aquí en adelante como matriz de Chip. Dicha matriz se exportó en archivos formato .ped y .map (formato de programa *plink*) con un módulo de complemento del GenomeStudio 2.0, y fue convertida a formato VCF mediante el programa *plink v1.9* (C. C. Chang et al., 2015), según los requerimientos de los programas a utilizar posteriormente.

2.3.1 Filtrado de matrices de SNPs de GBS y Chip según su calidad

2.3.1.1 FILTRADO SEGÚN EL NÚMERO DE DATOS PERDIDOS

Para analizar la proporción de datos perdidos por matriz en ambas matrices genotípicas de SNPs de 308 individuos, GBS y Chip, se utilizó la opción *--recode A* del programa *plink v1.9* (C. C. Chang et al., 2015). Para poder observar patrones que pudieran indicar algún sesgo o error, se visualizó la distribución de dichos datos faltantes con la función *missmap* del paquete de R Amelia (Honaker et al., 2011).

Se evaluó el porcentaje de datos perdidos por individuo mediante el programa *plink v1.9* (función *--missing*) con el objetivo de eliminar a aquellos individuos que presentaran un porcentaje elevado, mayor al 60% (función

--*mind*). Luego, se definió que se utilizaría hasta un 20% de datos perdidos y un MAF menor a 0,01 por SNP (funciones --*geno* y --*maf*).

2.3.1.2 FILTRO DE INDIVIDUOS POR HETEROCIGOSIS Y RELACIONES DE PARENTESCO

Dado que los SNPs son bialélicos, para evaluar la existencia de posibles contaminaciones entre las muestras en los diferentes pasos de genotipificación, en ambas matrices se calculó la heterocigosis observada por individuo y las relaciones genéticas inesperadas entre los mismos. En esta evaluación se prestó especial atención a aquellos individuos que pudieran presentar heterocigosis extremas y elevadas, fuera del rango de la distribución poblacional, y a aquellos individuos que evidenciaran una relación genética mayor a la esperada. Por lo tanto, fueron eliminados del análisis a aquellos individuos cuya similitud fue superior a la relación padre o madre/hijo y muy cercana a gemelos o clones. Este tipo de heterocigosis elevada y relaciones genéticas estarían vinculadas a algún tipo de contaminación en los pasos de genotipificación, ya que no son las esperadas en una población de polinización abierta constituida por familias de medios hermanos, siendo a lo sumo posible una relación genética esperada de hasta hermanos completos. El cálculo de heterocigosis observada se realizó mediante la opción --*het* del programa *plink 1.9*. El filtro según las relaciones de parentesco se aplicó mediante *plink 2.0* (Chang et al., 2015) que implementa la escala King (comando --*king-cutoff* 0,354), para eliminar aquellos individuos con un grado de similitud mayor al de padres e hijos, similar al de clones (se utilizó una línea de corte de 0,354, siendo el promedio entre relación padre hijos y clones. Escala King: 0,5 clones, 0,25 padres e hijos, 0,125 hermanos completos, 0,0625 medios hermanos).

Aquellos individuos que se eliminaron, en cualquiera de las dos matrices, fueron también eliminados en la otra para que fueran comparables los análisis posteriores entre ambas metodologías genómicas. Este filtrado se realizó mediante la herramienta *VCFtools* (Danecek et al., 2011).

Como una corroboración final del filtrado de las matrices, para cada metodología de genotipificación, se evaluaron las relaciones genéticas entre los individuos mediante una matriz de distancia considerando a los marcadores según la identidad por estado. Estas matrices se calcularon con la opción *snpgdsIBS* del paquete *SNPRelate* de R (Zheng et al., 2012). Así, ambas matrices se correlacionaron con una prueba de Mantel (Mantel, 1967; correlación punto a punto calculada con el coeficiente de correlación de Pearson y 999 permutaciones mediante la opción mantel del paquete vegan de R (Dixon, 2003) para evaluar su grado de significancia).

2.3.2 Imputación y unión de matrices de SNPs

Debido a la existencia de datos perdidos en ambas matrices genotípicas, obtenidas en la población de mejoramiento de *E. dunnii* de Ubajay, se llevó a cabo la imputación de éstas mediante el programa LinkImpute. Este procedimiento se aplicó tanto a las matrices filtradas por DL como no filtradas por esta propiedad. Luego

de la imputación, ambas matrices completas de GBS y Chip fueron fusionadas mediante la herramienta BCFtools (opción: *bcftools concat*; Li et al., 2009), para generar una tercera matriz conjunta de marcadores SNPs (de ahora en adelante denominada Chip-GBS o conjunta).

2.3.1 Análisis de genética de poblaciones a partir de las matrices genómicas (GBS, Chip, GBS-Chip)

2.3.1.1 COMPARACIÓN DE DISTRIBUCIÓN DE SNPS Y MAF EN LAS 3 MATRICES

Con el objetivo de comparar el desempeño de cada una de las metodologías genómicas aplicadas y las características de los SNPs evaluados por ambas conjuntamente, se evaluaron las distribuciones de los SNPs y sus frecuencias alélicas a lo largo del genoma de *E. grandis* mediante la función *summaryGenMap* del paquete de Synbreed de R (Wimmer et al., 2012). Se visualizaron las distribuciones de dichos SNPs, mediante el paquete de R CMplot (https://github.com/YinLiLin/R-CMplot). Se compararon tanto dichas distribuciones como las distribuciones de frecuencias alélicas mínimas de ambas metodologías de genotipificación y su matriz conjunta.

2.3.1.2 CÁLCULO DE DESEQUILIBRIO DE LIGAMIENTO

Para evaluar la magnitud del desequilibrio de ligamiento (DL) presente en la población de *E. dunnii* de Ubajay, y corroborar que sea bajo, como lo esperado para este tipo de poblaciones, se estimó el desequilibrio de ligamiento (DL) presente en cada cromosoma. Para ello, se estimó el promedio general a través el parámetro r^2, mediante la función *pairwiseLD* del paquete de *Synbreed* de R (Wimmer et al., 2012), para las 3 matrices de marcadores por separado (GBS, Chip, GBS-Chip). Por cada cromosoma, se obtuvieron los gráficos de distancia en pares de bases versus el r^2 entre pares de SNPs, mediante el paquete *ggplot2* de R (Wickham, 2016).

Por otro lado, con el objetivo de eliminar información redundante, obtener una estimación más certera de la estructura poblacional y aligerar la demanda computacional de los análisis estadísticos, para cada una de las matrices se eliminaron aquellos SNPs que estando en desequilibrio de ligamiento con otro marcador (r^2 mayor 0,2), poseían menor MAF. Este filtro se implementó mediante la función *--indep-pairwise* del programa *plink 1.9*, evaluando los *loci* presentes en ventanas de 2 Mb y desplazando dichas ventanas con un solapamiento entre ellas de 200 Kpb (10% del total de la ventana). Estas matrices filtradas se utilizaron para estimar los parámetros de diversidad, parentesco y estructura genética del siguiente apartado y para aplicar SG. Para el análisis de GWAS se utilizaron las matrices sin este filtro con el objetivo de tener más marcadores y cobertura en dicho análisis, y como una forma de validar las asociaciones encontradas si marcadores cercanos asocian a la vez a un mismo caracter.

2.3.1.3 ESTIMACIÓN DE PARÁMETROS DE ESTRUCTURA Y DIVERSIDAD GENÉTICA POBLACIONAL

El estudio de la diversidad genética utilizando enfoques multivariados se basa en encontrar variables sintéticas construidas como combinaciones alélicas lineales (es decir, por ejemplo: nueva variable = *a1* (alelo 1) + *a2* (alelo 2), donde a1 y a2 son coeficientes reales), las cuales reflejan lo mejor posible la variación genética entre los individuos. Algunos enfoques, como el análisis de componentes principales (PCA), se centran en estimar dicha diversidad entre individuos, describiendo la diversidad global (Jombart & Collins, 2015). Sin embargo, no describen detalladamente a la variabilidad genética poblacional, que se puede descomponer usando un modelo ANOVA multivariado estándar como: varianza total = (varianza entre grupos) + (varianza dentro de los grupos).

En el presente estudio se aplicó el Análisis Discriminante de Componentes Principales (*Discriminant Analysis of Principal Components* o DAPC; Jombart et al., 2010) del paquete de R *adegenet* 2.0.0 (Jombart, 2008), para evaluar la estructura genética poblacional. Este análisis, por el contrario, busca variables sintéticas, las funciones discriminantes, que muestran las diferencias entre los grupos mientras minimiza la variación dentro de los grupos. El DAPC requiere que se defina el número de grupos previamente. Para ello se transformaron los datos de los SNPs utilizando un PCA y se utilizó el algoritmo de agrupamiento *k-means* que encuentra el número de grupos (k) que maximiza la variación entre los mismos. Se corrieron sucesivos *k-means* con un número creciente de k, mediante la función *find.clusters* (*adegenet*) y se compararon diferentes soluciones de agrupamiento utilizando el Criterio de Información Bayesiano (BIC, de *Bayesian Information Criterion* también denominado Criterio Bayesiano de Schwarz o SBC; Schwarz, 1978). El agrupamiento óptimo fue elegido a través del valor más bajo de BIC.

Para este análisis de estructura poblacional se aplicó un submuestreo al azar de 800 SNPs para cada una de las tres matrices genotípicas filtradas por DL.

Posteriormente, se calculó el F_{ST} entre los grupos genéticos definidos por DAPC, para cada una de las tres matrices, mediante el módulo *populations* (opción: *--fstats*) del programa *Stacks* (Catchen et al., 2013). Para las estadísticas de todo el genoma como F_{ST}, una hipótesis común es si el valor particular en una región genómica es significativamente diferente del promedio de todo el genoma. Un enfoque común para la prueba de hipótesis es mediante el uso de permutación o remuestreo. De este modo, se calculó la significancia de cada valor de F_{ST} mediante el remuestreo *bootstrap* implementado en dicho módulo *populations* (*--fst_correction p_value -k --bootstrap_fst*). Este cálculo se utilizó con los parámetros por defecto, siendo estos un número de remuestreo de 10.000 veces y un p-valor menor a 0,05 para informar un valor de F_{ST} (Catchen et al., 2013).

Por otro lado, las tres matrices filtradas por DL se utilizaron para calcular los siguientes estadísticos de genética de poblaciones: proporciones alélicas *p* y *q*, Heterocigosis Esperada (*He*) y Observada (*Ho*) y el Contenido de información polimórfica (PIC). Se utilizó la función *popgen* del paquete *snpReady* de R (Granato et al., 2018). Dichos estadísticos se calcularon para la población total y para cada grupo genético detectado utilizando DAPC. Las diferencias significativas entre estadísticos a nivel poblacional obtenidos con las tres matrices se evaluaron mediante pruebas t pareadas (valor p <0,05).

4. Aplicación de metodologías genómicas para el mejoramiento molecular de *E. dunnii* mediante Mapeo por Asociación y Selección Genómica

2.4.1 Análisis de Asociación de Genoma Amplio (Genome Wide Association Study)

Para el análisis de GWAS se utilizó la herramienta *Genomic Association and Prediction Integrated Tool* (GAPIT 2.0; Lipka et al., 2012). GAPIT es un paquete bioinformático que opera en R, y utiliza el Modelo linear mixto comprimido (CMLM o *Compressed Mixed Linear Model*; Zhang et al., 2010) aplicando los algoritmos EMMA (*Efficient Mixed Model Association*, Kang et al., 2008) y P3D (*population parameters previously determined*; Zhang et al., 2010). De esta manera, reduce el tiempo computacional al estimar los componentes de la varianza y optimiza el rendimiento estadístico.

Se llevó a cabo un análisis de GWAS con GAPIT para cada matriz (GBS, Chip, Chip-GBS), donde al mismo tiempo el programa calculó e incorporó la matriz de parentesco (K) en cada análisis utilizando el método VanRaden (VanRaden, 2008). Además, se aplicó el nivel de compresión óptimo utilizando el algoritmo de agrupación predeterminado (promedio) y el tipo agrupamiento mediante parentesco (media).

Asimismo, se evaluó el grado de ajuste del modelo de GWAS al considerar la matriz K y además al incorporar al modelo la estructura genética mediante distinto número de las componentes principales (CP). Se evaluó dicho ajuste para cada uno de los caracteres fenotípicos, teniendo en cuenta que el grado de influencia de la estructura genética de la población sobre el análisis de GWAS varía de un rasgo fenotípico a otro. Por este motivo, se aplicó la opción *Model.selection = TRUE* que evalúa el ajuste de diferentes modelos basándose en el BIC para encontrar la cantidad óptima de CP. Se evaluaron entre 0 y 15 CP, donde 0 sólo considera la matriz K en el modelo.

Los gráficos de Manhattan fueron obtenidos con el paquete de R CMPlot (https://github.com/YinLiLin/R-CMplot). Además del umbral FDR brindado por GAPIT, se evaluaron tres umbrales para determinar marcadores asociados: -log (0,05/n) (Bonferroni, n: n° de marcadores), -log(1/n) (Wang et al., 2012; Yang et al., 2013) y -log(1e-4) (umbral *ad hoc*).

El CMLM es un MLM (*Mixed Linear Model* o Modelo Lineal Mixto) que incluye efectos fijos y aleatorios. La inclusión de individuos como efectos aleatorios permite incorporar información sobre las relaciones entre individuos. Esta información sobre las relaciones entre individuos es incorporada a través de la matriz de parentesco (K) basada en marcadores, implementada como la matriz de varianza-covarianza entre los individuos. Para mejorar el poder estadístico, una de las opciones es utilizar la matriz K conjuntamente con la estructura de la población o matriz Q (Structure, Pritchard et. al., 2000; o análisis de componentes principales, Price et al., 2006). Según la notación matricial de Henderson se describe de la siguiente manera:

$$\boldsymbol{Y = X\beta + Zu + e}$$

donde Y es el vector de los fenotipos observados; β es un vector que contiene efectos fijos, que incluyen el marcador genético, la estructura de la población (Q) y el intercepto; "*u"* es un vector de efectos genéticos aditivos aleatorios de QTL para individuos; X y Z son las matrices de diseño conocidas; y "*e"* es el vector de residuos. Se asume que los vectores "*u"* y "*e"* se distribuyen normalmente con una media nula y una varianza de:

$$\boldsymbol{Var}\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{e}\end{pmatrix} = \begin{pmatrix}\boldsymbol{G} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}\end{pmatrix}$$

donde $G = \sigma^2_a K$ siendo σ^2_a la variación genética aditiva y K la matriz de parentesco. Se asume una varianza homogénea para el efecto residual; es decir, $R = \sigma^2_e I$, donde σ^2_e es la varianza residual.

Además, GAPIT utiliza un análisis de agrupamiento para asignar individuos similares en grupos. Los elementos de la matriz de parentesco se utilizan como medidas de similitud en dicho análisis. El método de agrupamiento es por la media aritmética (p. Ej., *Unweighted Pair Group Method with Arithmetic Mean*, UPGMA). Una vez que los individuos se asignan en grupos, se utilizan estadísticas resumidas del parentesco entre y dentro de los grupos como elementos de una matriz de parentesco reducida. Este procedimiento se utiliza para crear una matriz de parentesco reducida para cada nivel de compresión.

En cada análisis de GWAS y por cada SNP se obtuvo un R^2 basado en la razón de probabilidad (R^2_{LR}) que proporciona una medida general del efecto de QTL en el mapeo de asociación con modelos mixtos (Sun et al., 2010). La diferencia entre el R^2 del modelo con el SNP y el R^2 del modelo sin el SNP brindó una aproximación de la variación porcentual explicada por cada SNP.

2.4.2 Genes próximos a los marcadores asociados

Para encontrar posibles genes candidatos, los marcadores asociados a SNP se utilizaron como puntos de referencia para la búsqueda de los genes cercanos que resultaron asociados para los 14 caracteres. Dicha caracterización se realizó en base a la información pública de la secuencia del genoma de *E. grandis* y su anotación (versión 2.0; http://phytozome.jgi.doe.gov, actualizado a febrero de 2016). Para ello, fueron evaluadas ventanas de 70 Kpb (35 Kpb río arriba y 35 Kpb río bajo respecto de la posición del marcador) y su tamaño elegido en base la distancia promedio de los SNPs de la Matriz Chip-GBS luego de que fuera filtrada por DL.

Para ello, se generó un archivo FASTA con las secuencias de las sondas del microarreglo del EUChip60k

(121 pb) y de los *loci* obtenidos a partir de datos de GBS (66 pb), que correspondían a los *loci* que presentaron SNPs asociados en las tres matrices (aquellos que superaron el umbral ad hoc de -log(1e-04)). Dichas secuencias fueron mapeadas con el programa *Bowtie2* (Langmead & Salzberg, 2012) con la opción de alineamiento --*very-sensitive*, y se transformó al archivo obtenido (SAM) a su binario BAM con la herramienta *samtools view* (Li et al., 2009) . Luego, se transformó el BAM al formato BED con la herramienta *Bedtools2* (Quinlan & Hall, 2010). Finalmente, se aplicó un *script* en *Perl* (desarrollado por J. Higgins de The Genome Analysis Centre, TGAC, Inglaterra, *comunicación personal*) diseñado para reportar los genes anotados en el genoma de referencia de *E. grandis* (en archivos en formato .gff3 o *General Feature Format 3*). Así, dicho *script* brinda información acerca de la posición donde mapearon los *loci* de GBS y sondas del Chip, y por cada uno de los genes encontrados, dentro de la ventana que le fue indicada, proporciona el nombre del gen, la posición de comienzo y fin en el genoma en pb, el sentido, y, si posee información, reporta la categoría de Gene Ontology a la que pertenece, y el nombre, símbolo y función de los genes de *Arabidopsis* relacionados.

Para describir los genes encontrados dentro de las ventanas conteniendo los marcadores asociados, se usó la categorización descripta por Myburg y col. (2014). Las categorías principales son: genes relacionados con la síntesis de celulosa y xilanos, genes relacionados con la síntesis de lignina, genes que codifican para peroxidasa y lacasas, genes de síntesis de terpenos, genes de receptores de la familia quinasa (SDRLK), genes que codifican para las familias de factores de transcripción MADS y K-box, entre otros.

2.4.3 Selección Genómica

Al aplicar la prueba de concepto de SG, se compararon tres métodos de evaluación diferentes: un ABLUP convencional y dos métodos genómicos, GBLUP estándar y ssGBLUP. Para cada método, se aplicó un modelo de SG simple con el objetivo de comparar las bondades de cada una de las matrices de SNPs generadas por las metodologías de GBS y Chip, así como la combinación de estas, todas filtradas por DL para eliminar los marcadores redundantes. Los modelos descriptos a continuación se aplicaron con el paquete de R *breedR* (Muñoz & Sánchez, 2014) usando la inferencia de máxima verosimilitud restringida (REML, Patterson & Thompson, 1971). Para cada modelo, se realizaron 10 validaciones cruzadas para todos los rasgos, donde en cada una se utilizó como Población de Validación (PV) un submuestreo al azar del 10% de la población total y el 90% de las muestras restantes como Población de entrenamiento (PE). En la Tabla 2.4.1 se detalla el número de árboles utilizados en la PE y la PV para cada carácter fenotípico y método predictivo. Todos los árboles con datos fenotípicos formaron parte de la PV al menos una vez. Para todos los modelos y caracteres fenotípicos los árboles que presentaron datos genotípicos fueron los mismos (280). Sin embargo, el número de individuos que presentaron ambos datos, fenotípicos y genotípicos, varío según el rasgo (ver Tabla 2.4.1).

Tabla 2.4.1: *Número de individuos utilizados en las Poblaciones de Entrenamiento y Validación en la prueba de concepto de Selección Genómica*. PE: número de árboles en población de entrenamiento y PV: número de árboles en población de validación para los modelos ABLUP-ssGBLUP y GBLUP; Características de crecimiento: dap6, dap11 y dap20: diámetro a la altura del pecho a los 6, 11 y 20 años, at6 y at20: altura total a los 6 y 20 años, for6: forma de fuste a los 6 años; Características de calidad de madera a los 20 años de edad: ir20: índice de rajado, extet20: extractivos etanólicos, exttot20: extractivos totales, klas20: lignina klason, lig20: lignina total, sg20: Siringilo/Guayacilo, cel20: celulosa, db20: densidad básica.

Caracter	$PE_{ABLUP-ssGBLUP}$	PE_{GBLUP}	$PV_{ABLUP-ssGBLUP}$	PV_{GBLUP}
dap6	1458	280	152	28
at6	1458	280	152	28
for6	1458	280	152	28
dap11	316	276	28	28
dap20	318	276	28	28
at20	318	276	28	28
ir20	318	276	28	28
extet20	269	232	28	28
exttot20	269	232	28	28
klas20	269	232	28	28
lig20	269	232	28	28
sg20	269	232	28	28
cel20	269	232	28	28
db20	269	232	28	28

ABLUP

El método convencional ABLUP utilizado, empleó el siguiente modelo mixto de árbol individual para predecir los Valores de Cría (VC) basados en pedigrí, para cada carácter fenotípico de cada árbol:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\alpha} + \boldsymbol{e}$$

donde, el vector $\mathbf{y}$ contiene los datos fenotípicos de la población de Ubajay completa antes de su raleo, tanto de los árboles no genotipados y como de los genotipados (N = 1458), para diámetro a la altura del pecho, altura total y forma de fuste (todos medidos a los 6 años), y en el caso del resto de los 11 caracteres, contiene los datos fenotípicos de los individuos de dicha población que no fueron raleados, y fueron los árboles genotipados (~ 280 individuos, el número varió levemente según el carácter medido, ver tabla 2.1.2); $\boldsymbol{\beta}$ es el vector de efectos fijos para la media general; $\boldsymbol{\alpha}$ es el vector de efectos genéticos aditivos aleatorios de los árboles individuales (es decir, los valores de cría); $\mathbf{X}$ y $\mathbf{Z}$ son matrices de incidencia que relacionan las observaciones ($\mathbf{y}$) con los efectos del modelo $\boldsymbol{\beta}$ y $\boldsymbol{\alpha}$, respectivamente. El vector $\mathbf{e}$ se distribuye como $\mathbf{e} \sim N(0, \mathbf{I}\,\sigma_e^2)$ y σ_e^2 es

la varianza residual. Finalmente, para el enfoque basado en el pedigrí, se asume que el vector $\boldsymbol{\alpha}$ se distribuye como $\boldsymbol{\alpha} \sim N(0, \mathbf{A}\,\sigma_\alpha^2)$ donde σ_α^2 es la varianza genética aditiva y $\mathbf{A}$ es la matriz de relaciones teóricas calculada a partir de la información del pedigrí. La matriz $\mathbf{A}$ se construyó a partir de la información de familias obtenida al momento del establecimiento del ensayo de *E. dunnii* en Ubajay, se asumió que no existía relación de parentesco entre madres.

GBLUP

Para el método GBLUP, en primer lugar se estimó a matriz $\mathbf{G}$ siguiendo el primer método propuesto por VanRaden (VanRaden, 2008):

$$\boldsymbol{G} = \frac{\boldsymbol{WW'}}{\mathbf{2}\sum_{\boldsymbol{k}} \boldsymbol{p_k}\,(\mathbf{1} - \boldsymbol{p_k})}$$

donde, $\mathbf{W}$ es la matriz centrada de las covariables SNP, y $\mathbf{pk}$ es la frecuencia alélica observada de los individuos genotipados para el marcador $\mathbf{k}$. Luego, se aplicó el mismo modelo mixto de árbol individual que para ABLUP con la única diferencia de que la matriz $\mathbf{A}$ del pedigrí fue sustituida por la matriz $\mathbf{G}$ estimada con marcadores. Por lo tanto, se asume que el vector $\boldsymbol{\alpha}$ se distribuye como $\boldsymbol{\alpha} \sim N(0, \mathbf{G}\,\sigma_g^2)$, donde σ_g^2 es la varianza genética aditiva, y $\mathbf{G}$ la matriz estimada por VanRaden.

ssGBLUP

El modelo para el método ssGBLUP también fue el mismo que para ABLUP, excepto que la matriz $\mathbf{A}$ fue sustituida por la matriz combinada de relaciones $\mathbf{H}$, basada en marcadores y pedigrí. Por lo tanto, el vector $\boldsymbol{\alpha}$ se distribuye $\boldsymbol{\alpha} \sim N(0, \mathbf{H}\,\sigma_\alpha^2)$. La inversa de la matriz combinada de relaciones ($\mathbf{H}^{-1}$) fue calculada según Misztal et al. (2009), Legarra et al. (2009), Aguilar et al. (2010) y Christensen & Lund (2010) como:

$$\boldsymbol{H^{-1}} = \boldsymbol{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\lambda}\left(\boldsymbol{G^{-1}} - \boldsymbol{A_{22}}^{\mathbf{-1}}\right) \end{bmatrix}$$

donde, $\boldsymbol{\lambda}$ escala las diferencias entre la información genómica y pedigrí, $\mathbf{A^{-1}}$ es la inversa de la matriz de relaciones en base al pedigrí, $\mathbf{G^{-1}}$ es la inversa de la matriz de relaciones basada en marcadores SNP, y $\mathbf{A_{22}}^{\mathbf{-1}}$ es la inversa de la matriz de relaciones basada en el pedigrí para los individuos genotipados ($\mathbf{A_{22}}$). El factor de ponderación $\boldsymbol{\lambda}$ se estableció en 1 para todos los rasgos según Cappa et al. (2019, 2017).

Un problema potencial del enfoque combinado es que $\mathbf{G}$ y $\mathbf{A_{22}}$ deben expresarse en la misma escala (Meuwissen et al., 2011). Sin embargo, $\mathbf{A_{22}}$ involucra relaciones de los individuos genotipados con respecto a las madres de la población base, y $\mathbf{G}$ corresponde a las relaciones entre los individuos de la población actual.

De este modo, la matriz **G** fue escalada, al igual que Cappa et al. (2019, 2017), según la ecuación 4 del trabajo de Christensen et al. (2012):

$$\boldsymbol{G_C = \beta G + \alpha}$$

donde **Gc** es la matriz **G** escalada y **β** y **α** son parámetros que fueron calculados resolviendo el siguiente sistema de ecuaciones: Promedio(diagonal(G)) β = Promedio(diagonal(A_{22})) y Promedio(G) β + α= Promedio(A_{22}).

Comparación de métodos

Con el objetivo de ver cual es el método que mejor predice los caracteres fenotípicos, los tres modelos se compararon en términos de precisión y sesgo de predicción, para cada una de las tres matrices y para los 14 rasgos fenotípicos.

Para comparar la exactitud de las metodologías ensayadas de ABLUP, GBLUP y ssGBLUP se utilizó la Exactitud Teórica (ET), calculada utilizando la varianza del error en la predicción (*prediction error variance* o PEV), propuesta como el mejor método de validación por Putz et al. (2018). Por lo tanto, la ET de los valores de cría predichos se calculó para todos los rasgos, con los tres métodos, y para cada una de las matrices genotípicas utilizando la siguiente expresión:

$$\boldsymbol{ET = \sqrt{1 - \frac{PEV}{\sigma_\alpha^2(1 + F_i)}}}$$

El **PEV** se calculó como los elementos diagonales de la inversa de la matriz de coeficientes a partir de cada ecuación de los modelos mixtos examinados (Gilmour et al., 1995), y **Fi** son los coeficientes de endogamia del árbol **i**.

Por otro lado, se utilizó un segundo método de comparación de los modelos. Dado que se desconoce el valor de cría "verdadero", los VC se calcularon utilizando la información de pedigrí y todos los árboles disponibles fueron considerados para obtener la mejor estimación del valor de cría de referencia. Luego, se calculó la habilidad predictiva (HP), es decir, la correlación entre el fenotipo observado y estimado, $r(y, \hat{y})$, para cada uno de los árboles disponibles, según Cappa et al. (2019) con la fórmula de Legarra et al. (2008).

Para cada rasgo, se calculó la precisión general como la media de las ET y la media de las HP de todas las iteraciones. Las diferencias significativas en forma pareada entre los diferentes enfoques ABLUP, ssGBLUP

y GBLUP para cada matriz genotípica y ambos métodos de validación (ET y HP), se evaluaron mediante pruebas *t* pareadas (valor p <0.05).

Además, se estimó mediante el modelo ABLUP la heredabilidad en sentido estricto ($\widehat{\boldsymbol{h}}^{\boldsymbol{2}}$) de cada carácter como:

$$\widehat{\boldsymbol{h}}^{\boldsymbol{2}} = \frac{\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\alpha}}^{\boldsymbol{2}}}{\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\alpha}}^{\boldsymbol{2}} + \widehat{\boldsymbol{\sigma}}_{\boldsymbol{e}}^{\boldsymbol{2}}}$$

donde $\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\alpha}}^{\boldsymbol{2}}$ es la varianza genética aditiva estimada y $\widehat{\boldsymbol{\sigma}}_{\boldsymbol{e}}^{\boldsymbol{2}}$ es la varianza estimada del error.

3 RESULTADOS

1. Caracterización Fenotípica de la Población de Mejoramiento de *E. dunnii*

3.1.1 Ajuste de los datos de los caracteres fenotípicos

Los datos fenotípicos disponibles de los catorce caracteres de interés medidos en la población de mejorameinto de *E. dunnii* de Ubajay, se evaluaron para su normalidad y se corrigieron según la necesidad, tal como se menciona en el apartado 2.1.2 de Materiales y Métodos. Los caracteres fenotípicos altura total a los 6 años (at6) y densidad básica a los 20 años (db20) fueron transformados mediante sus valores crudos al cuadrado, y dap20, ir20, extet20, exttot20 (diámetro a la altura del pecho, índice de rajado en rollizo, extractivos etanólicos y totales a los 20 años) fueron transformados con el logaritmo en base diez. Como ejemplo, en la figura 3.1.1 se puede observar la distribución de los datos de ir20 y su correspondiente gráfico cuantil-cuantil o *qqplot*, tanto para los datos crudos como después de corregir por el logaritmo en base diez (antes de estandarizar y eliminar el efecto de bloque). Los gráficos cuantil-cuantil muestran el efecto de la transformación aplicada a los datos de ir20. En la figura 3.1.2 se pueden observar las distribuciones de los datos fenotípicos luego de las correcciones para cada una de las características.

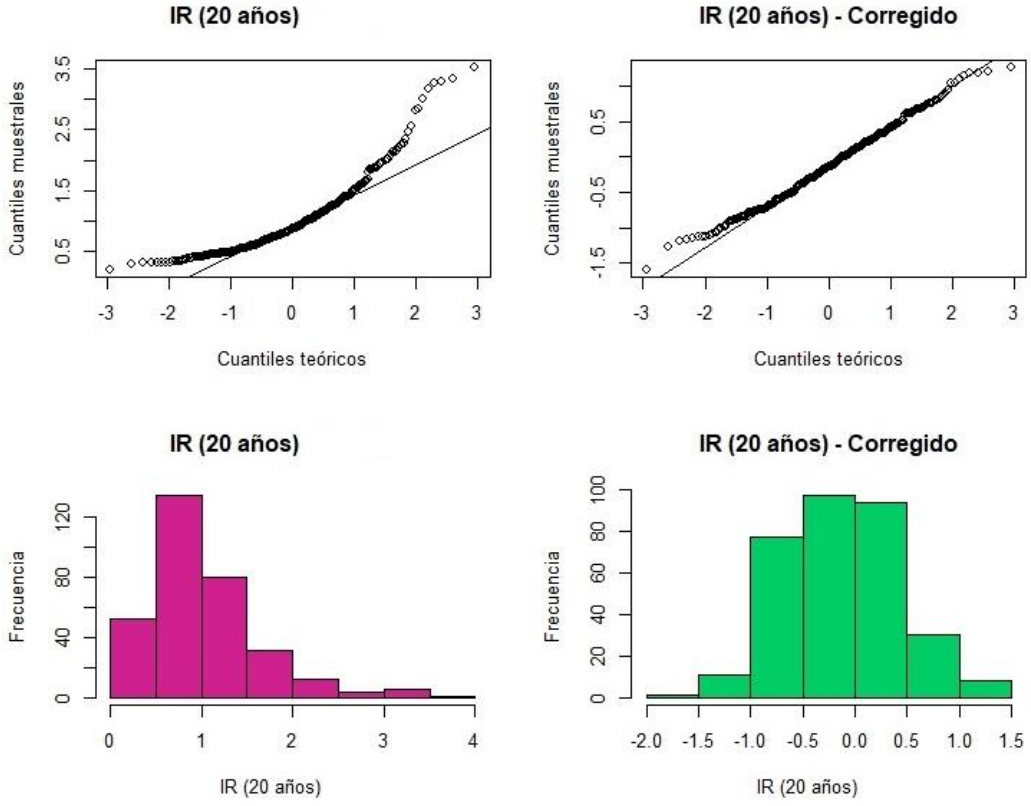


Figura 3.1.1: *Corrección de la distribución de datos fenotípicos del índice de rajado*. por log_{10}. Gráficos de la izquierda: datos sin corregir; Gráficos de la derecha: datos corregidos por log_{10}. Arriba: Gráficos cuantil-cuantil. Abajo: Histogramas de frecuencia graficando la distribución de los datos. IR: índice de rajado de rollizo a los 20 años.

Para conocer el comportamiento de las variables y entre ellas en toda la población, se calcularon las correlaciones entre todas las características. Este conocimiento es de gran utilidad para definir el mejoramiento de estas propiedades, por separado o de manera conjunta, según el comportamiento. A continuación, se describen las correlaciones entre los 14 caracteres estudiados, cuyos valores y significancias se pueden encontrar en la Figura 3.1.2.

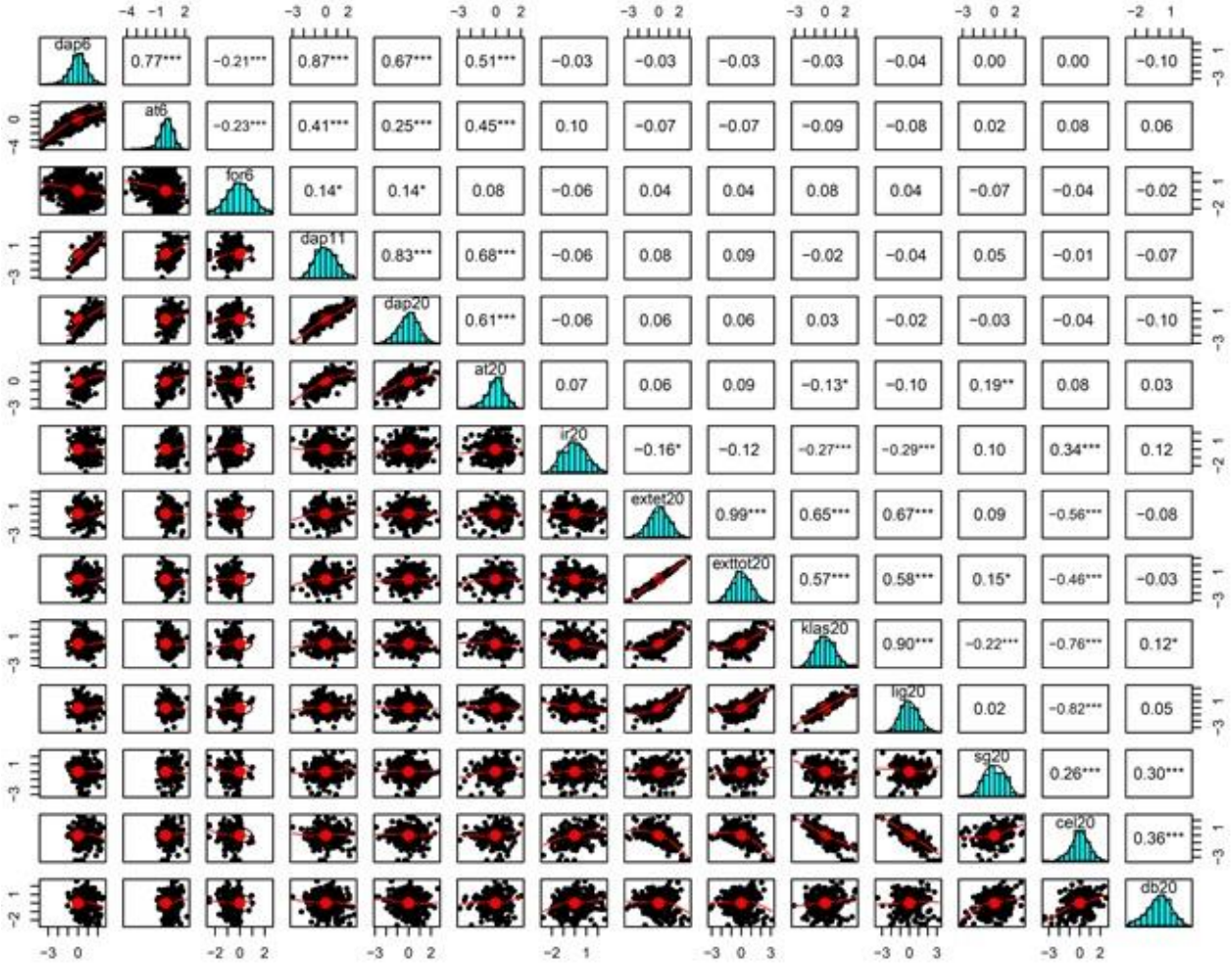


Figura 3.1.2: *Correlaciones de los datos fenotípicos corregidos*. En la diagonal se encuentran los histogramas de distribuciones de los datos. En la parte superior fuera de la diagonal se indican los coeficientes de correlaciones (r^2) con significancias (*) para p-valor < 0,05; (**) para p-valor < 0,01; (***) para p-valor < 0,001. En la parte inferior se observan los gráficos de dispersión de datos entre pares de caracteres fenotípicos. dap6: diámetro a la altura del pecho a los 6 años; for6: Forma de fuste a los 6 años; dap11: diámetro a la altura del pecho a los 11 años; at6: altura total a los 6 años; dap20: diámetro a la altura del pecho a los 20 años; at20: altura total a los 20 años; ir20: índice de rajado de rollizo a los 20 años; extet20: extractivos etanólicos a los 20 años, exttot20: extractivos totales a los 20 años; klas20: lignina klason a los 20 años; lig20: lignina total a los 20 años; sg20: relación siringilo- Guayacilo a los 20 años; cel20: celulosa total a los 20 años; db20: densidad básica a los 20 años.

Entre las correlaciones pareadas para cada uno de los caracteres se puede observar que se presentaron tanto negativas como positivas y muchas de ellas fueron significativas. El rango de valores fue entre -0,82 (entre cel20 y lig20; $P < 0,001$) y 0,99 (entre extet20y exttot20; $P < 0,001$). Esto se corresponde con lo esperado ya que las proporciones entre celulosa y lignina son inversamente proporcionales (a mayor contenido de lignina, se observa menor de celulosa). Por otro lado, la correlación positiva entre ambos extractivos se explica debido a que ambos son mediciones de los compuestos de la madera extraídos con distintos solventes, y extractivos totales incluye a extractivos etanólicos.

Se evidenciaron dos grupos de caracteres que presentan entre la mayoría de ellos correlaciones altas (>0,5) y muy significativas. Uno de los grupos fue el de los caracteres de crecimiento DAP a las distintas edades entre sí; y también at20 versus DAP a todas las edades. El otro gran grupo de correlaciones se evidenció entre los caracteres de calidad de madera, que fueron aquellos seis estimados mediante NIR, ir20 y db20.

Así, para el primer grupo, la mayor correlación fue evidenciada entre el mismo carácter, diámetro a la altura del pecho, pero medido a distintas edades de la población. Entre dap6 y dap11, la misma fue alta, positiva y muy significativa (0,87; $P < 0,001$). Esto concuerda con lo esperado, ya que son mediciones del mismo caracter a través del tiempo. También, la correlación entre dap11 y dap20 fue alta, positiva y significativa (0,83; $P < 0,001$), y levemente menor entre dap6 y dap20, siendo estas fueron mediciones más alejadas en el tiempo (0,67; $P < 0,001$) y podrían ocurrir otros fenómenos como competencia.

Por otra parte, las mediciones de altura total mostraron entre sí una correlación positiva y significativa pero mediana (at6 vs. at20: 0,45; $P < 0,001$). Estos valores sugieren que las mediciones tempranas de altura son útiles para estimar el comportamiento futuro, si bien otros efectos van apareciendo posiblemente de competencia. Sin embargo, no debe dejarse de lado, el error en la medición de la altura del árbol, cuando éstos son muy altos.

Se espera una tendencia a que los ejemplares que presentan mayor diámetro de fuste sean los más altos. De este modo, at6 mostró una alta correlación positiva y muy significativa con dap6 (0,77; $P < 0,001$), pero moderada y baja con dap11 y dap20 (0,41 y 0,25, respectivamente; ambas con $P < 0,001$). Por el contrario, at20 mostró correlaciones altas, positivas y significativas ($P < 0,001$) con todas las mediciones de diámetro a la altura del pecho a distintas edades (0,51, 0,68 y 0,61 para dap6, dap11 y dap20, respectivamente).

Con respecto al caracter de forma de fuste, el mismo evidenció una correlación baja negativa y significativa con los caracteres medidos a la misma edad de dap6 y at6, esto es de esperar, ya que a menor rectitud de fuste los árboles tienden a presentar menor altura. Luego, mostró correlaciones positivas bajas con los DAP medidos en otras edades (0,14; $P < 0,05$), y no presentó correlación con at20.

Dentro del otro gran grupo de correlaciones, correspondiente a los caracteres de calidad de madera, se observó, por ejemplo, un subgrupo con los caracteres de lig20, klas20, extet20 y exttot20 que mostraron correlaciones positivas y altas entre r = 0,57 (para klas20 y exttot20) y r= 0,90 (lig20 y klas20), todas ellas muy significativas (P < 0,001). Dichas correlaciones pueden ser explicadas porque, por ejemplo, lignina Klason es la porción de lignina insoluble en ácido y por lo tanto está contenida dentro de la medición del caracter lignina total.

Asimismo, estos caracteres presentaron correlaciones muy significativas pero negativas con cel20. Así, se evidenció una alta correlación negativa entre cel20 y klas20 (-0,76; P < 0,001), al igual que entre cel20 y lig20, como se mencionó anteriormente. Al mismo tiempo, cel20 presentó correlación negativa con ambos extractivos, alta con extet20 (-0,56; P < 0,001) y mediana con exttot20 (-0,46; P < 0,001), lo que se espera debido a las correlaciones positivas entre ambos tipos de lignina y extractivos.

Respecto a sg20, que es la proporción de monómeros que constituyen a la lignina, si bien mostró una correlación baja, negativa y muy significativa con klas20 (-0,22; P < 0,001), evidenció una correlación despreciable y no significativa con lig20. Asimismo, la correlación con cel20 fue baja pero positiva y muy significativa (0,26; P < 0,001). Además, presentó una correlación baja con exttot20 (0,15; P < 0,05).

El caracter de db20 mostró una correlación positiva, mediana y significativa con cel20 y sg20, y positiva y baja con klas20 (0,12; P < 0,05).

El índice de rajado mostró correlación mediana positiva con cel20 (0,34; P < 0,001), y correlación baja, negativa y significativa con ambas mediciones de lignina (lig20: -0,29; P < 0,001; klas20: -0,27; P < 0,001), y con extet20 (-0,16; P < 0,05). Aunque las correlaciones fueron moderadas o bajas, al seleccionar individuos con menor contenido de celulosa (o mayor contenido de lignina) en la madera, se estarían seleccionando ejemplares con menor índice de rajado. Contar con el conocimiento de estas correlaciones entre caracteres es de gran interés en el mejoramiento genético, ya que permiten tener una idea de que caracteres podrían ser mejorados en forma conjunta simultáneamente y cuáles serían los que se encontrarían afectados en dicha selección.

Así en el caso de *E. dunnii*, uno de sus problemas tecnológicos es el alto índice de rajado de su madera, que hace que no sea apreciada para uso sólido, como en el caso de *E. grandis* u otras especies. Orientar el programa de mejoramiento de *E. dunnii* hacia individuos con menor IR permitiría complementar a *E. grandis* en zonas donde éste último se comporta marginalmente debido a su baja tolerancia a heladas. Sin embargo, esto generaría un detrimento en el contenido de celulosa y un aumento en el de lignina, no deseado para la industria de pasta celulósica.

Para ello, el programa de mejoramiento podría ser orientado en forma paralela hacia la mejora con destino para uso sólido o a la industria de la pasta celulósica. Para esta industria la especie preferida es *E. globulus*, pero su distribución geográfica en Argentina lo acota a zonas costeras del sudeste bonaerense, de gran costo. La selección de *E. dunnii* de alto rendimiento pulpable permitiría una mejora sustantiva de la calidad industrial y su plantación en el norte de la provincia de Buenos Aires, sur de Santa Fe y Entre Ríos, en radios inferiores a los 200 km de la industria manufacturera.

2. Desarrollo de una Metodología de Genotipificación Masiva para *E. dunnii*

3.2.1 Evaluación de enzimas y rango de selección de tamaño

Para garantizar la mejor combinación de enzimas en la generación de fragmentos para ddRADseq, y de acuerdo con las simulaciones *in silico* de la digestión del genoma *E. grandis v2.0*, se detectó que el par de enzimas SphI-MboI generó 2.499.866 fragmentos (Figura 3.2.1a, área gris), de los cuales 248.275 tenían ambas terminaciones de sitios de corte de enzimas (fragmentos tipo AB y BA, datos no mostrados). El par de enzimas PstI-MspI produjo 1.090.783 fragmentos, es decir menos de la mitad que SphI-MboI (Figura 3.2.1b, área gris), y de los cuales, sólo 174.771 contuvieron el par esperado de extremos (AB + BA).

Además, se cuantificaron los fragmentos presentes según la selección (manual o automática) de acuerdo con los tamaños adecuados para la técnica de secuenciación. En la Tabla 3.2.1 se muestran los fragmentos de AB + BA predichos generados para ambas combinaciones de enzimas y diferentes rangos de selección de tamaño (desde 270 a 420 pb para la selección de tamaño manual; y de 285 a 415 pb para la selección de tamaño automatizada).

Tabla 3.2.1. *Simulaciones de cortes enzimáticos in silico del genoma de E. grandis v2.0.* Subpoblaciones de fragmentos obtenidos de acuerdo con el tipo de selección de tamaño manual o automatizada.

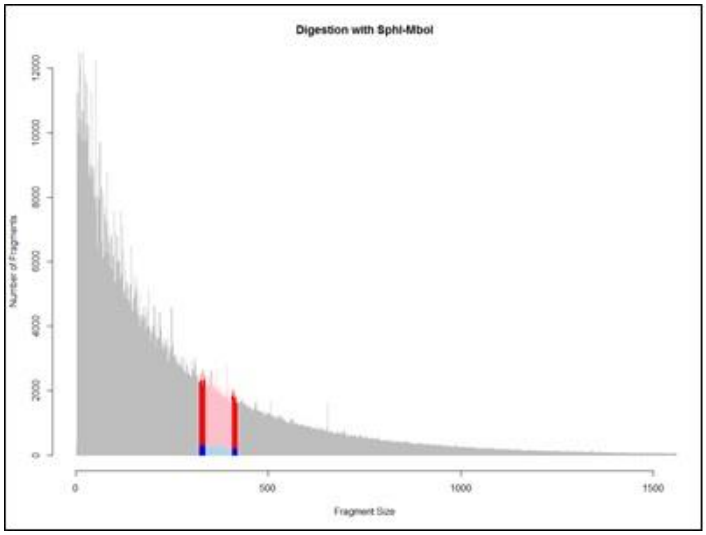
Enzimas	Ventana	Selección Manual				Selección Automatizada	
		Tamaño de inserto-Ventanas de 100 o 150 pb (media)	**Tamaño de fragmentos en gel (Protocolo 1)**	**Tamaño hipotético de fragmentos en gel (Protocolo 2; media)**	**N° de fragmentos predichos (Ventana de 100 o 150 pb)**	**Tamaño de inserto 70 pb (1 o 2 wells)**	**N° de fragmentos predichos (Ventana de 100 o 150 pb)**
SphI-MboI	**100 o 70 pb**	**270-370 (320)**	400-500	350-450 (400)	**28107**	**285 - 355 (1)**	**19184**
		320-420 (370)	450-550	400-500 (450)	**24508**	**335 - 405 (1)**	**17317**
		370-470 (420)	500-600	450-550	**19347**	**385 - 455 (1)**	**12906**
	150 pb	**220-370 (295)**	350-500	300-450	**45655**	**225 - 365 (2)**	**~ selección manual[a]**
		270-420 (345)	400-550	350-500	**39122**	**265 - 415 (2)**	**~ selección manual[a]**
PstI-MspI	**100 o 70 pb**	**270-370 (320)**	400-500	350-450	**13102**	**285 - 355 (1)**	**9137**
		320-420 (370)	450-550	400-500 (450)	**12026**	**335 - 405 (1)**	**8359**
		370-470 (420)	500-600	450-550	**10940**	**385 - 455 (1)**	**7595**
	150 pb	**220-370 (295)**	350-500	300-450	**20749**	**225 - 365 (2)**	**~ selección manual[a]**
		270-420 (345)	400-550	350-500	**18826**	**265 - 415 (2)**	**~ selección manual[a]**

[a] Dado que para seleccionar una fracción de 150 pb usando la Selección de tamaño automatizada en SAGE ELF sería necesario recolectar dos pocillos de elución contiguos de 70 pb cada uno, el número final de fragmentos predichos sería similar para ambos tipos de selecciones, manual y automatizada.

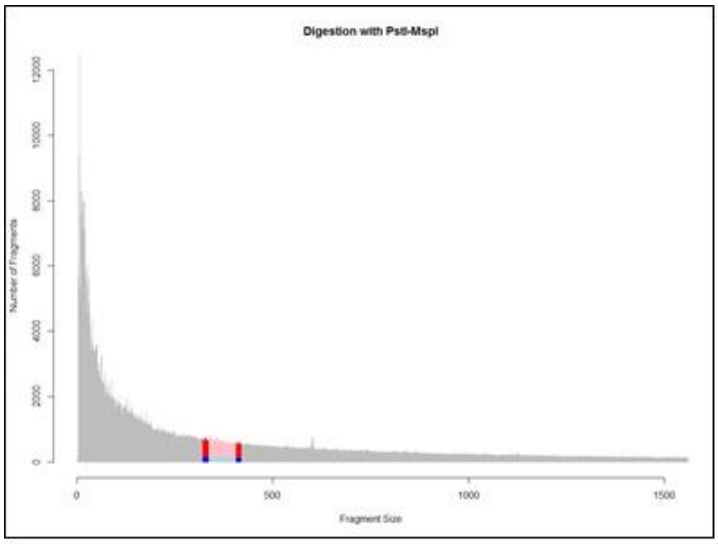
Debido a que la combinación de enzimas SphI-MboI para eucalipto mostró mayor número de fragmentos en las ventanas de 100 y 70 pb, este par fue seleccionado para desarrollar la técnica del genotipado masivo mediante ddRADseq. Sin embargo, se evaluó también el desempeño *in silico* de las otras combinaciones de enzimas posibles y se detallan posteriormente. Específicamente, se eligió un tamaño promedio de población de fragmentos de ADN de 370 pb. Este tamaño de fragmento promedio brinda una superposición mínima entre lecturas de PE de 150 pb (P1). Dentro de dicho rango (320 a 420 pb) se obtuvieron 24.508 fragmentos AB + BA con la selección manual en P1 (en realidad, el tamaño del fragmento de la biblioteca fue de 450 a 550 pb, incluidos los adaptadores y cebadores, Figura 3.2.1a, azul + celeste). Por otro lado, la selección automática recuperó 17.317 fragmentos AB + BA, para el mismo tamaño promedio, pero con un rango entre 335 y 405 pb en P2 (Figura 3.2.1a, área celeste). Considerando una secuenciación PE, el número de *loci* de ddRADseq que se obtuvo fue de 24.508 × 2 = 49.036 para la selección de tamaño manual y de 17.317 × 2 = 34.634 para la selección de tamaño automatizada (ventana de 70 pb, debido a la restricción del equipo) para la combinación de enzimas SphI-MboI.

El par enzimático PstI-MspI recuperó 12.026 fragmentos de ADN AB + BA en una ventana de selección de tamaño manual de 100 pb (entre 320 y 420 pb; Figura 3.2.1b, área azul + celeste), mientras que proporcionó 8.359 fragmentos en una ventana de selección de tamaño automática de 70 pb (entre 335 y 405 pb; Figura 3.2.1b, área celeste). Nuevamente, para la secuenciación de PE, dichos fragmentos correspondieron a 12.026 × 2 = 24.052 *loci* ddRADseq secuenciados predichos para la selección de tamaño manual y 8.359 × 2 = 16.718 *loci* ddRADseq secuenciados predichos para la selección de tamaño automatizada.

Otras combinaciones de selección de tamaño con PstI-MspI produjeron un número de fragmentos predichos similar al de SphI-MspI (24.508). Por ejemplo, la combinación de enzimas PstI-MspI con una selección de tamaño promedio de 345 pb y un ancho de ventana de 150 pb recuperó 18.826 fragmentos predichos, y la misma podría ser aplicada para evaluar otra porción genómica y encontrar nuevos marcadores. Sin embargo, el equipo utilizado para la selección de tamaño (SAGE ELF) brinda un ancho de ventana de 70 pb y permite recuperar varias subpoblaciones de fragmentos de ADN de diferentes rangos al mismo tiempo y, en consecuencia, una selección de tamaño con un ancho de 150 pb solo se puede hacer recolectando dos pocillos de elución o por selección manual. Esto último haría más laborioso el protocolo de ddRADseq, por lo que se eligó el par enzimático SphI-MspI para aplicar a *E. dunnii*.



(**a**)

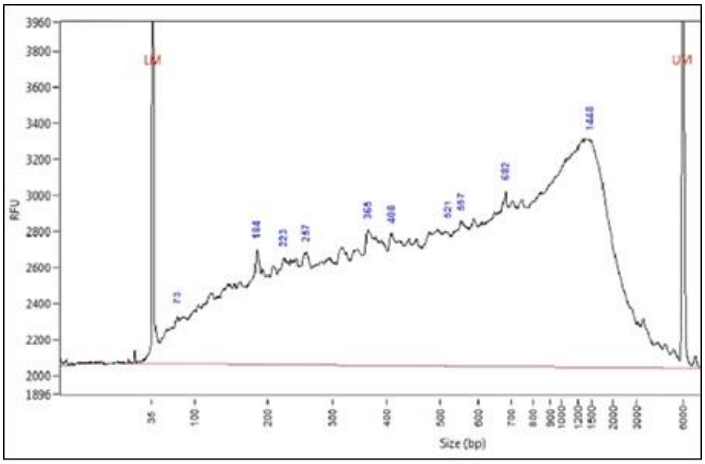


(**b**)

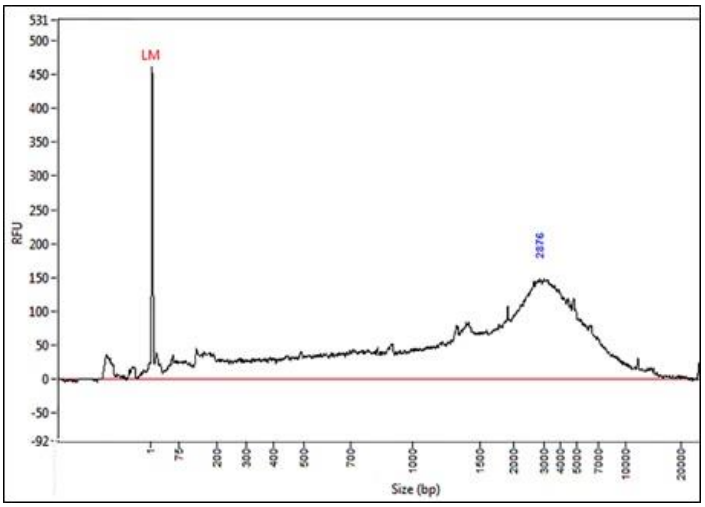
Figura 3.2.1. *Histogramas de simulaciones in silico (frecuencia versus tamaño del fragmento).* (a) Digestión *in silico* con SphI-MboI. (b) Digestión *in silico* con PstI-MspI. Fragmentos totales obtenidos en la digestión (gris), subpoblación de fragmentos obtenidos por selección de tamaño manual (área coloreada total), subpoblación de fragmentos obtenidos por selección automática de tamaño (área rosa + celeste) y subpoblación de fragmentos AB = BA seleccionados, que serían amplificados y secuenciados (manual: azul + celeste; automático: área celeste).

Para validar los resultados hallados por el análisis *in silico* basado en la secuencia del genoma de *E. grandis* y corroborarlo en *E. dunnii*, se realizó el análisis de la digestión *in vitro* para las combinaciones enzimáticas de SphI-MboI y PstI-MspI. En la Figura 3.2.2, se observan los patrones de digestión de ADN obtenidos

mediante una corrida electroforética capilar en el equipo Fragment Analyzer. Este equipo utiliza un intercalante fluorescente para la detección de ADN y otorga gráficos de Unidades de fluorescencia relativa (RFU o *relative fluorescence units*) versus tamaño de fragmentos en pares de bases. Así, este estudio, mostró que la combinación de enzimas SphI-MboI presentó un patrón de fragmentos más homogéneo (Figura 3.2.2a) respecto del otro par de enzimas. De acuerdo con los resultados de las simulaciones *in silico*, estas enzimas dieron frecuencias más altas de fragmentos de menor tamaño dentro del rango de selección que las obtenidas con la combinación PstI-MspI (Figura 3.2.2b).



(a)



(b)

Figura 3.2.2.: *Digestiones in vitro de ADN genómico de E. dunnii.* Análisis en Fragment Analyzer. Eje de abscisas: tamaño de fragmentos de ADN en pares de bases; Eje de ordenadas: Señal de fluorescencia detectada en RFU; (a) SphI-MboI. (b) PstI-MspI.

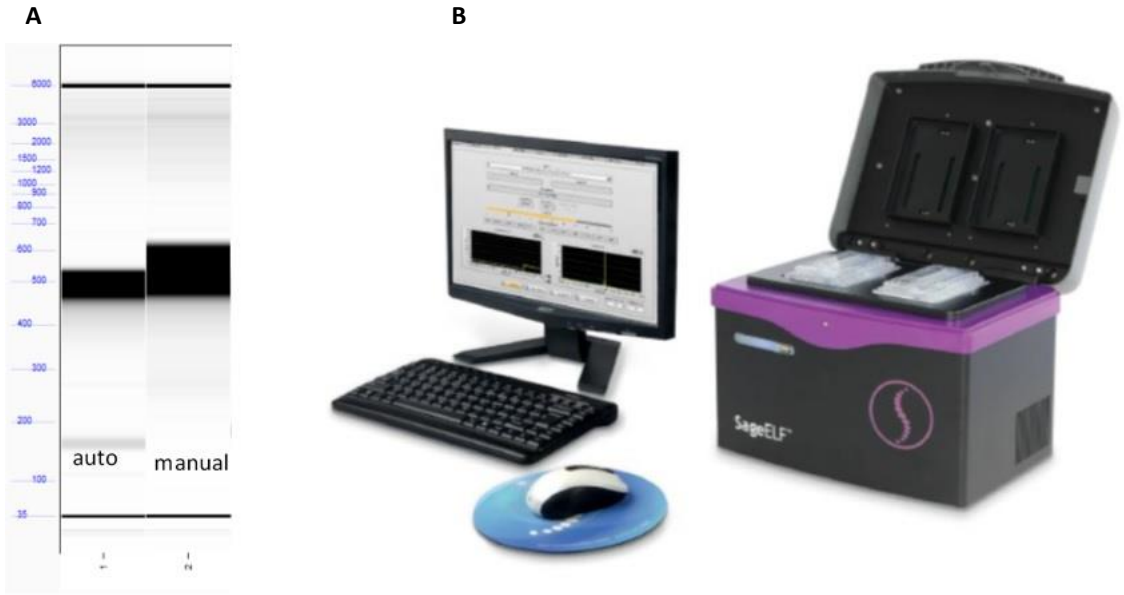


Figura 3.2.3: *Tipos de selección de tamaño de fragmentos de ADN: Automático vs. manual*. A. Comparación de selección de tamaño manual y automatizado (Protocolo 1 y 2), perfiles de genotecas en Fragment Analyzer. B. Equipo SAGE ELF.

3.2.2 Desarrollo del Protocolo 1: análisis de las muestras A y B

A partir de la secuenciación MiSeq, se obtuvieron 1.984.145 y 2.294.900 lecturas PE de 151 pb para las muestras A y B, respectivamente. La calidad general de lectura, de acuerdo con la visualización en FastqC (Andrews, 2010), fue lo suficientemente alta para los análisis posteriores. Al filtrar por calidad con *process_radtag.pl* se obtuvieron muestras que retuvieron más del 96% de las lecturas, promediando unas 2.066.064,5 secuencias.

El 82% de las lecturas pudo ser ubicado en el genoma de referencia de *E. grandis* mediante el uso de Bowtie2 (Langmead & Salzberg, 2012) empleando los parámetros predeterminados como alineador. Se identificaron un total de 77.885 *loci* ddRADseq para la muestra A y 71.395 *loci* ddRADseq para la muestra B, mediante el componente *ref_map.pl* de Stacks. Estos resultados mostraron una cobertura media de 24.16 × de lecturas y se usaron para construir un catálogo de *loci*. Luego de un filtrado posterior por calidad (con el módulo *rxstacks*) se retuvieron 41.834 *loci* ddRADseq. Este resultado fue en el orden de magnitud esperado de acuerdo con la simulación *in silico* (49.016 = 2 *loci* en 24.508 fragmentos con secuenciación PE). Dentro de estos *loci* ddRADseq, 9.299 fueron polimórficos en total (es decir, tenían al menos un SNP, en heterocigosis dentro de cada muestra, y polimórficos entre ellas) y albergaron 19.525 SNP (una media de 2,1 SNPs por

locus). Además, ambas muestras compartieron 7.346 de estos *loci* ddRADseq con 15.792 SNPs (Tabla 3.2.2). Sacando ventaja de la disponibilidad de las secuencias, se buscaron también SSR, detectándose 4.246 en todas las secuencias, de los cuales 420 SSR fueron comunes entre A y B y presentaron diferentes motivos de repetición, siendo 16 de estos SSR polimórficos (Tabla 3.2.2).

Además de la búsqueda de marcadores tomando como referencia al genoma de *E. grandis*, se llevó adelante el análisis sin genoma de referencia, que permitirá encontrar SNPs más específicos de *E. dunnii* o de secuencias que no estén completas del genoma disponible de *E. grandis*. El mismo fue realizado con el módulo *denovo_map.pl* implementado en Stacks (Catchen et al., 2011) y permitió obtener un mayor número de *locus* ddRADseq (aproximadamente el doble: 156.013 y 135.501 para la muestra A y B, respectivamente) y marcadores polimórficos respecto del análisis con referencia. En este caso, el catálogo definitivo conservó 125.432 *loci*. Dentro de estos *loci* ddRADseq *de novo*, 18.951 fueron polimórficos con 33.313 SNP en total (1,8 SNPs promedio por *loci*), así como 1.366 SSR. Finalmente, las muestras compartieron 14.423 *loci*, 25.778 SNPs y 55 SSR polimórficos (Tabla 3.2.2).

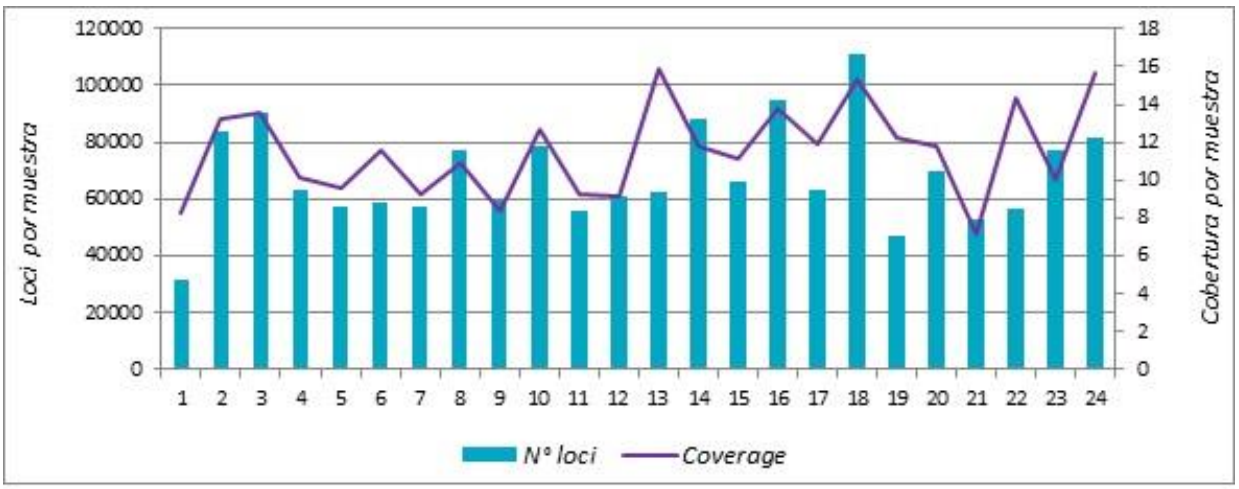
Tabla 3.2.2. *Número de loci ddRADseq y marcadores (SNPs y SSRs) identificados en las muestras A y B.* Con referencia: marcadores descubiertos con el análisis que involucró genoma de referencia de *E. grandis*; *de novo:* marcadores descubiertos sin genoma de referencia. Total: total de marcadores/*loci* descubiertos; Compartido: marcadores/*loci* compartidos entre ambas muestras. SSRs: número total de SSRs descubiertos. SSRs Polim: Número de SSRs polimórficos (incluye SSRs heterocigotas dentro de una sola de las dos muestras).

Análisis	**Total**			**Compartido**			
	SNPs	**loci**	**SSRs**	**SNPs**	**loci**	**SSRs**	**SSRs Polim.**
con referencia	19.525	9299	4246	15,792	7346	420	16
de novo	33.313	18.951	7717	25.778	14.423	1366	55

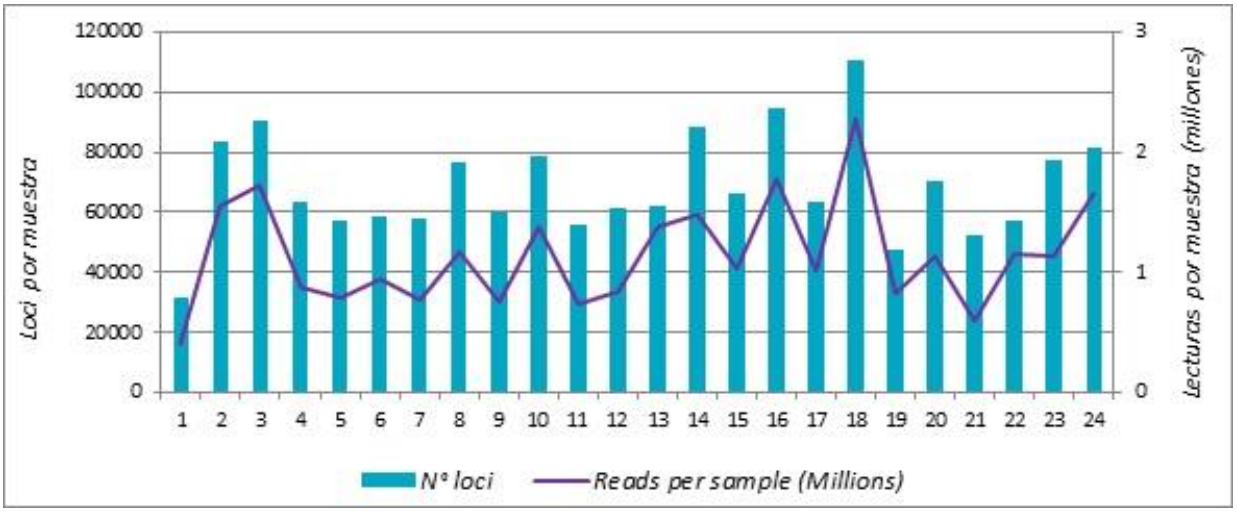
Los motivos dinucleótidos de SSR (AG / GA> AT / TA> TC / CT) fueron los más frecuentes en ambos casos (SSR descubiertos con referencia (16 SSR) o *de novo* (55 SSR)), seguidos de tetra y trinucleótidos (aproximados 15: 5: 1 respectivamente). Al menos 30 SSR fueron polimórficos en estado heterocigosis (20 con análisis *de novo*). De acuerdo con el análisis con referencia, los SSR polimórficos se distribuyeron en todos los cromosomas, excepto en los cromosomas 3 y 9.

3.2.3 Prueba piloto de escalado a 24 muestras (Protocolo 2)

La demultiplexación del grupo de 24 genotecas secuenciadas en la plataforma NextSeq permitió recuperar 27.400.302 lecturas PE de buena calidad, con una media de 1.141.679,25 lecturas PE por muestra. Este número varió de 404.702 para la muestra 1 a 2.280.731 para la muestra 18, con una desviación estándar de 440.542,6 y un coeficiente de variación (CV) de 0,39 (Figura 3.2.4a; Anexo Tabla 7.1.1). De estas lecturas, una media de 82,39% mapeó con éxito contra el genoma de *E. grandis*. El número medio de *loci* ddRADseq por muestra fue de 68.622, duplicando el valor esperado de acuerdo con la predicción *in silico* (34.634). Este número de *loci* también varió entre 31.733 y 110.951 por muestra, un cuarto de las cuales (6) mostraron más de 80.000 *loci* (Figura 3.2.4). Esta variación del número de *loci* por muestra presentó una correlación más alta con el número de lecturas por muestra (r^2: 0,8742) que con la cobertura media por muestra (r^2: 0,3654), siendo la cobertura total promedio de 11,56 × (d.e.: 2.44, Anexo Tabla 7.1.1). Se identificaron 138.624 SNPs en 62.487 *loci* polimórficos. Después de aplicar filtros de MAF 0,05 y 20% de los datos faltantes, se obtuvieron 16.371 SNP distribuidos en 9.466 *loci* ddRADseq, promediando 1,73 SNP por *locus*. De estos SNPs, 15.950 se ubicaron a lo largo de los 11 cromosomas del genoma de *E. grandis* (Figura 3.2.5), mientras que el resto se ubicó en los *scaffolds* y, por lo tanto, no se consideraron en los análisis posteriores.



(**a**)



(**b**)

Figura 3.2.4: *Número de loci en 24 muestras secuenciadas en NextSeq*: Número de *loci* por muestra en comparación con: (a) cobertura (×) media por muestra y (b) número de lecturas por muestra.

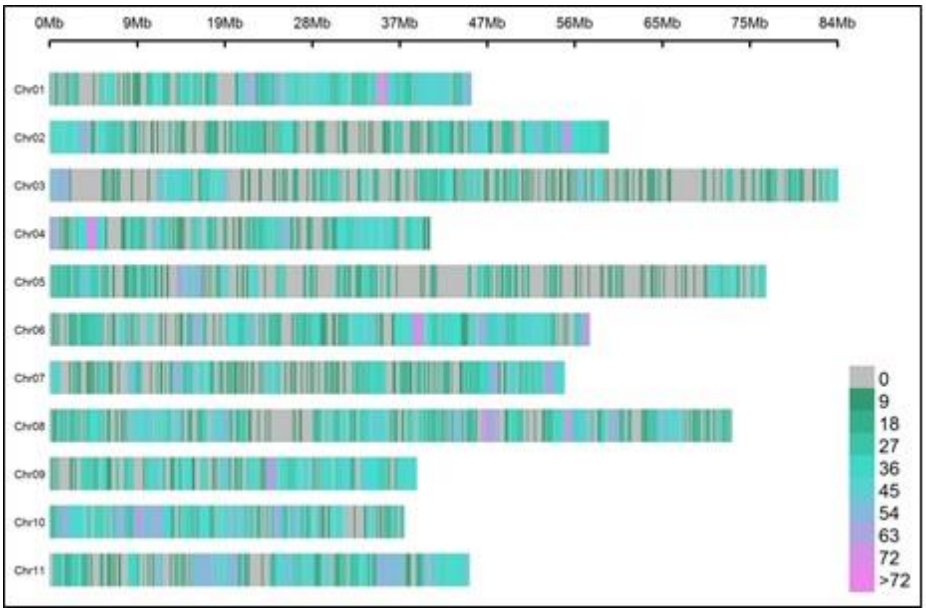


Figura 3.2.5: *Distribución de 15.950 SNP en los 11 cromosomas del genoma de referencia de E. grandis, secuenciados en plataforma NextSeq (ventana de 1 Mb).* Se identifican los cromosomas en filas y las distancias en Mb en las columnas.

3.2.4 Evaluación de la robustez - Comparación de plataformas de secuenciación

Para comparar la robustez de las técnicas de secuenciación según el equipamiento y el potencial de cada uno, las 23 genotecas fueron secuenciadas utilizando MiSeq y NextSeq (la genoteca completa de toda la población se realizó en un NextSeq y el ajuste y puesta a punto se analizó en un MiSeq). El conjunto de 23 bibliotecas, que se secuenciaron usando MiSeq (Illumina Inc.) en baja profundidad, presentó una cobertura media de 4,49 ×, con un rango entre 3,94 y 5,32 ×. Del número total de 138.403 lecturas PE que se obtuvieron por muestra, el 85% mapeó contra el genoma de referencia de *E. grandis*. Posteriormente, las 23 muestras secuenciadas en ambas plataformas, NextSeq y MiSeq, (46 genotecas) rindieron 158.996 SNPs en 294.212 *loci* sin filtrar, con una media de 16.807,63 *loci* por genoteca. Sin embargo, después de los filtrarlos por calidad con el módulo *rxstacks*, y por MAF inferior a 0,05 y 20% de datos perdidos, quedaron un total de 1.051 SNP en 702 *loci* ddRADseq. Estos datos fueron utilizados para construir la matriz de distancia genética y el respectivo dendrograma (Figura 3.2.6). Todas las réplicas de individuos se agruparon de a pares, con un coeficiente de disimilitud inferior a 0,05. Esta mínima diferencia entre réplicas podría explicarse por el 20% de datos perdidos no imputados (principalmente en los datos de MiSeq, debido a la baja cobertura de secuenciación), la tasa de error esperada de la secuenciación y las diferencias entre los equipos de secuenciación. Por otro lado, los medios hermanos (es decir, las muestras de las familias N° 222, 247, 262) tuvieron los coeficientes de disimilitud más bajos (por debajo de 0,17), de acuerdo con las relaciones cercanas esperadas dentro de las familias.

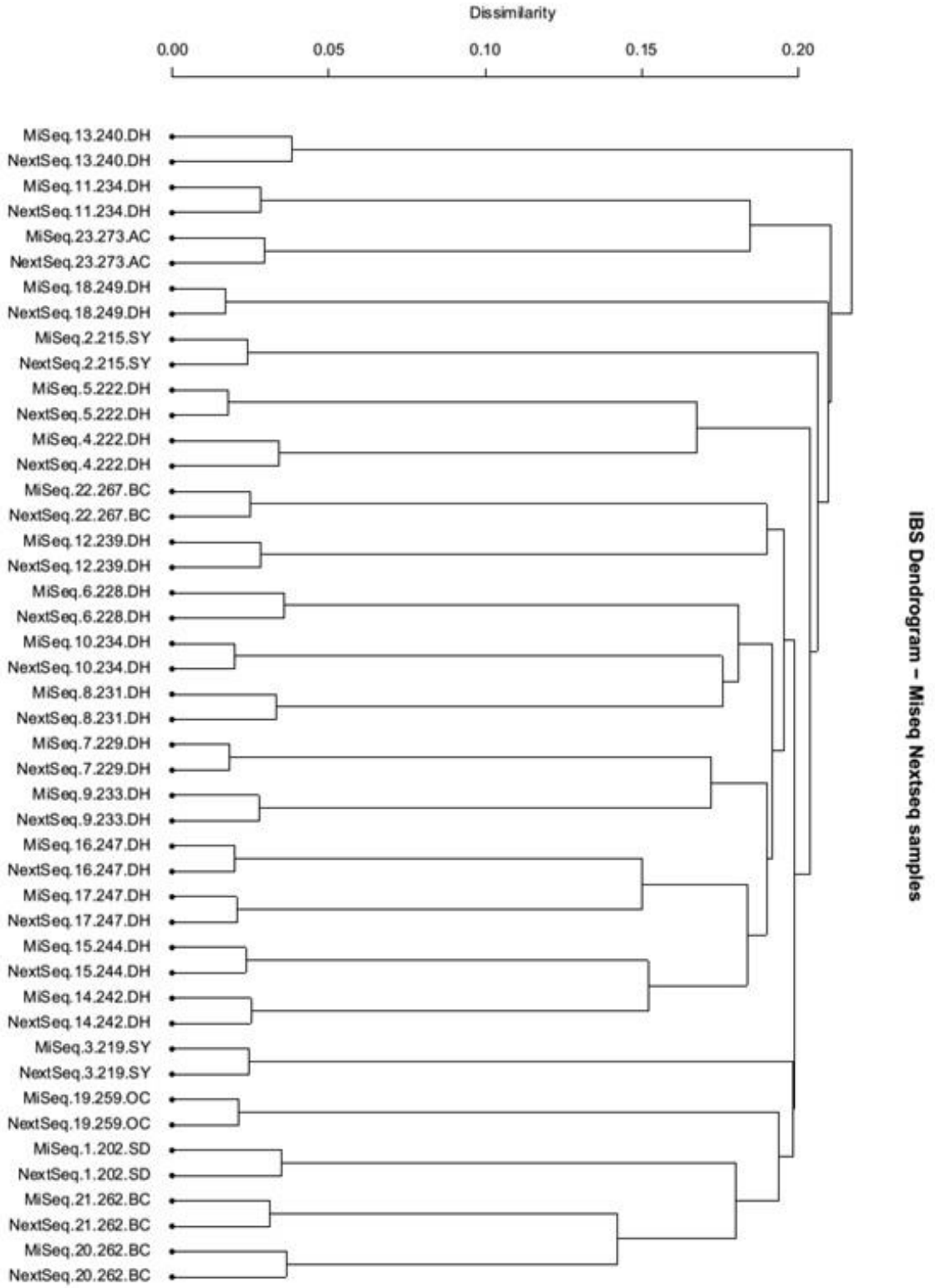


Figura 3.2.6. *Dendrograma para el conjunto de datos MiSeq y NextSeq de 23 muestras de E. dunnii.* Se utilizaron 1.051 SNPs con un MAF del 0,05 y datos perdidos <20% sin imputar. Cada uno de los 23 individuos tiene dos conjuntos de datos SNP de ddRADseq: un conjunto secuenciado en un MiSeq y otro secuenciado en un NextSeq 500. Se utilizó el índice de identidad por descendencia y se utilizó una agrupación jerárquica para el agrupamiento.

3. Caracterización Genotípica de la Población de Mejoramiento de *E. dunnii*

3.3.1 Obtención de datos de ddRADSeq optimizado para población de mejoramiento de E. dunnii

Para la generación de los SNPs mediante ddRADseq/GBS para toda la población de *E. dunnii*, se generaron 13 pooles de genotecas de 24 muestras mediante el protocolo 2. Cada pool o grupo presentó una concentración final entre ~ 5 y 10 ng/µl medida en Qubit 2.0, y en el rango de pesos moleculares esperados (~ 510 pb, Figura 3.3.1), con una leve variación entre ellos. Los perfiles entre pooles fueron similares en tamaño y concentración, sugiriendo que la variabilidad entre muestras no va a ser muy importante, y por lo tanto se espera que haya menos datos perdidos.

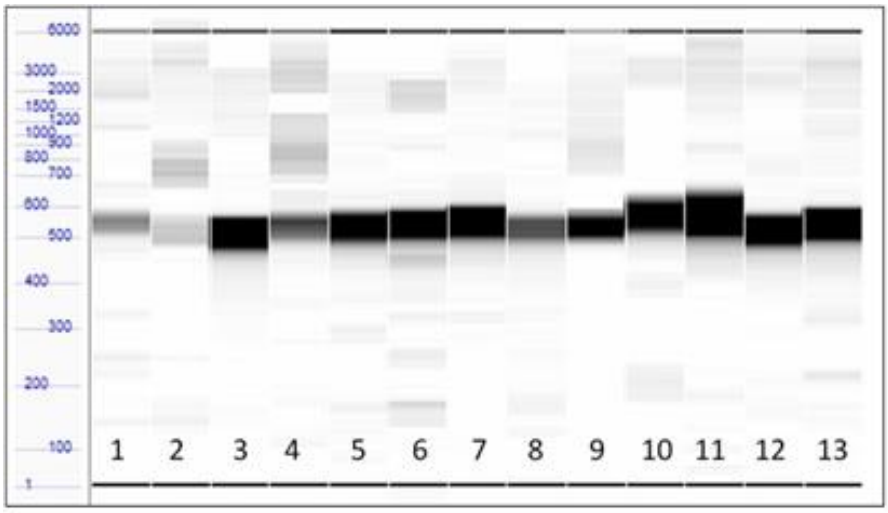


Figura 3.3.1. *Evaluación de calidad de genotecas finales.* Electroforesis capilar en Fragment Analyzer de los pooles del 1 al 13 (uno por calle) de 24 genotecas de ddRADseq de cada uno. Eje de ordenadas: marcador de peso molecular con rango de 1 a 6.000 pb.

En la tabla 3.3.1 se detalla el número de lecturas (secuencias de 75 pb) y la cantidad de Giga-bases (Gb: número de lecturas por 75 pb) esperados según los reactivos de secuenciación de Illumina (*150-cycle high output kit NextSeq*) y el número de lecturas y Gb que fueron obtenidos luego de la secuenciación de las 312 muestras o genotecas contenidas en los 13 pooles (de 24). Además, se especifica el porcentaje de lecturas y Gb recuperados después del filtrado por calidad de secuencias aplicado por el mismo equipo de secuenciación y luego del filtrado por calidad de secuencias específico para datos de ddRADseq, aplicado por el programa Stacks, que deja secuencias con sitio de corte de las enzimas utilizadas, entre otros filtros.

Tabla 3.3.1. *Número de lecturas y Giga bases obtenidos para las genotecas de ddRADSeq de la población de Ubajay*. Lecturas Single End (SE) y Giga bases (Gb) totales. Esperado: esperado según los reactivos de secuenciación de Illumina. Obtenido: obtenido en la secuenciación. Filtrado por calidad NextSeq (PF): número de lecturas y Gb que pasaron el filtrado (PF) de calidad aplicado por NextSeq según índice de Phred. % PF: porcentaje de lecturas o Gb que pasaron dicho filtro. Filtrado Stacks: número de lecturas y Gb que pasaron el filtrado de calidad de Stacks (*process_radtags*). % final: porcentaje final de número de lecturas y Gb que pasaron el filtro, con respecto a los valores obtenidos inicialmente.

150-ciclos de alto rendimiento NextSeq	**Secuenciación de población de *E. dunnii***					
	Esperado	**Obtenido**	**Filtrado por calidad NextSeq (PF)**	**% PF**	**Filtrado Stacks**	**% final**
Lecturas SE totales (millones)	800	823,0	767,0	95,9	629,8	78,6
Gb totales	60	61,7	57,6	95,9	47,2	78,7

La secuenciación de los 13 pooles de genotecas en el equipo NextSeq brindó mayor número de lecturas o Gb neto (823 millones de lecturas SE o 61,7 Gb) que lo esperado según especificaciones de los reactivos de secuenciación comerciales (800 millones de lecturas SE o 60Gb), como se detalla en la Tabla 3.3.1. Luego, el 95,9 % de las lecturas y Gb pasaron el filtrado por calidad realizado en primera instancia por el equipo de secuenciación (lecturas con una media de calidad mayor a 30 de índice de *Phred*, lo que equivale a una probabilidad del 99,9% de que las bases nucleotídicas adjudicadas sean correctas; 767,0 millones de lecturas SE o 57,6 Gb, Tabla 3.3.1). En base a este resultado se puede afirmar que la corrida de secuenciación fue buena, presentando un nivel muy bajo de errores. El promedio de lecturas por *pool* o grupo de muestras fue de 48.448.533 (cercano al esperado de 61,5 millones, 800 millones para 13 *pooles*), y este número varió entre 25.361.852 para el *pool* 2 y 70.681.108 para el *pool* 3 como se puede observar en la figura 3.3.2. Esto se debe a que el *pool* 2 tuvo menor concentración en ng/µl y que se reflejó en una menor cantidad de lecturas con respecto a los otros pooles, a pesar de haber mezclado los 13 pooles en cantidades equimolares (en nM).

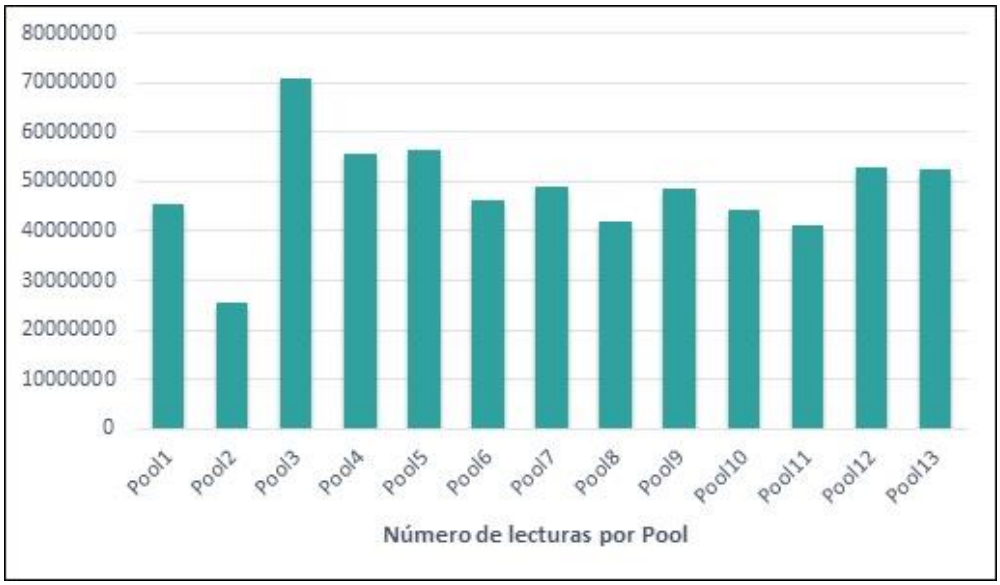


Figura 3.3.2. *Número de lecturas SE por pool de muestras obtenidos para los 13 pooles de genotecas ddRADseq*. En el eje de las abscisas se detallan los grupos o pooles de muestras y en el de las ordenadas se indica el número de lecturas por pool de muestras.

Por último, luego del filtrado por calidad de secuencias y demultiplexado de muestras (separación de los archivos fastq R1 y R2, forward y reverse, de cada pool de muestras, en 24 archivos cada uno) realizado para cada pool con *process_radtags*, el número de lecturas SE total que se obtuvo para las 312 genotecas fue de 629,8 millones, siendo un 78,6 % de las lecturas y Gb iniciales los que superaron esta limpieza (Tabla 3.3.1). Cabe destacar que los secuenciadores NGS de Illumina necesitan que las genotecas a procesar presenten una cierta diversidad nucleotídica para que la calidad de las secuencias sea óptima. Como ya se mencionó anteriormente, las genotecas de representación reducida como las de ddRADseq/GBS o de amplicones, presentan menor diversidad nucleotídica que aquellas obtenidas para resecuenciación de genomas. Por ello, viendo los resultados de este último filtrado, se puede afirmar que la calidad de los pooles o grupos de genotecas estuvo dentro del rango de lo esperado para esta metodología.

En base a estos resultados, luego de todos los filtros por calidad de secuencias, finalmente se obtuvo un promedio de lecturas SE por muestra de 2.018.689, similar al esperado (2,56 millones o 1,28 millones de lecturas PE). No obstante, como se observa en la Figura 3.3.3, se evidenció una gran variación del número de lecturas entre muestras (224.684 a 6.233.016; d.e.: 974903,95). Esta variación es esperada y principalmente debida a la variación del número de lecturas dentro de pooles de muestras, como ya se vió en los resultados de la puesta a punto de la metodología de ddRADSeq en el apartado anterior. Como ejemplo, dentro del pool 1 hubo muestras con un mínimo de 603.538 y un máximo de 3.555.038 de lecturas (promedio: 1.892.247,58; d.e: 689.788). Sin embargo, esta variación también se ve influenciada por la variación lecturas entre pooles.

Esta variación en el número de lecturas por muestra puede llevar a la obtención de un gran porcentaje de datos perdidos.

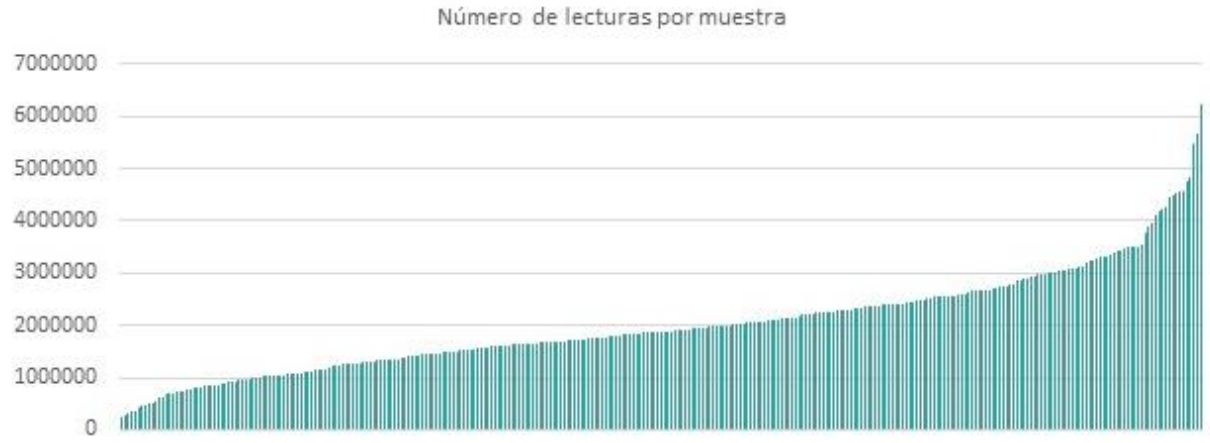


Figura 3.3.3. *Número de lecturas SE por muestras obtenidos para los 312 individuos de E. dunnii mediante las genotecas de GBS (ddRADseq).* En eje de las abscisas se indican las muestras de los árboles y en el de las ordenadas el número de lectura por muestra.

Luego de implementar la búsqueda de *loci* y SNPs en los datos de secuencias de ddRADseq/GBS a través del ensamblado *de novo* mediante el módulo *denovo_map.pl* del software Stacks v1.48 (Catchen et al., 2013), se obtuvieron en total 591.823 SNPs para la población, dentro de 210.957 *loci*, conteniendo en promedio 2,8 SNPs cada 66 pb. Debido a que dicho número de SNPs presentó una gran proporción de datos perdidos, después de aplicar los filtrados de calidad para la definición de *loci* y SNPs (componentes *rx_stacks.pl* y *populations*), se encontraron 55.338 SNPs en 22.629 *loci* polimórficos, recordando que los mismos fueron definidos con 6 reads de profundidad, una verosimilitud mayor a -20 y un MAF de 0,01 y compartidos por el 50% de los individuos de la población de mejoramiento.

Por otra parte, la búsqueda de SNPs en la población utilizando genoma de referencia, mapearon un ~80% de las secuencias contra el genoma de referencia, a partir de las cuales se encontraron 530.885 SNPs en 195.010 *loci*. Esta cantidad de SNPs y *loci* fue un 10 y un 8% menor a la encontrada en el análisis *de novo*, respectivamente. Así, al igual que en el análisis *de novo*, luego de aplicar los filtros de calidad de *loci* y SNPs, quedaron 42.058 SNPs en 16.123 *loci* polimórficos, también cada uno definido por al menos 6 lecturas, considerando sólo los SNPs con una verosimilitud mayor a -20, un MAF mayor de 0,01 y una presencia de estos en al menos el 50 % de los individuos (50% de datos perdidos). Este último número de SNPs y *loci* son un 24 y un 29% menor que el obtenido en el análisis *de novo*. Esto último puede ser debido a que el ~20% de las lecturas que no mapean contra dicha referencia no se utilizaron para la búsqueda de marcadores en el análisis con genoma de referencia.

En base a un trabajo previo con estrategias de imputación para datos ddRADseq (Merino, 2018), se aplicó un filtro a los SNPs dejando sólo aquellos que contenían hasta un 20% de datos perdidos (80% de individuos con datos genotípicos en cada *locus*). Aplicando este criterio, el número de SNPs se redujo a 9.691 SNPs en 4.730 *loci* polimórficos con el análisis *de novo* y a 6.671 en 2.900 *loci* polimórficos con el análisis con genoma de referencia. Cabe destacar que el número de SNPs disminuye drásticamente aplicando filtros que contemplen un máximo de datos perdidos por SNPs, para ambas metodologías de búsqueda de marcadores (con y sin referencia). Esto puede explicarse debido a la variación en el número de secuencias obtenidas por individuo. Por este motivo, y para no perder muchos datos en el filtrado, se decidió evaluar la proporción de datos perdidos por individuo, con el objetivo de eliminar aquellos individuos que presentaran una gran cantidad, previamente a un filtro más restrictivo de los SNPs.

Para los análisis posteriores de filtrado para MA y SG, se eligió utilizar únicamente la matriz de ddRADseq/GBS obtenida utilizando al genoma de referencia de *E. grandis* v2.0, la que presentó 42.058 SNPs en 16.123 regiones o *loci* polimórficas (MAF mayor de 0,01 y un 50% de datos perdidos). De ahora en adelante se hará referencia a esta matriz como la matriz de GBS.

3.3.2 Obtención de datos de EUChip60K para población de mejoramiento de E. dunnii

Luego del análisis realizado para la designación alélica de los individuos utilizando el módulo de genotipificación del programa GenomeStudio, se observó que todas las muestras (308) fueron genotipadas con éxito, superando el umbral del 90% del *Call Rate.* De esta manera, se obtuvo la matriz de Euchip60k que fue luego filtrada según los criterios del apartado siguiente.

3.3.3 Filtrado de matrices de SNPs por calidad

3.3.3.1 FILTRADO SEGÚN CANTIDAD DE DATO PERDIDO POR INDIVIDUO

Matriz de GBS

Con el objetivo de visualizar la presencia de datos perdidos por individuo y por SNP, y considerar la posible exclusión de individuos de la matriz básica de datos de GBS, se realizaron gráficos de distribución de los SNP en relación con los individuos (Figura 3.2.6). Cabe destacar que este tipo de gráficos sólo muestran la distribución de los datos perdidos y no da información sobre si los datos son polimórficos o no. La matriz analizada es la matriz de GBS del apartado 3.3.1, obtenida a partir del análisis con genoma de referencia de las 308 muestras y que presentó 42.058 SNPs (MAF mayor de 0,01 y hasta un 50% de datos perdidos por SNPs).

Se observó que dicha matriz presentó una proporción total de datos genotipados de 0,663 y de datos faltantes de 0,337 (Figura 3.3.4; información calculada con *plink v1.9* y graficada con *Amelia*, como se describe en Materiales y Métodos).

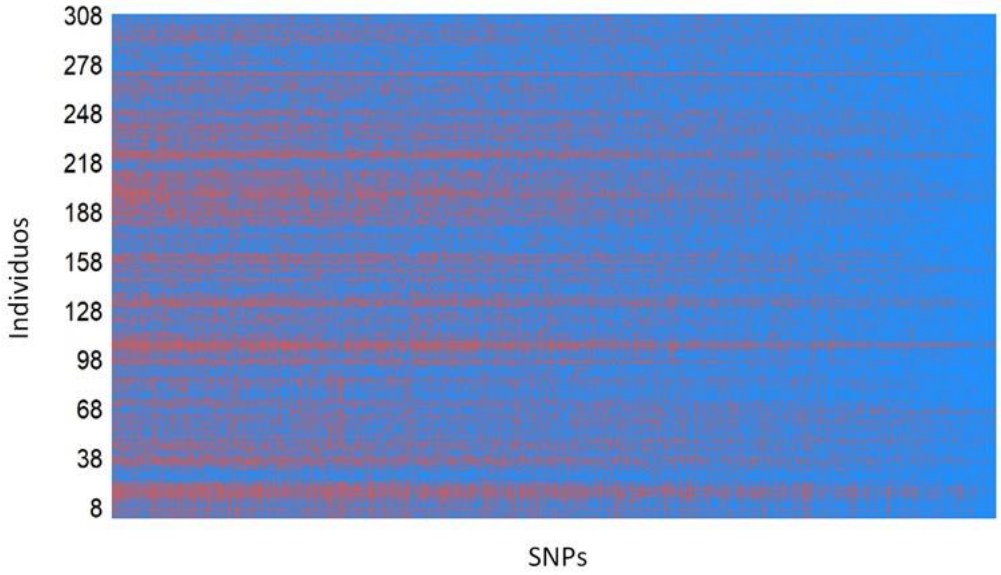


Figura 3.3.4: *Proporción de datos perdidos de la matriz original de GBS*. La matriz corresponde 308 individuos (eje abscisas) genotipados con 42.058 SNPs (eje ordenadas). Rojo: SNPs con datos perdidos (33,7%). Celeste: SNPs informativos, con datos genotipados (66,3%).

Como se puede observar en la Figura 3.3.4, algunas muestras presentaron gran cantidad de datos faltantes (líneas rojas que cruzan el gráfico en forma horizontal). Por el contrario, no se observaron SNPs con gran cantidad de datos perdidos (ausencia de líneas rojas verticales).

La proporción de datos perdidos por individuo varió entre un mínimo de 0,13 a 0,87 con una media de 0,34 (d.e.: 0,15). La distribución de datos perdidos fue graficada en la figura 3.3.5a. A partir de esta información, se decidió eliminar de los análisis posteriores a los 18 individuos que presentaron más del 60% de datos perdidos (*plink v1.9*, Figura 3.3.5b), quedando la matriz compuesta por 290 individuos.

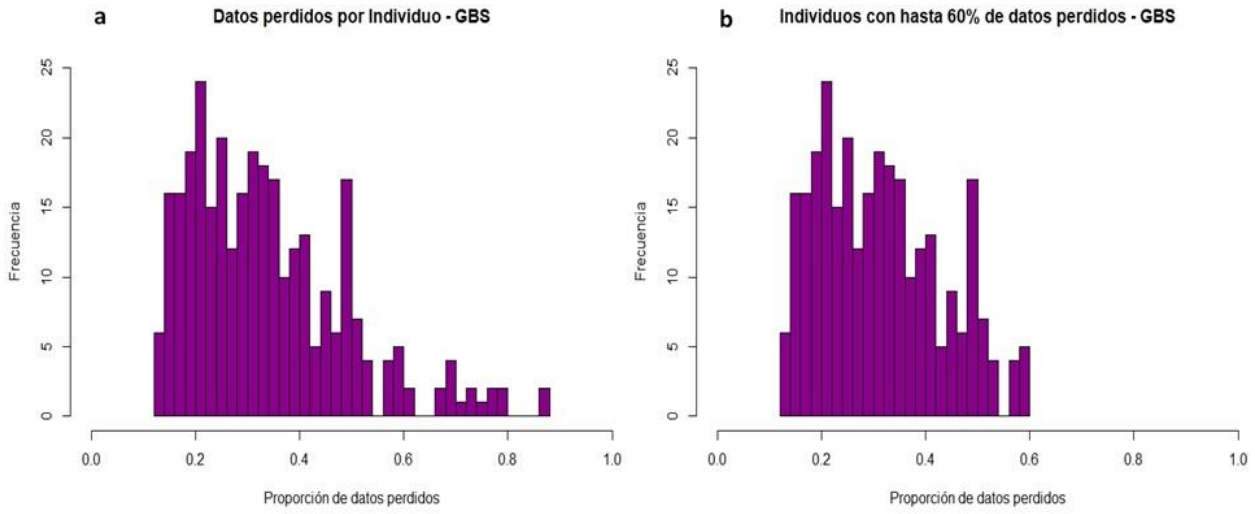


Figura 3.3.5: *Histogramas de frecuencia de la proporción de datos perdidos por individuo genotipado con GBS*. **a.** Matriz de 42.058 SNPs y 308 individuos. **b.** Matriz con 290 individuos, filtrando a los individuos que presentaron más de un 60% de datos perdidos.

Matriz de Chip

Al igual que con la matriz de GBS, y con el objetivo de visualizar la presencia de datos perdidos por individuo y por SNP, y considerar la posible exclusión de individuos de la matriz básica de datos del Chip, se realizaron gráficos de distribución de dichos datos perdidos de los SNP con relación a los individuos (Figura 3.3.6). La matriz analizada es la mencionada en el apartado anterior 3.3.2, con 64.639 SNPs para 308 individuos. Para dicha matriz se observó una proporción total de datos genotipados de 0,89 y de datos faltantes de 0,11 (Figura 3.3.6, información calculada con plink v1.9 y graficada con *Amelia* como se describe en MyM). Cabe mencionar que en este análisis se incluyeron todos los SNPs, sin remover los que tenían menos del 20% de datos perdidos. Como puede observarse, una proporción de SNPs (líneas verticales rojas) presentó datos perdidos. Dichos SNPs son aquellos de baja calidad que fueron reportados como datos perdidos luego de aplicar el filtro técnico sugerido por el grupo de investigación que desarrolló el Chip (Silva-Junior et al., 2015).Sin embargo, al contrario de lo observado para GBS, no se observaron individuos con gran cantidad de datos perdidos (ausencia de líneas rojas horizontales).

Figura 3.3.6: *Proporción de datos perdidos de la matriz original del Chip.* La matriz corresponde 308 individuos (eje abscisas) genotipados con 64.639 SNPs (eje ordenadas). Rojo: SNPs con datos perdidos (11%). Celeste: SNPs informativos, con datos genotipados (89%).

Asimismo, al evaluar la proporción de datos perdidos por individuo en los datos del Chip, la misma varió entre un mínimo de 0,10 a 0,14 con una media de 0,11 (d.e.: 0,006). La distribución de datos perdidos fue graficada en la figura 3.3.7. Dado que no se observaron individuos con gran cantidad de datos perdidos, no se eliminó a ninguno por esta causa.

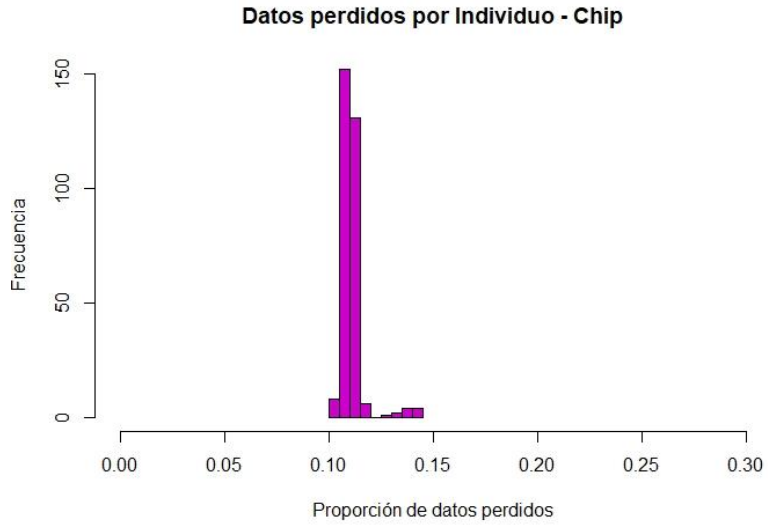


Figura 3.3.7: *Histogramas de frecuencia de la proporción de datos perdidos por individuo genotipado con el Chip*. Matriz de 64.639 SNPs y 308 individuos.

3.3.3.2 Filtrado de individuos según heterocigosis y relaciones de parentesco

Para descartar individuos que hayan sufrido de un error técnico de genotipado (ie, los SNPs por su característica bialélica no permiten distinguir si hubiese ocurrido en algún paso contaminación de muestras,

salvo la observación del exceso de heterocigosis), se evaluó el grado de heterocigosis elevada y un apartamiento severo de la relación de parentesco teórica esperada, como ser del tipo de clones.

Matriz de GBS

Al calcular la heterocigosis observada de los individuos a partir de la matriz del GBS con 290 individuos, los valores variaron entre 0,10 y 0,18, presentando una media de 0,16 (d.e.: 0,01). Al ver la distribución de dicha heterocigosis en la Figura 3.3.8, no se observaron individuos fuera de rango o con heterocigosis extremas, que podrían evidenciar algún tipo de contaminación en el proceso de genotipado.

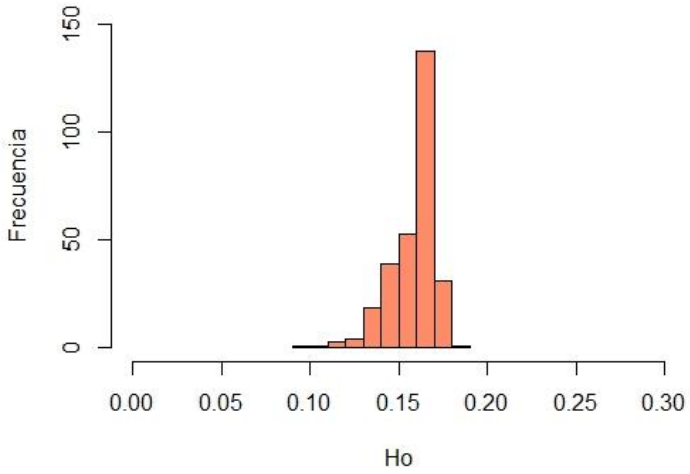


Figura 3.3.8: *Histograma de frecuencias de Heterocigosis observada por individuo genotipado con GBS.* Ho: Heterocigosis observadas. N° de individuos genotipados: 290.

Luego, se evaluaron las relaciones de parentesco considerando el coeficiente King mediante *plink 2.0* (C. C. Chang et al., 2015) para ver si existían pares de individuos con un grado de similitud mayor al de padres e hijos, similar al de clones, y así removerlos. Este análisis permitió evidenciar la presencia de un par de individuos con relación similar a clones (DHAC.246.4 y DHAC.272.17 origen.madre.bloque). Ambos fueron de familias diferentes (familia 246 y 272), lo que refleja algún tipo de error en el proceso de manejo y etiquetado de plantines, en la colecta de hojas o contaminación en el de genotipado, por lo que fueron ambos removidos del procedimiento posterior de análisis.

Matriz de Chip

Al calcular la heterocigosis observada de los individuos a partir de la matriz del Chip, los valores variaron entre 0,21 y 0,58, presentando una media de 0,29 (d.e.: 0,04). Al ver la distribución de dichas heterocigosis en la Figura 3.3.9, se pudo observar que 10 individuos presentaban valores de heterocigosis (entre 0,40 y 0,58)

alejados de la distribución poblacional esperada. Como ya fue mencionado, dichos individuos podrían estar reflejando algún tipo de contaminación al momento de manipulación de las muestras, por lo que se tuvieron en cuenta como candidatos a ser removidos de ambas matrices.

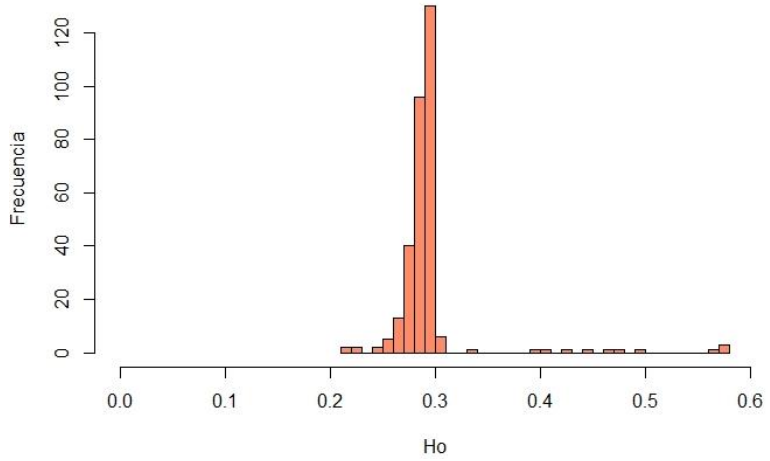


Figura 3.3.9: *Histograma de frecuencias de Heterocigosis observada por individuo genotipado con Chip*. Ho: Heterocigosis observadas. N° de individuos genotipados: 308.

Al seguir evaluando posibles contaminaciones, se evaluaron las relaciones de parentesco y se encontraron 10 individuos con un grado de similitud mayor al de padres e hijos, similar al de clones (Tabla 3.3.2). Como ejemplo, los cuatro individuos DHAC.227.13, DHAC.237.17, DHAC.228.14 y OC.256.18 presentaron relación cercana a la de clones. Dichas similitudes se evidenciaron entre individuos de familias diferentes (familias 227, 237, 228 y 256) e incluso orígenes australianos diferentes (DHAC y OC), lo que refleja algún tipo de error en el proceso de manejo y etiquetado de plantines, en la colecta de hojas, extracción de ADN o contaminación en el genotipado. De este modo, para tomar una decisión final con respecto a que individuos eliminar y no tener en cuenta en los análisis posteriores, se observó que, de los 10 individuos con similitud cercana a clones, 8 coincidieron con aquellos que presentaron heterocigosis observada elevadas. Esto último reafirmó la existencia de algún tipo de contaminación en el proceso de genotipado. A su vez, entre los 2 restantes, uno de ellos fue uno de los integrantes del par que presentó una relación de parentesco similar a clones con la metodología de GBS. Esta evidencia confirmó que, en dicho par, alguno de los individuos sufrió algún tipo de error en el proceso de manejo y etiquetado de plantines, en la colecta de hojas o extracción de ADN (Tabla 3.3.2). Por lo tanto, estos 10 individuos se eliminaron de ambas matrices de datos para que no arrastrar el error a los análisis posteriores.

Asimismo, explorando aún con más profundidad los datos, se observó que los 8 individuos con mayor *Ho* y relación cercana a clones presentaron un mayor porcentaje de datos perdidos. Además, se observó que estos mismos individuos se encontraban en la misma fila en la placa de ADNs liofilizados utilizada en la

genotipificación mediante el microarreglo EUChip60k. Esto confirmó la existencia de algún tipo de contaminación en el proceso de genotipado.

Tabla 3.3.2. *Individuos que presentaron relación de parentesco cercana a clones y heterocigosis elevada*. Individuo: Origen, Familia, bloque; Porcentaje (%) de datos perdidos; *Ho*: Heterocigosis observada; Relación de parentesco ~clon: matriz en la que presentaron relación de parentesco cercana a clones; Filtrado: individuos que fueron finalmente eliminados de ambas matrices.

Individuo	% dato perdido	Ho	Relación de parentesco ~clon	Filtrado
BCUN.263.20	19,46	0,41	----	----
DHAC.228.14	19,29	0,57	Chip	Si
DHAC.237.17	19,18	0,58	Chip	Si
OC.256.18	19,12	0,58	Chip	Si
DHAC.227.13	19,07	0,58	Chip	Si
DHAC.249.19	17,65	0,42	----	----
OC.253.17	17,39	0,50	Chip	Si
BCUN.261.13	16,54	0,40	----	----
OC.254.6	15,14	0,47	Chip	Si
DHAC.243.1	14,98	0,48	Chip	Si
BCUN.267.13	12,94	0,34	----	----
SD.201.11	8,79	0,44	Chip	Si
DHAC.272.17	2,38	0,29	Chip y GBS	Si
DHAC.243.8	2,38	0,29	Chip	Si

3.3.3.3 MATRICES FINALES

Finalmente, como conclusión del apartado anterior, para los análisis posteriores (ej.: GWAS, SG) se eliminaron de ambas matrices originales aquellos 18 individuos que presentaron más del 60% de datos perdidos en la matriz de GBS, y los 10 que presentaron una relación de parentesco cercana a clones en ambas matrices, además de una heterocigosis muy elevada y mayor porcentaje de datos perdidos en la matriz del chip (Filtrado, Tabla 3.3.2). Dicho filtrado se realizó con VCFtools, quedando compuestas ambas matrices de 280 individuos en total. Luego, ambas matrices fueron filtradas por un 20% de datos perdidos por SNP y un MAF de 0,01 mediante *plink 1.9*.

De este modo, la matriz de GBS con 280 individuos con hasta un 20% de datos perdidos por SNP y un MAF de 0,01 presentó 8.170 SNPs y un total de 13% de datos perdidos (Figura 3.3.10). Este número fue mayor al de la matriz de GBS con los mismos filtros obtenida anteriormente desde el programa Stacks (6.671 SNPs, apartado 3.3.1), ya que a esta última no se la había filtrado por aquellos individuos con gran proporción de datos perdidos. Por lo que se concluye que al realizar dichos filtrados posteriores se logró recuperar un gran número de marcadores útiles para los análisis posteriores.

Figura 3.3.10: *Proporción de datos perdidos de la matriz final de GBS*. La matriz corresponde 280 individuos (eje ordenadas) genotipados con 8.170 SNPs (eje abscisas). Rojo: SNPs con datos perdidos (13%). Celeste: SNPs informativos, con datos genotipados (87%).

A su vez, la matriz del Chip de 280 individuos, con hasta 20% de datos perdidos por SNP y un MAF de 0,01, presentó 19.045 SNPs y un total de 3% de datos perdidos (Figura 3.3.11).

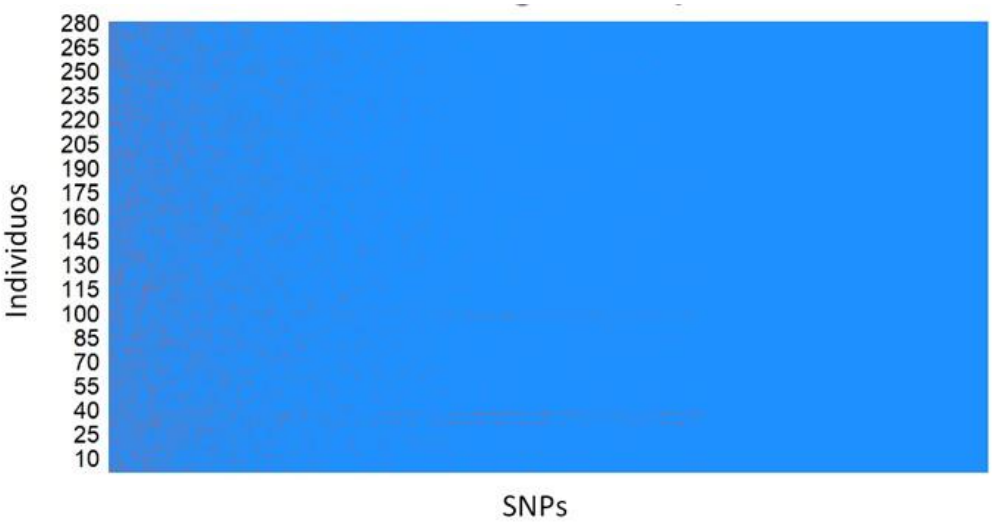


Figura 3.3.11: *Proporción de datos perdidos de la matriz final del Chip*. La matriz corresponde 280 individuos (eje ordenadas) genotipados con 19.045 SNPs (eje abscisas). Rojo: SNPs con datos perdidos (3%). Celeste: SNPs informativos, con datos genotipados (97%).

Finalmente, se calcularon las correlaciones de las distancias genéticas calculadas con cada matriz de datos y se correlacionaron mediante una prueba de Mantel. De esta manera se observó que la correlación fue r = 0,69 (Significancia de 0,001).

3.3.1 Imputación y unión de matrices de SNPs

El programa utilizado para imputar, LinkImpute, hace una estimación de la precisión de la imputación antes de realizar el proceso, submuestreando los datos existentes, borrándolos e imputándolos (Money et al., 2015). Para la matriz completa de GBS (8.170 SNPs) la precisión de la imputación fue de 0,8949, logrando imputar la totalidad de los datos perdidos, que alcanzaban ~13% del total de datos. Del mismo modo, fue imputada la matriz del Chip (19.045 SNPs), que presentaba un 3% de datos perdidos, con una precisión de la imputación de 0,8443.

Finalmente, se unieron ambas matrices con *BCFtools*, creando la matriz de Chip-GBS que presentó un total de 27.215 SNPs.

3.3.1 Análisis de genética de poblaciones a partir de las matrices genómicas (GBS, Chip, GBS-Chip)

3.3.1.1 Comparación de distribución de SNPs y MAF en las 3 matrices

Luego de la imputación, y debido a que la misma puede modificar las frecuencias alélicas de los SNPs, las tres matrices (8.170, 19.045 y 27.215 SNPs para GBS, Chip y Chip-GBS, respectivamente), fueron nuevamente filtradas por MAF, dando un número de SNPs de 8.011, 19.008 y 27.019 para GBS, Chip y Chip-GBS respectivamente (Tabla 3.3.3).

El número promedio de SNPs por cromosoma fue de 712, 1728 y 2.440 para GBS, Chip y Chip-GBS respectivamente, distribuidos a lo largo de los 11 cromosomas (Figura 3.3.12, Tabla 3.3.3). El Chip mostró en promedio 2,4 veces más SNPs por cromosoma que GBS, según lo esperado, ya que su matriz presentó 2,3 veces más SNPs. La matriz de SNPs de Chip-GBS resultó ser la sumatoria de ambas matrices que la componen. El número mínimo de SNPs por cromosoma fue de 572 (cromosoma 9, GBS), 1.106 (cromosoma 1, Chip) y 1.732 (cromosoma 9, Chip-GBS), donde la metodología de GBS y la matriz de Chip-GBS coincidieron en el cromosoma 9 (39 Mb) siendo éste el de menor tamaño del genoma de *E. grandis*. El número máximo de SNPs fue de 1.059 (GBS), 2.518 (Chip) y 3.577 (Chip-GBS), coincidiendo en el cromosoma 8 y siendo éste uno de los cromosomas de mayor longitud de *E. grandis*, como era de esperar. El resto de los cromosomas variaron en orden con respecto al número de SNP entre ambas metodologías y su conjunto (Figura 3.3.12, Tabla 3.3.3), pero siempre reflejaron relación entre la cantidad de marcadores y el tamaño relativo del cromosoma.

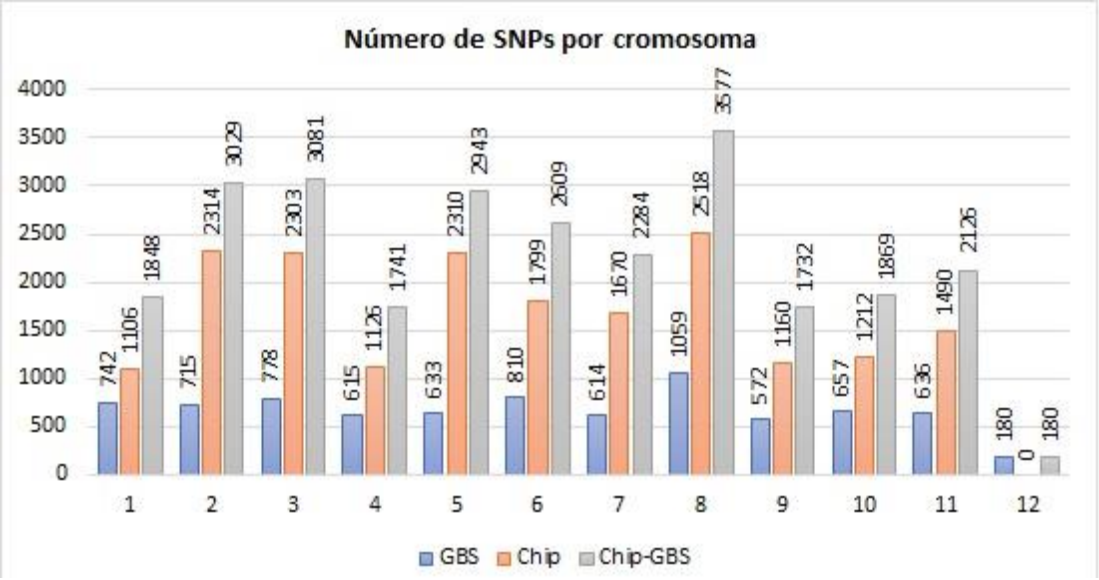


Figura 3.3.12. *Número de SNPs por cromosoma de la población de mejoramiento de E. dunnii en los 11 cromosomas de E. grandis.* 1 al 11: cromosomas; 12: *Scaffolds* del genoma de referencias de *E. grandis* 2.0. Azul: GBS; Naranja: Chip; Gris: Chip-GBS.

Al observar los valores de la distancia entre el primer SNPs y el último en cada cromosoma presentes en la tabla 3.3.3 (Distancia entre SNPs extremos (Mb)), se puede ver que GBS presenta las distancias máximas en los cromosomas 1, 3, 5, 6 y 7, y en el resto de los cromosomas, estas distancias corresponden a SNPs del Chip. Al comparar ambas matrices con la conjunta, esta última mostró en algunos casos distancias entre SNPs extremos levemente mayores a las observadas en ambas matrices por separado. Esto se puede observar en los promedios de dichas distancias, donde para GBS fue de 55,12, para Chip fue de 54,94 y para Chip-GBS de 56,34. Esto último sugiere que ambas metodologías presentan patrones complementarios, donde una cubre regiones no cubiertas por la otra. Además, respecto a la sumatoria de las distancias de los SNPs estremos de todos los cromosomas (total), se observa que los números se acercan al tamaño del genoma de *E. grandis* de 640 Mb, siendo la matriz que presenta mayor cobertura la de Chip-GBS, con un valor de 619,71 Mb. A pesar de que la estimación del tamaño del genoma de *E. dunnii* es de ~530 Mb, los marcadores polimórficos obtenidos en el presente trabajo se encuentran distribuidos ampliamente respecto del genoma de *E. grandis*, como se puede observar en la figura 3.3.13.

Asimismo, al ver las distancias promedio entre SNPs contiguos (tabla 3.3.3), las mayores distancias las presentó la matriz de GBS (promedio: 75.831 pb, variando entre 56.949 y 119.976 pb), como se esperaba, debido a que esta matriz presenta menor número total de SNPs. La matriz del Chip mostró en promedio menos de la mitad de esta distancia (promedio: 32.315 pb, variando entre 27.725 y 37.199 pb) y la matriz conjunta mostró en promedio una distancia de menos de un tercio que la de GBS (23.079 pb, variando entre 20.705 y 27.113 pb). Al ver que la distancia promedio entre SNPs de la matriz conjunta fue menor que las de las matrices

que la conforman, se puede deducir que los SNPs de GBS y Chip presentan un patrón intercalado en la matriz conjunta, lo que se confirma con la inspección visual de la dicha matriz ordenada por cromosoma y posición. Estos últimos resultados dan una idea de que ambas metodologías genómicas por separado presentan marcadores genotipados en distintas regiones del genoma.

Tabla 3.3.3. *Número de SNPs y distancias entre los mismos para las 3 matrices genotípicas.* Crom.: Cromosomas del 1 al 11 del genoma de *E. grandis*, el cero corresponde a los *Scaffolds* que no pudieron ser ensamblados a ningún cromosoma en la referencia; N° de SNP: Número de SNPs por cromosoma; Distancia entre SNPs extremos (Mb): Distancia entre el primer SNPs y el último en cada cromosoma, en Mega pares de bases; Distancia promedio entre SNPs contiguos en pares de bases; Prom.: promedio de las medias de todos los cromosomas; Máx.: Máximas de las medias de todos los cromosomas; Mín.: Mínimas de las medias de todos los cromosomas; D.e.: Desvío estándar; Total: sumatoria de todos los cromosomas.

	GBS			**Chip**			**Chip-GBS**		
Crom.	**N° de SNPs**	**Distancia entre SNPs extremos (Mb)**	**Distancia promedio entre SNPs contiguos**	**N° de SNPs**	**Distancia entre SNPs extremos (Mb)**	**Distancia promedio entre SNPs contiguos**	**N° de SNPs**	**Distancia entre SNPs extremos (Mb)**	**Distancia promedio entre SNPs contiguos**
0	180			-	-	-	180		
1	742	**44,72**	60.355	1.106	40,22	36.403	1.848	44,78	24.245
2	715	59,17	82.866	2.314	**64,13**	27.725	3.029	64,13	21.179
3	778	**83,51**	107.472	2.303	79,70	34.621	3.081	83,51	27.113
4	615	40,02	65.187	1.126	**41,85**	37.199	1.741	41,85	24.051
5	633	**75,82**	119.976	2.310	74,69	32.348	2.943	75,82	25.773
6	810	**57,26**	70.779	1.799	53,75	29.892	2.609	57,36	21.993
7	614	**54,48**	88.869	1.670	52,29	31.332	2.284	54,59	23.914
8	1.059	71,73	67.799	2.518	**74,04**	29.416	3.577	74,04	20.705
9	572	38,30	67.071	1.160	**39,00**	33.650	1.732	39,00	22.531
10	657	37,36	56.949	1.212	**39,24**	32.406	1.869	39,25	21.010
11	636	43,97	69.242	1.490	**45,38**	30.475	2.126	45,38	21.354
Prom.	712	**55,12**	75.831	1.728	**54,94**	32.315	2.440	**56,34**	23.079
Máx.	1.059	83,51	119.976	2.518	79,70	37.199	3.577	83,51	27.113
Mín.	572	37,36	56.949	1.106	39,00	27.725	1.732	39,00	20.705
D.e.	137	16,09	20.108	551	15,62	2.956	642	15,98	2.110
Total	8.011	**606,62**	/	19.008	**604,29**	/	27.019	**619,71**	/

Respecto a la visualización de las distribuciones de los SNPs a lo largo del genoma de *E. grandis* (paquete de R CMplot), se observó que ambas metodologías presentaron una distribución amplia en el genoma, pero que los SNPs de GBS mostraron un patrón más disperso, como se esperaba, debido a las características de la metodología y, como ya se mencionó, a que el número de SNPs fue menor que la mitad con respecto a la cantidad de SNPs evaluados por el Chip (Figura 3.3.13).

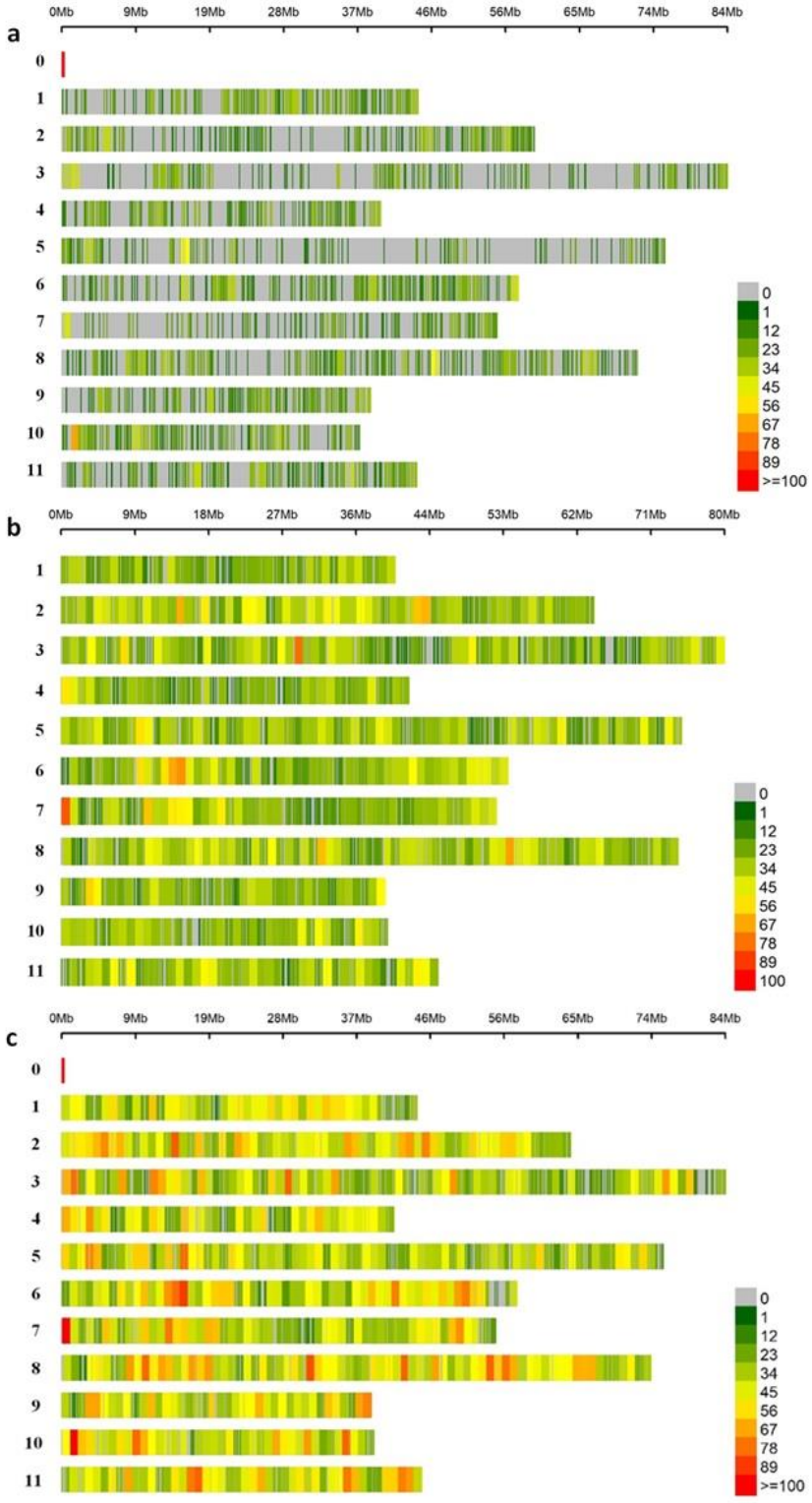


Figura 3.3.13. *Distribución de SNPs de E. dunnii en los 11 cromosomas de E. grandis*. SNPs en ventanas de 1Mb. Cromosomas del 1 al 11 graficados en horizontal, el 0 corresponde a los *Scaffolds* del genoma de referencia de *E. grandis*, ausentes en matriz de Chip. **a**. Distribución de SNPs de GBS (8.011); **b**. Distribución de SNPs de Chip (19.008); **c**. Distribución de SNPs de Chip-GBS (27.019).

Para analizar los MAF obtenidos por cada metodología, se graficaron sus distribuciones. Al ver el número de SNPs según sus distintos valores de MAF en la figura 3.3.14, GBS presentó una gran proporción de alelos con un MAF menor a 0,1. Por otra parte, los SNPs del Chip tienden a presentar una mayor proporción de MAF con valores mayores a los que fueron descubiertos por GBS (Figura 3.3.14). La distribución de los MAF de los SNPs de la matriz de Chip-GBS mostró un patrón intermedio al que evidenciaron ambas matrices que la componen.

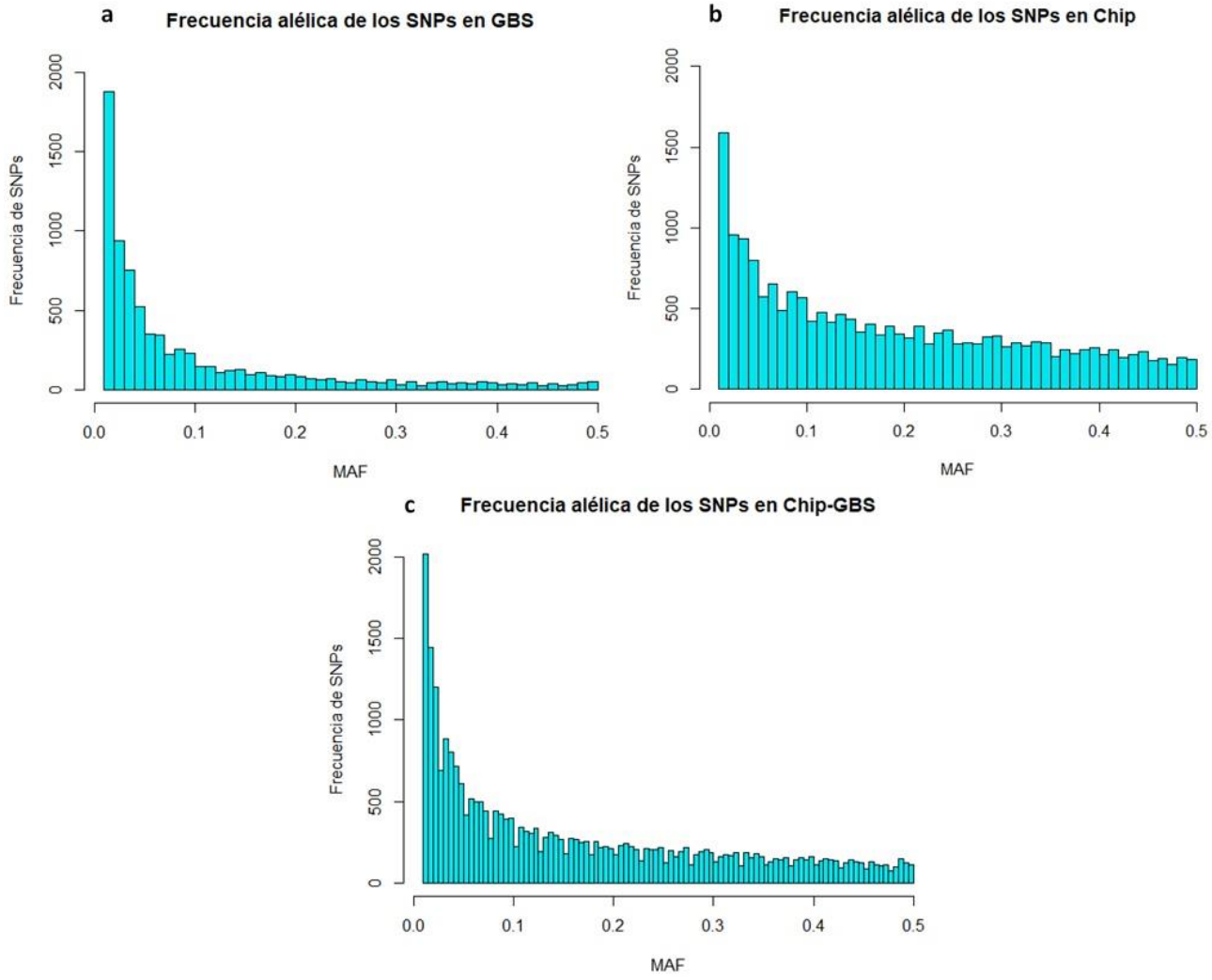


Figura 3.3.14. *Distribución de frecuencias alélicas de los SNPs.* **a**: GBS; **b**: Chip; **c**: Chip-GBS. Matrices con 20% de datos perdidos, filtro de MAF 0,01 e imputadas mediante el algoritmo *LD-kNNi*.

3.3.1.2 CÁLCULO DE DESEQUILIBRIO DE LIGAMIENTO

El desequilibrio de ligamiento (r^2) promedio para todos los cromosomas fue bajo y similar para las tres matrices, siendo de un r^2 de 0,0073 para GBS, 0,0074 para el Chip y 0,0071 para Chip-GBS. Se observó un menor valor promedio de DL en la matriz conjunta explicado por la mayor cantidad de SNPs presentes en dicho análisis. El valor intermedio fue para la matriz de GBS, que a pesar de ser la de menor densidad de marcadores, esta matriz presenta algunos marcadores dentro del mismo *locus*.

En la figura 3.3.15 se observa que el cromosoma 3 mostró el menor DL para las tres matrices (r^2: 0,0068, 0,0067 y 0,0065, para GBS, Chip y Chip-GBS, respectivamente), y coincide con que dicho cromosoma presenta la mayor longitud (83,5 Mb) en el genoma de *E. grandis*. En la figura 3.3.16 se grafican los valores del r^2 a lo largo del cromosoma tres, calculado con las tres matrices. Con respecto al mayor DL, este lo presentó el cromosoma cuatro para GBS (r^2: 0,0082), y el nueve para Chip y Chip-GBS (r^2: 0,0085 y 0,0080, respectivamente). Esta coincidencia en las últimas dos matrices se explica porque el cromosoma nueve es el de menor longitud (39 Mb) de *E. grandis*. El gráfico de los valores del r^2 a lo largo del cromosoma nueve para las tres matrices se puede observar en la figura 3.3.16. A pesar de que para GBS el cromosoma cuatro fue el que presentó el mayor DL, este fue seguido por el cromosoma nueve (r^2 medio= 0,0081). Sin embargo, el cromosoma cuatro mostró un valor intermedio al calcular el DL mediante la matriz del Chip (0,0075), que podría ser explicado por las diferencias en la designación de SNPs entre las metodologías de genotipificación.

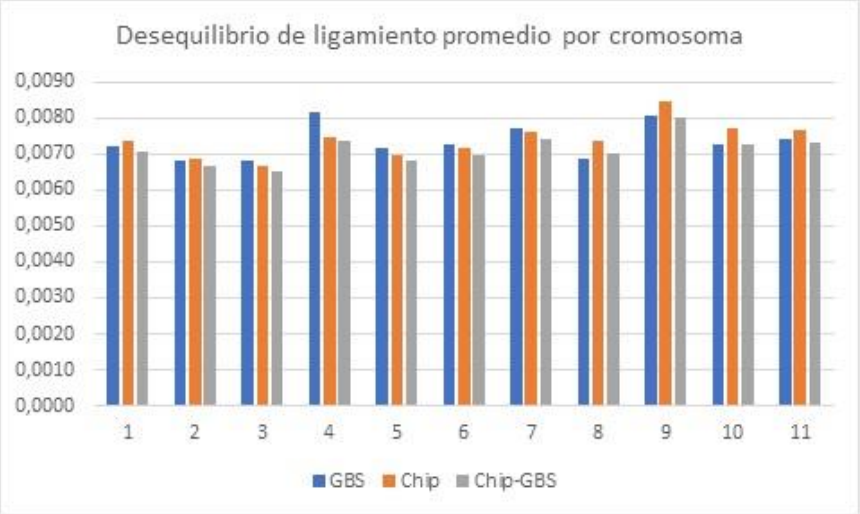


Figura 3.3.15. *Desequilibrio de ligamiento de E. dunnii en los 11 cromosomas de E. grandis*. Eje abscisas: cromosomas del 1 al 11; Eje ordenadas: r^2 promedio por cromosoma. Referencia: Azul: matriz de GBS; Naranja: matriz de Chip; Gris: matriz de Chip-GBS.

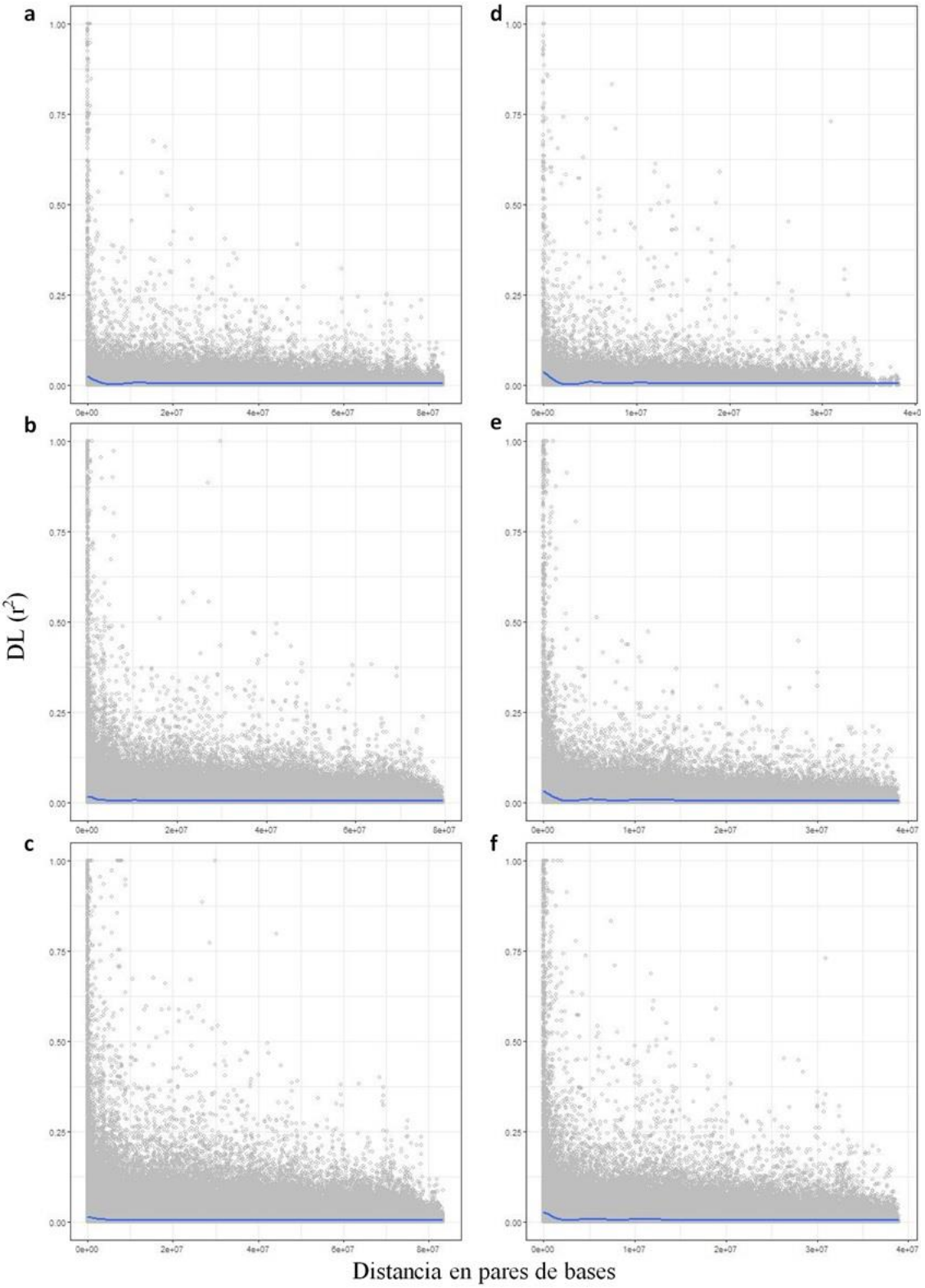


Figura 3.3.16. *Desequilibrio de ligamiento a lo largo de los cromosomas con mayor y menor r^2*. DL en valores de r^2 en función de la distancia cromosómica. Referencia: Cromosoma 3 con el menor valor promedio de r^2 en gráficos de la izquierda, siendo calculado con: **a** GBS, **b**. Chip, **c**. Chip-GBS; Cromosoma 9 con uno de los mayores valores promedio de r^2 en gráficos de la derecha, siendo calculado con: **a** GBS; **b**. Chip; **c**. Chip-GBS.

Posteriormente, con el objetivo de obtener estimaciones de parámetros y estructura genético poblacionales más precisos, se filtraron aquellos SNPs de menor MAF de los pares que presentaban un DL con r^2 mayor a 0,2 (ventanas de 2 Mb). De este modo, las matrices filtradas por DL presentaron 4.848 (GBS), 13.385 (Chip) y 17.611 (Chip-GBS) SNPs. Presentaron un promedio de 428, 1.217 y 1.588 SNPs por cromosoma, para GBS, Chip y Chip-GBS respectivamente, manteniendo el orden de número creciente de SNPs entre las matrices. Luego de este filtrado, por lo tanto, los DL promedio (r^2) disminuyeron, tanto los valores por cromosoma, como el promedio de todos los cromosomas. Dichos valores variaron de 0,0056 a 0,0062 y una media de 0,0058 para GBS, de 0,0062 a 0,0071 y una media de 0,0067 para el Chip, y de 0,0060 a 0,0067 con una media de 0,0063 para Chip-GBS. Nuevamente, GBS presenta los menores valores y la matriz conjunta valores intermedios. Por último, respecto de las distancias entre los SNPs promedio por cromosoma, se observó un promedio de 125.423 pb de distancias entre SNPs contiguos para GBS, de 45.177 pb para el Chip y de 35.490 pb para la matriz conjunta. Estos últimos valores observados fueron mayores a los de las matrices sin filtrar por DL presentados en el apartado anterior, y era previsible al disminuir el número de marcadores por cromosoma.

3.3.1.3 ESTIMACIÓN DE PARÁMETROS DE ESTRUCTURA Y DIVERSIDAD GENÉTICA POBLACIONAL

La información generada mediante análisis de DAPC respecto de la estructura genética poblacional concordó cuando se analizaron las tres matrices por separado como insumos para el análisis (filtradas por DL según el apartado anterior). El método DAPC identificó dos grupos genéticos determinados por el menor valor de BIC (Figura 3.3.17, a, b y c), difiriendo sólo en la asignación de 4 individuos a los grupos genéticos de pertenencia, entre las 3 matrices evaluadas. El grupo 1 fue el de menor representación (entre un 14,6% y 16% para Chip y Chip-GBS, respectivamente) y estuvo compuesto por 43 individuos para GBS, 45 para Chip y 41 para Chip-GBS. Como se observa en la figura 3.3.17 (d, e y f) este grupo fue el mismo para las 3 matrices, y fue conformado en su totalidad por la mayoría de los 52 individuos de procedencia local (origen/procedencia 5) de Oliveros Santa Fe (plantaciones comerciales), que son de un origen australiano conocido en Moleton, NSW. Siete a 11 individuos de dicha procedencia local formaron parte del grupo dos, según la matriz a considerar. Así, el grupo dos fue el de mayor representación, correspondiendo al resto de la población de mejoramiento. Este grupo fue conformado por 237 individuos para GBS (84,6% de la población total), 235 para el Chip (83,9%) y 239 para la matriz conjunta (85,4%), variando sólo en la incorporación o no de algunos individuos de procedencia local (SD; Figura 3.3.17 d, e y f).

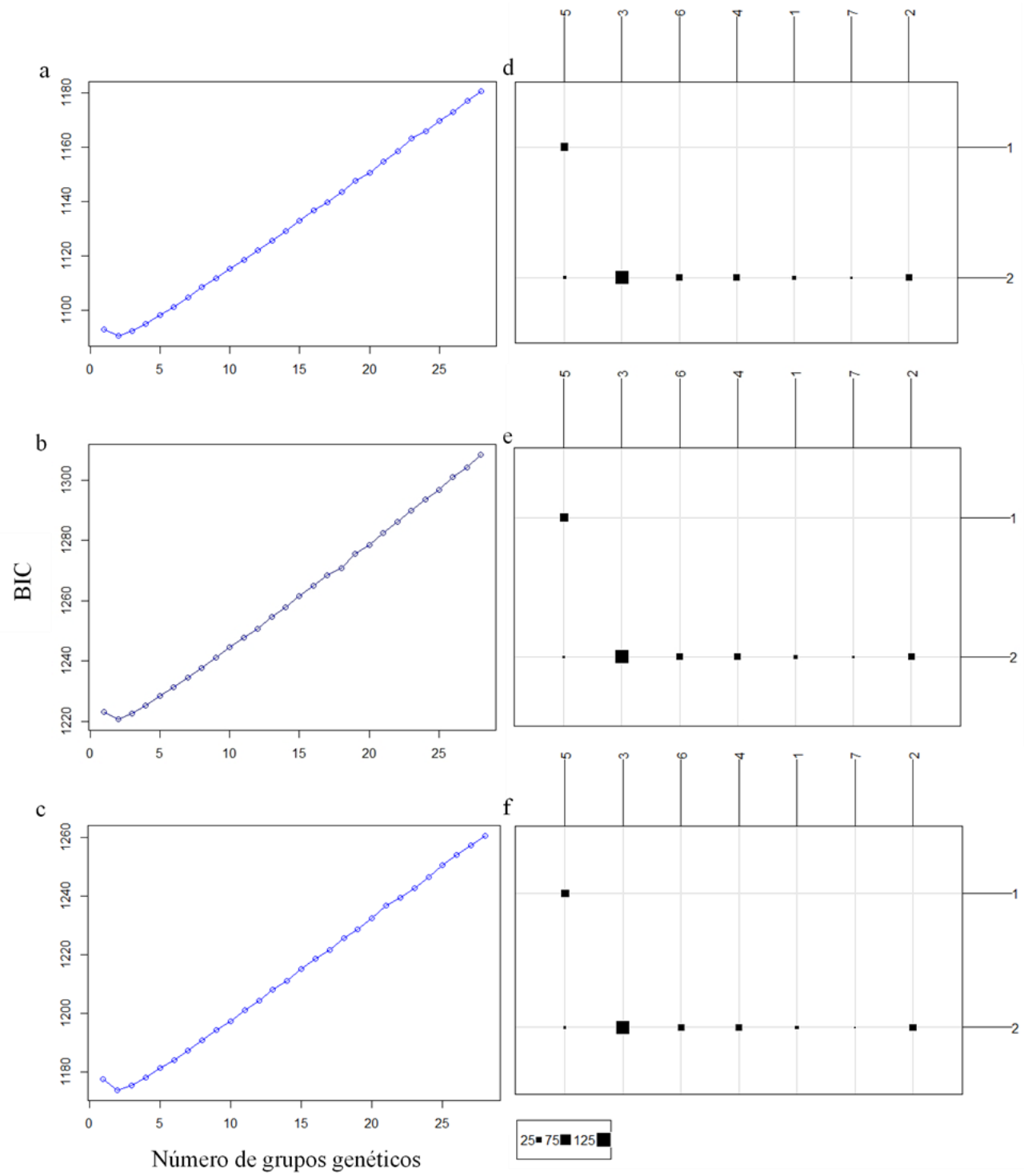


Figura 3.3.17. *Estructura genética poblacional inferida por DAPC.* Referencias: a, b y c: Gráfico de valores de BIC versus números de grupos genéticos a partir de GBS, Chip y Chip-GBS, respectivamente; d, e y f: Correspondencia entre los 7 orígenes australianos y 2 grupos genéticos inferidos por DAPC para individuos de la población de mejoramiento de *E. dunnii*, obtenidos a partir de GBS, Chip y Chip-GBS, respectivamente. Eje horizontal: Orígenes de australianos de Nueva Gales del Sur y/o plantaciones de procedencia: 1: Acacia Creek; 2: Boomi Creek; 3: Dearth Horse Track Region; 4: Oaky Creek, NSW; 5: árboles seleccionados en plantaciones comerciales (Oliveros Santa Fe, Argentina) del origen australiano de Moleton; 6: South Yabra S. F.; 7: Unumgar S.F. Eje Vertical: grupos genéticos inferidos 1 y 2.

Las tres matrices presentaron en común siete de los individuos del origen local asignados al grupo dos, seis de ellos pertenecientes a dos familias completas (familia 10 con dos individuos y familia 11 con cuatro) y uno perteneciente a la familia 12 (SD.212.13). La matriz de GBS además asignó al grupo dos a otro individuo de la familia 12 (SD.212.7) y uno de la familia 3 (SD.203.7). Por otro lado, la matriz del Chip no asignó al grupo dos ningún individuo aparte de los comunes a las tres matrices. Por último, la matriz de Chip-GBS asignó a la familia 12 completa (SD.212.13, SD.212.17, SD.212.7 y SD.212.16) al grupo dos y también al individuo SD.205.18 de la familia cinco.

Sin embargo, a través del módulo *populations* del programa *Stacks* (v1.48), el F_{ST} entre los grupos genéticos inferidos por DAPC mostró valores bajos de diferenciación entre los mismos y similares para las tres matrices, siendo de 0,0148 (p-valor < 0.05) para la matriz de GBS, de 0,0155 (p-valor < 0.05) para la matriz de Chip y 0,0148 (p-valor < 0.05) para la conjunta. Esto evidencia que la población no presenta una estructura genética muy marcada entre ambos grupos.

En la tabla 3.3.4 podemos observar los estadísticos de genética de poblaciones con cada metodología de genotipificación, calculados para la población completa y para los grupos genéticos definidos por DAPC. Los valores observados de *p*, *He*, *Ho* y PIC fueron mayores, más del doble, para la matriz del Chip (0,20, 0,28, 0,29 y 0,23, respectivamente) con respecto a los obtenidos para GBS (0,11, 0,14, 0,13 y 0,12, respectivamente, Tabla 3.3.4), siendo estas diferencias significativas. Estos resultados concuerdan con los observados en el apartado anterior, donde la matriz de GBS presentó mayor proporción de MAF con valores menores a 0,1 que la matriz del Chip. Esto último, ratifica que la matriz del Chip proporciona información de polimorfismos más comunes en la población que la matriz de GBS, siendo que estas metodologías evalúan distintas regiones genómicas con diferentes niveles de variabilidad. Así, con la matriz conjunta se obtuvieron valores intermedios (0,18, 0,25, 0,26 y 0,21, para *p*, *He*, *Ho* y PIC respectivamente) según se esperaba debido a que la misma fue compuesta por ambas matrices, aunque dichos valores fueron cercanos a los presentados por la matriz del Chip, lo que podría ser explicado debido al mayor número de SNPs del Chip respecto de GBS.

En todos los casos anteriores, se observó que el grupo genético uno mostró valores levemente inferiores a los de la población total (ej. *Ho* para GBS: 0,13; Chip: 0,27; Chip-GBS: 0,24), a diferencia del grupo dos que presentó los mismos que la totalidad (Tabla 3.3.4). Esto puede ser debido a que el grupo uno presentó una cantidad de individuos 6,5 veces menor que el grupo dos y que fue conformado por individuos procedentes de un solo origen, llevando a índices de diversidad menores por dicho motivo. Además, dichos individuos podrían haber reducido su variabilidad al haber sido los seleccionados de una plantación comercial.

Tabla 3.3.4. *Parámetros de diversidad genética poblacional para las 3 matrices*. Pob. Total: parámetros calculados para la totalidad de la población; Grupo 1 y Grupo 2: parámetros calculados para el grupo genético 1 y 2 definidos por el análisis de DAPC; p: Frecuencia alélica media del alelo minoritario; q: Frecuencia alélica media del alelo mayoritario; He: Heterocigosis esperada; Ho Heterocigosis observada; PIC: Información del contenido polimórfico (*Polymorphic information content*).

Parám. Pob.	GBS			Chip			Chip-GBS		
	Pob. Total	Grupo 1	Grupo 2	Pob. Total	Grupo 1	Grupo 2	Pob. Total	Grupo 1	Grupo 2
p	0,11	0,11	0,11	0,20	0,20	0,20	0,18	0,18	0,18
q	0,89	0,89	0,89	0,80	0,80	0,80	0,82	0,82	0,82
He	0,17	0,14	0,17	0,28	0,25	0,28	0,25	0,22	0,25
Ho	0,15	0,13	0,15	0,29	0,27	0,29	0,26	0,24	0,26
PIC	0,14	0,12	0,14	0,23	0,20	0,23	0,21	0,18	0,21

4. Aplicación de metodologías genómicas para el mejoramiento molecular de *E. dunnii* mediante Mapeo por Asociación y Selección Genómica

3.4.1 Análisis de Asociación de Genoma Amplio (Genome Wide Association Study)

Los análisis de asociación genotipo (genotipado amplio con GBS, Chip, GBS-Chip) y fenotipo (14 caracteres), se llevaron adelante aplicando un modelo lineal mixto, como se mencionó en Materiales y Métodos.

Como es sabido, en las poblaciones, las relaciones genéticas entre individuos generan asociaciones que no se deben al DL sino a las relaciones existentes de la historia evolutiva de la población, denominadas asociaciones espurias. Por lo tanto, para disminuir esta situación, se ensayaron los modelos estadísticos incorporando la matriz de parentesco y/o la matriz de estructura genética en el modelo estadístico elegido.

De esta manera, se analizó individualmente cada modelo estadístico CMLM para todas las características medidas y con las tres matrices genotípicas (GBS, Chip, GBS-Chip) incluyendo la matriz de parentesco K, y de estructura genética (esta última variando el número de Componentes Principales de 0 a 15).

El modelo que mejor ajustó según el mínimo valor de BIC, fue el que sólo incluyó la matriz K (es decir "0" CP) para todas las características, excepto para forma de fuste, pero que se consideró despreciable la diferencia de 1: -224 vs -225). Dichos valores se muestran en la Tabla 3.4.1, donde se detallan, a modo de ejemplo, los valores de BIC obtenidos con la matriz de Chip-GBS y la inclusión de los distintos CP para todas las características.

Tabla 3.4.1. *Valores de BIC para los 16 modelos evaluados en los 14 fenotipos con matriz de Chip-GBS.* N° CP: número de componentes principales, dap: diámetro a la altura del pecho, at: altura total, for: forma de fuste, ir: índice de rajado, extet: extractivos etanólicos, exttot: extractivos totales, klas: lignina klason, lig: lignina total, sg: Siringilo/Guayacilo, cel: celulosa, db: densidad básica (números corresponden a las edades de medición). Se encuentran resaltados en negrita los valores de BIC de los modelos que mejor ajustan.

N° CP	dap6	at6	for6	dap 11	dap 20	at20	ir	extet	exttot	klas	lig	sg	cel	db
0	**-282**	**-220**	-225	**-384**	**-372**	**-327**	**-393**	**-333**	**-334**	**-319**	**-320**	**-329**	**-331**	**-340**
1	-285	-221	**-224**	-387	-375	-330	-396	-335	-336	-321	-321	-331	-333	-343
2	-288	-224	-226	-390	-377	-332	-397	-336	-337	-322	-323	-332	-334	-345
3	-291	-225	-228	-393	-380	-335	-399	-337	-338	-323	-325	-335	-336	-347
4	-294	-228	-229	-396	-383	-337	-402	-340	-341	-325	-327	-337	-338	-348
5	-296	-230	-232	-398	-386	-340	-404	-342	-343	-328	-329	-340	-341	-351
6	-299	-233	-234	-401	-389	-343	-406	-345	-346	-331	-332	-342	-344	-351
7	-302	-235	-237	-403	-391	-345	-408	-347	-347	-333	-334	-344	-346	-354
8	-304	-238	-239	-406	-394	-348	-409	-350	-350	-334	-336	-347	-348	-357
9	-307	-241	-239	-409	-397	-352	-412	-352	-353	-335	-338	-349	-351	-358
10	-310	-243	-242	-411	-399	-352	-415	-353	-353	-338	-340	-349	-353	-360
11	-312	-246	-244	-413	-401	-353	-417	-354	-354	-340	-343	-352	-356	-363
12	-314	-249	-246	-416	-404	-356	-417	-357	-356	-343	-345	-353	-358	-363
13	-317	-249	-249	-418	-407	-359	-418	-359	-359	-346	-348	-356	-360	-366
14	-320	-252	-252	-421	-410	-362	-420	-360	-360	-344	-345	-359	-356	-369
15	-322	-253	-255	-424	-413	-365	-421	-363	-364	-347	-347	-361	-359	-368

Con el objetivo de poder comparar las bondades de ambas metodologías genómicas, en el próximo apartado se describen dos análisis de asociación por separado para cada una de las características analizadas y con cada matriz por separado (GBS y Chip).

3.4.1.1 Análisis de Asociación de Genoma Amplio (Genome Wide Association Study) con Matriz de GBS

En primer lugar, se realizó un análisis de GWAS considerando los SNP provenientes sólo la matriz de GBS sin filtrar por DL (8.170 SNPs). Con este análisis se encontró un total de siete marcadores asociados con siete de las 14 características evaluadas (for6, dap20, exttot20, lig20, cel20, sg20 y db20), de los cuales uno fue compartido entre lig20 y cel20, y dos estuvieron asociados a sg20. El criterio utilizado para definir marcadores asociados fue que superaran el umbral *ad hoc* de -log(1E-04). Estos resultados se pueden observar en la Tabla 3.4.2, donde se detalla el nombre del SNP y cromosoma donde se ubicó, el número de individuos con fenotipo medido, la proporción de la variación fenotípica explicada por el SNP (diferencia de R^2) y los umbrales

superados por el SNP. Además, en la tabla 3.4.2 se informó que los p-valores de tres de los siete SNPs superaron el umbral de -log(1/n), pero ninguno superó el de Bonferroni. Debido a que ningún marcador superó el umbral de FDR < 0,05, se adoptó el criterio utilizado por Kainer et al. (2019) (que consideró como significativos SNPs con un FDR < a 0,1 y reportó también aquellos < a 0,2). De este modo, dos de los siete SNPs de GBS asociados presentaron valores de FDR inferiores a 0,2 (para lig20 y sg20), y uno inferior a un FDR de 0,1 (para for6), dándole más robustez a dichas asociaciones. Estos últimos SNPs se resaltan en negrita en la columna correspondiente a los valores de FDR de la tabla 3.4.2. Los siete SNPs asociados se distribuyeron en 4 cromosomas: l 2, 3, 4 y 8.

Tabla 3.4.2. *SNPs de la matriz de GBS asociados para los 14 caracteres analizados*. Fenotipo: Carácter fenotípico asociado; SNP: nombre del SNP (EuBR: Chip, número: GBS). Cromosoma: N° de Cromosoma; Posición: posición en pb del SNP en el cromosoma; p-valor: p-valor presentado por el SNP; MAF: Frecuencia alélica mínima del SNP; N° obs: número de individuos con fenotipo; Diferencia de R^2: diferencia de R^2 entre el modelo que contempla el SNP y el que no lo contempla, este valor refleja la proporción de la variación fenotípica explicada por SNP; FDR: valor de FDR; Efecto: efecto del SNP; Umbral: Umbral superado por el SNP. En negrita: valores de FDR menores a 0,2; dap20: diámetro a la altura de pecho a los 20 años; for6: forma de fuste a los 6 años; exttot20: Extractivos totales; lig20: lignina; sg20: siringilo/guayacilo; cel20: celulosa; db20: densidad básica. Se destacan en negrita los valores de FDR menores a 0,2.

	Fenotipo	SNP	Cromosoma	Posición	p-valor	MAF	N° obs	diferencia R^2	FDR	Efecto	Umbral
1	for6	8074_34	2	35977170	7,59E-06	0,11	280	0,074	**0,06**	0,30	[1/n]
2	dap20	21975_49	4	3122855	8,35E-05	0,01	276	0,058	0,48	0,98	ad hoc
3	exttot20	48478_24	8	43190754	9,48E-05	0,03	232	0,066	0,45	0,91	ad hoc
4	lig20	13753_15	3	18501528	1,73E-05	0,02	232	0,077	**0,14**	-1,06	[1/n]
	cel20		3	18501528	5,27E-05	0,02	232	0,071	0,43	1,03	ad hoc
5	sg20	47543_33	8	36839043	1,44E-05	0,03	232	0,079	**0,12**	0,942	[1/n]
6		11288_34	2	56443464	6,36E-05	0,02	232	0,066	0,26	1,34	ad hoc
7	db20	19771_30	4	11899166	9,64E-05	0,08	232	0,066	0,79	0,77	ad hoc

En la figura 3.4.1 se muestran los siete SNPs de la matriz de GBS asociados, en forma conjunta en un gráfico de Manhattan para las 14 características de productividad, calidad y propiedades fisicoquímicas de la madera evaluadas.

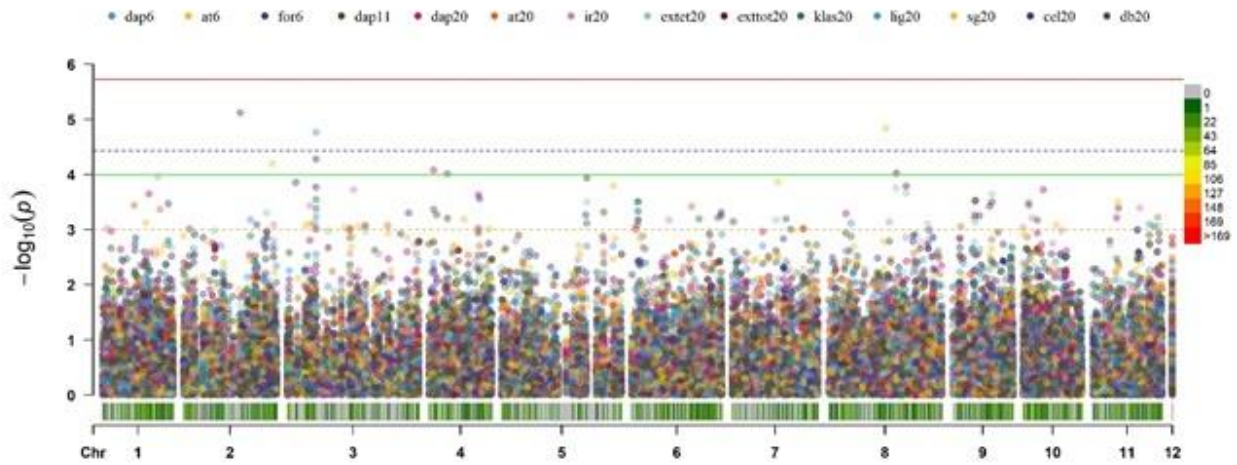


Figura 3.4.1: *Gráfico de Manhattan de GWAS para matriz de GBS para los 14 caracteres analizados*. Eje de abscisas: Número de cada cromosoma y Posición en pares de bases de cada SNP (Crom 12: *Scaffolds*). Eje de ordenadas: -log10(p): logaritmo negativo en base 10 de los p-valores de cada SNP; Umbral naranja punteado: *ad hoc* de -log(1E-03); verde: *ad hoc* de -log(1E-04); azul punteado: -log(1/n); y rojo: Bonferroni. Dap: diámetro a la altura de pecho a los 6, 11 y 20 años; at: altura total a los 6 y 20 años; for6: forma de fuste a los 6 años; ir20: índice de rajado a los 20 años; extet20: Extractivos etanólicos; exttot20: Extractivos totales; klas20: lignina klason; lig20: lignina; sg20: siringilo/guayacilo; cel20: celulosa; db20: densidad básica.

El porcentaje de variación fenotípica explicada por cada SNP asociado fue bajo (promedio= 7%), y varió entre 5,8 y 7,9% para dap20 y sg20, respectivamente. Esto significa que no se encontraron SNPs de QTLs con efectos mayores, aunque estos valores son los esperados para caracteres cuantitativos como los evaluados.

El MAF de los siete marcadores asociados varió de 0,01 a 0,11, siendo sólo dos de ellos superior a 0,05, lo que justifica el haber utilizado un filtro de SNPs con MAF hasta de 0,01 en vez de hasta 0,05, ya que este resultado sugiere que con un filtrado de SNPs con valores mayores de MAF no se hubiesen detectado la gran mayoría de las asociaciones.

3.4.1.2 Análisis de Asociación de Genoma Amplio (Genome Wide Association Study) con Matriz de Chip

Por otro lado, se realizó un análisis de GWAS considerando sólo los SNPs provenientes de la matriz del Chip sin filtrar por DL (19.045 SNPs). Con este análisis se encontró un total de 13 marcadores asociados para diez de las 14 características evaluadas (dap6, dap20, at20, ir20, extet20, exttot20, klas20, cel20, sg20 y db20), de los cuales uno fue compartido entre extet20 y exttot20, dos estuvieron asociados a dap20 y tres a klas20. El

criterio utilizado para definir marcadores asociados fue el mismo que para GBS, siendo aquellos que superaran el umbral *ad hoc* de -log(1E-04). Estos resultados se pueden observar en la Tabla 3.4.3, donde, al igual que en la tabla de GBS, se detalla el nombre del SNP y cromosoma donde se ubicó, el número de individuos con fenotipo medido, la proporción de la variación fenotípica explicada por del SNP (diferencia de R^2) y los umbrales superados por el SNP. Además, en la tabla 3.4.3 se informa en la columna de nombre "umbral" que seis de los p-valores de los 13 SNPs superaron el umbral de -log(1/n), pero ninguno superó el de Bonferroni. Asimismo, uno de los 13 SNPs asociados, vinculado con sg20, presentó un FDR con un valor inferior a 0,1, y otro SNP asociado a exttot20 un valor igual a 0,2, dándole más robustez a las asociaciones. Estos últimos dos SNPs se resaltan en negrita en la columna correspondiente a los valores de FDR de la tabla 3.4.3. Los 13 SNPs asociados se distribuyeron en todos los cromosomas, con excepción del cromosoma dos, evidenciando una mayor dispersión a lo largo del genoma con respecto a los encontrados con GBS.

Tabla 3.4.3. *SNPs de la matriz de Chip asociados para los 14 caracteres analizados*. Fenotipo: Carácter fenotípico asociado; SNP: nombre del SNP (EuBR: Chip, número: GBS). Cromosoma: N° de Cromosoma; Posición: posición en pb del SNP en el cromosoma; p-valor: p-valor presentado por el SNP; MAF: Frecuencia alélica mínima del SNP; N° obs: número de individuos con fenotipo; Diferencia de R^2: diferencia de R^2 entre el modelo que contempla el SNP y el que no lo contempla, este valor refleja la proporción de la variación fenotípica explicada por SNP; FDR: valor de FDR; Efecto: efecto del SNP; Umbral: Umbral superado por el SNP. En negrita: valores de FDR menores a 0,2; Dap: diámetro a la altura de pecho a los 6 y 20 años; at: altura total a los 20 años; ir20: índice de rajado a los 20 años; extet20: Extractivos etanólicos; exttot20: Extractivos totales; klas20: lignina klason; lig20: lignina; sg20: siringilo/guayacilo; cel20: celulosa; db20: densidad básica. Se destacan en negrita los valores de FDR menores a 0,2.

	Fenotipo	SNP	Cromosoma	Posición	p-valor	MAF	N° obs	diferencia R^2	FDR	Efecto	Umbral
1	dap6	EuBR05s51176000A3B1C1D30E1	5	51176000	1,24E-05	0,28	280	0,071	0,24	-0,36	[1/n]
2	dap20	EuBR03s2637451A2B0C1D40E0	3	2637451	7,04E-05	0,22	276	0,059	0,67	0,46	ad hoc
3		EuBR07s10247657A1B0C1D40E0	7	10247657	9,92E-05	0,16	276	0,056	0,67	-0,49	ad hoc
4	at20	EuBR05s25410255A3B0C0D40E0rep1	5	25410255	1,23E-05	0,42	276	0,072	0,23	-0,28	[1/n]
5	ir20	EuBR08s68220660A3B1C1D60E1	8	68220660	4,79E-05	0,31	276	0,059	0,88	0,40	ad hoc
6	extet20	EuBR11s17351865A2B0C1D60E1	11	17351865	1,75E-05	0,01	232	0,077	0,33	1,77	[1/n]
	exttot20			17351865	1,03E-05			0,083	**0,20**	1,83	[1/n]
7	klas20	EuBR09s4271448A4B0C1D40E1	9	4271448	5,79E-05	0,15	232	0,066	0,38	0,52	ad hoc
8		EuBR04s29699828A2B0C1D30E1	4	29699828	7,14E-05	0,07	232	0,064	0,38	0,77	ad hoc
9		EuBR04s29197264A3B0C0D40E0	4	29197264	8,05E-05	0,13	232	0,063	0,38	0,45	ad hoc
10	sg20	EuBR06s47777412A1B1C1D60E1	6	47777412	4,81E-06	0,14	232	0,088	**0,09**	0,63	[1/n]
11	cel20	EuBR01s34250726A1B1C1D60E1	1	34250726	1,74E-05	0,06	232	0,081	0,33	0,91	[1/n]
12		EuBR10s12580608A3B0C1D40E1	10	12580608	4,55E-05	0,12	232	0,073	0,43	0,64	ad hoc
13	db20	EuBR05s57414738A3B0C1D40E1	5	57414738	4,00E-05	0,16	232	0,074	0,76	0,64	ad hoc

En la figura 3.4.2 se muestran los 13 SNPs de la matriz de Chip asociados, en forma conjunta en gráfico de Manhattan para los 14 caracteres de productividad, calidad y propiedades fisicoquímicas de la madera evaluadas.

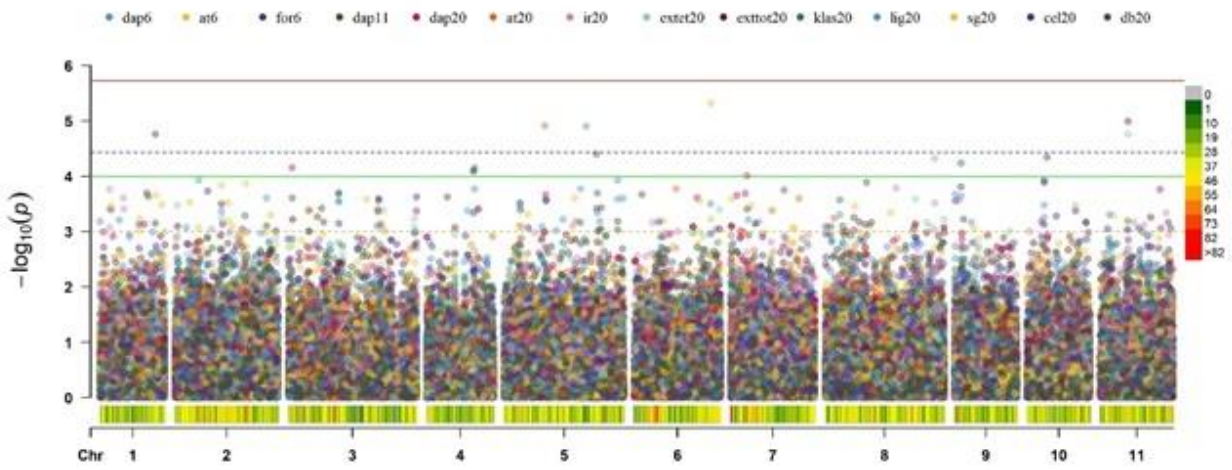


Figura 3.4.2: *Gráfico de Manhattan de GWAS para matriz de Chip para los para los 14 caracteres analizados*. Eje de abscisas: Número de cada cromosoma y Posición en pares de bases de cada SNP. Eje de ordenadas: -log10(p): logaritmo negativo en base 10 de los p-valores de cada SNP. Umbral naranja punteado: *ad hoc* de -log(1E-03); verde: *ad hoc* de -log(1E-04); azul punteado: -log(1/n); y rojo: Bonferroni. Dap: diámetro a la altura de pecho a los 6, 11 y 20 años; at: altura total a los 6 y 20 años; for6: forma de fuste a los 6 años; ir20: índice de rajado a los 20 años; extet20: Extractivos etanólicos; exttot20: Extractivos totales; klas20: lignina klason; lig20: lignina; sg20: siringilo/guayacilo; cel20: celulosa; db20: densidad básica.

El porcentaje de variación explicada por cada SNP asociado varió entre 5,6 y 8,8%, siendo estos valores extremos, al igual que con GBS, para dap20 y sg20, respectivamente. Además, el promedio también fue del 7%, lo que ratifica el hecho de que no se encontraran SNPs de QTLs con efectos mayores por la naturaleza cuantitativa de los caracteres bajo estudio, además de las características particulares de la población.

El MAF de los siete marcadores asociados varió de 0,01 a 0,42, siendo sólo uno de ellos inferior a 0,05, siendo estos valores superiores a los encontrados para GBS, seguramente por el tamiz que tiene el chip y la fuente de SNPs que se utilizaron para su desarrollo.

Comparación de Asociación GWAS de GBS vs Chip: De los resultados individuales luego del análisis de los SNPs provenientes de ambas metodologías de genotipificación, se observó que se detectaron marcadores asociados a los caracteres de dap20, exttot20, cel20, sg20 y db20 con ambas técnicas. Esto sugiere que los datos fenotípicos para estas características reúnen las condiciones suficientes para ser detectados por ambas distribuciones de SNPs y que los caracteres están gobernados por muchos genes. Sin embargo, ninguna de estas características fenotípicas presentó una región genómica de asociación común entre ambos análisis de GWAS, siendo que cada una evidenció marcadores asociados en distintos cromosomas para un mismo carácter,

cómo era de esperar, considerando la complementariedad que se observó y describió previamente. Asimismo, para algunos caracteres se encontraron SNPs asociados exclusivamente con una de las metodologías y no así con la otra, tal fue el caso de los SNPs asociados a for6 y lig20 con GBS y de los asociados a dap6, at20, ir20, exttot20 y klas20 con el Chip. Por último, en ninguno de los dos análisis se encontraron asociaciones con dap11 y at6.

3.4.1.3 ANÁLISIS DE ASOCIACIÓN DE GENOMA AMPLIO (*GENOME WIDE ASSOCIATION STUDY) EN MATRIZ CONJUNTA*

Finalmente, se realizó un último análisis de GWAS considerando la matriz conjunta de Chip-GBS sin filtrar por DL (27.019 SNPs), con el objetivo de evaluar si, al aumentar la cantidad de marcadores en un mismo análisis, surgían las mismas asociaciones y/o nuevas. En este análisis se encontraron un total de 19 marcadores asociados para 12 de las 14 características evaluadas (excepto para at6 y dap11). El criterio utilizado para definir marcadores asociados fue el mismo que para GBS, ie. aquellos que superaron el umbral *ad hoc* de -log(1E-04). De este modo, el número de marcadores asociados por característica fenotípica fue de uno para dap6, at20, for6, ir20, lig20, extet20 y db20, dos para exttot20 y tres para dap20, cel20, klas20 y sg20. Estos resultados se pueden observar en la Tabla 3.4.4, donde, como en los análisis de GWAS anteriores, se detalla el nombre del SNP y cromosoma donde se ubicó, el número de individuos con fenotipo medido, la proporción de la variación fenotípica explicada por del SNP (diferencia de R^2) y los umbrales superados por el SNP. Se destacan en negrita los valores de FDR menores a 0,2.

Tabla 3.4.4. *SNPs de la matriz de Chip-GBS asociados para los 14 caracteres analizados.* Fenotipo: Carácter fenotípico asociado; SNP: nombre del SNP (EuBR: Chip, número: GBS). Crom: N° de Cromosoma; Posición: posición en pb del SNP en el cromosoma; p-valor: p-valor presentado por el SNP; MAF: Frecuencia alélica mínima del SNP; N° obs: número de individuos con fenotipo; R^2 sin SNP: R^2 del modelo sin contemplar el SNP; R^2 con SNP:; Diferencia de R^2: diferencia de R^2 entre el modelo que contempla el SNP y el que no lo contempla, este valor refleja la proporción de la variación fenotípica explicada por SNP; FDR: valor de FDR; Efecto: efecto del SNP; Umbral: Umbral superado por el SNP. En negrita: valores de FDR menores a 0,2; Dap: diámetro a la altura de pecho a los 6, 20 años; at: altura total a los 20 años; for6: forma de fuste a los 6 años; ir20: índice de rajado a los 20 años; extet20: Extractivos etanólicos; exttot20: Extractivos totales; klas20: lignina klason; lig20: lignina; sg20: siringilo/guayacilo; cel20: celulosa; db20: densidad básica. Se destacan en negrita los valores de FDR menores a 0,2.

	Fenotipo	SNP	Crom	Posición	p-valor	MAF	N° obs	diferencia R^2	FDR	Efecto	Umbral
1	dap6	EuBR05s5117600	5	51176000	1,14E-05	0,28	280	0,071	0,31	-0,36	[1/n]
2	dap20	EuBR03s2637451	3	2637451	7,03E-05	0,22	276	0,059	0,62	0,46	ad hoc
3		EuBR07s10247657	7	10247657	9,08E-05	0,16	276	0,059	0,62	-0,49	ad hoc
4		21975_49	4	3122855	9,18E-05	0,01	276	0,059	0,62	0,98	ad hoc
5	at20	EuBR05s25410255	5	25410255	1,30E-05	0,42	276	0,072	0,23	-0,28	[1/n]
6	for6	8074_34	2	35977170	6,92E-06	0,11	280	0,074	**0,19**	0,30	[1/n]
7	ir20	EuBR08s68220660	8	68220660	4,64E-05	0,31	276	0,059	0,96	0,44	ad hoc
8	lig20	13753_15	3	18501528	1,58E-05	0,02	232	0,076	0,43	-1,04	[1/n]
	cel20				4,86E-05		232	0,072	0,44	1,04	ad hoc
9	cel20	EuBR01s34250726	1	34250726	2,10E-05	0,06	232	0,079	0,44	0,90	[1/n]
10		EuBR10s12580608	10	12580608	4,40E-05	0,12	232	0,073	0,44	0,64	ad hoc
11	kla20	EuBR09s4271448	9	4271448	4,45E-05	0,15	232	0,068	0,53	0,53	ad hoc
12		EuBR04s29197264	4	29197264	7,70E-05	0,13	232	0,064	0,53	0,45	ad hoc
13		EuBR04s29699828	4	29699828	7,81E-05	0,07	232	0,063	0,53	0,77	ad hoc
14	extet20	EuBR11s17351865	11	17351865	1,57E-05	0,01	232	0,078	0,42	1,85	[1/n]
	exttot20				9,18E-06		232	0,085	0,25	1,85	[1/n]
15	exttot20	28672_21	5	53723552	9,06E-05	0,01	232	0,065	0,96	-1,43	ad hoc
16	sg20	EuBR06s47777412	6	47777412	5,59E-06	0,14	232	0,087	**0,15**	0,63	[1/n]
17		47543_33	8	36839043	1,77E-05	0,03	232	0,077	0,24	0,95	[1/n]
18		11288_34	2	56443464	7,77E-05	0,02	232	0,065	0,70	1,33	ad hoc
19	db20	EuBR05s57414738	5	57414738	4,19E-05	0,16	232	0,074	0,76	0,64	ad hoc

Entre los SNPs asociados en este caso, se encontraron seis marcadores de GBS y 13 del Chip y en todos los cromosomas: uno para los cromosomas 1, 6, 7, 9, 10 y 11, dos para los cromosomas 2, 3 y 8, tres para el 4 y cuatro para el 5 (Tabla 3.4.4).

Con respecto a los SNPs de GBS, cinco coincidieron con aquellos siete encontrados en el análisis que contempló sólo la matriz de GBS, y se encontró un nuevo marcador para exttot20 en el cromosoma cinco (SNP 28672_21). Los dos SNPs que no presentaron asociación en la matriz conjunta y que habían sido encontrados con la de GBS fueron el 48478_24 en el cromosoma 8 asociado a exttot20 y 19771_30 en el cromosoma 4 asociado a db20, sin embargo, presentaron valores muy bajos de p-valor cercanos al umbral -log(E-04) (0,00021 y 0,00012, respectivamente). Por otra parte, los SNPs del Chip asociados a los distintos caracteres fenotípicos en la matriz conjunta fueron los mismos que se encontraron en el análisis de GWAS con la matriz del Chip en forma individual. Estas diferencias entre el análisis del conjunto e individuales se explican por una disminución en los valores de p-valores de los SNPs en el GWAS con la matriz de Chip-GBS, por lo que los SNPs de GBS que no se encontraron en la matriz conjunta y si en la individual, aumentaron levemente su p-valor y no superaron el umbral de *ad hoc* de -log(1E-04), pero si superaron el umbral de *ad hoc* de -log(1E-03) (datos no mostrados). Asimismo, de los 19 SNPs, nueve superaron el umbral -log(1/n) y ninguno el de Bonferroni y/o un FDR significativo (< 0,05). Es interesante destacar que, según el criterio adoptado por Kainer et al. (2019) que tomó como significativos SNPs con un FDR < a 0,1 y reportó también aquellos < a 0,2, dos de los 19 marcadores presentaron FDR menor a 0,2 (Tabla 3.4.4). Estos fueron el SNP de GBS 8074_34 asociado a for6 (FDR = 0,19) y el SNP del Chip EuBR06s47777412 asociado a sg20 (FDR = 0,15), lo que refuerza a dichas asociaciones. Estos SNPs fueron los que también presentaron menor valor de FDR en los análisis independientes, los cuales fueron de un valor de FDR de 0,06 para el SNP 8074_34 con la matriz GBS y de 0,09 para el SNP EuBR06s47777412 con la matriz del Chip.

A continuación, se pueden observar los 19 SNPs asociados con la matriz de Chip-GBS a través de las figuras 3.4.3 y 3.4.4. La figura 3.4.3 que presenta en forma conjunta los gráficos de Manhattan para las características de crecimiento (dap, at y for) e índice de rajado en rollizo y en la figura 3.4.4 para las características químicas de calidad de la madera. Estos gráficos se presentan en forma separada para poder tener una definición más clara de cada SNP asociado.

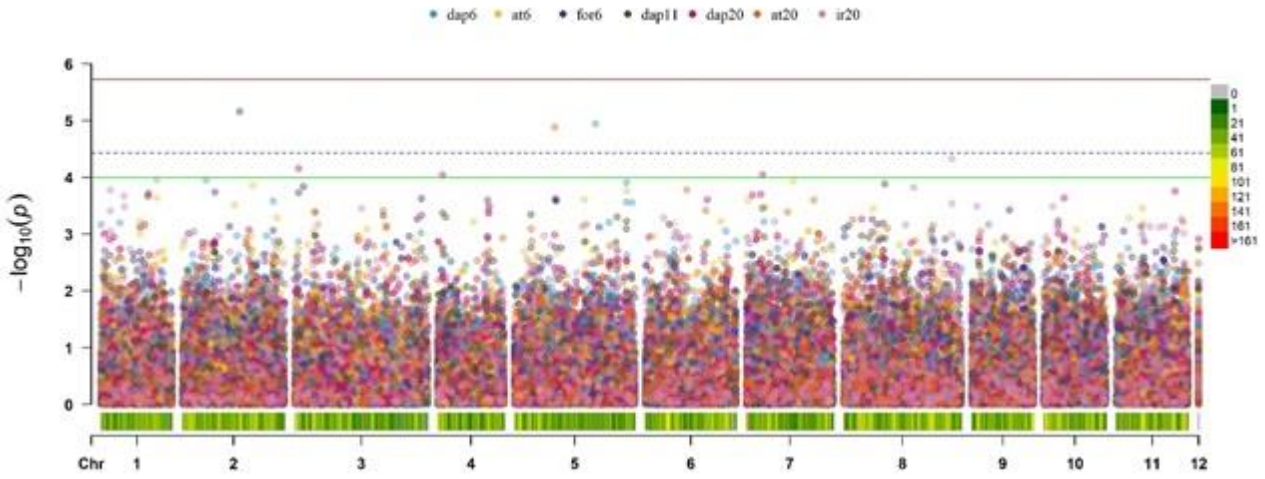


Figura 3.4.3: *Gráfico de Manhattan de GWAS utilizando SNPs provenientes de la matriz de Chip- GBS para 7 caracteres de productividad y calidad de madera: Crecimiento (Diámetro a la altura del pecho y altura total), Forma del fuste e Índice de rajado de la madera*. Eje de abscisas: Número de cada cromosoma y Posición en pares de bases de cada SNP (Crom 12: scaffolds); Eje de ordenadas: -log10(p): logaritmo negativo en base 10 de los p-valores de cada SNP; Dap: diámetro a la altura de pecho a los 6, 11 y 20 años; at: altura total a los 6 y 20 años; for6: forma de fuste a los 6 años; ir20: índice de rajado en rollizo a los 20 años. Umbral verde: *ad hoc* de -log(1E-04); azul punteado: -log(1/n); y rojo: Bonferroni.

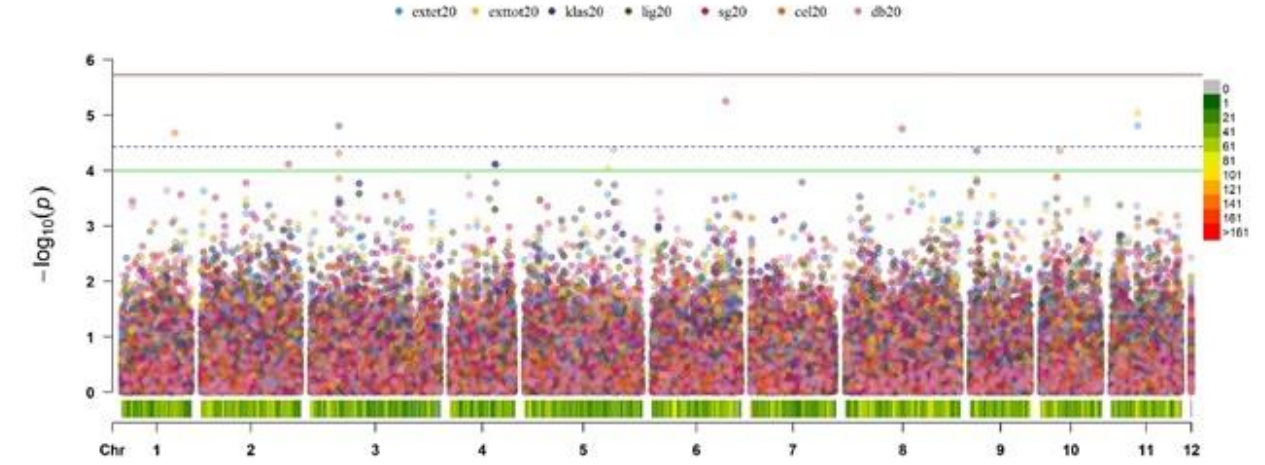


Figura 3.4.4: *Gráfico de Manhattan de GWAS utilizando SNPs provenientes de la matriz de Chip-GBS para siete caracteres fisicoquímicos de madera evaluadas*. Eje de abscisas: Número de cada cromosoma y Posición en pares de bases de cada SNP (Crom 12: scaffolds); Eje de ordenadas: -log10(p): logaritmo negativo en base 10 de los p-valores de cada SNP; extet20: Extractivos etanólicos; exttot20: Extractivos totales; klas20: lignina klason; lig20: lignina; sg20: siringilo/guayacilo; cel20: celulosa; db20: densidad básica. Umbral verde: *ad hoc* de -log(1E-04); azul punteado: -log(1/n); y rojo: Bonferroni.

Además, al igual que en los análisis por separado, el porcentaje promedio de variación explicada por SNP asociado fue bajo de 7%, variando entre 5,9 y 8,7%, lo que podría ser explicado por la naturaleza cuantitativa de los caracteres bajo estudio.

Los MAF de dichos SNPs variaron entre 0,01 y 0,42, siendo seis de los mismos menores a un MAF de 0,05 (31,5%), de los cuales cinco fueron SNPs de GBS (Tabla 3.4.4), lo que coincide con los análisis por separado.

Comparando con los análisis de GWAS anteriores, para todos los caracteres y como se mencionó en el apartado 3.4.1.1, se observa que los SNPs asociados con la matriz de GBS no se encuentran en las mismas regiones genómicas que con la matriz del Chip. A su vez, como se puede observar en la siguiente figura (Figura 3.4.5) que presenta los gráficos de Manhattan obtenidos para celulosa para cada una de las tres matrices, al realizar el análisis con ambas matrices en forma conjunta, se observa a grandes rasgos que los resultados son la sumatoria de ambos análisis por separado, aunque presentando variaciones en las magnitudes de los p-valores. Estas diferencias en los p-valores son debidas al aumento del número de SNPs implicados en el análisis conjunto.

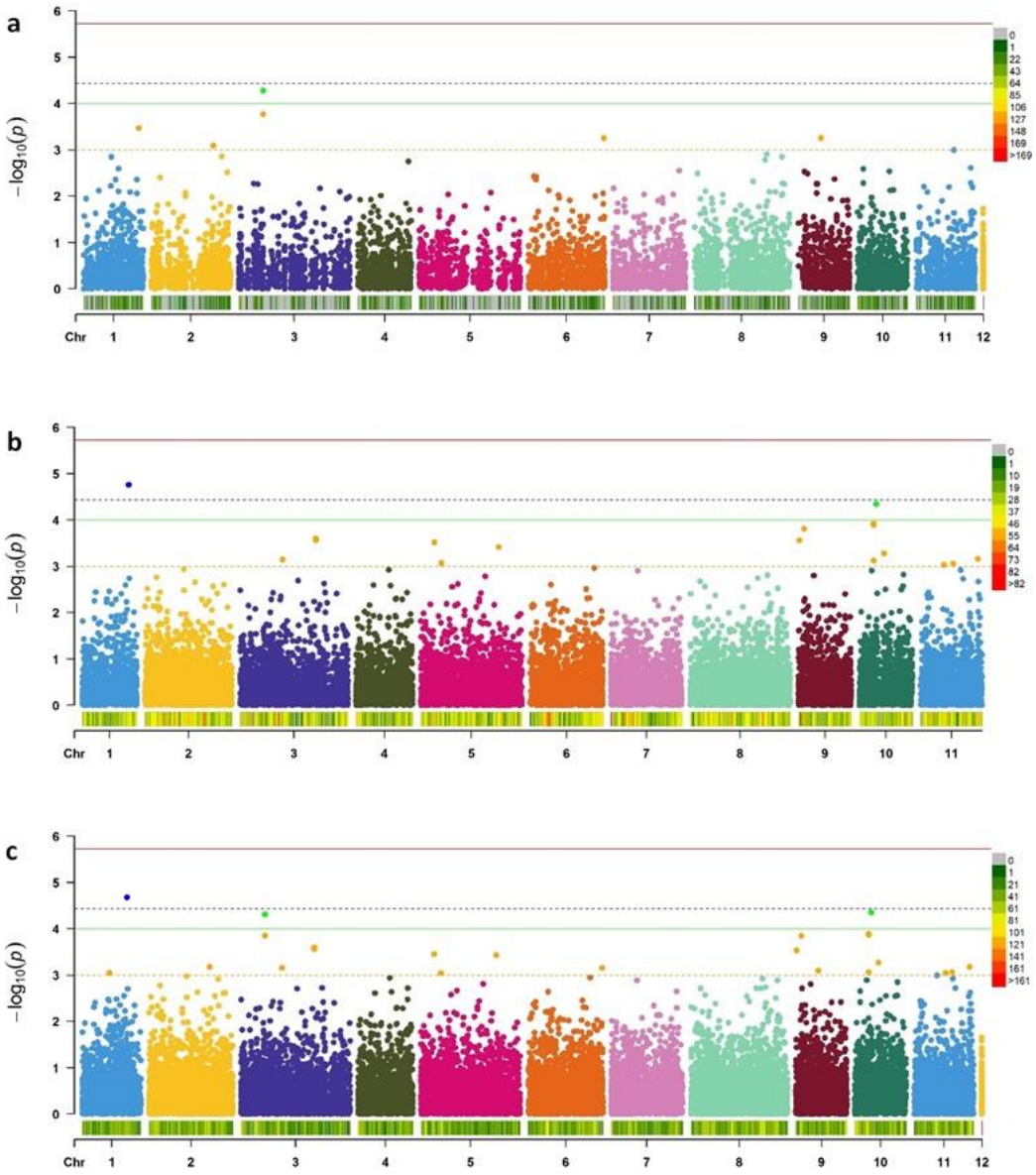


Figura 3.4.5: *Gráficos de Manhattan de GWAS para Celulosa utilizando SNPs provenientes de las matrices de: a) GBS, b) Chip, c) Chip-GBS.* Eje de abscisas: Número de cada cromosoma y Posición en pares de bases de cada SNP (Crom 12: scaffolds); Eje de ordenadas: -log10(p): logaritmo negativo en base 10 de los p-valores de cada SNP. Umbral naranja punteado: *ad hoc* de -log(1E-03); verde: *ad hoc* de -log(1E-04); azul punteado: -log(1/n); y rojo: Bonferroni.

Es interesante destacar que en los tres análisis de GWAS algunos marcadores fueron compartidos por algunas características fenotípicas que presentaron correlaciones significativas entre sí, según los resultados obtenidos en el apartado 3.1.1. Así, viendo como ejemplo los resultados del análisis conjunto, el marcador de GBS 13753_15 ubicado en el cromosoma 3 presentó asociación tanto como único SNP para lignina (p-valor: 1,58E-05, figura 3.3.4) como en tercer lugar para celulosa (p-valor: 4,86E-05, figura 3.4.4), y mostró un efecto opuesto (-1,04 y 1,04, respectivamente), como era de esperar según la alta correlación negativa y significativa

presentada entre ambos caracteres fenotípicos (ver apartado 3.1.1; figura 3.1.2; r^2 = -0,82; p-valor < 0,001). Asimismo, el porcentaje de la varianza explicada por cada carácter fue similar, de 7,6% para lig20 y 7,2% para cel20.

Del mismo modo, el SNP EuBR11s17351865 del Chip en el cromosoma 11 manifestó asociación con los caracteres de Extractivos totales y Etanólicos, que se encontraron altamente correlacionados y en forma positiva entre sí (ver apartado 3.1.1, figura 3.1.2, r^2 = 0,99, p-valor < 0,001), por lo cual el efecto del SNP para ambos caracteres fue de 1,85 (ambos puntos del cromosoma 11 en gráficos de Manhattan de la figura 3.4.4 y tabla 3.4.3). También el porcentaje de la varianza explicada fue similar, de 7,8% para extet20 y de 8,5% para exttot20.

De estas dos últimas observaciones surgió la necesidad de inspeccionar un umbral de asociación menos restrictivo (*ad hoc* de -log(1E-03)) con la posibilidad de que aparecieran falsos positivos. De esta manera, se reportaron 256 SNPs asociados y 42 de ellos compartidos entre al menos dos características fenotípicas cada uno. Dichas observaciones se detallan en Anexo 7.2.

3.4.1.4 GENES PRÓXIMOS A LOS MARCADORES ASOCIADOS

Los 21 SNPs asociados significativos según umbral *ad hoc* de -log(1E-04) fueron alineados contra el genoma de referencia de *E. grandis v2.0* y se exploraron ventanas de 70 Kpb alrededor de ellos (es decir a 35 Kpb río arriba y río abajo de distancia de ellos), para detectar los genes cercanos descriptos. Además, se exploró a menos de 500 Kpb para la existencia de genes catalogados por Myburg et al. (2014) involucrados en las distintas rutas metabólicas de la síntesis de celulosa, xilanos, fenilpropanoides, terpenos, lacasas, peroxidasas, entre otros. Diecinueve SNPs de Chip-GBS fueron asociados para las 12 de las 14 características, de los cuales 13 también fueron asociados con la matriz del Chip y seis con la de GBS, más dos SNPs asociados sólo con matriz de GBS. Como resultado de esta búsqueda se encontraron 100 genes descriptos en el genoma de *E. grandis*, que en la denominación se indican con el sufijo "Eucgr." a menos de 35 Kpb; y 1538 (50 de los cuales están catalogados en las rutas mencionadas) a menos de 500 Kpb que se describen posteriormente. Además, se obtuvieron los términos de Gene Ontology (GO), y los genes, símbolos y funciones en Arabidopsis. A continuación, en la tabla 3.4.6 se resume para cada marcador a que características fenotípicas se encontró asociado, los genes encontrados en dichas ventanas y su función determinada en Arabidopsis. En la tabla también se menciona el número de genes catalogados según Myburg et al. (2014) correspondientes a las vías metabólicas mencionadas y a menos de 500 Kpb.

Tabla 3.4.6. *Genes de E. grandis encontrados alrededor de los SNPs asociados a las 14 características fenotípicas.* Carac.: Características fenotípicas asociadas, p-valor del SNP mayor a 1E-04; Matriz: Matrices de SNP con las cuales se encontró asociación mediante GWAS con umbral *ad hoc* de -log(1E-04); Nombre SNP: Nombre del SNP según GBS y Chip; Crom.: Número de cromosoma de *E. grandis*; Dap: diámetro a la altura de pecho a los 6, 11 y 20 años; at: altura total a los 6 y 20 años; for6: forma de fuste a los 6 años; ir20: índice de rajado a los 20 años; extet20: Extractivos etanólicos; exttot20: Extractivos totales; klas20: lignina klason; lig20: lignina; sg20: siringilo/guayacilo; cel20: celulosa; db20: densidad básica; Genes cercanos en ventanas de 70 Kpb (distancia menor a 35 Kpb del SNP): Genes de *E. grandis*: Nombres de los genes de *E. grandis*; Función de *Arabidopsis*: función relativa al genoma de *Arabidopsis*; Genes a menos de 500 Kpb según categorías de Myburg et al. (2014): genes a menos de 500 Kpb según Myburg et al. (2014).

Carac.	Matriz	SNP	Nombre SNP	Crom.	Genes cercanos en ventanas de 70 Kpb (distancia menor a 35 Kpb del SNP)		Genes a menos de 500 Kpb según categorías de Myburg et al. (2014)
					Genes de *E. grandis*	Función de *Arabidopsis*	
dap6	Chip y Chip-GBS	1	EuBR05s51176000	5	Eucgr.E03046	cysteine-rich RLK (RECEPTOR-like protein kinase) 29	1 gen a 196 Kpb dentro de genes predichos de enzimas sintetizadoras de terpenos
					Eucgr.E03047	Receptor-like protein kinase-related family protein	
					Eucgr.E03049	cysteine-rich RLK (RECEPTOR-like protein kinase) 29	
					Eucgr.E03050	RING/U-box superfamily protein	
					Eucgr.E03051	cysteine-rich RLK (RECEPTOR-like protein kinase) 29	
dap20	Chip y Chip-GBS	2	EuBR03s2637451	3	Eucgr.C00046		3 genes a una distancia de 345 Kpb dentro de genes únicos de *Eucalyptus* detectados con dominio interpro
	Chip y Chip-GBS	3	EuBR07s10247657	7	Eucgr.G00581	emp24/gp25L/p24 family/GOLD family protein	
					Eucgr.G00582	cytochrome P450, family 72, subfamily A, polypeptide 8	
					Eucgr.G00583		
					Eucgr.G00584	cytochrome P450, family 72, subfamily A, polypeptide 8	
	GBS y Chip-GBS	4	21975_49	4	Eucgr.D00190	Transducin/WD40 repeat-like superfamily protein	
at20	Chip y Chip-GBS	5	EuBR05s25410255	5	Eucgr.E01900	SBP (S-ribonuclease binding protein) family protein	3 genes a una distancia de 400 Kpb dentro de genes SDRLK
					Eucgr.E01901	zinc transporter 7 precursor	
					Eucgr.E01902	SBP (S-ribonuclease binding protein) family protein	
for6	GBS y Chip-GBS	6	8074_34	2	Eucgr.B01917	Ribosomal protein S21e	3 genes: 2 dentro de genes únicos de *Eucalyptus* detectados con dominio interpro, 1 gen predicho de enzimas sintetizadoras de terpenos a 200Kpb
					Eucgr.B01918		
					Eucgr.B01919	Heavy metal transport/detoxification superfamily protein	
					Eucgr.B01920	Actin-binding FH2 (formin homology 2) family protein	
					Eucgr.B01921	Transducin/WD40 repeat-like superfamily protein	
					Eucgr.B01924	Actin-binding FH2 (formin homology 2) family protein	
					Eucgr.B01927	Transducin/WD40 repeat-like superfamily protein	

Carac.	Matriz	SNP	Nombre SNP	Crom.	Genes cercanos en ventanas de 70 Kpb (distancia menor a 35 Kpb del SNP)		Genes a menos de 500 Kpb según categorías de Myburg et al. (2014)
					Genes de *E. grandis*	Función de Arabidopsis	
ir20	Chip y Chip-GBS	7	EuBR08s68220660	8	Eucgr.H04783	signal responsive 1	8 genes: 6 dentro de genes únicos de Eucalyptus detectados con dominio interpro, 1 gen de la síntesis de celulosa y xilanos a 54 Kpb, 1 gen predicho de enzimas sintetizadoras de terpenos
					Eucgr.H04784	glycine-tRNA ligases	
					Eucgr.H04785	Nucleotide-sugar transporter family protein	
					Eucgr.H04786	TLD-domain containing nucleolar protein	
					Eucgr.H04787	annexin 2	
lig20 - cel20	GBS y Chip-GBS	8	13753_15	3	Eucgr.C01156	P-loop containing nucleoside triphosphate hydrolases superfamily protein	6 genes: 3 dentro de genes únicos de *Eucalyptus* detectados con dominio interpro, 2 genes CESA (de la síntesis de celulosa y xilanos) a 180 (Crom. 3) y 240 Kpb (Crom. 1), 1 gen SDRLK a 343 Kpb
					Eucgr.C01157	Heat shock protein DnaJ, N-terminal with domain of unknown function (DUF1977)	
					Eucgr.C01158	Peptidase S24/S26A/S26B/S26C family protein	
					Eucgr.C01159	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein	
cel20	Chip y Chip-GBS	9	EuBR01s34250726	1	Eucgr.A02343		
					Eucgr.A02344	villin 2	
					Eucgr.A02345	DTW domain-containing protein	
					Eucgr.A02346		
					Eucgr.A02347	Adenine nucleotide alpha hydrolases-like superfamily protein	
					Eucgr.A02348	magnesium-chelatase subunit chlH, chloroplast, putative /Mg-protoporphyrin IX chelatase, putative (CHLH)	
					Eucgr.A02349	Protein of unknown function (DUF288)	
					Eucgr.A02350		
					Eucgr.A02351		
					Eucgr.A02352	Insulinase (Peptidase family M16) family protein	
cel20	Chip y Chip-GBS	10	EuBR10s12580608	10	Eucgr.J01141	phosphatidylglycerolphosphate synthase 1	
					Eucgr.J01142	glutathione transferase lambda 2	
					Eucgr.J01143		
					Eucgr.J01144	LisH/CRA/RING-U-box domains-containing protein	
					Eucgr.J01148	SET domain-containing protein	
					Eucgr.J01149	Histone superfamily protein	
					Eucgr.J01150	hercules receptor kinase 1	

Carac.	Matriz	SNP	Nombre SNP	Crom.	Genes cercanos en ventanas de 70 Kpb (distancia menor a 35 Kpb del SNP)		Genes a menos de 500 Kpb según categorías de Myburg et al. (2014)
					Genes de *E. grandis*	Función de *Arabidopsis*	
klas20	Chip y Chip-GBS	14	EuBR09s4271448	9	Eucgr.I00199	Pectin lyase-like superfamily protein	6 genes (entre 250 y 460 Kpb): 5 dentro de genes únicos de *Eucalyptus* detectados con dominio interpro, 1 gen CESA (de la síntesis de celulosa y xilanos)
					Eucgr.I00200	photosystem I reaction center subunit PSI-N, chloroplast, putative / PSI-N, putative (PSAN)	
					Eucgr.I00201	SWIM zinc finger family protein	
					Eucgr.I00202		
					Eucgr.I00203	LIM domain-containing protein	
					Eucgr.I00204	LIM domain-containing protein	
					Eucgr.I00205	NAD(P)H: plastoquinone dehydrogenase complex subunit O	
klas20	Chip y Chip-GBS	15	EuBR04s29197264	4			
	Chip y Chip-GBS	16	EuBR04s29699828	4	Eucgr.D01601	Disease resistance protein (TIR-NBS class)	
					Eucgr.D01602	Dihydrolipoamide succinyltransferase	
					Eucgr.D01603	germin-like protein 9	
extet20- exttot20	Chip y Chip-GBS	11	EuBR11s17351865	11	Eucgr.K01423		2 genes únicos de Eucalyptus detectados con dominio interpro a 121 y 356 Kpb
					Eucgr.K01424	photosystem II reaction center PSB28 protein	
					Eucgr.K01425	Leucine-rich repeat transmembrane protein kinase family protein	
					Eucgr.K01426	indole-3-acetic acid inducible 11	
					Eucgr.K01427	Plant protein of unknown function (DUF946)	
					Eucgr.K01428	Nuclear transport factor 2 (NTF2) family protein with RNA binding (RRM-RBD-RNP motifs) domain	
					Eucgr.K01429	Tetratricopeptide repeat (TPR)-like superfamily protein	
					Eucgr.K01430	translocase of outer membrane 22-V	
					Eucgr.K01431	plasmodesmata-located protein 2	
exttot20	GBS y Chip-GBS	12	28672_21	5	Eucgr.E03257	dormancy-associated protein-like 1	
					Eucgr.E03258	protease-related	
					Eucgr.E03259	Mitochondrial substrate carrier family protein	
					Eucgr.E03260	squamosa promoter binding protein-like 5	
					Eucgr.E03261	Dof-type zinc finger DNA-binding family protein	
	GBS	13	48478_24	8	Eucgr.H03047	Lactate/malate dehydrogenase family protein	
					Eucgr.H03048	F-box family protein	
					Eucgr.H03049	F-box family protein	
					Eucgr.H03050	F-box family protein	
					Eucgr.H03052	F-box family protein	

Carac.	Matriz	SNP	Nombre SNP	Crom.	Genes cercanos en ventanas de 70 Kpb (distancia menor a 35 Kpb del SNP)		Genes a menos de 500 Kpb según categorías de Myburg et al. (2014)
					Genes de *E. grandis*	Función de *Arabidopsis*	
sg20	Chip y Chip-GBS	17	EuBR06s47777412	6	Eucgr.F03941	Integrase-type DNA-binding superfamily protein	14 genes: 6 dentro de genes únicos de Eucalyptus detectados con dominio interpro, 5 genes de la vía de síntesis de fenilpropanoides (~200 Kpb), 3 genes de la síntesis de celulosa y xilanos (DUF258, HEX y Susy a 100 Kpb)
					Eucgr.F03942		
					Eucgr.F03943	S15/NS1, RNA-binding protein	
					Eucgr.F03944	RNI-like superfamily protein	
					Eucgr.F03945	Protein kinase superfamily protein	
					Eucgr.F03947	Integrase-type DNA-binding superfamily protein	
					Eucgr.F03948	SUPPRESSOR OF AUXIN RESISTANCE 3	
	GBS y Chip-GBS	18	47543_33	8	Eucgr.H02700		
					Eucgr.H02701	F-box and associated interaction domains-containing protein	
					Eucgr.H02702	F-box and associated interaction domains-containing protein	
	GBS y Chip-GBS	19	11288_34	2	Eucgr.B03667		
					Eucgr.B03668		
					Eucgr.B03669	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	
					Eucgr.B03670	FAR1-related sequence 4	
					Eucgr.B03671	Leucine-rich repeat (LRR) family protein	
					Eucgr.B03672	Putative glycosyl hydrolase of unknown function (DUF1680)	
					Eucgr.B03673		
db20	Chip y Chip-GBS	20	EuBR05s57414738	5	Eucgr.E03380	Plant protein of unknown function (DUF247)	5 genes: 2 genes únicos de Eucalyptus detectados con dominio interpro, 2 genes de la vía de síntesis de fenilpropanoides (a 441 y 425 Kpb), 1 gen predicho de enzimas sintetizadoras de terpenos (crom. 4 a 455 Kpb).
					Eucgr.E03382	Plant protein of unknown function (DUF247)	
	GBS	21	19771_30	4	Eucgr.D00641		
					Eucgr.D00642	sensitive to freezing 6	
					Eucgr.D00643	Protein kinase superfamily protein	
					Eucgr.D00644	alpha/beta-Hydrolases superfamily protein	
					Eucgr.D00645	alpha/beta-Hydrolases superfamily protein	

A continuación, se detallan los principales genes detectados (ver tabla 3.4.6), en particular asociados a diamétro a la altura de pecho, altura total, forma de fuste, índice de rajado, contenido de lignina total y Klason, celulosa, relación S:G, extractivos y densidad básica de la madera:

Diámetro a la altura del pecho: A modo de ejemplo, uno de los tres marcadores asociados a diámetro a la altura del pecho a los 20 años, y también a dap11 con umbral menos restrictivo, fue el SNP EuBR07s10247657 en el cromosoma 7, el cual presentó cuatro genes dentro de la ventana de 70 Kpb, dos de ellos con la misma función y uno sin función descripta:

- ✓ Eucgr.G00581: relacionado con la Familia de proteínas emp24/gp25L/p24 y GOLD.
- ✓ Eucgr.G00582 y Eucgr.G00584: relacionados con las proteínas Citocromo P450.
- ✓ Eucgr.G00583: sin función descripta.

Altura total: El SNP EuBR05s25410255 asociado a at20 y ubicado en el cromosoma 5, presentó tres genes cercanos de los cuales dos, genes Eucgr.E01900 y Eucgr.E01902, se encontraron relacionados con la familia de proteínas de unión a ribonucleasa S o SBP (*S-ribonuclease binding protein*).

Forma o rectitud de fuste: Se buscaron genes cercanos para el único marcador asociado para Forma de fuste obtenido en el análisis de GWAS con la matriz conjunta y la de GBS. Dicho SNP fue el 8074_34 en el cromosoma 2 y evidenció genes cercanos. Como ejemplo, dos de ellos, Eucgr.B01920 y Eucgr.B01924, presentaron la misma anotación funcional relacionada con proteínas de la familia FH2 (formina homología 2) de unión a actina (Tabla 3.4.6).

Índice de Rajado: Se buscaron genes cercanos para el SNP EuBR08s68220660 del chip en cromosoma 8 asociado para índice de rajado en el análisis conjunto y con la matriz del Chip. Dicho SNP presentó cinco genes cercanos, entre los cuales se encontró, por ejemplo, el gen Eucgr.H04787 con función descripta de Arabidopsis para la proteína Anexina 2 (AT5G65020.1, ANNAT2; Tabla 3.4.6).

Extractivos de la madera: Se buscaron los genes cercanos a los tres SNPs asociados a extractivos etanólicos y totales, EuBR11s17351865, 28672_21, 48478_24, y se encontraron nueve genes para el primero y cinco para los dos últimos. Todos los genes, salvo uno, presentaron anotación funcional respecto de Arabidopsis, como por ejemplo el gen Eucgr.K01428 que presentó función de Proteína de la familia del factor de transporte nuclear 2 (*Nuclear transport factor 2 (NTF2) family protein with RNA binding (RRM-RBD-RNP motifs) domain*), y el gen Eucgr.E03260 que estuvo relacionado con el factor de transcripción *squamosa promoter binding protein-like 5*.

Lignina total, lignina Klason y Contenido de Celulosa: Se buscaron genes cercanos (ventanas de 70 Kpb) para los seis marcadores asociados a los caracteres fenotípicos de lig20, kla20 y cel20. Cada marcador presentó un número variable de genes dentro de dichas ventanas (Tabla 3.4.6). Ejemplos de ellos son:

- El SNP 13753_15 de GBS del cromosoma 3, que presentó asociación para lig20, cel20, y para klas20 y extet20 con umbral menos restrictivo, presentó 4 genes cercanos cada uno con función de Arabidopsis descripta en el genoma de *E. grandis*.
- El SNP EuBR09s4271448 del Chip en el cromosoma 9, que presentó asociación para klas20, y también a lig20 y cel20 con umbral menos restrictivo, presentó siete genes cercanos. De estos genes, seis tenían función de Arabidopsis descripta en el genoma de *E. grandis*, dos de ellos la misma (*LIM domain-containing protein*).
- El SNP EuBR01s3425072 del Chip en el cromosoma 1, que presentó asociación con cel20, y también con db20 (umbral menos restrictivo), presentó 10 genes cercanos. Como ejemplo, presento cercanía al gen Eucgr.A02344 con anotación funcional relacionada a la proteína Villina 2.

Relación Siringilo/Guayacilo de lignina: Para el carácter fenotípico de sg20 se buscaron genes cercanos para los tres SNPs que se encontraron asociados en las tres matrices. Estos SNPs fueron el 11288_34 de GBS, el EuBR06s47777412 del chip y el 47543_33 de GBS en cromosomas 2, 6 y 8 respectivamente. A modo de ejemplo, el marcador 11288_34 de GBS presentó siete genes cercanos, cuatro de los cuales tuvieron función descripta en Arabidopsis (Tabla 3.3.6). El SNP 47543_33 de GBS presentó tres genes cercanos, dos de los cuales presentaron función anotada, siendo esta la misma para ambos (*F-box and associated interaction domains-containing protein*, Tabla 3.4.6).

Densidad básica de la madera: De los dos SNPs asociados a db20, EuBR05s57414738 (Cromosoma 5) presentó dos genes cercanos y 19771_30 (Cromosoma 4) presentó cinco genes a su alrededor. Entre las funciones de estos últimos genes se encontró uno relacionado a proteínas de la superfamilia de las quinasas (*Protein kinase superfamily protein*), y dos genes a proteínas de la superfamilia de las hidrolasas (*alpha/beta-Hydrolases superfamily protein*). Estos dos últimos genes coinciden con el trabajo de Lamara et al. (2016) donde encontraron entre los nueve dominios de proteínas enriquecidos asociados a densidad de madera (seca al aire: *air-dry wood density*) enzimas del metabolismo de carbohidratos como las glicosil hidrolasas (ver apartado de discusión).

Genes relacionados con categorías según Myburg et al. (2014)

Los 100 genes descriptos en el genoma de *E. grandis* que se encontraron dentro las ventanas de 70kb alrededor de los SNPs asociados, se contrastaron con los genes descriptos en el genoma de *E. grandis* como

específicos de la producción de biomasa lignocelulósica y metabolitos secundarios y aceites según la anotación del genoma llevada a cabo por Myburg et al. (2014). Dichos autores catalogaron a los genes en categorías según su función en las vías de síntesis de Terpenos, celulosa y xylanos, genes SDRLK (*S-domain receptor like kinase* o Receptor de la subfamilia del dominio S tipo quinasa), de la ruta biosíntesis de fenilpropanoides, factores de transcripción MAD, genes con el motivo de secuencia KBOX, genes únicos de eucaliptos detectados con dominio interpro (base de datos de familias, dominios y sitios funcionales de proteínas), peroxidasas y lacasas, entre otros. Entre estos 100 genes, no se encontró coincidencia con ninguno de los genes descriptos por Myburg et al. (2014).

Sin embargo, cuando se amplió la búsqueda de genes a ventanas de 1 Mb alrededor de los SNPs asociados, se encontraron 1538 genes, dentro de los cuales 50 coincidieron con genes dentro de las categorías propuestas por Myburg et al, (2014). A continuación, se detallan los genes cercanos encontrados para cada carácter y dentro de las categorías descriptas (ver tabla 3.4.6).

Diámetro a la altura del pecho a los seis años (dap6): Para esta característica se encontraron 38 genes en el cromosoma 5 próximos a un SNP del Chip. Uno de estos genes se encontró dentro de las categorías descriptas en Myburg et al. (2014) (a una distancia del SNP de 196 Kpb) dentro de la correspondiente a genes predichos de enzimas sintetizadoras de terpenos.

Diámetro a la altura del pecho a los 20 años (dap20): Para esta característica se encontraron 135 genes en los cromosomas 4, 3, 7 próximos a un SNP de GBS y 2 del Chip. Tres genes se encontraron dentro de las categorías descriptas en Myburg et al. (2014) según (a una distancia del SNP de 345 Kpb), uno dentro de los 968 genes únicos de eucaliptos detectados con dominio interpro.

Altura total a los 20 años (at20): Para altura total se encontraron 53 genes próximos a un SNP del Chip, en el cromosoma 5. Tres se encontraron dentro de las categorías descriptas en Myburg et al. (2014) y dentro de los denominados genes SDRLK (~400 Kpb).

Forma de fuste a los 6 años (for6): Para esta característica se encontraron 55 genes en el cromosoma 2, próximos al SNP de GBS. Tres de estos genes se encontraron dentro de las categorías descriptas en Myburg et al. (2014) según: dos dentro de los 968 únicos de eucaliptos detectados con dominio interpro y un gen dentro de los predichos de enzimas sintetizadoras de terpenos a 200 Kpb.

Índice de rajado en rollizo a los 20 años (ir20): Para índice de rajado se encontraron 83 genes próximos al SNP del Chip en el cromosoma 8. Entre estos genes, ocho se encontraron dentro de las categorías descriptas en Myburg et al. (2014) según: seis dentro de los 968 únicos de eucaliptos detectados con dominio interpro,

uno de la síntesis de celulosa y xilanos (DUF264(a 54 Kpb) y un gen predicho de enzimas sintetizadoras de terpenos (a 184 Kpb).

Celulosa y lignina total (cel20 y lig20): Para esta característica se encontraron 234 genes próximos a un SNP de GBS (cromosoma 3, SNP compartido por ambas características) y dos del Chip (cromosoma 1 y 10, sólo asociado a cel20), cinco de dichos genes se encontraron dentro de las categorías descriptas en Myburg et al. (2014) según: tres dentro de los 968 únicos de eucaliptos detectados con dominio interpro, dos genes CESA (ruta de síntesis de celulosa y xilanos) en el cromosoma 3 y 1 (a 180 y 240 Kpb de distancia, respectivamente) y un gen SDRLK próximo (343 Kpb).

Lignina Klason (klas20): Se encontraron 96 genes en las proximidades de tres SNPs del Chip. Seis de estos genes se encontraron dentro de las categorías descriptas en Myburg et al. (2014) (entre 250 y 460 Kpb) según: cinco dentro de los 968 únicos de eucaliptos detectados con dominio interpro y un gen DUF249 dentro de los involucrados en la biosíntesis de celulosa y xilanos.

Extractivos etanólicos y totales (extet20 y exttot20): Se encontraron 174 genes próximos al SNP del Chip, en el cromosoma 11, compartido por ambas características y a los dos SNPs de GBS, en el cromosoma 5 y 8. De estos genes dos fueron reconocidos entre los 968 únicos de eucaliptos detectados con dominio interpro, a 121 y 356 Kpb.

Relación Siringilo/Guayacilo (sg20): Para esta característica se encontraron 246 genes dentro de la ventana de 1 Mb, de los cuales 148 genes fueron próximos a los dos SNP a partir de GBS en el cromosoma 2 y 8; y 98 genes próximos al SNP del Chip, del cromosoma 6. Dentro de estos 246 genes, 14 se encontraron dentro de las categorías descriptas en Myburg et al. (2014), de los cuales seis pertenecieron a la categoría de los 968 únicos de eucaliptos detectados con dominio interpro, cinco a la vía de síntesis de fenilpropanoides (CCR2 (a 78 Kpb), HCT1, HCT2, HCT3, HCT4 (todos a ~200 Kpb)), tres a la categoría de síntesis de celulosa y xilanos (DUF258, HEX (*Hexoquinasa)*, Susy (*Sucrose synthase* o sacarosa sintasa, a 100 Kpb)).

Densidad básica de la madera a los 20 años (db20): Para esta característica se encontraron 130 genes próximos a 2 SNP. Entre ellos, 59 cercanos al SNP de GBS (cromosoma 4) y 71 al SNP proveniente del Chip (cromosoma 5). Cinco genes se encontraron dentro de las categorías descriptas en Myburg et al. (2014) según: dos dentro de los 968 únicos de eucaliptos detectados con dominio interpro, dos en la vía de biosíntesis de fenilpropanoides (COMT30, COMT31 (a 441 y 425 Kpb), y un gen predicho de enzimas sintetizadoras de terpenos (crom. 4 a 455 Kpb).

3.4.2 Selección Genómica

En este apartado se presentan en primer lugar las predicciones de las características obtenidas con el BLUP tradicional que utiliza la matriz de relaciones genéticas teóricas mediante el conocimiento de pedigrí denominada ABLUP, así como el cálculo de las heredabilidades en sentido estricto para cada carácter fenotípico.

Además, las predicciones tradicionales (ABLUP) se compararon con las predicciones obtenidas mediante las metodologías que incluyen datos genómicos en la construcción de las matrices de relaciones (o relaciones genéticas realizadas) (GBLUP) y también mediante metodologías mixtas, que involucran matrices de relaciones compuestas por datos de pedigrí y de relaciones calculadas en base al genotipado (H/ssGBLUP), en el caso que los árboles genotipados sean una proporción del total analizado para una característica determinada.

Dichas comparaciones, se realizaron en base a los cálculos de la Exactitud Teórica (ET, evaluación de la varianza del error en la predicción) y Habilidad Predictiva (HP, correlación entre el fenotipo observado y estimado) de cada modelo. Asimismo, se analizan las ET y HP considerando el sistema de genotipado utilizado, es decir, derivadas de las matrices de GBS, Chip y Chip-GBS.

Exactitud teórica

En la tabla 3.4.7 se detallan los valores de las medias y desvíos estándares de la Exactitud Teórica obtenidos para los 14 caracteres mediante los modelos de ABLUP, ssGBLUP y GBLUP. Para estas dos últimas metodologías dicha tabla expone las medias y d.e de las ET obtenidas para cada una de las tres matrices genotípicas disponibles.

En la misma tabla se presentan las pruebas t pareadas (valor p <0.05) entre los diferentes enfoques ABLUP, ssGBLUP y GBLUP dentro de cada matriz genotípica. Las diferencias significativas se señalan de la siguiente manera: todos los valores de medias de ET de ABLUP presentan un superíndice “a” y dentro de cada matriz para las medias de ET en HBLUP (ssGBLUP) o GBLUP presentan un superíndice “a” si no existen diferencias significativas con ABLUP, “b” si existen diferencias significativas y “c” si el valor de GBLUP es significativamente diferente a ABLUP y HBLUP.

Tabla 3.4.7: *Exactitud teórica calculada para los tres modelos predictivos BLUP según las tres matrices genómicas.* ET-Chip, ET-GBS y ET-Chip-GBS: Exactitudes teóricas calculadas con las matrices del Chip, GBS y Chip-GBS, respectivamente; A_{Medias}, H_{Medias} y G_{Medias}: medias de las 10 repeticiones de validación cruzada de ABLUP, H/ssGBLUP y GBLUP, respectivamente, para cada carácter; d.e.: Desvíos estándares de la ET para cada modelo; $\widehat{h}^2$: heredabilidad en sentido estricto calculada en base a ABLUP; Características de crecimiento: dap6, dap11 y dap20: diámetro a la altura del pecho a los 6, 11 y 20 años, at6 y at20: altura total a los 6 y 20 años, for6: forma de fuste a los 6 años; Características de calidad de madera a los 20 años de edad: ir20: índice de rajado, extet20: extractivos etanólicos, exttot20: extractivos totales, klas20: lignina klason, lig20: lignina total, sg20: Siringilo/Guayacilo, cel20: celulosa, db20: densidad básica; "a", "b", "c": superíndices indicando si existen diferencias significativas entre modelos dentro de cada matriz genotípica, misma letra: no hay diferencias, letras diferentes: existen diferencias significativas.

Caracter	$\widehat{h}^2$	ET	ET-Chip		ET-GBS		ET-Chip-GBS	
		Amedias (d.e.)	Hmedias (d.e.)	Gmedias (d.e.)	Hmedias (d.e.)	Gmedias (d.e.)	Hmedias (d.e.)	Gmedias (d.e.)
dap6	0,24	0,341[a] (0,015)	0,435[b] (0,018)	0,324[a] (0,028)	0,442[b] (0,021)	0,323[a] (0,029)	0,425[b] (0,018)	0,309[c] (0,03)
at6	0,35	0,373[a] (0,01)	0,452[b] (0,017)	0,12[c] (0,06)	0,461[b] (0,017)	0,162[c] (0,076)	0,442[b] (0,016)	0,103[c] (0,031)
for6	0,26	0,347[a] (0,01)	0,425[b] (0,017)	0,144[c] (0,16)	0,429[b] (0,017)	0,118[c] (0,03)	0,414[b] (0,016)	0,115[c] (0,046)
dap11	0,30	0,215[a] (0,034)	0,307[b] (0,03)	0,308[b] (0,024)	0,342[b] (0,029)	0,339[b] (0,025)	0,298[b] (0,028)	0,297[b] (0,023)
dap20	0,33	0,228[a] (0,022)	0,293[b] (0,035)	0,307[b] (0,024)	0,347[b] (0,028)	0,355[b] (0,027)	0,287[b] (0,034)	0,299[b] (0,025)
at20	0,08	NA[a]	0,144[a] (0,062)	0,166[a] (0,054)	0,174[a] (0,07)	0,16[a] (0,058)	0,153[a] (0,063)	0,145[a] (0,053)
ir20	0,73	0,306[a] (0,025)	0,361[b] (0,024)	0,343[b] (0,023)	0,372[b] (0,028)	0,353[b] (0,028)	0,362[b] (0,026)	0,345[c] (0,025)
extet20	0,43	0,236[a] (0,033)	0,353[b] (0,022)	0,331[c] (0,02)	0,333[b] (0,022)	0,301[c] (0,025)	0,347[b] (0,025)	0,324[c] (0,022)
exttot20	0,31	0,204[a] (0,037)	0,336[b] (0,022)	0,305[c] (0,019)	0,307[b] (0,022)	0,261[c] (0,027)	0,33[b] (0,025)	0,297[c] (0,021)
klas20	0,45	0,242[a] (0,029)	0,377[b] (0,026)	0,363[b] (0,026)	0,372[b] (0,027)	0,362[b] (0,027)	0,368[b] (0,026)	0,354[b] (0,026)
lig20	0,53	0,253[a] (0,027)	0,382[b] (0,024)	0,366[b] (0,024)	0,388[b] (0,027)	0,376[b] (0,027)	0,37[b] (0,025)	0,355[b] (0,025)
sg20	0,49	0,248[a] (0,03)	0,348[b] (0,022)	0,363[b] (0,022)	0,36[b] (0,021)	0,376[b] (0,025)	0,347[b] (0,021)	0,364[b] (0,022)
cel20	0,10	NA[a]	0,284[b] (0,036)	0,275[b] (0,031)	0,344[b] (0,033)	0,336[b] (0,029)	0,283[b] (0,034)	0,273[b] (0,029)
db20	0,39	0,23[a] (0,023)	0,3[b] (0,023)	0,324[c] (0,022)	0,327[b] (0,02)	0,357[c] (0,025)	0,297[b] (0,021)	0,322[c] (0,02)
promedio	/	0,269 (0,025)	0,343 (0,027)	0,289 (0,038)	0,357 (0,027)	0,299 (0,033)	0,337 (0,027)	0,279 (0,028)

La heredabilidad en sentido estricto varió entre 0,08 y 0,73, siendo la más baja para altura total a los 20 años y la más alta para índice de rajado de rollizo (Tabla 3.4.7).

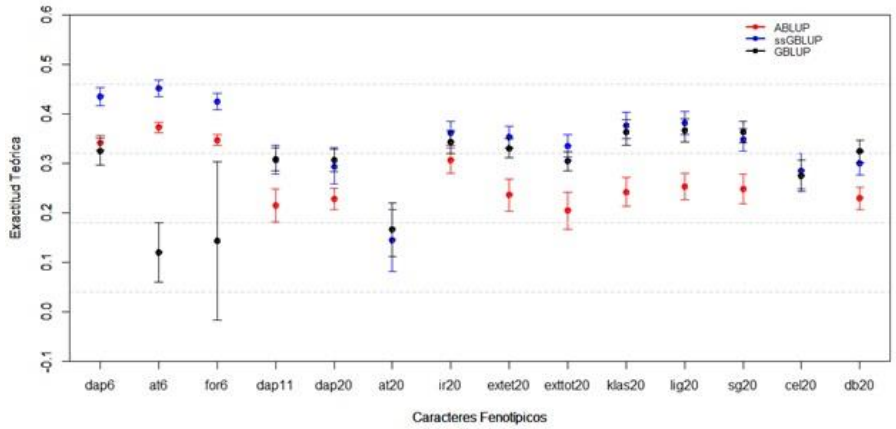
El promedio de Exactitud Teórica de las predicciones convencionales mediante ABLUP para todos los caracteres fue de 0,269, y obtuvo un máximo de 0,373 y un mínimo de 0,204 para altura total a los 6 años y

extractivos totales a los 20 años, respectivamente. Para los caracteres con menor heredabilidad (at20 y cel20, con heredabilidades de 0,08 y 0,1, respectivamente) las ET no pudieron ser calculadas (Tabla 3.4.7).

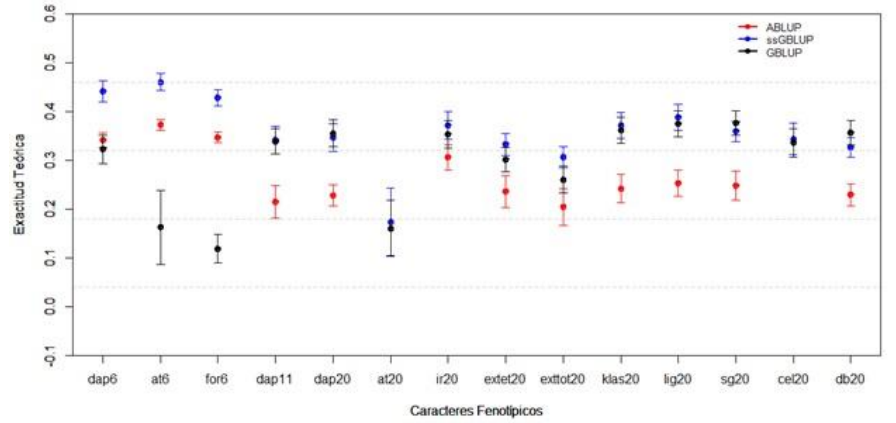
Con el objetivo de ver cómo influía la cantidad de datos fenotípicos analizados y la heredabilidad de cada característica, se realizaron diversas correlaciones. Se observó una correlación positiva entre la ET estimada mediante ABLUP y el número de individuos que presentaron datos fenotípicos por carácter ($R^2 = 0,805$). En resumen, se observó que a mayor heredabilidad del carácter y mayor cantidad de datos fenotípicos medidos se obtiene mayor ET.

Posteriormente, se compararon las ET del ABLUP respecto de los BLUP calculados con la inclusión de enfoques genómicos. A grandes rasgos y para las tres matrices, los enfoques genómicos (ssGBLUP y GBLUP) proporcionaron ET promedio más altas para todos los caracteres (Figura 3.4.8, resaltado en Tabla 3.4.7) que el enfoque ABLUP tradicional basado en pedigrí (promedio entre rasgos = 0,269). Las ET de GBLUP y ssGBLUP para los caracteres con menor heredabilidad pudieron ser calculadas, a diferencia que con el enfoque ABLUP.

a.



b.



c.

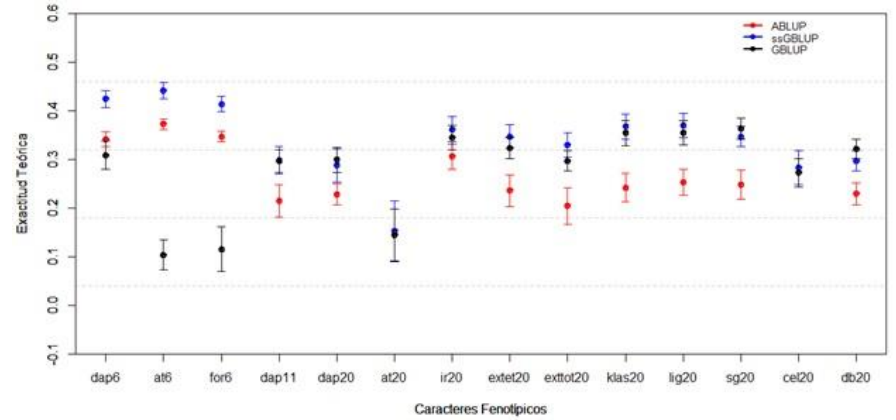


Figura 3.4.8: *Exactitud Teórica de cada modelo y para cada característica calculadas mediante las tres matrices*: Matrices: **a:** Chip, **b:** GBS y **c:** Chip-GBS. En el eje de las ordenadas, se muestran las Medias y desvíos estándares para cada carácter y cada metodología luego de 10 validaciones cruzadas de ssGBLUP y GBLUP obtenidas con matriz del Chip, y ABLUP para los 14 caracteres descriptos en el eje de abscisas. Características de crecimiento: dap6, dap11 y dap20: diámetro a la altura del pecho a los 6, 11 y 20 años, at6 y at20: altura total a los 6 y 20 años, for6: forma de fuste a los 6 años; Características de calidad de madera a los 20 años de edad: ir20: índice de rajado, extet20: extractivos etanólicos, exttot20: extractivos totales, klas20: lignina klason, lig20: lignina total, sg20: Siringilo/Guayacilo, cel20: celulosa, db20: densidad básica.

ssGBLUP presentó valores significativamente mayores de ET comparado con ABLUP para todos los caracteres y con las tres matrices, excepto para altura total a los 20 años. Esto último, podría ser debido a que la medición de altura a dicha edad presenta menor precisión que al medirla en otras edades, lo que fue evidenciado también por la baja heredabilidad obtenida para esta característica.

En general, también GBLUP presentó valores significativamente mayores de ET para diez características fenotípicas comparado con ABLUP. Por otro lado, para at6 y for6 los valores de ET para GBLUP fueron significativamente menores que para ABLUP, para dap6 sólo fue significativamente menor con la matriz de Chip-GBS y, por último, para at20 no presentaron diferencias significativas con ninguna de las tres matrices. Las diferencias observadas para los caracteres medidos a los 6 años se presentan de acuerdo con lo esperado ya que el modelo de GBLUP sólo contempló datos de individuos genotipados y fenotipados (~280), y ABLUP incluyó mayor cantidad de individuos con datos fenotípicos (~1520, ver tabla 2.4.1 en apartado 2.4.3 en Materiales y Métodos), lo que mejora el valor de ET.

En este sentido, como ya se mencionó, se puede concluir que las relaciones de parentesco "realizadas" o calculada con los índices de relaciones (Van Raden) y utilizando los SNP mejoraron las ET en comparación con las relaciones basadas en pedigrí.

Comparando la ET de los modelos de ssGBLUP vs GBLUP (considerando las tres matrices), la mayoría de las características no presentaron diferencias significativas (ocho de las 14). De este modo, ssGBLUP presentó valores significativamente mayores para las características de dap6, at6, for6, extet20 y exttot20 (e ir20 sólo para la matriz de Chip-GBS). Estas diferencias entre las ET de los modelos genómicos podrían ser explicadas por una mayor cantidad de datos fenotípicos implicados en el modelo de ssGBLUP respecto de GBLUP. Aunque ssGBLUP presentó una ET significativamente menor para db20 respecto de GBLUP, ssGBLUP presentó una mayor ET general.

Habilidad predictiva

Por otro lado, también se evaluó la Habilidad Predictiva para comparar los modelos predictivos y el desempeño de cada matriz genotípica. Con relación a la HP promedio del modelo ABLUP para todos los caracteres, ésta fue de 0,163, y varió en los 14 caracteres entre un 0,032 (altura total a los 20 años) y 0,278 (índice de rajado de rollizo). Estos valores mínimos y máximos son coincidentes con los valores mínimos y máximos de heredabilidad, lo que concuerda con la alta correlación observada entre HP y h^2 ($R^2 = 0,795$).

En la Tabla 3.4.8 y Figura 3.4.9 se puede observar que el valor de HP promedio fue levemente mayor para el modelo de ABLUP (0,163) que aquellos valores promedio obtenidos para ssGBLUP y GBLUP (0,145 y 0,111 para Chip, 0,138 y 0,105 para GBS, 0,146 y 0,113 para Chip-GBS, respectivamente). Esto es lo contrario

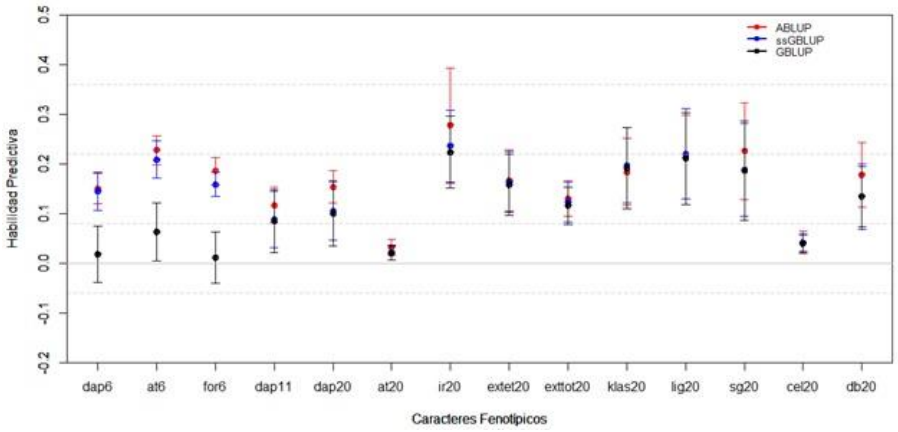
a lo observado con la ET, aunque se mantuvo la relación de valores mayores observados en ssGBLUP con respecto a GBLUP, en particular en aquellos caracteres que presentaban mayor cantidad de datos fenotípicos (aquellos medidos a los 6 años).

En la misma tabla se presentan las pruebas t pareadas (valor p <0.05) entre los diferentes enfoques ABLUP, ssGBLUP y GBLUP dentro de cada matriz genotípica. Las diferencias significativas se señalan de la siguiente manera: todos los valores de medias de HP de ABLUP presentan un superíndice “a” y, dentro de cada matriz, las medias de HP en HBLUP (ssGBLUP) o GBLUP presentan un superíndice “a” si no existen diferencias significativas con ABLUP, “b” si existen diferencias significativas y “c” si el valor de GBLUP es significativamente diferente a ABLUP y HBLUP.

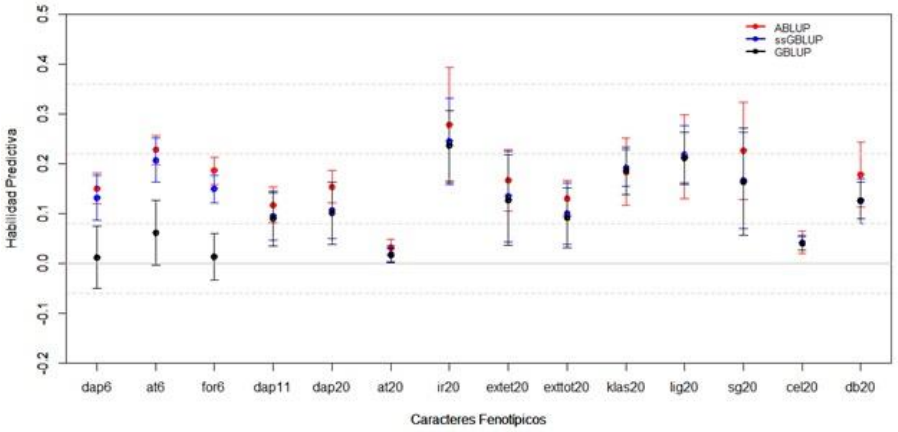
Tabla 3.4.8: *Habilidad predictiva calculada para los tres modelos predictivos BLUP según las tres matrices genómicas*. HP-Chip, HP-GBS y HP-Chip-GBS: Exactitudes teóricas calculadas con las matrices del Chip, GBS y Chip-GBS, respectivamente; A_{Medias},H_{Medias} y G_{Medias}: medias de las 10 repeticiones de validación cruzada de ABLUP, H/ssGBLUP y GBLUP, respectivamente, para cada carácter; d.e.: Desvíos estándares para cada modelo; $\widehat{h}^2$: heredabilidad en sentido estricto calculada en base a ABLUP; Características de crecimiento: dap6, dap11 y dap20: diámetro a la altura del pecho a los 6, 11 y 20 años, at6 y at20: altura total a los 6 y 20 años, for6: forma de fuste a los 6 años; Características de calidad de madera a los 20 años de edad: ir20: índice de rajado, extet20: extractivos etanólicos, exttot20: extractivos totales, klas20: lignina klason, lig20: lignina total, sg20: Siringilo/Guayacilo, cel20: celulosa, db20: densidad básica. "a", "b", "c": superíndices indicando si existen diferencias significativas entre modelos dentro de cada matriz genotípica, misma letra: no hay diferencias, letras diferentes: existen diferencias significativas.

Caracter	$\widehat{h}^2$	**HP**	**HP-Chip**		**HP-GBS**		**HP-Chip-GBS**	
		Amedias (d.e.)	**Hmedias (d.e.)**	**Gmedias (d.e.)**	**Hmedias (d.e.)**	**Gmedias (d.e.)**	**Hmedias (d.e.)**	**Gmedias (d.e.)**
dap6	0,24	0,15[a] (0,031)	0,145[a] (0,038)	0,018[b] (0,056)	0,131[a] (0,045)	0,012[b] (0,063)	0,144[a] (0,04)	0,017[b] (0,06)
at6	0,35	0,228[a] (0,029)	0,209[a] (0,037)	0,063[b] (0,059)	0,207[a] (0,044)	0,061[b] (0,065)	0,212[a] (0,036)	0,068[b] (0,056)
for6	0,26	0,186[a] (0,027)	0,159[b] (0,024)	0,011[c] (0,052)	0,149[b] (0,028)	0,013[c] (0,047)	0,16[b] (0,025)	0,011[c] (0,05)
dap11	0,30	0,117[a] (0,036)	0,088[a] (0,057)	0,085[a] (0,063)	0,094[a] (0,048)	0,09[a] (0,055)	0,09[a] (0,057)	0,086[a] (0,063)
dap20	0,33	0,154[a] (0,032)	0,105[b] (0,058)	0,1[b] (0,066)	0,106[b] (0,057)	0,101[b] (0,063)	0,105[b] (0,06)	0,101[b] (0,067)
at20	0,08	0,032[a] (0,016)	0,021[a] (0,015)	0,02[a] (0,014)	0,018[b] (0,017)	0,016[b] (0,013)	0,021[a] (0,016)	0,02[a] (0,014)
ir20	0,73	0,278[a] (0,116)	0,236[a] (0,073)	0,224[a] (0,072)	0,245[a] (0,087)	0,236[a] (0,071)	0,257[a] (0,065)	0,246[a] (0,063)
extet20	0,43	0,167[a] (0,062)	0,164[a] (0,061)	0,158[a] (0,062)	0,134[a] (0,091)	0,127[a] (0,091)	0,159[a] (0,066)	0,153[a] (0,066)
exttot20	0,31	0,13[a] (0,036)	0,123[a] (0,04)	0,116[a] (0,038)	0,1[a] (0,062)	0,092[a] (0,06)	0,119[a] (0,043)	0,113[a] (0,041)
klas20	0,45	0,184[a] (0,068)	0,197[a] (0,076)	0,191[a] (0,082)	0,192[a] (0,037)	0,186[a] (0,048)	0,199[a] (0,072)	0,193[a] (0,079)
lig20	0,53	0,214[a] (0,085)	0,22[a] (0,091)	0,211[a] (0,093)	0,219[a] (0,058)	0,211[a] (0,052)	0,22[a] (0,082)	0,212[a] (0,085)
sg20	0,49	0,226[a] (0,098)	0,188[a] (0,094)	0,187[a] (0,1)	0,167[a] (0,097)	0,164[a] (0,108)	0,186[a] (0,093)	0,185[a] (0,098)
cel20	0,10	0,042[a] (0,023)	0,042[a] (0,018)	0,039[a] (0,017)	0,041[a] (0,015)	0,04[a] (0,013)	0,042[a] (0,017)	0,04[a] (0,016)
db20	0,39	0,178[a] (0,065)	0,134[a] (0,066)	0,134[a] (0,061)	0,125[b] (0,045)	0,126[b] (0,037)	0,133[a] (0,061)	0,134[a] (0,056)
promedio	/	0,163 (0,052)	0,145 (0,053)	0,111 (0,06)	0,138 (0,052)	0,105 (0,056)	0,146 (0,052)	0,113 (0,058)

a.



b.



c.

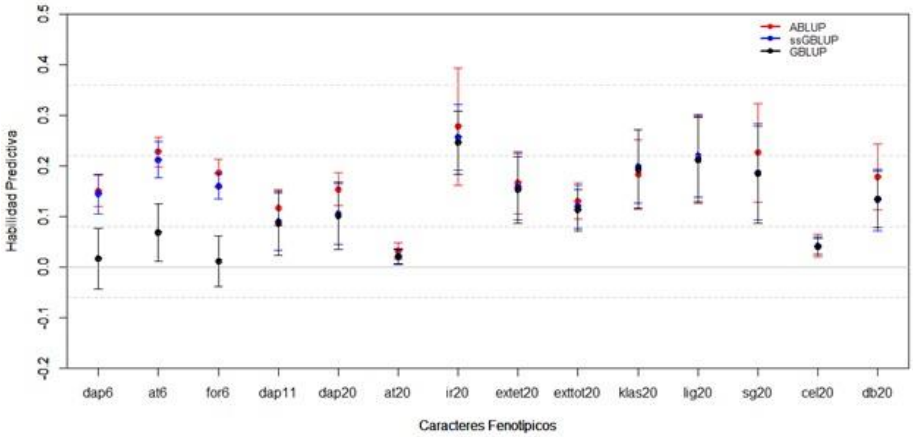


Figura 3.4.9: *Habilidad predictiva de cada modelo y para cada característica calculadas mediante las tres matrices*: Matrices: **a:** Chip, **b:** GBS y **c:** Chip-GBS. En el eje de las ordenadas, se muestran las Medias y desvíos estándares para cada carácter y cada metodología luego de 10 validaciones cruzadas de ssGBLUP y GBLUP obtenidas con matriz del Chip, y ABLUP para los 14 caracteres descriptos en el eje de abscisas. Características de crecimiento: dap6, dap11 y dap20: diámetro a la altura del pecho a los 6, 11 y 20 años, at6 y at20: altura total a los 6 y 20 años, for6: forma de fuste a los 6 años; Características de calidad de madera a los 20 años de edad: ir20: índice de rajado, extet20: extractivos etanólicos, exttot20: extractivos totales, klas20: lignina klason, lig20: lignina total, sg20: Siringilo/Guayacilo, cel20: celulosa, db20: densidad básica.

Por un lado, ABLUP presentó valores de HP significativamente mayores que ssGBLUP para las características fenotípicas de: for6, dap20, at20 (sólo con GBS) y db20 (sólo con GBS)

Por otro lado, ABLUP presentó valores de HP significativamente mayores que GBLUP para las características fenotípicas de: dap6, at6, for6, dap20, at20 (sólo con GBS) y db20 (sólo con GBS). El hecho de que GBLUP haya presentado menores valores de HP es de acuerdo con lo esperado, ya que, como ya se mencionó, el modelo de GBLUP sólo contempló datos de individuos genotipados y fenotipados (~280), y ABLUP y ssGBLUP incluyeron mayor cantidad de individuos con datos fenotípicos (~1520, ver tabla 2.4.1 en apartado 2.4.3 en Materiales y Métodos), lo que mejora las predicciones de los modelos.

A pesar de que ABLUP presentó valores mayores de HP, sólo cuatro características fenotípicas fueron significativamente diferentes a los valores del modelo de ssGBLUP y seis a los de GBLUP.

En general, se puede concluir que los modelos que incluyen datos genómicos son prometedores a la hora de ser aplicados en los programas de mejoramiento genético, en particular en el de *E. dunnii*, ya que presentan HP similares respecto de ABLUP. Además, las ET de los modelos genómicos fueron superiores a las de ABLUP. De este modo, pueden ser utilizados para generar una clasificación de los individuos de *E. dunnii* según las prioridades del programa de mejoramiento, por ejemplo, para seleccionar individuos con menor índice de rajado y mayor crecimiento. Así, a los fines prácticos, estos modelos podrían ser aplicados en el Huerto Semillero Clonal derivado de la población de *E. dunnii* estudiada para la elección de los mejores padres de la próxima generación (semillas), y luego aplicarlos a los hijos en una edad temprana (plantines ya genotipados), contando sólo con los datos genotípicos (y los modelos de ssGBLUP y GBLUP) para elegir cuales de ellos llevar a campo.

Comparación del desempeño de metodologías genotipificación para Selección Genómica

Respecto a los valores promedio de ET (ssGBLUP y GBLUP), se observó que la matriz de GBS presentó los mayores valores para todos los caracteres, al contrastar los resultados de las predicciones a partir de las tres matrices genotípicas. A estos valores le siguieron las Es promedio obtenidas con la matriz del Chip y luego la de Chip-GBS (Tabla 3.4.7).

Al comparar los valores de HP obtenidos con las tres matrices, la matriz conjunta fue la que presentó los mayores valores promedio de HP en comparación con las otras dos matrices, mientras que GBS presentó el menor promedio de las tres (Tabla 3.4.8).

Con el objetivo de comparar los desempeños de las metodologías de genotipificación de GBS y Chip, por un lado, se compararon los valores de ET obtenidos con la matriz de GBS versus la matriz del Chip, tanto para

el modelo de ssGBLUP como GBLUP. Estas comparaciones se realizaron en base a pruebas t pareadas, cuyos valores de significancia se pueden observar en la tabla 3.4.9. De este modo, se observó que para ocho de las 14 características no hubo diferencias significativas. Para dap11, dap20, cel20 y db20, GBS presentó ET significativamente mayores respecto del Chip tanto para ssGBLUP como para GBLUP. La matriz del Chip generó ET significativamente mayores para ssGBLUP y GBLUP sólo para las características fenotípicas de exttot20 y extet20, esta última sólo para GBLUP.

Por otro lado, se compararon los desempeños de las metodologías de genotipificación de GBS y Chip mediante los valores de HP obtenidos para ssGBULP y GBLUP, no encontrándose diferencias significativas entre valores de HP para ninguna de las características fenotípicas y en ninguno de los modelos predictivos.

Tabla 3.4.9: *Significancias de las diferencias entre GBS y Chip respecto de los valores de Exactitud Teórica y Habilidad Predictiva para los modelos ssGBLUP y GBLUP.* Se muestran los valores de significancia de pruebas t pareadas entre valores obtenidos con matriz de GBS versus Chip para: Exactitud teórica para ssGBLUP y GBLUP, Habilidad Predictiva para ssGBLUP y GBLUP. Características de crecimiento: dap6, dap11 y dap20: diámetro a la altura del pecho a los 6, 11 y 20 años, at6 y at20: altura total a los 6 y 20 años, for6: forma de fuste a los 6 años; Características de calidad de madera a los 20 años de edad: ir20: índice de rajado, extet20: extractivos etanólicos, exttot20: extractivos totales, klas20: lignina klason, lig20: lignina total, sg20: Siringilo/Guayacilo, cel20: celulosa, db20: densidad básica. Se resaltan en letra negrita los valores significativos ($p < 0,05$).

	Exactitud Teórica		**Habilidad Predictiva**	
	ssGBLUP	GBLUP	ssGBLUP	GBLUP
dap6	0,456	0,902	0,480	0,834
at6	0,261	0,184	0,875	0,935
for6	0,612	0,636	0,392	0,940
dap11	**0,018**	**0,012**	0,812	0,870
dap20	**0,002**	**0,001**	0,945	0,981
at20	0,327	0,830	0,731	0,603
ir20	0,376	0,387	0,808	0,723
extet20	0,052	**0,009**	0,408	0,389
exttot20	**0,008**	**0,001**	0,337	0,300
klas20	0,696	0,947	0,869	0,868
lig20	0,571	0,426	0,968	0,991
sg20	0,229	0,226	0,634	0,626
cel20	**0,001**	**0,000**	0,916	0,965
db20	**0,015**	**0,006**	0,720	0,734

4 DISCUSIÓN

1. Desarrollo de una Metodología de Genotipificación Masiva para *E. dunnii*

Las metodologías de secuenciación de ADN asociadas al sitio de restricción se están convirtiendo en las estrategias más populares para la generación de datos genómicos para una variedad de aplicaciones relacionadas con el mejoramiento genético de cultivos y especies arbóreas (Parchman et al., 2018). Sin embargo, para el género *Eucalyptus*, el uso de métodos derivados de RADseq es escaso (Grattapaglia et al., 2011; Aguirre et al., 2019; Durán et al., 2018). Hasta la fecha, el fácil acceso al microarreglo o chip comercial de SNP (EUChip60K) ha llevado a los investigadores a utilizarlo en el análisis del género (Cappa et al., 2019; Durán et al., 2018; Müller et al., 2019; Suontama et al., 2019), en lugar de los métodos derivados de RADseq. Sin embargo, algunas especies están poco representadas en este chip con respecto a *E. grandis* (que tiene mayor representación por su importancia económica en el país en donde se desarrolló el microarreglo y mayor distribución en el mundo), como por ejemplo *E. cladocalyx* que presentó sólo el ~ 6% de los SNPs informativos (~ 3,9 mil SNPs en 480 individuos, Ballesta et al., 2020). Por esta razón, las estimaciones de las frecuencias alélicas de la población y las relaciones genéticas entre los individuos pueden verse afectadas para dichas especies (Albrechtsen et al., 2010; Bajgain et al., 2016; Li & Kimmel, 2013). Las metodologías basadas en RADseq y GBS tienen el potencial de evitar este tipo de sesgo (Poland et al., 2012), pero la experiencia en *Eucalyptus* reportada hasta la fecha no es alentadora. De hecho, Durán et al. (2018) aplicaron GBS en 500 individuos de *E. globulus* y solo obtuvieron 2.597 SNP polimórficos entre ellos. El bajo número de marcadores sugiere que el protocolo debería mejorarse para este género a fin de obtener suficientes marcadores con gran cobertura del genoma para realizar estudios a nivel poblacional, como mapeo por asociación y selección genómica, entre otros. Además, existen inconvenientes técnicos asociados con GBS y sus protocolos derivados, ya que éstos solo pueden enriquecer poblaciones de fragmentos de ADN secuenciados que están por debajo de ~ 350 pb. También, con dichos protocolos se generan altos niveles de datos perdidos. La metodología de RADseq, por otro lado, involucra más pasos y equipamiento, así como mayores cantidades de ADN inicial, y muestra una alta variación de cobertura entre *loci* (revisados en Poland & Rife, 2012, y Scheben et al., 2017).

En este trabajo se describe la optimización de un protocolo derivado de ddRADseq para el genotipado de *E. dunnii*. A diferencia de estos últimos métodos, ddRADseq usa dos enzimas y reduce el subconjunto de fragmentos muestreados, lo que permite una mayor reproducibilidad, una mayor cobertura por *locus*, tamaños de fragmentos a secuenciar más grandes y una definición de SNPs más confiable (Peterson et al., 2012; Scheben et al., 2017). Aunque la proporción de alelos nulos aumenta en comparación con RADseq (Andrews et al., 2016), todas las características mencionadas hacen que el genotipado con ddRADseq sea una estrategia más apropiada.

Con esto en mente, se desarrolló un protocolo ddRADseq modificado (P1) de bajo costo y optimizado para especies de plantas. Este protocolo se configuró con un número pequeño de muestras (solo dos individuos), y luego se puso a punto (P2) para poder aplicarlo a poblaciones de una forma más simple y factible.

Ventajas del Protocolo 1: Al igual que en el protocolo de Peterson et al. (2014), el paso de ligación del P1 involucró adaptadores universales, y la PCR se realizó antes de agrupar las muestras. Por lo tanto, la selección del tamaño de los fragmentos a secuenciar resultó ser el último paso del protocolo. El desarrollo del P1 implicó el análisis de diferentes combinaciones de enzimas y tres concentraciones y proporciones de los adaptadores. Además, en lugar de los *primers* con índices de Illumina de 6 pb de longitud, se utilizaron 16 *primers forward* y 96 *primers reverse* con un índice de 8 pb de longitud (Lange et al., 2014). Este cambio permitió una mayor multiplexación de muestras, no solo para la construcción de las genotecas, sino también para la secuenciación (hasta 1.536 muestras). Asimismo, se cambió la selección automática, normalmente utilizada en ddRADseq, por una selección manual en gel de agarosa. Todas estas modificaciones permitieron que P1 fuera fácilmente aplicable en un número muy bajo de muestras de *E. dunnii* y con un costo mínimo. Por lo tanto, esta estrategia puede extrapolarse a otras especies de plantas, convirtiéndose en una herramienta atractiva para laboratorios con presupuesto limitado.

Ventajas del Protocolo 2: A pesar de su potencial utilidad para establecer un protocolo ddRADseq en cualquier especie de planta, P1 involucró el manejo de cada muestra de forma independiente hasta casi el final del protocolo, lo que impide su uso para una gran cantidad de muestras. En este sentido, se propuso el P2 como una ampliación del P1. La adición de 24 adaptadores con códigos de barras en el paso de ligación permitió la agrupación de muestras y la aplicación de la selección de tamaños de fragmentos antes de la PCR, como el protocolo original de ddRADseq (Peterson et al., 2012). Además, el uso de diferentes longitudes de códigos de barras (4 a 9 pb, Poland et al., 2012) permitió evitar los errores de fase en la secuenciación (*phishing error*) al comienzo de las lecturas. Dichos *barcodes* se aplicaron en los protocolos de GBS y MiddRADseq (Elshire et al., 2011; Yang et al., 2016), pero no se consideraron en el de ddRADseq original (Peterson et al., 2012). Para este P2 ampliado, también se utilizó la selección automática de tamaño, lo que disminuye la posibilidad de contaminación cruzada y aumenta la precisión y consistencia al aplicar el protocolo para más de un grupo o *pool* de muestras (Andrews et al., 2016), como se informó en el protocolo de ddRADseq original (Peterson et al., 2012).

4.1.1 Consideraciones del Material vegetal, extracción y cuantificación de ADN

El primer paso y quizás el más crítico en cada método RADseq es obtener buena calidad, cantidad e integridad de ADN. Incluso ddRADseq tiene este requisito de ADN de alta calidad (Guo et al., 2018; Scheben et al., 2017). Por lo tanto, la extracción de ADN debió realizarse con un método que garantice la integridad del

ADN, y esta integridad fue verificada posteriormente (por ejemplo, utilizando un espectrofotómetro de tipo Nanodrop®). El ADN debió cuantificarse mediante un método sensible como Qubit® (Thermofisher). Por ejemplo, si el ADN se degradó, o si la cantidad fue insuficiente, los resultados podrían generar un CV más alto entre los números de lecturas obtenidos para cada muestra. Con el protocolo de extracción de ADN de CTAB, se pudo alcanzar la integridad y cantidad de ADN requerida (Inglis et al., 2018; Ver apartado 7.1.1 a 7.1.3 de anexos). Sin embargo, las altas concentraciones de ADN de buena calidad no siempre son fáciles de lograr en todas las especies. En este sentido, la cantidad inicial de ADN necesaria para el protocolo fue algo a considerar. Mientras que algunos protocolos derivados de ddRADseq dependen de una gran cantidad de material de partida (por ejemplo, 1000 ng en Kess et al., 2015), el protocolo propuesto requirió de cantidades mínimas (solo 150 ng). El CV obtenido para las primeras 24 muestras analizadas con el P2 fue más bajo (0,39), pero en el mismo orden de magnitud, que los reportados para otros enfoques ddRADseq (por ejemplo, 0,42 en Scaglione et al., 2015; 0,47 en Kess et al., 2015). Además, el CV obtenido estuvo claramente influenciado por la muestra *E. dunnii*-1.202.SD, que tuvo el menor número de lecturas secuenciadas y, por lo tanto, la menor cantidad de marcadores genotipados (Ver apartado 7.1.1 a 7.1.3 de anexos).

4.1.2 Evaluación de enzimas y rango de selección de tamaño del protocolo de ddRADseq

Con respecto a los criterios para la selección de enzimas, algunos autores han propuesto seleccionar enzimas para una especie determinada basándose únicamente en la predicción *in silico*, mientras que otros han sugerido el uso de enzimas universales (por ejemplo, después de hacer una evaluación *in silico* de muchas enzimas). Por ejemplo, Yang et al. (2016) informaron el uso de un único par de enzimas universal, AvaII-MspI, para todas las angiospermas, incluyendo *Eucalyptus*. Según los resultados de la presente tesis, la evaluación de combinaciones de enzimas (una de corte frecuente y otra de corte raro) a través del análisis *in silico* e *in vitro* fue un paso esencial para optimizar ddRADseq en nuevas especies (por ejemplo como Wang et al., 2017). Debido a la ausencia de un genoma de referencia de *E. dunnii*, se utilizó el genoma de referencia de una especie del mismo género (*E. grandis*) para la predicción *in silico*. Sin embargo, si la especie en estudio careciese de un genoma de referencia (o una especie que puede tomarse como referencia debido a su proximidad), la predicción *in silico* también se podría hacer en base a información sobre el contenido de GC y el tamaño del genoma (Lepais & Weir, 2014).

Otro paso crítico que debió ajustarse fue la ventana de la selección de tamaño. Primero, según estudios previos (Kess et al., 2015; Scaglione et al., 2015), si la selección de tamaño para un protocolo derivado de RADseq se realizara en gel, y con un marcador de peso molecular de 100 pb, es aconsejable el uso de rangos múltiplos de 50 o 100 pb para minimizar los errores de escisión manual. Por lo tanto, se evaluaron (*in silico*) ventanas de 100 o 150 pb en un rango entre 220 y 470 pb de fragmentos de ADN de interés (rango de tamaños en el gel de 350 a 600 pb) para la selección manual en P1. Por otro lado, al utilizar un equipo de selección de

tamaño automático, la amplitud de la ventana fue determinada por la capacidad del equipo utilizado en este estudio (es decir, en Sage ELF 2%, la ventana de tamaños de fragmentos a seleccionar fue de alrededor de 70 pb, sagescience.com). Este último tamaño de ventana fue comparable a la selección de tamaño "ancho" (72 pb) aplicada en el protocolo ddRADseq original (Peterson et al., 2012).

Por otro lado, los tamaños finales de los fragmentos de la genoteca no deberían ser demasiado pequeños (es decir, <200 pb) para evitar la superposición de las secuencias PE, lo que daría como resultado una sobreestimación de los SNPs al analizarlos de acuerdo con las estrategias de llamado de variantes, con o sin genoma de referencia (*de novo*). Esto se debe a que Stacks v1.48 considera las lecturas PE como *loci* independientes (es decir, el software no ensambla en contigs aquellas lecturas que se superponen sólo en los extremos, Catchen et al., 2013). Sin embargo, los fragmentos tampoco deberían ser demasiado largos (es decir, > 800 pb), porque generarían una baja calidad de bases en la secuenciación PE de Illumina (Tan et al., 2019).

En comparación con MiddRADseq (Yang et al., 2016), aquí también se utilizó la predicción *in silico* para evaluar la selección de tamaño. Sin embargo, en vez de una ventana de 300 pb (400–700 pb), en el presente trabajo se seleccionó una ventana más estrecha de 70 o 100 pb (en el rango de 320–420 pb en la selección manual y una media de 370 pb para la selección automatizada). En una publicación reciente, Kess et al. (2015) también informó el uso de un tamaño de ventana de 300 pb. No obstante, la elección de rangos más reducidos evita un posible sesgo de amplificación por PCR, que aumentaría al usar fragmentos con diferentes longitudes, mientras que disminuiría la cantidad y la calidad de los datos (DaCosta & Sorenson, 2014; Quail et al., 2008). Además, con una ventana más estrecha serían necesarias menos lecturas por muestra para alcanzar una cobertura media óptima por *locus*.

4.1.3 Optimización y aplicación del protocolo de ddRADseq en E. dunnii

En términos del número de *loci* ddRADseq generados con P1, se obtuvieron un 50% más de *loci* por muestra que los *loci* predichos (74.640 *loci* promedio por muestra versus 49.016 *loci* esperados). Además, después de filtrar el catálogo mediante *rxstacks*, se obtuvieron 41.834 ddRADseq *loci*, siendo este un número más cercano al predicho esperado, con solo un 15% menos de *loci*. Por otro lado, al usar P2, se obtuvo una media de 68.622,38 *loci* ddRADseq por muestra. Este resultado duplicó la predicción *in silico* esperada (34.634). De acuerdo con Scaglione et al. (2015), este fenómeno puede deberse a la posibilidad estocástica de cada individuo de producir *loci* que estén fuera del rango esperado (Catchen et al., 2013). Es decir, el número obtenido de *loci* ddRADseq presentó variabilidad entre muestras, mostrando una correlación más alta con el número de lecturas por muestra (r^2: 0,8742) que con la cobertura media por muestra (r^2: 0,3654). Además, también deberían considerarse las diferencias en los tamaños del genoma entre las especies (640Mb para *E. grandis* versus

530Mb para *E. dunnii*; Grattapaglia & Bradshaw Jr., 1994), y las estructuras de los genomas, ya que solo alrededor del 82% de las lecturas de *E. dunnii* se ubicaron con éxito en el genoma de referencia de *E. grandis*.

Con respecto a la metodología de selección de tamaño, se aplicó la selección tanto manual como automática. El P1, contempló la escisión manual en geles en agarosa para alcanzar una metodología de bajo costo, como en Scaglione et al. (2015) y MiddRADseq (Yang et al., 2016), mientras que el P2 consideró el uso del dispositivo SAGE ELF. En la mayoría de las publicaciones, los investigadores usan Pippin-prep como el método automático de elección (por ejemplo, Peterson et al., 2012). Sin embargo, el equipo SAGE ELF fue más fácil de configurar y proporcionó una recuperación de ADN más estricta y más alta en comparación con BluePippin (Heavens et al., 2015). Como se esperaba, al comparar el método de selección de tamaño manual (P1) y el método automático (P2) se vió que el método automático recuperó un menor número de *loci* (74.640 vs 68.622 *loci*, respectivamente). Dicho menor número de *loci* podría deberse al rango de tamaño más estrecho utilizado en P2.

Otro punto crítico para considerar en la optimización de la metodología fue la concentración de adaptadores. En este trabajo, se probaron diferentes concentraciones (datos no mostrados), y finalmente se eligieron concentraciones similares a las reportadas en Scaglione et al. (2015). Algunos protocolos evaluaron diferentes concentraciones de adaptadores por titulación, cómo en GBS y ddRADseq (Elshire et al., 2011; Peterson et al., 2012). Sin embargo, este procedimiento no sería necesario para especies con genomas por debajo de 20 Gb, como el eucalipto (Yang et al., 2016), ya que es posible utilizar un exceso de adaptadores para su ligación adecuada con los fragmentos de ADN de interés, como se utilizó en los P1 y P2. Por otro lado, el adaptador Y utilizado para el extremo con el sitio de restricción de la enzima de corte frecuente, generó genotecas ddRADseq donde el Adaptador 1 y el Adaptador 2 están en extremos opuestos de cada fragmento amplificado. Este tipo de construcción de genotecas reduce la complejidad de la porción del genoma a secuenciar (Peterson et al., 2012; Peterson et al., 2014; Poland et al., 2012; Yang et al., 2016). Dado que el P1 desarrollado solo requirió un par de adaptadores por par enzimático, se evitó así una inversión sustancial de fondos al comienzo del ensayo.

Respecto del P2, éste incluyó adaptadores con códigos de barras, como en los protocolos originales ddRADseq y MiddRADseq (Peterson et al., 2012; Yang et al., 2016) y este agregado simplificó los pasos posteriores. Por ejemplo, se pudieron agrupar muchas muestras en la misma genoteca, reduciendo así el número de PCRs simultáneas a una sola. Además, se utilizaron específicamente 24 códigos de barras de diferente longitud diseñados para el protocolo *two-enzyme* GBS (Poland et al., 2012), y el mismo concepto fue utilizado en el protocolo de GBS original (Elshire et al., 2011) y en MiddRADseq (Yang et al., 2016), pero no en el de ddRADseq original (Peterson et al., 2012). El uso de códigos de barras con diferente longitud evita el error de fase (baja calidad de secuencia). Este error ocurre cuando todas las bases al comienzo de las lecturas

son iguales en todos los *clusters* (cada *cluster* genera una lectura de secuenciación Illumina) debido al sitio de restricción. En el P1, este problema fue resuelto utilizando al menos el 5% de PhiX, como se describió en el protocolo de Peterson et al. (2014), o mezclando las genotecas ddRADseq en la misma corrida de secuenciación NGS con otros tipos de bibliotecas con mayor variabilidad de nucleótidos en las primeras bases.

En la etapa de PCR, se puede lograr un nivel adicional de multiplexación mediante el uso de índices en ambos *primers*, los cuales permiten la inclusión de más genotecas en la misma corrida de secuenciación. Esta es una de las principales singularidades de los protocolos P1 y P2. Ambos protocolos, utilizaron en el paso de PCR los índices duales desarrollados por Lange et al. (2014) dentro de los *primers forward* (16) y *reverse* (96). Estos índices combinatorios permitieron multiplexar hasta 1.536 muestras/genotecas en una misma corrida NGS. En este sentido, estos protocolos no están limitados por el número de índices Illumina, como en otros métodos ddRADseq (Peterson et al., 2012; Peterson et al., 2014; Yang et al., 2016). Secuenciadores de menor rendimiento como el MiSeq (Illumina Inc.) brindan de 24 a 30 millones de lecturas por lo que sólo admiten 24 genotecas ddRADseq a una profundidad óptima. Por el contrario, el uso de secuenciadores de mayor rendimiento permite mayor nivel de multiplexación y así la reducción del costo por muestra/genoteca. Con el P2, se podrían multiplexar hasta 36.864 muestras (24 códigos de barras × 1536 cebadores duales) ampliando enormemente la capacidad de análisis y reduciendo el costo. Por ejemplo, este número de genotecas ddRADseq se podrían secuenciar a baja profundidad en un secuenciador NovaSeq, que ofrece un número máximo de lecturas de 20 mil millones (para una celda de flujo S4 dual ejecutada en el Sistema NovaSeq 6000, Illumina Inc.).

Debido a la ausencia de un genoma de referencia para *E. dunnii*, se decidió trabajar con el programa Stacks (Catchen et al., 2013) en ambas estrategias, *de novo* y con referencia (llamados *ref_map.pl* y *denovo_map.pl*, respectivamente, en el software), para comparar los resultados obtenidos con el P1. Por lo tanto, se pudieron aplicar ambos análisis para identificar SNP y SSR con alta precisión después de aplicar filtros bioinformáticos y de calidad estrictos. Como se esperaba, el análisis *de novo* recuperó más *loci* y marcadores ddRADseq, ya que todas las lecturas se consideraron para la identificación de los marcadores, en contraste con el análisis con referencia, que solo consideró las lecturas que mapearon contra el genoma de *E. grandis* (82% de las lecturas). Esta misma estrategia de análisis se implementó a la población completa de mejoramiento de *E. dunnii*, a la cual se le aplicó el P2. En este sentido, también se observó el mismo porcentaje promedio de mapeo contra el genoma de *E. grandis* para los 308 individuos, y por lo tanto la misma tendencia en la obtención de la cantidad de *loci*, siendo mayor el número obtenido con el análisis sin genoma de referencia. Ambos protocolos lograron una cobertura óptima (10–20 ×; Andrews et al., 2016) y, en consecuencia, se pueden utilizar de manera eficiente para descubrir *loci de novo*. Sin embargo, esta estrategia requiere criterios y parámetros más estrictos al definir los *loci*, debido a la mayor cantidad de falsos positivos obtenidos con este método (Rochette &

Catchen, 2017). Por lo tanto, para análisis posteriores (Prueba piloto de P2 y aplicación a la población completa), se consideraron los SNPs descubiertos por el análisis con la referencia de *E. grandis*. No obstante, estos resultados son alentadores para explorar estos SNP exclusivos de *E. dunnii* y poder descubrir variantes propias.

Entonces, en base a los análisis con genoma de referencia, usando la información de las muestras A y B generadas con el P1, se identificaron 7.346 SNPs compartidos entre las dos muestras. Al aplicar P2 en 24 individuos de *E. dunnii*, a pesar de que se identificó un total de 138.624 SNPs, después de descartar los marcadores con un alto porcentaje de datos perdidos y un MAF menor a 0,05, 15.950 SNPs pasaron el filtro. Estos marcadores mostraron una distribución homogénea en los 11 cromosomas. El límite de datos perdidos por SNP del 20% fue aplicado en base a un trabajo previo con estrategias de imputación para datos ddRADseq (Merino, 2018), para utilizar como referencia estricta. Finalmente, al utilizar el P2 para la población completa de Ubajay (308 individuos de *E. dunnii*), se decidió aumentar la profundidad de 3 a 6 lecturas de ddRADseq utilizadas para definir un *locus*, para que sea más robusta la definición *loci* heterocigotas (Rochette & Catchen, 2017). Además, se disminuyó el valor de corte de MAF de 0,05 a 0,01, como se discutirá más adelante. El total de SNPs identificados para la población completa fue de 530.885 SNPs, y luego de los filtros (hasta 20% de datos perdidos y MAF> 0,01 por SNP) quedaron 8.170 SNPs, y del mismo modo, mostraron una distribución en los 11 cromosomas del genoma de referencia de *E. grandis*. Dicho número final de SNPs fue menor al compartido por las 24 muestras de la prueba piloto, debido a que al aumentar la cantidad individuos a analizar, disminuye la proporción de *loci* compartidos entre los mismos por la posibilidad estocástica de cada individuo de producir *loci* que están fuera del rango esperado. Esto último genera la gran cantidad de datos perdidos esperada para este tipo de metodologías de representación reducida del genoma (Catchen et al., 2013; Scaglione et al., 2015). Sin embargo, esta cantidad de SNPs (8.170) fue tres veces mayor que la reportada para 500 individuos de *E. globulus* (2.597 SNP; Durán et al., 2018).

Respecto del número de SNPs por *locus* identificados utilizando ambos protocolos, se observó que la media del P1 fue de 1,95 SNPs dentro de 145 pb entre los individuos A y B de *E. dunnii* (es decir 1 SNP cada 74 pb) y de 1,73 SNP y 2,7 SNPs dentro de 66 pb en 24 y 308 individuos, respectivamente, para el P2 (1 SNP cada 38 y 24 pb, respectivamente). Sin embargo, la diferencia entre P1 y P2 podría deberse a la disparidad en el número de individuos analizados con cada protocolo (2 individuos frente a 24 o 308, respectivamente). Por lo tanto, P2 generó frecuencias de polimorfismo (o densidad de SNPs por *locus*) más altas. Aunque no hay información reportada para *E. dunnii*, estas frecuencias están en el mismo rango que las observadas para otras especies del género *Eucalyptus*. De esta manera, Hendre et al. (2011) informaron 1 SNP cada 65 pb en intrones y cada 108 pb en exones en *E. camaldulensis*, mientras que Külheim et al. (2009) detectaron dentro de genes (exones e intrones) 1 SNP por cada 33 pb, 31 pb, 16 pb y 17 pb para *E. nitens*, *E. globulus*, *E. camaldulensis* y *E. loxophleba*, respectivamente.

La compatibilidad de los SNP obtenidos entre plataformas y la robustez en la identificación y llamado de los SNPs derivados de protocolos ddRADseq son críticos, y es un área poco estudiada hasta ahora. Solo un trabajo (Campbell et al., 2017) comparó el desempeño entre HiSeq y NextSeq para la identificación de SNPs derivados de ddRADseq en un género de mariposas. En el presente trabajo, se secuenciaron las mismas 23 muestras con las plataformas de secuenciación MiSeq y NextSeq. Como se esperaba, las genotecas NextSeq (P2) presentaron más *loci* que aquellas secuenciadas en MiSeq, debido a la mayor profundidad de lectura por *locus* (4,49 × vs. 11,56 ×, respectivamente) y la menor generación de datos faltantes. Ambas plataformas de secuenciación lograron una alta calidad de datos, y según el informe de FastqC, se observó que más del 96% de las lecturas generadas pasaron los filtros de calidad *Stacks*. Por todo lo antedicho, los bajos valores de coeficiente de disimilitud entre las réplicas de las muestras (0,05) confirmaron una alta confiabilidad, a pesar de las diferencias entre las construcciones de las genotecas y las plataformas de secuenciación (además de haber sido secuenciadas con plataformas y profundidades diferentes, se aplicó selección de tamaños manual y automática a las genotecas secuenciadas en MiSeq y NextSeq, respectivamente).

Entre las bondades del P1 se destaca que puede usarse cuando se aplica en un laboratorio una metodología GBS/ddRADseq por primera vez en una especie. Su bajo costo se basa principalmente en el uso de adaptadores universales para cada enzima, como los utilizados por Peterson et al. (2014), el uso de cebadores con 1.536 combinaciones de índices duales y la implementación de una selección de tamaño final por electroforesis en gel de agarosa. Además, dependiendo del enfoque de la investigación, las secuencias generadas para un pequeño número de muestras (al menos dos) podrían ser suficientes para obtener información de nuevos marcadores. Es interesante observar que, si bien el costo por SNP genotipado en una matriz o una técnica derivada de NGS disminuye cuando aumenta el número de SNP interrogados, no todos los estudios genómicos se basan en el genotipado de una gran cantidad de marcadores. Debido al balance de costos, muchos estudios poblacionales, principalmente relacionados con el manejo de la diversidad genética, la conservación y la evolución, dan prioridad a aumentar el número de individuos muestreados, en lugar de agregar más marcadores. Un buen ejemplo de esto es el uso de secuencias para la identificación de SSRs especie específicos, y más aún, de SSRs polimórficos y en estado heterocigosis. Los métodos RADseq implican NGS y, en consecuencia, las lecturas pueden usarse para diseñar cebadores. Estos SSR podrían usarse para la genotipificación de una población mediante el uso una estrategia de menor rendimiento como la electroforesis capilar fluorescente. De esta manera, mediante el P1 se pudieron identificar a través del programa MISA (Thiel et al., 2003) 420 SSRs, 16 de los cuales fueron polimórficos entre las muestras A y B. Mediante dicho análisis utilizando la referencia, se vio además su distribución en el genoma y su presencia en casi todos los cromosomas. Un análisis más completo de la identificación de SSR utilizando datos MiddRADseq se puede encontrar también en la publicación de Qin et al. (2017).

En resumen, después de establecer el protocolo inicial P1 para la especie de interés, en este caso *E. dunnii*, el P2 se pudo utilizar para un mayor número de muestras y aplicarlo a la población completa de la especie (308 individuos), del programa de mejoramiento de INTA. La incorporación de adaptadores con códigos de barras diseñados permitió agrupar 24 muestras en la misma biblioteca y simplificar los pasos siguientes en el protocolo. Al igual que con el protocolo ddRADseq original, el enfoque descripto aquí se puede usar con más de una combinación de enzimas de restricción diferentes, y así obtener otras subpoblaciones de fragmentos y representación de otras porciones del genoma.

2. Caracterización Genotípica de la Población de Mejoramiento de *E. dunnii*

4.2.1 Aplicación del microarreglo EUChip60K en E. dunnii

Respecto de la aplicación del microarreglo EUChip60K (Chip) a toda la población de *E. dunnii*, la cantidad de SNPs polimórficos obtenidos (19.045 SNPs, hasta 20% de datos perdidos y MAF> 0,01 por SNP) fue similar a la esperada según se reportó en el trabajo del desarrollo de dicha plataforma (Silva-Junior et al., 2015) donde se informaron 17.014 SNP con MAF > 0,01 para esta especie. Al igual que en Resende et al. (2017b), menos del 50% de los 64 mil SNPs disponibles en el Chip fueron polimórficos para la población de Ubajay (alrededor del 30% de los SNPs). Este valor fue acorde a lo esperado debido a la naturaleza multiespecie del Chip, en el cual no todos los SNPs fueron informativos al aplicarlo en una especie que no fue la principal que se consideró para la construcción del Chip (Silva-Junior et al., 2015).

Otros trabajos reportaron un porcentaje similar o mayor de SNPs, debido a que lo aplicaron a otras especies más representadas en el Chip y a un mayor número de individuos. Tal es el caso del trabajo de Cappa et al. (2019), quienes obtuvieron 33.398 SNPs (MAF > 0,01) para 999 árboles de híbridos entre *E. grandis* × *E. urophylla* y *E. grandis* × *E. camaldulensis*. Otro ejemplo es el trabajo de Suontama et al. (2019) que aplicaron el Chip a 691 individuos de *E. nitens* y reportaron 12.236 SNPs (MAF > 0,01, filtro de SNPs con $r^2 > 0,9$). *E. nitens* tuvo la misma representación que *E. dunnii* en el desarrollo del EUChip60K (5% de los árboles; Silva-Junior et al., 2015), por lo cual es comparable con el presente trabajo en el que se obtuvo mayor cantidad de marcadores polimórficos.

4.2.2 Comparación de las metodologías de genotipado en E. dunnii

Al comparar la metodología del Chip con la de ddRADseq (o GBS), la primera presentó una proporción mucho menor de datos perdidos, conforme a lo esperado debido a las cualidades de cada metodología. Sin embargo, la desventaja de la gran cantidad de datos perdidos obtenida con las metodologías de GBS puede ser superada mediante la aplicación de métodos de imputación, como fue observado al aplicar este tipo de genotipificación en especies forestales, como por ejemplo en *Picea glauca* (El-Dien et al., 2015).

Así, en el presente trabajo se aplicó el algoritmo imputación LD-kNNi, implementado en LinkImpute, a ambas matrices de datos genotípicos. Se obtuvieron precisiones altas de asignación de genotipos a los datos faltantes (0,8949 para la matriz de GBS con 8.170 SNPs; y de 0,8443 para la matriz del Chip con 19.045 SNPs). El programa utilizado mostró mayor velocidad, precisión comparable, y menor sesgo en las estimaciones de frecuencia de alelos al ser comparado con varios métodos de imputación de genotipos, entre ellos los más utilizados (Money et al., 2015). Asimismo, un algoritmo basado también en el uso de la información del genotipo vecino más cercano, kNN-Fam, mostró junto con el algoritmo de EM (Expectation Maximization) las mayores precisiones al imputar datos de GBS para SG en *Picea glauca* (El-Dien et al, 2015) al compararlos con los métodos de imputación por la media y por descomposición de valor singular (*singular value decomposition*). Por otro lado, Rutkoski et al. (2013) compararon cuatro métodos de imputación para su aplicación en selección genómica, y observaron que el método de regresión aleatoria forestal (*random forest regression*) produjo una precisión superior, seguida por el método kNNi, y el de menor precisión fue el de imputación por la media. Además, concluyeron que incluir marcadores con una gran proporción de datos faltantes casi siempre condujo a mayores precisiones en SG, incluso cuando no se conoce el orden de los marcadores (Rutkoski et al., 2013). Al igual que en el presente trabajo, en otro estudio en *Eucalyptus* (*E. cladocalyx*) también se aplicó el algoritmo de LD-kNNi, implementado en el programa TASSEL 5.2 (*Trait Analysis Association, Evolution and Linkage*; Bradbury et al., 2007), lo cual les permitió imputar datos y aplicar SG en dicha especie no modelo (Ballesta et al., 2020).

Debido a la gran cantidad de datos perdidos obtenidos con la metodología de GBS (ddRADseq), se decidió utilizar un filtrado de SNPs con MAF< a 0,01. Generalmente, los SNPs raros con baja frecuencia de alelos minoritarios (menor a 0,05) causan falsos positivos en los análisis de mapeo por asociación, especialmente si se utiliza un bajo número de muestras y rasgos que no tienen una distribución normal. En la mayoría de los trabajos aplicación de GWAS en especies forestales fue utilizado un MAF igual a 0,05 como mínimo (Lee et al., 2017), debido al poder limitado de los modelos estadísticos (revisado por Du et al., 2018). Sin embargo, una práctica recomendada es no eliminarlos ya que muchas variantes genéticas causales son raras, aunque igualmente se deben interpretar con precaución (http://www.zzlab.net/GAPIT/gapit_help_document.pdf). Al igual que en la presente tesis, algunos autores utilizaron una frecuencia alélica menor como filtro, cómo por ejemplo Lamara et al. (2016), que aplicaron un análisis de mapeo por asociación en base a genes candidato utilizando un MAF > a 0,003. Resende et al. (2017a) usaron SNPs con MAF > a 0,01, y Fahrenkrog et al. (2017) utilizaron MAF > a 0,003835 para SNPs encontrados con captura de exones. Los autores de este último trabajo afirmaron que al incluir alelos de tan baja frecuencia en las pruebas de asociación de un solo marcador detectaron una gran cantidad de asociaciones que no se encontraron al usar solo SNP comunes, superando la desventaja de aumentar el número de pruebas estadísticas. Además, destacaron las ventajas de los enfoques de genotipado flexibles, como la captura de secuencias y GBS, que permiten la inclusión de variantes de baja

frecuencia en el análisis de asociación de todo el genoma. Además, SNPs con un MAF > a 0,01 también fueron utilizados en trabajos de SG en *Eucalyptus* (Cappa et al., 2019; Suontama et al., 2019).

Aunque, en el desarrollo del EUChip60k se buscó minimizar el sesgo hacia SNP más comunes (Silva-Junior et al., 2015), al comparar las distribuciones de MAF obtenidas en ambos sistemas de genotipificación, se observó que los SNPs del Chip tienden a presentar frecuencias alélicas mayores o más comunes que los descubiertos por GBS en la población de *E. dunnii*. Esta misma tendencia fue observada en otros trabajos que compararon ambos tipos de metodologías de genotipado en la misma población de individuos. Negro et al. (2019) observaron que la distribución de MAF fue diferente entre ambas tecnologías aplicadas en maíz, donde los microarreglos utilizados mostraron una distribución bastante uniforme, y GBS presentó un exceso de alelos raros con una distribución de MAF en forma de "L", lo que significa que los valores de MAF estaban sesgados hacia valores bajos. Los autores justifican estas diferencias debido a que los microarreglos de maíz que utilizaron de 50K y 600K se desarrollaron en base a la secuenciación de un número reducido de individuos, 27 y 30 líneas respectivamente, mientras que GBS en 247 líneas, permitiendo un mayor descubrimiento de alelos raros. En el caso de *E. dunnii*, el desarrollo del EUChip60K consideró sólo 12 individuos de la especie de los 240 del genero utilizados, lo que justifica el patrón de MAF observado, similar al evidenciado en maíz.

En concordancia con la distribución de MAF, para *E. dunnii* se observó una heterocigosis (H_e) promedio menor con GBS (0,17) que con el microarreglo (0,28). Consistentemente, Negro et al. (2019) observaron la misma tendencia de valores en maíz, donde la H_e para GBS fue de 0,27 y para los microarreglos de 50K y 600K fue 0,35 y 0,34, respectivamente. La heterocigosis observada para *E. dunnii* (*Ho:* 0,29) con los datos de la matriz de Chip-GBS fue similar a la encontrada en *E. cladocalyx* con el EUchip60k (*Ho:* 0,22; Ballesta et al., 2020). Asimismo, fue más baja que la estimada para un huerto semillero de *E. dunnii* mediante nueve marcadores SSR (*Ho:* 0,66; Zelener et al., 2005), aunque esta gran diferencia podría ser debida a la naturaleza de los marcadores.

Con respecto a la distribución de los SNPs a lo largo del genoma, se observó que los SNPs obtenidos a partir del análisis de GBS presentaron un patrón más disperso y menos homogéneo que los del microarreglo. Esta misma diferencia entre distribuciones en el genoma fue observada en maíz, donde los SNPs provenientes de GBS presentaron un patrón de mayor densidad de SNPs en las regiones teloméricas, y el microarreglo de 50K exhibió una distribución más uniforme, mientras que el microarreglo de 600K mostró mayor densidad de marcadores en las regiones pericentroméricas (Negro et al., 2019). Asimismo, para *E. dunnii* no se encontraron SNPs comunes a ambas metodologías de genotipificación, concordando con un estudio en *E. globulus* (Durán et al., 2018) donde se observó que de 2.597 SNPs obtenidos con la metodología de GBS (Elshire et al., 2011) sólo 24 SNPs fueron comunes a los 13.669 SNPs polimórficos presentados por el EUChip60K. A partir de

estas observaciones, se puede concluir que ambos métodos evalúan distintas regiones genómicas, siendo estas complementarias.

Con respecto al DL (r^2) promedio para todos los cromosomas observado en *E. dunnii*, este fue bajo y similar para las tres matrices (r^2 de 0,0073, 0,0074 y 0,0071 para GBS, Chip y Chip-GBS, respecto). En especies de la misma sección Maidenaria, como *E. globulus* se observaron valores más altos, con un r^2 de 0,09 (Ballesta et al., 2018; Cappa et al., 2013). Esto sugiere un muy bajo DL en *E. dunnii*, y la necesidad de una gran cantidad de marcadores para estudios de GWAS, como los utilizados en el presente estudio. Más en detalle, el cromosoma 3 mostró el menor DL para las tres matrices (r^2= 0,0068, 0,0067 y 0,0065, GBS, Chip y Chip-GBS, respectivamente). Esta observación coincide con que dicho cromosoma presenta la mayor longitud (83,5 Mb) en el genoma de *E. grandis*. El cromosoma 3 de eucalipto tiene la métrica de expresión génica promedio más baja de cualquiera de los cromosomas de eucalipto, lo que favorece la hipótesis de que proviene de una estructura cromosómica ancestral fusionada (es decir, múltiples telómeros y centrómeros en un cromosoma) que suprime la expresión génica, recombinación y/o reordenamiento sucesivo (Myburg et al., 2014).

A pesar de las diferencias entre las distribuciones y frecuencias de SNPs provenientes de ambas metodologías, la estructura genética evaluada en la población de *E. dunnii* resultó similar independientemente de la matriz genotípica utilizada, ya que utilizando DAPC sólo dos individuos difirieron en la pertenencia a los dos grupos genéticos evidenciados detectados. Dicha correspondencia entre la estructura poblacional obtenida con las matrices de GBS y Chip también fue observada por Negro et al. (2019) en maíz y por Elbasyoni et al. (2018) para trigo de invierno. Dicha estructura genética poblacional detectada mediante el análisis de DAPC en la población de *E. dunnii*, refiere a dos poblaciones con escasa diferenciación genética entre ellas, aunque significativa. Esta podría ser explicada por causa del raleo previo de los individuos en la plantación de acuerdo con su crecimiento y forma, y también, por la composición de los distintos orígenes y procedencias geográficas de las semillas que integraron el ensayo, lo que conduce a tener frecuencias alélicas específicas (Alqudah et al., 2019). En particular el grupo genético que se diferenció pobremente del resto de la población está constituido por casi todos los individuos de la procedencia de plantaciones de Oliveros, Santa Fe, siendo semillas que provenían de una polinización abierta en la plantación local con una base genética más estrecha, adaptada a la zona, en contraste con las semillas de orígenes australianos de poblaciones naturales. En este caso, se esperaba que dicho grupo de individuos presentasen frecuencias alélicas particulares, por lo cual ignorar la corrección de la estructura de la población en el análisis de GWAS conduciría a tener asociaciones espurias entre el genotipo y el rasgo de interés. Sin embargo, la estructura genética poblacional detectada para esta población fue muy baja (0,0148) aunque significativa. Tal como sugirieron Yu et al. (2006), las relaciones de parentesco son capaces de capturar la estructura genética subyacente salvo en los casos en que existiese una obvia diferencia regional (Cappa et al, 2013). En el caso de la población de *E. dunnii* estudiada, la misma proviene de semillas de una región geográfica muy acotada, que se corresponde con la

distribución de la especie, y donde se comprueba un flujo génico alto y por lo tanto escasa diferenciación genética.

En cultivos agrícolas como el trigo, Bajgain et al. (2016), utilizaron y compararon ambos métodos de genotipificación de alto rendimiento (GBS y Chip). En dicho estudio, el enfoque GBS superó al Chip por su menor costo y tiempo de obtención de los datos, y por su capacidad de proporcionar una cobertura más amplia del genoma de trigo, incluida la del genoma D a menudo mal representado en los Chips. Para *E. dunnii*, si bien el Chip presentó una cobertura más amplia, sí se observaron diferencias de cobertura y distribución de los SNPs entre las metodologías y con respecto al genoma de *E. grandis*. La genotipificación basada en Chips de SNP, sin embargo, requiere un menor conocimiento computacional y recursos para procesar los datos. En conclusión, ambos métodos son poderosos medios para estudiar el genoma y proporcionar la resolución suficiente para llevar a cabo estudios de asociación (Bajgain et al., 2016).

3. Aplicación de metodologías genómicas para el mejoramiento molecular de *E. dunnii* mediante Mapeo por Asociación y Selección Genómica

La presente tesis describe el primer trabajo en el que se genotipificó con GBS (ddRADseq) y un microarreglo (EUChip60k) a una población de mejoramiento de *E. dunnii*, y en la cual se aplicó GWAS y SG para 14 caracteres de interés forestal. En particular es la primera vez que se aplican estás metodologías genómicas para encontrar asociaciones genéticas y predicciones respecto del índice de rajado de la madera, característica de suma importancia para el uso sólido.

4.3.1 Análisis de Asociación de Genoma Amplio (Genome Wide Association Study) en E. dunnii de Ubajay.

El tamaño de la población experimental es una consideración importante a la hora de evaluar el poder del GWAS para detectar asociaciones verdaderas. Al ir incrementando el número de individuos, se elevan las chances de detectar QTLs con efectos menores con una frecuencia aceptable dentro de la población (Alqudah et al., 2019). El tamaño de la población de *E. dunnii* estudiada (280 individuos) se encuentra sobre el límite inferior del número de individuos utilizados hasta ahora en GWAS en especies forestales, que oscila entre 303 y 1694 árboles según lo revisado por Du et al. (2018). Sin embargo, diversos estudios en otras especies de plantas aplicaron GWAS en tamaños poblacionales menores, como por ejemplo 270 variedades de arroz por Wu et al. (2015), 250 líneas de trigo por Bajgain et al. (2015), 237 variedades de manzano por Lee et al. (2017) y 247 híbridos F1 (primera generación filial) de maíz por Negro et al. (2019). Cabe mencionar el alto DL que poseen las especies autógamas permite que el tamaño poblacional sea menor, sin embargo, tanto las variedades de manzano como el maíz, su DL es sensiblemente menor. Aunque es sabido que las colecciones de

germoplasma más grandes proporcionan más potencia en el análisis, en la práctica, se necesitan como mínimo de 100 a 500 individuos (J. A. Rafalski, 2010). Independientemente del método estadístico utilizado para el GWAS, disponer de un número considerable de individuos a analizar es importante para poder detectar también *loci* de efecto menor (Du et al., 2018).

El número de marcadores moleculares evaluados en este trabajo con las tres matrices, fue mayor al utilizado en otros trabajos que aplicaron mapeo por asociación en especies forestales, como 7.434 SNPs para *Picea glauca* (Lamara et al., 2016), o 7.680 DArT por Cappa et al. (2013) para *Eucalyptus globulus*, como antecedente del grupo de trabajo. Sin embargo, debido al bajo DL de este tipo de poblaciones se sugirió la utilización de mayor densidad de marcadores. Al aplicar la matriz conjunta de Chip-GBS, el número de marcadores alcanzado (27.019 SNPs) se encuentra en el mismo orden de magnitud que otros trabajos de GWAS en especies forestales, como 24.806 SNPs (EuCHIP60K) para una población híbrida de *E. grandis* × *E. urophylla* (Resende et al., 2017a) y 29.233 SNPs para 334 individuos no relacionados de *Populus trichocarpa* mediante un microarreglo (Porth et al., 2013). Sin embargo, algunos trabajos aplican GWAS con un número mucho mayor de SNPs, como Kainer et al. (2019), que obtuvieron 2,39 millones de SNPs mediante la resecuenciación a baja profundidad de los genomas de 468 individuos de *Eucalyptus polybractea*, analizándolos para rasgos relacionados con terpenos y biomasa.

El análisis de GWAS puede ser realizado mediante varios programas estadísticos siendo los más importantes y utilizados: TASSEL (Bradbury et al., 2007) que es el más común y libre, GenStat (de no libre disponibilidad, https://genstat.kb.vsni.co.uk/), PLINK (C. C. Chang et al., 2015) y GAPIT (*Genomic Association and Prediction Integrated Tool*; Lipka et al., 2012) que permiten analizar gran cantidad de caracteres fenotípicos y genotípicos. Sin embargo, TASSEL y GAPIT fueron más robustos que PLINK al comparar el mismo análisis variando el número de marcadores a utilizar (Yan et al., 2019). GAPIT es uno de los muchos programas en R que permiten aplicar GWAS, y presenta muchas ventajas, ya que reduce el tiempo computacional sin disminuir el poder estadístico (Alqudah et al., 2019). Éste se utilizó para los tres análisis de GWAS en el presente estudio. El mismo implementa un Modelo Lineal Mixto (MLM) que incluye la matriz de parentesco como el propuesto por Yu et al. (2006). Sin embargo, GAPIT utiliza una matriz de parentesco grupal calculada a partir de individuos agrupados mediante el denominado MLM comprimido (CMLM) propuesto por Zhang et al. (2010). Dicho modelo es más eficiente desde el punto de vista computacional. Además, la resolución de un MLM utilizando el enfoque tradicional de máxima probabilidad restringida es computacionalmente intensiva, debido a que en un análisis de GWAS el número típico de puntos de datos genotípicos supera los cientos de millones. Por lo tanto, GAPIT también utiliza el algoritmo de modelo mixto eficiente de asociación (*efficient mixed model association* o EMMA) (Kang et al., 2008) que reduce esta carga computacional del MLM. Al igual que GAPIT, el paquete de R EMMA (Kang et al., 2008) y el programa

TASSEL utilizan el algoritmo EMMA para GWAS. TASSEL además también implementa CMLM más el algoritmo P3D. Lipka et al., (2012) compararon estas tres herramientas, y obtuvieron resultados idénticos para GWAS. Sin embargo, GAPIT resulta ser siete veces más rápido que TASSEL y 180 veces más que EMMA (Lipka et al., 2012).

Un aspecto importante de los análisis de GWAS, es la consideración en los modelos estadísticos de la estructura genética y/o relaciones de parentesco, para disminuir la aparición de falsos positivos (Cappa et al., 2013). Para controlar asociaciones espurias a través de la contabilidad de múltiples niveles de parentesco, Yu et al. (2006) desarrollaron un enfoque de modelo mixto a través de una matriz de parentesco por pares llamada matriz de parentesco (K). Ésta puede calcular la relación entre pares de individuos utilizando información genotípica. El alto valor de las relaciones entre los individuos indica una alta similitud genética, donde por ejemplo los individuos de la misma región geográfica pueden ser agrupados en un mismo grupo.

A la hora de elegir qué tipo de análisis se iba a realizar para calcular la estructura poblacional para incluir en el modelo lineal mixto utilizado en GWAS, en el trabajo de Zhao et al. (2007) se compararon los métodos bayesianos utilizados en el programa STRUCTURE y el agrupamiento mediante Análisis de Componentes Principales (PCA). Ambas metodologías agruparon de la misma manera a los individuos. La mayoría de los estudios utilizan ambos métodos (STRUCTURE y PCA) para confirmar sus resultados (revisado en Alqudah et al., 2019). Sin embargo, a pesar de que el programa STRUCTURE es ampliamente utilizado, demanda de mucha capacidad computacional y tiempo para el análisis de datos. No obstante, existe la limitación en la definición del número de clústeres y a cómo asignar individuos en grupos (Alqudah et al., 2019) y esto puede conllevar a eliminar asociaciones válidas.

La aplicación de métodos de modelos mixtos para corregir la estructura de la población utilizando la matriz PCA o K, es comúnmente una herramienta para controlar la estructura de la población en cultivos agronómicos, mientras que una combinación (Q + K) de estos enfoques parece ser la más poderosa. En algunos casos, controlar la estructura de la población utilizando Q + K puede conducir a una sobrecorrección y luego a perder información y resultados significativos (Alqudah et al., 2019). Zhao et al. (2007) destacaron que cualquier método que elimine efectivamente el sesgo o falsos positivos dados por la existencia de estructura poblacional, también eliminará efectivamente los verdaderos positivos que están fuertemente correlacionados con dicha estructura. Teniendo en cuenta que en general la estructura familiar captura cantidades sustanciales de variación causadas por la estructura de la población, incluir la estructura de la población solo sería necesario en casos donde existen grupos con diferencias genéticas bien definidas (Cappa et al., 2013).

Varios trabajos de asociación, y en particular en forestales, evaluaron y eliminaron asociaciones espurias incorporando la matriz de parentesco y/o estructura poblacional estimada por PCA como covariables en el

MLM, al igual que en el presente estudio (Lamara et al., 2016; Porth, et al., 2013). Del mismo modo que en este trabajo para *E. dunnii*, Wu et al. (2015) aplicaron GWAS en arroz utilizando el programa GAPIT y evaluaron distintos modelos CMLM considerando diferente número de componentes de PCA viendo que el mejor ajuste (analizado con BIC) fue el que no consideraba a los componentes del PCA. En dicho trabajo se encontró un mayor número de marcadores asociados por carácter (largo del mesocotilo), 99 SNPs localizados en regiones intergénicas y posiciones diferentes de 36 genes anotados relacionados (Wu et al., 2015). No obstante, la mayor cantidad de asociaciones que obtuvieron puede estar relacionada a que utilizaron de una gran cantidad de SNPs (1.019.883). Sin embargo, cada población de mejoramiento y carácter a estudiar es particular, debido a su propio LD, He, etc. combinado con la disponibilidad de barrido genómico.

El umbral de los estudios de mapeo por asociación utilizando una gran cantidad de marcadores SNP sigue siendo un tema debatido. Nakagawa (2004) sugirió que los procedimientos estándar y ajustados de Bonferroni deberían abandonarse debido a que es muy conservador y reduce el poder estadístico (Nakagawa, 2004; Wu et al., 2015). Esto es consistente con los resultados obtenidos, ya que en ninguno de los tres análisis de GWAS realizados en *E. dunnii* se detectaron SNPs que superaran dicho umbral. Por lo tanto, se recomienda el uso de la tasa de descubrimiento falso o FDR (False Discovery Rate) (Storey & Tibshirani, 2003) como una mejor referencia estadística para establecer el umbral de los *loci* asociados.

Por otro lado, a pesar de que ninguno de los SNPs asociados en la presente tesis haya superado el nivel de significancia de 0,05 de FDR, en otros trabajos fueron reportados como asociados SNPs que presentaron un FDR < 0,20 (Kainer et al., 2019; Lamara et al., 2016). En el presente trabajo los SNPs que presentaron valores de FDR menores a 0,2 en los análisis de GWAS fueron tres con la matriz de GBS, uno para el Chip (con uno igual a 0,2) y dos para la matriz de Chip-GBS. Asimismo, en un trabajo donde se compara la eficiencia de distintos programas estadísticos para asociación, como Plink, TASSEL y GAPIT, se vio que GAPIT no presentó marcadores que superaran dicha corrección (significancia de 0,05 de FDR) para las pruebas múltiples en comparación con los otros programas (Yan et al., 2019), y puede deberse a su modelo estadístico alternativo. En esta última publicación también sugieren que el establecimiento de un umbral de valor p o q para SNP significativos debería ser específico para cada programa. Por este motivo, en el presente trabajo se utilizó un umbral *ad hoc* (-log (1 E-4)) para definir SNPs asociados a cada carácter fenotípico. Para *E. dunnii* también se tuvo en consideración el umbral -log(1/n) (Wang et al., 2012; Yang et al, 2013), que fue aplicado en dos trabajos en soja (Sonah et al., 2013; Torkamaneh & Belzile, 2015), como una forma de reafirmar asociaciones válidas, que aunque los SNPs no superaran un FDR de 0,05, gran porcentaje de las asociaciones reportadas en el presente trabajo sí superaron este último umbral (37,5%, 42,8% y 42,8% para GBS, Chip y Chip-GBS respectivamente).

En el presente trabajo, se encontraron marcadores asociados que fueron compartidos entre diferentes características fenotípicas (SNP 13753_15 de GBS compartido por lignina y celulosa totales, y el SNP EuBR11s17351865 del Chip compartido entre extractivos totales y etanólicos), y esta situación también fue observada por otros autores. Como ejemplo, Rollins et al. (2013) combinando DArT y SSR y mediante el análisis de QTL en Cebada, demostraron que los genes relacionados con la fecha de *heading* tenían efectos pleiotrópicos sobre los rasgos relacionados con el rendimiento y la biomasa. Asimismo, en soja se encontró una correlación negativa entre el contenido de aceite y proteína, y que las mismas regiones genómicas estaban asociadas con ambos rasgos, siendo los alelos favorables para un rasgo desfavorables para el otro (Sonah et al., 2013). En forestales, en particular en *Picea glauca*, Lamara et al. (2016) encontraron muchos de los SNPs asociados a estadíos tempranos de desarrollo de la madera también asociados a estadíos más avanzados (38,4%). En particular, en un estadío temprano de la madera, el ángulo de microfibrilla y el módulo de elasticidad compartieron el 43% de sus marcadores asociados, concluyendo que muchos de los genes que controlan un carácter también controlan a su correlacionado (Lamara et al., 2016).

En este trabajo, se encontraron varias características fenotípicas correlacionadas significativamente entre sí, como: lignina total, lignina Klason, extractivos totales, etanólicos y celulosa total; relación de monómeros de lignina de siringilo/guayacilo con lignina Klason, celulosa y extractivos totales; densidad básica de la madera y celulosa total; diámetro a la altura de pecho y altura total a distintas edades; índice de rajado y celulosa. Estas correlaciones fenotípicas observadas tuvieron relación con los SNPs asociados compartidos entre las características fenotípicas (SNP 13753_15 de GBS compartido por lignina y celulosa totales, y el SNP EuBR11s17351865 del Chip compartido entre extractivos totales y etanólicos).

Porth et al. (2013), en el trabajo analizado de *P. trichocarpa*, encontraron que la mayoría de las asociaciones detectadas eran con marcadores con una baja frecuencia de alelos minoritarios (0,05 <MAF < 0,2). Asimismo, no detectaron una correlación entre MAF y la varianza fenotípica explicada por los SNPs (R^2, relacionado con acción aditiva, define la varianza fenotípica explicada por el marcador), encontrando grandes variaciones en los efectos alélicos para los marcadores de bajo MAF. Esto es concordante con los resultados del presente trabajo, donde tampoco se observó una correlación entre MAF y R^2 de los SNPs asociados, siendo que, por ejemplo, los dos SNPs que presentaron los valores más altos de R^2 en el GWAS con la matriz conjunta (Chip-GBS) tenían un MAF de 0,01 y R^2 de 8,5% (SNP asociado a exttot20, cromosoma 11, EuBR11s17351865) y un MAF de 0,14 y R^2 de 8,7% (SNP asociado a sg20, cromosoma 6, EuBR06s47777412). Dichos valores de variación fenotípica explicada por cada marcador fueron bajos y entre 5,9 y 8,7% (promedio de 7%). Sin embargo, se encontraron dentro de un rango y promedios similares a los encontrados en *E. globulus* por Cappa et al. (2013), que observaron entre 4,02% a 13,76% de la varianza explicada por marcador para los caracteres de diámetro a la altura de pecho, densidad básica de la madera (estimada por Pilodyn), lignina total y Klason, relación Siringilo/Guayacilo y extractivos, con, por ejemplo, promedios de 7,27% para DAP y 5,64% para la

relación Siringilo/Guayacilo de lignina. Estos pequeños valores de R^2 son consistentes con la arquitectura genética de los caracteres cuantitativos que están controlados por muchos *loci*.

Por ejemplo, Wang et al. (2015) utilizaron variantes de baja frecuencia y descubrieron un gen del factor de transcripción TCP vital para el desarrollo del zarcillo en el pepino (*Cucumis sativus L.*). Xing et al. (2015) encontraron un SNP de baja frecuencia que puede aumentar moderadamente el rendimiento y reducir la altura del tallo en el maíz. Las especies leñosas no han experimentado reducciones en la diversidad genética debido su bajo nivel de domesticación respecto de los cultivos agronómicos, y las variantes raras son abundantes en este tipo de genomas, pero menos manejables en GWAS. Estas abundantes variantes raras son realmente importantes para explicar la falta de heredabilidad de los rasgos complejos (Resende et al., 2017), como la baja heredabilidad observada para el rasgo de altura total a los 20 años. Los alelos beneficiosos raros generalmente se emplean en el mejoramiento de árboles, que una vez que se demostró útil, se someten a un barrido selectivo y se fijan en todos los principales cultivares (Du et al., 2018). Por su parte, para la detección de alelos raros y con gran efecto sobre QTL, se han desarrollado estrategias con distinto poder estadístico y que involucran el barrido genómico de fenotipos extremos. Se denominan de genotipado selectivo, e implican la evaluación fenotípica de una gran población y el genotipado restringido de los extremos de la distribución fenotípica. No obstante, Xing & Xing (2009) concluyeron que sólo se detecta variantes comunes. Sin embargo, basado en estudios llevados adelante en humanos se concluye que, en la práctica, es rentable usar un diseño de muestreo extremo, que logra una potencia similar al análisis de QTL en la mayoría de las situaciones, excepto cuando el efecto del alelo es relativamente pequeño y se propone que para maximizar el poder del análisis, puede valer la pena explorar las proporciones desequilibradas del uso de caso-control (Li et al., 2019). De esta manera, en soja, Yan et al. (2017) pudieron identificar y validar QTL de efectos mayores para el peso del grano, utilizando esta estrategia de muestreo extremo selectivo.

La formación de madera es un proceso complejo de desarrollo involucrando miles de genes (McCann & Carpita, 2008), la mayoría de los cuales tienen funciones desconocidas (Mewalal et al., 2014). Mejorar las propiedades de la madera para el bioprocesamiento y la producción de nuevos productos biológicos, sin afectar negativamente el crecimiento y la defensa, requerirá una comprensión más completa de los principales factores transcripcionales y metabólicos de la biosíntesis de la madera (Mizrachi & Myburg, 2016) en árboles adultos con fenotipos relevantes para el medio ambiente y la industria.

4.3.2 Comparación entre metodologías de GBS y Chip en el desempeño para la Asociación de Genoma Amplio (Genome Wide Association Study)

Proporcionalmente con el número de SNPs brindado por cada metodología, se encontró un mayor número de asociaciones con el microarreglo que con GBS, los cuales se presentaron en diferentes regiones del genoma.

La diferencia de distribución y densidad de SNPs observada entre ambas genotipificaciones podría explicar la falta de colocalización de ambas técnicas. Esto también fue observado por Negro et al. (2019) al comparar GBS (400K) con el microarreglo de 600K (500K informativos), señalando una alta complementariedad entre las regiones que presentaron QTLs o SNPs asociados entre GBS y el microarreglo, siendo algunas regiones de QTLs comunes a ambas metodologías pero no así los SNPs asociados. Dichos autores explicaron estos resultados no sólo por la diferencia de la distribución y densidad de SNP a lo largo del genoma brindada por cada metodología, sino porque además observaron, a una escala más fina, que los SNPs del chip de 600K y GBS podrían marcar regiones genómicas cercanas pero diferentes alrededor de los genes. Los SNPs de dicho microarreglo fueron seleccionados principalmente dentro de las regiones de codificantes de genes, mientras que los SNPs de GBS se encontraron en su mayoría en regiones codificantes, pero también gran parte en regiones reguladoras de genes. Ambas tecnologías capturaron haplotipos diferentes en las mismas regiones genómicas, siendo estas últimas regiones altamente recombinogénicas (Negro et al., 2019). Este patrón de QTLs encontrado en maíz, es similar a lo evidenciado aquí en *E. dunnii* al comparar ambas técnicas, y podría explicar que los marcadores asociados no colocalicen debido a la diferencia de distribución y densidad de SNPs, tanto por el bajo DL presente en *E. dunnii* como por el menor número de marcadores o menor cobertura genómica utilizados en el presente estudio. Aunque el genoma de maíz (2,3 Gb) es más de 4,3 veces mayor que el de *E. dunnii* (530 Mb), dichos autores genotiparon con una cantidad 30 veces mayor de SNPs. Asimismo, si se aumentara la densidad de marcadores genotipados en la población de *E. dunnii* se esperarían encontrar un porcentaje de regiones donde colocalicen SNPs asociados tanto de GBS como del microarreglo.

4.3.3 Genes próximos a los marcadores asociados

Los marcadores identificados en este estudio representan asociaciones genéticas novedosas con las 14 características fenotípicas no reportadas previamente en otras especies de *Eucalyptus*. Lee et al. (2017) obtuvieron resultados similares al ver que los genes asociados no coincidieron con precisión al comparar con otros resultados de QTL y GWAS descubiertos previamente. Dichos autores explican los resultados debido a varias razones. Una de ellas es que es probable que haya diferencias en las poblaciones analizadas, siendo este punto crucial en el presente trabajo porque es el primero que aplica búsqueda de QTLs en la especie, *E. dunnii*. Asimismo, los trabajos que utilizan poblaciones de mapeo biparental, se limitan a las variaciones presentes en los parentales y su generación F1, mientras que GWAS usa varios germoplasmas, donde las variaciones pueden ser distintas y mucho más amplias. Por otro lado, es difícil encontrar coincidencia en los marcadores, regiones o genes encontrados utilizando el enfoque GWAS si los QTL estaban dominados por un alelo raro y presente sólo en individuos particulares (Lee et al., 2017). Ninguno de los SNPs del microarreglo EUChip60k asociados en el presente trabajo coinciden, por ejemplo, con los encontrados con el mismo microarreglo en el trabajo de Müller et al. (2019) donde aplicaron GWAS a híbridos de *E. grandis* × *E. urophylla* para las características de diámetro a la altura del pecho y altura total. Esto puede ser debido a que los SNPs del microarreglo que

resultaron polimórficos para *E. dunnii*, no lo son para dichos híbridos. Sin embargo, pueden tratarse de regiones genómicas compartidas o próximas.

Los genes que se localizaron a una distancia menor a 35 Kpb (ventanas de 70 Kpb) de los marcadores asociados no se encontraron entre los descriptos por otros trabajos de asociación en *Eucalyptus* (Cappa et al., 2013; Müller et al., 2019; Müller et al., 2017) y tampoco entre los genes descriptos según las categorías de Myburg et al. (2014) para distintas vías metabólicas relacionadas con el desarrollo de la madera en *Eucalyptus*.

Sin embargo, se encontraron similitudes con otros trabajos de asociación en especies agronómicas y forestales con respecto a las anotaciones funcionales de dichos genes cercanos (a menos de 35 Kpb) a los marcadores asociados, como se describe a continuación para cada característica fenotípica evaluada. Estos genes podrían ser posibles genes candidato. No obstante, estudios más profundos se requieren para estas afirmaciones. Por tentativo que sea, estas funciones similares proporcionan una validación biológica indirecta de las asociaciones que encontramos en nuestro estudio.

Diámetro a la altura del pecho: Los dos genes Eucgr.G00582 y Eucgr.G00584 que presentaron función Citocromo P450, cercanos al marcador EuBR07s10247657 en el cromosoma 7, podrían estar involucrados en la síntesis de lignina, según se detalla a continuación. En el genoma de *Arabidopsis* hay 244 genes del citocromo P450 (y 28 pseudogenes), siendo una de las mayores familias de genes en las plantas. Los P450 frecuentemente comparten menos del 20% de identidad y catalizan reacciones extremadamente diversas que conducen a los precursores de macromoléculas estructurales como la lignina, la cutina, la suberina y la esporopollenina, o están involucradas en la biosíntesis o el catabolismo de hormonas y moléculas de señalización, de pigmentos, odorantes, sabores, antioxidantes, aleloquímicos y compuestos de defensa, y en el metabolismo de los xenobióticos (Bak et al., 2011). Fahrenkrog et al. (2017) en GWAS en *Populus deltoides* encontraron un único SNP en el gen F5H3 (POPTR_0005s11950, Potri.005G117500) asociado con la relación S:G, que codifica para ferulato - 5 - hidroxilasa (F5H), una monooxigenasa dependiente del citocromo P450 (P450) que funciona en la rama de la ruta de la lignina que conduce a la síntesis de monómeros de Siringilo de lignina (Weng & Chapple, 2010). En un estudio de GWAS en *E. globulus* (Cappa et al., 2013), también un marcador asociado a S:G se localizó muy próximo a una F5H, siendo una enzima clave implicada en la síntesis del alcohol monolignol sinapílico y, en última instancia, los restos de lignina. Por lo tanto, F5H afecta la división entre los dos monolignoles principales, coniferil y sinapil alcoholes.

Altura total: El SNP EuBR05s25410255 asociado a at20 y ubicado en el cromosoma 5, presentó tres genes cercanos de los cuales dos, genes Eucgr.E01900 y Eucgr.E01902, se encontraron relacionados con la familia de proteínas de unión a ribonucleasa S o SBP (*S-ribonuclease binding protein*). Este tipo de proteína también fue encontrada por Petzold et al. (2018) quienes determinaron en forma experimental un listado de proteínas

que estaban asociadas al desarrollo de la madera de álamo (*P. trichocarpa*), particularmente involucradas en Interacciones de proteína-proteína e interacciones de proteína-ADN relevantes para el desarrollo de plantas leñosas. Entre el listado encontró el gen Potri.012G119200 de álamo que es similar al gen AT5G45100 (SBP2) de *Arabidopsis*, y que coincide con el gen Eucgr.E01900 de *E. grandis* relacionado al mismo gen AT5G45100.1.

Forma o rectitud de fuste: Dos genes interesantes de abordar, Eucgr.B01920 y Eucgr.B01924, que se encontraron dentro de las ventanas para forma de fuste, son aquellos que presentaron la misma de anotación funcional para la proteína de la familia FH2 (formina homología 2) de unión a actina, ambos con homología al gen AT3G05470.1 de *Arabidopsis*. Las forminas son intermediarias en las cascadas de transducción de señales que afectan la reorganización del citoesqueleto. El control del crecimiento celular y la polaridad depende de un citoesqueleto de actina dinámico que tiene la capacidad de reorganizarse en respuesta a estímulos ambientales y de desarrollo (Deeks et al., 2002). Estos dos genes podrían ser posibles genes candidato que influirían en la forma final del tronco de *E. dunnii*. En este sentido, Kainer et al. (2019) en GWAS en *E. polybractea* encontraron un gen (Eucgr.J00015, cromosoma 10) que también presentó relación con Actina (anotación Actina-4) asociado al carácter fenotípico AT, siendo que at6 y for6 presentaron una correlación baja negativa, pero muy significativa para *E. dunnii* (r^2 = -0,23; $p < 0,001$).

Índice de Rajado: El gen Eucgr.H04787, dentro de la ventana del SNP EuBR08s68220660 del microarreglo en el cromosoma 8, mostró función relacionada con anexina 2 de *Arabidopsis* (AT5G65020.1, ANNAT2). Las anexinas son una familia multigénica de proteínas de unión a membrana dependientes de calcio que juegan un papel importante en la señalización de las células vegetales. Son proteínas multifuncionales, cuya función en las plantas no se comprende completamente. Wang et al. (2018) vieron que ANN1 y ANN2 juegan un papel importante en el transporte de azúcar post-floema a la punta de la raíz. La inmunolocalización de una proteína anexina de guisante condujo a la sugerencia de que las anexinas pueden desempeñar un papel en el desarrollo de las células del floema (Clark et al., 1992). Lamara et al. (2016) realizaron un análisis de enriquecimiento de dominios de proteínas Pfam en genes significativamente asociados para varios caracteres en árboles de abeto blanco (*Picea glauca*), donde anexina mostró asociación para el ancho de anillo en madera tardía. Esta concordancia reforzaría la evidencia encontrada en el gen Eucgr.H04787 asociado a ir20 en *E. dunnii*, siendo este un posible gen candidato que influiría en la varianza fenotípica del carácter, al estar asociado al desarrollo de la madera (floema y anillos de crecimiento).

Dicho SNP EuBR08s68220660 del chip asociado a ir20, también es cercano al gen Eucgr.H04785 con función designada como proteína de la familia de transportadores de nucleótido-azúcares (AT5G65000.1), y podría estar involucrado en el proceso de síntesis de xilano. Los nucleótido-azúcares se encuentran involucrados en la síntesis de xilano y su transporte a través de la membrana del aparato de Golgi. Nucleótido-

azúcares, como UDP-Xyl, UDP-GlcA y UDP-Araf, son los donantes de azúcar a las glicosiltransferasas para la síntesis del esqueleto beta- 1,4-xilano y las sustituciones de glicosilo. Un análisis genético en *Arabidopsis* ha demostrado que las mutaciones en los transportadores de UDP-xilano causan un defecto en la síntesis de xilano (Zhong et al., 2019, Figura 4.3.1). Relacionado con la síntesis de polisacáridos, Porth et al. (2013) encontraron en *P. trichocarpa*, entre los marcadores asociados a ángulo de la microfibrilla de madera, genes relacionados a procesos como la activación e interconversión de monosacáridos mediados por la actividad de la enzima de interconversión de nucleótido – azúcares.

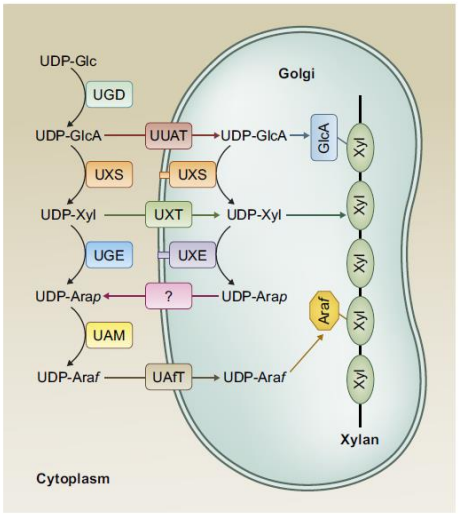


Figura 4.3.1. *Esquema de la conversión de nucleótido-azúcares involucrados en la síntesis de xilano y su transporte a través de la membrana de Golgi*. UAfT: transportador UDP-Araf; UAM: Mutasa UDP-Ara; UGD: UDP Glc 6-dehidrogenasa; UGE: UDP-glucosa 4-epimerasa; UUAT: transportador de UDP-ácido uronico; UXE: Epimerasa UDP-Xil 4-; UXS: UDP-Xil sintasa; UXT: transportador UDP-Xil. Xylan/Xyl: xilano o hemicelulosa. Cytoplasm: citoplasma. Fuente: Zhong et al. (2019).

Contenido de Lignina total, lignina Klason y Celulosa: Como ejemplo, uno de los genes encontrados, el Eucgr.I00199 de *E. grandis*, ubicado cercano al SNP EuBR09s4271448 del microarreglo en el cromosoma 9 asociado al contenido de lignina Klason, presentó similitud con el gen de *Arabidopsis* AT4G33440.1, cuya anotación se relaciona a una Proteína de la superfamilia similar a la pectina liasa. La pectina liasa es una enzima liasa que rompe las moléculas de pectina como una endoglucanasa específica de enlaces α (1->4). La pectina es un heteropolisacárido estructural contenido en las paredes celulares primarias (Voragen et al., 2009). Un gen en el cromosoma 4 relacionado con la síntesis de pectinas fue encontrado por Müller et al. (2019) al aplicar GWAS a híbridos *E. grandis* × *E. urophylla* vinculado a un SNP del microarreglo EUChip60k (EuBR04s17531959) asociado con diámetro a la altura del pecho. Este gen codifica para una galacturonosiltransferasa 4 (GAUT4, Eucgr.D00963, AT5G47780) involucrada a la biosíntesis de pectina y xilanos en las paredes celulares (de Godoy et al., 2013; Bryan et al., 2016).

Para el EuBR01s3425072 del microarreglo en el cromosoma 1 asociado a celulosa, se encontró, entre sus diez genes cercanos en una ventana de 70 Kpb, al gen Eucgr.A02344 con función Villina 2 vinculado al gen ATVLN2 de *Arabidopsis*. Las Villinas son proteínas que agrupan, cortan y nuclean la actina. Los genes ATVLN (Villina de *Arabidopsis*) se expresan en todos los órganos, con niveles de expresión elevados en ciertos tipos de células. Presentan menor expresión en los tejidos meristemáticos lo que sugiere que no es importante para la división celular sino para la diferenciación celular y, más específicamente, en los tejidos que están compuestos de células alargadas (Klahre et al., 2000). Los complejos de subunidades catalíticas de celulosa sintasa son ensamblados en el aparato de Golgi, y luego transportados hacia la membrana plasmática en un proceso mediado por actina. Se observó que la interrupción de la actina por un fármaco despolimerizante produce alteraciones en la localización en la membrana plasmática de dichos complejos de celulosa sintasa y una deposición alterada en la pared secundaria (Zhong et al., 2019). Por este motivo, variaciones en el gen Eucgr.A02344 con función Villina 2, podrían influir directamente sobre la actina y la interrupción de actina sobre la celulosa sintasa, lo que podrían estar estrechamente relacionadas a la variación en el contenido de celulosa total, siendo un posible gen candidato a considerar en el mejoramiento de *E. dunnii.*

Extractivos de la madera: Entre los genes cercanos a los tres SNPs asociados a extractivos etanólicos y totales, EuBR11s17351865, 28672_21, 48478_24, y se encontraron nueve genes para el primero y cinco para los dos últimos. El gen Eucgr.E03260 cercano al marcador 28672_21 de GBS estuvo relacionado con el factor de transcripción *squamosa promoter binding protein-like 5 o SPL.* Este tipo de proteínas representan una familia específica de factores de transcripción en plantas (Li et al., 2018). Muchos de los genes SPL son reguladores importantes para varios procesos diversos de desarrollo de las plantas, incluido el cambio a la fase vegetativa, la arquitectura de la planta, la biosíntesis de antocianinas, biosíntesis y señalización de giberelina, embriogénesis somática, y en respuesta de la planta al estrés. Por ejemplo, se encuentran 16 genes SPL en *A. thaliana* (Cardon et al., 1999), 19 en arroz (Xie et al., 2006; Yang et al., 2008) y 28 en *P. trichocarpa* (Li y Lu, 2014).

Relación Siringilo/Guayacilo de lignina: Uno de los siete genes cercanos al SNP 11288_34 de GBS en el cromosoma 2 asociado a sg20 (Eucgr.B03672), en *Arabidopsis* fue descripto como una Glucosil hidrolasa putativa de función desconocida (DUF1680). La función de este tipo de enzimas es la hidrólisis de cualquier enlace O-glicosil. La hidrólisis de los enlaces (1,3) -b-D-glucosídicos en (1,3) -b-D-glucanos es clave en el metabolismo de los carbohidratos y la organización de la pared celular (Lopez-Casado et al., 2008). Lamara et al. (2016) en *Picea glauca* encontraron entre los nueve dominios de proteínas enriquecidos asociados a densidad de madera (*air-dry wood density*) enzimas del metabolismo de carbohidratos como las glicosil hidrolasas. Esta correspondencia entre sg20 y db20 podría explicarse en base a la correlación positiva media y muy significativa evidenciada en la presente tesis (R^2: 0,30; p< 0,001; Figura 3.2.1).

Müller et al. (2019) al aplicar GWAS con EUChip60K a híbridos *E. grandis* × *E. urophylla* encontraron tres SNPs asociados a enzimas tipo Glucosil hidrolasa. Dos de los SNPs (EuBR06s6100971, EuBR04s17486529) se asociaron con Diámetro a la altura del pecho ubicados en el cromosoma 6 en el gen Eucgr.F00486 (AT5G42100) y en el cromosoma 4 en el gen Eucgr.D00955, respectivamente, y otro SNP asociado a altura total (EuBR05s70210869) en el cromosoma 5 en el gen Eucgr.E04103 (AT1G61820). Aunque los SNPs de dicho trabajo asociaron a otras características fenotípicas, en el presente trabajo se observó una correlación positiva baja pero significativa entre sg20 y at20, ambos a la misma edad de medición (R^2: 0,19; p< 0,01; Figura 3.2.1), lo que explicaría el haber encontrado enzimas de la misma familia relacionadas a la característica fenotípica de sg20. Asimismo, el primer SNP en el gen Eucgr.F00486 encontrado por Müller et al. (2019) codifica una glucano endo-1,3-b-glucosidasa. El segundo SNP se ubica en un gen (Eucgr.D00955 / AT4G17180) que codifica una proteína de la familia 17 de O-glicosil hidrolasas y esta asociación también fue vista por Du et al. (2016) en *Populus* (Potri.018G000900) para volumen de fuste (calculado en base a diámetro y altura). El tercer SNP se encontró en el gen Eucgr.E04103 (AT1G61820), que codifica una b-glucosidasa 46 (BGLU46), que también es un tipo de Glucosil hidrolasa, que puede estar involucrada en la lignificación, hidrolizando glucósidos de monolignol (Escamilla-Treviño et al., 2006). Este último tipo de Glucosil hidrolasa es el que podría estar involucrado en la asociación encontrada en el presente estudio, influyendo en la relación siringilo/guayacilo de la composición de lignina, lo que se refuerza con la correlación negativa baja pero significativa entre sg20 y klas20 (R^2: −0,22; p< 0,001; Figura 3.2.1). Asimismo, Porth et al. (2013) en *Populus trichocarpa* encontraron genes involucrados en el reensamblaje de la pared celular, en particular relacionados con glucósido hidrolasas, asociados a la composición química, propiedades físicas y ultraestructurales de la madera.

Adicionalmente, dos de los siete genes encontrados en el cromosoma 6 alrededor del SNP EuBR06s47777412 presentaron función según *Arabidopsis* de proteínas de la super familia de las integrasas de ADN (*Integrase-type DNA-binding superfamily protein*). Este tipo de proteína también fue encontrada por Petzold et al. (2018) en el listado de proteínas que evaluaron y encontraron asociadas al desarrollo de la madera de álamo (*P. trichocarpa*).

<u>*Densidad básica de la madera:*</u> De los dos SNPs asociados a densidad básica a los 20 años, EuBR05s57414738 (Cromosoma 5) presentó dos genes cercanos y 19771_30 (Cromosoma 4) presentó cinco genes a su alrededor. Entre las funciones de estos últimos genes se encontró uno relacionado a proteínas de la superfamilia de las quinasas (*Protein kinase superfamily protein*), y dos genes a proteínas de la superfamilia de las hidrolasas (*alpha/beta-Hydrolases superfamily protein*). Petzold et al. (2018) también encontraron proteínas de la superfamilia de las quinasas y de las glicosil hidrolasas entre el listado de proteínas que evaluaron y estuvieron asociadas al desarrollo de la madera de álamo (*P. trichocarpa*).

Genes relacionados con categorías según Myburg et al. (2014)

Finalmente, como se mencionó anteriormente, los 100 genes descriptos en el genoma de *E. grandis* alrededor de los SNPs asociados (a menos de 35 Kpb) no se encontraron entre los genes descriptos en el genoma de *E. grandis* como específicos de la producción de biomasa lignocelulósica, metabolitos secundarios y aceites según Myburg et al. (2014).

Por este motivo, se amplió la búsqueda de genes a ventanas de 1 Mpb alrededor de los SNPs asociados, para evaluar si se encontraban genes cercanos que coincidieran con los genes descriptos según categorías propuestas por Myburg et al. (2014). Estas ventanas permiten detectar genes que están a menos de 500 Kpb, y podrían constituir una validación independiente de las asociaciones encontradas. Además, considerando que 1cM es equivalente a 618 Kpb en *E. grandis* (Grattapaglia et al., 2015), estas asociaciones podrían ser interesantes para el mejoramiento, y con bajo potencial de perder el DL durante el proceso de mejora. Sin embargo, la comparación entre el genoma de *E. grandis* y la resecuenciación de ejemplares de *E. globulus* reveló que la considerable diferencia de tamaño entre los genomas de *E. grandis* (640 Mpb) y *E. globulus* (530 Mpb, igual tamaño y especie más cercana a *E. dunnii*) se debe en gran medida a la suma de muchas pequeñas inserciones/deleciones (InDels) y la presencia de transposones ampliamente distribuidos en todo el genoma (Myburg et al., 2014). Por lo tanto, es razonable especular que una diferencia de 500 Kpb entre el marcador y el gen con la variación causal podría ser mucho más pequeña en el genoma de *E. dunnii*. En este sentido, en un trabajo de GWAS en *E. globulus* se reportaron posibles genes candidato a una distancia de 1 Mpb del marcador asociado (Cappa et al., 2013). Asimismo, en un trabajo de GWAS en 612 individuos de *E. grandis* y *E. globulus* se evaluaron ventanas de 600 Kpb (corresponde aproximadamente a una distancia de recombinación de 1,2 cM, (Petroli et al., 2012) alrededor de 243 marcadores DArT y SNPs asociados ($p < 0,05$) a características de crecimiento, calidad de madera, dentro de las cuales pudieron identificar 4.000 genes relacionados (Villalba, 2016). Estos números concuerdan con los 1.538 genes dentro de ventanas de 1Mb de los 21 marcadores asociados de *E. dunnii*, siendo el número de marcadores menor, pero más amplias las ventanas.

De este modo, se encontraron 50 genes descriptos dentro de las categorías mencionadas (Myburg et al., 2014). Como ejemplo, se puede destacar el gen CESA de la síntesis de celulosa y xilanos (DUF264 (a 54 Kb) y el gen predicho de enzimas sintetizadoras de terpenos (a 184 Kb) cercanos al marcador del Chip del cromosoma 8 asociado a índice de rajado en rollizo. Otros genes para destacar son: el gen CESA a 184 Kpb del SNP del cromosoma 3 asociado tanto a lignina como a celulosa total, y otro gen CESA a 240 Kpb del SNP (cromosoma 1) asociado a celulosa. Estos 50 genes se presentan como posibles genes candidatos para las características estudiadas en la población de *E. dunnii* evaluada. Asimismo, como se mencionó anteriormente, estudios más profundos se requieren para estas afirmaciones.

4.3.1 Aplicación de Selección Genómica en E. dunnii de Ubajay.

En el presente estudio se comparó el desempeño de las metodologías de GBS, Chip y su combinación al aplicar distintos modelos de Selección Genómica en una población de mejoramiento de *E. dunnii*. Con este objetivo, fueron evaluados tres métodos de SG, el ABLUP tradicional, el GBLUP estándar y el ssGBLUP. Luego de una revisión bibliográfica, se concluyó que el presente trabajo fue el primero en comparar GBS y un microarreglo (EUChip60k), en la aplicación de SG en especies forestales y, el primero en aplicar ssGBLUP y GBS en *E. dunnii*.

Los tres métodos se compararon en términos de exactitud teórica y habilidad predictiva, al igual que en el trabajo de Cappa et al. (2019), que fue el primero que comparó ssGBLUP y GBLUP para la predicción de valores de cría en el mejoramiento genético forestal, en particular en el género *Eucalyptus* con el microarreglo EUChip60k. Otros trabajos también aplicaron SG en *Eucalyptus* utilizando el microarreglo EUChip60k (Durán et al., 2017; Müller et al., 2017; Suontama et al., 2019), pero ninguno de ellos aplicando el método de genotipificación de GBS.

En coincidencia con Cappa et al. (2019), los modelos que incluyeron datos genómicos mostraron ETs promedio más altas que con ABLUP. Sin embargo, para *E. dunnii*, la superioridad de las predicciones genómicas sobre las predicciones de pedigrí tradicionales fue más evidente para el enfoque de ssGBLUP que para el enfoque GBLUP (con las tres matrices), siendo lo contrario a lo observado por Cappa et al. (2019) en híbridos de *E. grandis* × *E. urophylla* y *E. grandis* × *E. camandulensis*. Cappa et al. (2019) observaron igual desempeño de ssGBLUP y GBLUP en los caracteres que presentaban el mismo número de individuos fenotipados y genotipados. Esto mismo fue observado en el presente trabajo para las características estimadas mediante NIR, ir20, dap11, dap20 y at20, que presentaban números semejantes en ambos tipos de datos por individuo, y las diferencias entre las ET de los modelos de ssGBLUP y GBLUP fueron mínimas (siendo la ET para ssGBLUP significativamente mayor respecto de la ET para GBLUP sólo para las características de extet20 y exttot20). Sin embargo, todos los modelos de GBLUP aplicados en *E. dunnii*, para los diferentes caracteres, contuvieron una menor proporción de datos fenotípicos (ya que utiliza los individuos que presentan tanto datos fenotípicos como genotípicos) implicados que en los modelos de ssGBLUP, variando entre un ~15% menos de datos para los caracteres estimados mediante NIR (extet20, exttot20, lig20, klas20, sg20, cel20 y db20) hasta un ~420% menos de datos para los caracteres de crecimiento medidos a los 6 años (dap6, at6, for6). Según Cappa et al. (2019), esta información fenotípica adicional con respecto a los datos genotípicos, le brindaría mayor exactitud teórica al modelo de ssGBLUP. Por lo tanto, para las características de crecimiento medidas a los 6 años en *E. dunnii*, las ET de ssGBLUP fueron significativamente mayores que las ET para GBLUP al aplicar cualquiera de las tres metrices genotípicas. Además, el método ssGBLUP generó habilidades predictivas significativamente mayores para estas últimas características mencionadas y con todas las matrices,

siendo un enfoque mejor que GBLUP en esta población, consistente con lo observado por Cappa et al. (2019). Asimismo, también se han reportado resultados similares aplicando ssGBLUP en animales (Christensen et al., 2012; Guo et al., 2015) y cultivos agronómicos como el trigo (Ashraf et al., 2016; Pérez-Rodríguez et al., 2017), por lo que demuestra ser un método genómico valioso para el mejoramiento forestal (Cappa et al., 2019).

Es esperado que los enfoques genómicos funcionen mejor que los enfoques basados en pedigrí porque utilizan información de relaciones de parentesco más precisas (Cappa et al., 2019). En la presente tesis, al observar la habilidad predictiva de los tres modelos evaluados, el que presentó la más elevada fue aquel que sólo utilizó la información basada en pedigrí (ABLUP) respecto de los enfoques genómicos (GBLUP y ssGBLUP). No obstante, sólo para cuatro características fenotípicas las HP de ABLUP fueron significativamente diferentes a los valores del modelo de HBLUP y seis a los de GBLUP. Esta tendencia también fue señalada en un gran número de trabajos en especies forestales (Beaulieu et al. , 2014; Cappa et al., 2019; El-Dien et al., 2015; Lenz et al., 2017; Resende et al., 2017). Que ABLUP supere a GBLUP en habilidad predictiva es un resultado común para los rasgos de crecimiento en el género *Eucalyptus*. Esto probablemente se deba a una variación aditiva sobreestimada por el enfoque ABLUP, ya que este no puede desligarse de la variación no aditiva, que posee gran influencia al igual que el efecto aditivo, como por ejemplo en el carácter de altura (Cappa et al., 2019; El-Dien et al., 2016; Muñoz et al., 2014). Sólo en algunos trabajos GBLUP presentó mejor desempeño que ABLUP, sin embargo en muchos de ellos se encontró que el pedigrí presentaba inconsistencias (Kainer et al., 2018; Müller et al., 2017; Tan et al., 2017).

Müller et al. (2017) utilizaron la plataforma EUChip60K y aplicaron GBLUP y otros cinco métodos Bayesianos de SG, los cuales mostraron habilidades predictivas para poblaciones de *E. benthamii* (n =505) y *E. pellita* (n =732), similares a los encontrados en la presente tesis. Dichos autores observaron para *E. pellita* HPs que fueron de 0,44 para diámetro a la altura del pecho y 0,34 para altura total a los 3,5 años. Por otro lado, para *E. benthamii* observaron habilidades predictivas de 0,16 para DAP y valores cercanos a cero para AT, ambos a los 5 años, lo que posiblemente fue explicado por una menor presencia de variación genética aditiva para estos rasgos en dicha población y en particular en la especie, en comparación con *E. pellita*. Al contrastar dichos resultados con los reportados en la presente tesis para *E. dunnii* en la población de Ubajay, se observó que los valores de habilidad predictiva para dap presentaron valores menores (HP con Chip: 0,018, 0,085 y 0,10 para DAP a los 6, 11 y 20 años, respectivamente, con GBLUP) pero también cercanos (HP con Chip: 0,145, 0,088 y 0,10, respectivamente, con ssGBLUP) a los observados por Müller et al. (2017) para *E. benthamii* (HP: 0,16). Para altura total, al igual que en *E. benthamii*, también se evidenciaron valores de HP cercanos a cero en *E. dunnii*, siendo los de *E. dunnii* de 0,06 y 0,016 para at6 y at20 con GBLUP, pero superiores con ssGBLUP (0,207 para at6 y 0,018 para at20).

Durán et al. (2017) aplicaron GBLUP y otros tres modelos Bayesianos de SG en *E. globulus* mediante el

uso del microarreglo EUChip60k, en una población clonal con un número de individuos (310) muy cercano al utilizado en la población de *E. dunnii* (280). En dicho trabajo evaluaron los caracteres de volumen de tronco y densidad de madera, y obtuvieron mediante GBLUP valores de HP de 0,78 y 0,63, respectivamente. Al comparar dichos resultados con los obtenidos para *E. dunnii*, estos últimos presentaron HP mucho menores que en la población de *E. globulus* (HP densidad básica: 0,134 con matriz del Chip). Una de las características que explicaría que las HP sean mayores en *E. globulus*, es que las relaciones de parentesco eran más estrechas y consideraban un menor número de familias (40 familias de hermanos completos y 13 familias de medios hermanos, producidas cruzando 23 padres). Es sabido que al aumentar el tamaño de la población de entrenamiento, aumenta la precisión de la predicción, pero también al aumentar las relaciones entre la población de PE y PV, aumenta la precisión (Durán et al., 2017).

En estudios preliminares, GBLUP fue aplicado en una población de *E. dunnii* implantada en Sudáfrica, también utilizando el EUChip60k (Naidoo et al., 2018). En este trabajo se analizaron 9.102 marcadores SNP en 840 descendientes de 89 familias de medios hermanos, y aplicaron GBLUP en cinco caracteres fenotípicos. Entre sus resultados obtuvieron valores de HP para altura del árbol de 0,33, para diámetro a la altura del pecho de 0,38 y para densidad básica de la madera de 0,51. Sin embargo, para la población de Ubajay se evidenciaron valores mucho menores, y esto podría ser debido a que fueron ambientes diferentes, quizás orígenes diferentes y un número de individuos genotipados tres veces menor respecto a la mencionada. Por otro lado, Jones et al., (2019), investigaron si los datos de diferentes ensayos podrían combinarse para mejorar la precisión del modelo SG en *E. dunnii*, y observaron que realmente las precisiones mejoraban, como por ejemplo para diámetro a la altura del pecho aumentó un 86% y para altura del árbol en un 290% (0,18 a 0,72). Esto sugiere que para el programa de mejoramiento de INTA de *E. dunnii* podría aplicarse SG considerando más ensayos de la red, lo que podría aumentar la precisión de la estimación de valores de cría de los árboles a seleccionar.

Un elemento clave que puede influir sobre el desempeño de la SG es la elección de la plataforma de genotipificación (Elbasyoni et al., 2018). Como se mencionó anteriormente, las metodologías de genotipificación por secuenciación brindan una gran cantidad de marcadores moleculares, pero con una alta tasa de datos faltantes. Por otro lado, los microarreglos proporcionan una proporción muy baja de datos faltantes, pero pueden presentar un sesgo en las frecuencias alélicas y no permiten el descubrimiento de alelos propios de la población a evaluar (Albrechtsen et al., 2010; Bajgain et al., 2016; Li & Kimmel, 2013). En el presente trabajo, al comparar las matrices de GBS y el EUChip60k mediante las ET de los modelos genómicos, se observaron valores muy similares, siendo significativamente mayores sólo para cuatro características (dap11, dap20, cel20 y db20) con la matriz de GBS y para dos (exttot20 para HBLUP y GBLUP y extet20 sólo para GBLUP) con la del Chip. Por el contrario, al comparar las HP ninguna de ellos fue significativamente diferente. Además, para ET y HP ambas matrices evidenciaron la misma tendencia al obtener los mayores valores en los modelos de ssGBLUP y para aquellos caracteres fenotípicos medidos en la población completa de Ubajay, con

mayor proporción de individuos no genotipados que genotipados. Un desempeño similar al del presente trabajo fue observado por Elbasyoni et al. (2018) al contrastar ambos tipos de plataformas de genotipado para SG en trigo, si bien se trata de un cultivo autógamo, con DL mayores. Dicho trabajo es uno de los pocos que comparó el desempeño de estas metodologías de genotipado para SG en plantas. Allí, fueron genotipadas 299 accesiones de trigo duro de invierno (*Triticum aestivum L.*) mediante GBS y un microarreglo. Dichos autores observaron que la matriz de GBS, tanto imputando el 10% de los datos perdidos (10.775 SNPs) como el 50% (39.674 SNPs), presentó similar e incluso mayor precisión de la predicción genómica que el microarreglo (19.515 SNPs) para todos los caracteres agronómicos predichos, dependiendo del porcentaje de datos perdidos imputados de la matriz de GBS de partida (Elbasyoni et al., 2018).

Los métodos de validación de precisión juegan un papel crítico en la comparación de los modelos de SG (Cappa et al., 2019; Putz et al., 2018). En el presente trabajo, se utilizaron dos métodos de validación de precisión: la exactitud teórica tradicional, derivada de la varianza del error de la predicción; y las habilidades predictivas, derivadas de la correlación entre las predicciones y el modelo basado en pedigrí. Usando la exactitud teórica tradicional, ssGBLUP se desempeña mejor en general (promedio entre rasgos = 0,357, 0,343 y 0,337 para GBS, Chip y Chip-GBS) seguido de GBLUP (promedio entre rasgos = 0,299, 0,289 y 0,279 para GBS, Chip y Chip-GBS) y ABLUP (promedio entre rasgos = 0,269). Sin embargo, cuando se utilizó la habilidad predictiva para medir el desempeño de los modelos, ABLUP mostró los valores más altos (promedio entre rasgos = 0,278) seguido de ssGBLUP (promedio entre rasgos = 0,146, 0,145 y 0,138 para Chip-GBS, Chip y GBS) y GBLUP (promedio entre rasgos = 0,113, 0,111 y 0,105 para Chip-GBS, Chip y GBS) y el orden de las matrices genotípicas se invirtió. Existen otros métodos de validación basados en la correlación entre el valor genómico predicho y las diferentes versiones de los fenotipos corregidos, y estos fueron utilizados en estudios empíricos con árboles forestales (Beaulieu et al., 2014; Resende et al., 2012; Tan et al., 2017). Sin embargo, el desarrollo de un método de validación de la precisión "perfecto" es actualmente necesario, pero es un tema muy debatido en actualmente en el mejoramiento de animales (Misztal, 2016).

En general, se puede concluir que los modelos que incluyen datos genómicos son prometedores a la hora de ser aplicados en los programas de mejoramiento genético, en particular en el de *E. dunnii*, ya que presentan HP similares respecto de ABLUP. Además, las ET de los modelos genómicos fueron superiores a las de ABLUP. De este modo, pueden ser utilizados para generar una clasificación de los individuos de *E. dunnii* según las prioridades del programa de mejoramiento, por ejemplo, para seleccionar individuos con menor índice de rajado y mayor crecimiento. Así, a los fines prácticos, estos modelos podrían ser aplicados en el Huerto Semillero Clonal, derivado de la población de *E. dunnii* estudiada, para la elección de los mejores padres de la próxima generación (semillas), y luego aplicarlos a los hijos en una edad temprana (plantines ya genotipados), contando sólo con los datos genotípicos (y los modelos de HBLUP y GBLUP) para elegir cuales de ellos llevar a campo.

4. Perspectivas

4.4.1 Análisis de Asociación de Genoma Amplio (Genome Wide Association Study)

Las variaciones fenotípicas complejas, como la formación de madera en poblaciones arbóreas longevas, implican una serie de procesos biológicos dinámicos organizados de manera precisa y cuantitativa, incluida la regulación transcripcional y traduccional y el flujo de intermediarios metabólicos de diversas vías bioquímicas (Mizrachi & Myburg, 2016). En el presente trabajo, al utilizar el genoma de *E. grandis*, sólo evaluamos los SNPs de *E. dunnii* que se encuentran en regiones comunes con *E. grandis*. Tal diseño específico conduce a la ausencia de algunas de las variantes causales precisas y no puede detectar ciertas las señales genéticas o mutaciones raras de rasgos complejos (Alqudah et al., 2019). Como perspectiva, un análisis a realizar en el futuro inmediato, es un análisis de GWAS con los SNPs encontrados en el análisis *de novo* de secuencias de GBS con *E. dunnii*, que brindará la posibilidad de encontrar nuevas variantes asociadas a los fenotipos estudiados. Gong y col. (2017) observaron una mayor contribución a la varianza fenotípica explicada aplicando InDels (14,7%) en comparación con los loci de SNPs (5%), para crecimiento y las propiedades de la madera en *Populus tomentosa.* De esto se desprende una de las perspectivas futuras del presente trabajo, siendo ésta la búsqueda de InDels en los datos de GBS, ya que dichos marcadores representan un sistema de marcadores más efectivo para MAS. Asimismo, sería de interés incorporar a los análisis de GWAS SSRs encontrados en los datos de GBS lo que aumentaría la probabilidad de encontrar variantes asociadas a los caracteres estudiados en el presente trabajo. Otra perspectiva es obtener una mayor resolución al analizar los efectos alélicos (haplotipo) en lugar de SNPs para comprender la arquitectura de la variación genética que afecta los fenotipos complejos (Mizrachi & Myburg, 2016).

4.4.2 Selección Genómica

Entre los factores que podrían aumentar las precisiones al aplicar SG, se encuentra el uso de métodos bayesianos que explotan información previa de que regiones genómicas particulares y contribuyen con efectos más grandes a la variación del rasgo. Los enfoques GBLUP y ssGBLUP asumen una arquitectura de rasgos poligénicos complejos y consideran que todas las regiones genómicas rastreadas por los marcadores tienen la misma contribución a la construcción del carácter (Cappa et al., 2019). Por el contrario, el análisis bayesiano podría ser más eficiente para los rasgos que involucran *loci* con mayores efectos conocidos sobre el fenotipo (Fernando et al., 2014). Sin embargo, según lo evidenciado en la presente tesis y en otros trabajos de especies forestales al aplicar modelos de mapeo por asociación para el crecimiento y los rasgos de calidad de la madera en especies forestales (Cappa et al., 2013; Müller et al., 2019; Porth et al., 2013; Resende et al., 2017a) estos caracteres están controlados por una gran cantidad de *loci* con un efecto pequeño, de modo que el modelo infinitesimal debería ajustarse adecuadamente.

Asimismo, en la presente tesis, con el objetivo de evaluar el desempeño de las metodologías de genotipificación, se utilizaron modelos de SG que sólo involucraron datos fenotípicos y genotípicos. Posteriormente, con el objetivo de aplicar dichas metodologías en el programa de mejoramiento de *E. dunnii* del INTA, se evaluará el desempeño de dichas plataformas de genotipado implementando modelos más complejos, que también consideren efectos espaciales o de ambientes heterogéneos y/o efectos genéticos de competencia entre individuos. Dichos modelos fueron aplicados por el presente grupo de trabajo, y se observó que minimizan la varianza del error y aumentan la exactitud de la precisión (Cappa et al., 2017). Por otro lado, se podrían incorporar datos genotípicos y fenotípicos de más sitios de la red de ensayos de *E. dunnii* que aumentarían las precisiones en las predicciones, según fue visto por Jones et al., (2019), en la misma especie, al combinar datos de diferentes ensayos. Otras de las variables que podrían aumentar las precisiones al aplicar SG, es la utilización de la matriz de SNPs obtenidos con el análisis *de novo* con datos de las secuencias de GBS, así como incorporar InDels y SSRs.

5 CONCLUSIONES

- ✓ Se desarrolló y optimizó un protocolo de ddRADseq reproducible y aplicable a especies sin recursos genómicos disponibles. Permitió generar miles de marcadores moleculares, incluyendo variantes exclusivas y de baja frecuencia en la población bajo estudio y analizar regiones genómicas complementarias a las evaluadas mediante un sistema comercial. El protocolo desarrollado permite su escalado directo a un elevado número de muestras. Por primera vez y a partir de este trabajo, esta metodología genómica es aplicada en el género *Eucalyptus* y es accesible a nivel nacional también para otras especies.
- ✓ Hasta la fecha, este es el primer trabajo que compara las tecnologías ddRADseq y EUChip60K para una especie de interés forestal y que presenta bajo desequilibrio de ligamiento. Estas metodologías resultaron complementarias para la estrategia de Mapeo por Asociación y similares para Selección Genómica.
- ✓ El Mapeo por asociación permitió identificar posibles genes y marcadores útiles para el programa de mejoramiento de *E. dunnii*. Se encontraron 7 SNPs de ddRADSeq, 13 SNPs EUChip60K y 19 SNPs del conjunto total asociados ($p < 0,001$) con las 14 características de interés forestal estudiadas, localizando alrededor de 100 genes de importancia biológica en el genoma de *E. grandis* público. Estos genes se encuentran a menos de 0,05 cM de cada marcador y las anotaciones funcionales de varios de ellos, podrían vincularse al fenotipo. Además, se ubicaron a menos de 0,8 cM, 50 genes predichos de *E. grandis* involucrados en rutas metabólicas de la síntesis de celulosa, xilanos, fenilpropanoides, terpenos, lacasas, peroxidasas, reflejando el potencial y las bondades de las metodologías evaluadas. Toda esta información nutrirá a los análisis futuros para filtrar marcadores cuando se utilicen otras metodologías de selección con herramientas genómicas.
- ✓ Para aplicar Selección genómica en el mejoramiento *E. dunnii* es importante tener en cuenta la distribución y número de marcadores, la heredabilidad de los caracteres, el tamaño poblacional, las relaciones entre individuos y la diversidad.
- ✓ La estrategia de Selección genómica evaluando exhaustivamente dos modelos de predicción para 14 características mostró que es relevante su empleo desde el punto de vista operativo en un programa de mejoramiento forestal para acortar los tiempos requeridos para la selección. Para esta población en general, las exactitudes teóricas de los modelos que contuvieron información genómica (ssGBLUP y GBLUP) superaron a la del modelo convencional (ABLUP). Comparando las dos metodologías de genotipado, ddRADSeq fue superior para cuatro de las cinco características que difirieron significativamente.
- ✓ Respecto a la aplicación de Selección Genómica, ambas matrices genotípicas en base a SNPs mostraron similar habilidad predictiva para la predicción de los valores de cría.

- ✓ Este trabajo permitió evaluar la potencialidad de la metodología de Selección Genómica. En ese sentido, se podrían emplear los modelos genómicos generados sobre plantines relacionados genéticamente para predecir tempranamente los mejores individuos.

6 REFERENCIAS BIBLIOGRÁFICAS

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., & Lawlor, T. J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science*, *93*(2), 743–752. https://doi.org/10.3168/jds.2009-2730

Aguirre, N., Filippi, C., Zaina, G., Rivas, J., Acuña, C., Villalba, P., … Marcucci Poltri, S. (2019). Optimizing ddRADseq in Non-Model Species: A Case Study in Eucalyptus dunnii Maiden. *Agronomy*, *9*(9), 484. https://doi.org/10.3390/agronomy9090484

Albrechtsen, A., Nielsen, F. C., & Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution*, *27*(11), 2534–2547. https://doi.org/10.1093/molbev/msq148

Alqudah, A. M., Sallam, A., Stephen Baenziger, P., & Börner, A. (2019). GWAS: Fast-Forwarding Gene Identification in Temperate Cereals: Barley as a Case Study- A review. *Journal of Advanced Research*. https://doi.org/10.1016/j.jare.2019.10.013

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews. Genetics*, *17*(2), 81–92. https://doi.org/10.1038/nrg.2015.28

Andrews, S. (2010). Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Retrieved December 27, 2019, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Ashraf, B., Edriss, V., Akdemir, D., Autrique, E., Bonnett, D., Crossa, J., … Jannink, J.-L. (2016). Genomic Prediction using Phenotypes from Pedigreed Lines with No Marker Data. *Crop Science*, *56*(3), 957–964. https://doi.org/10.2135/cropsci2015.02.0111

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*(10), 1–7. https://doi.org/10.1371/journal.pone.0003376

Bajgain, P., Rouse, M. N., Bulli, P., Bhavani, S., Gordon, T., Wanyera, R., … Pumphrey, M. O. (2015). Association mapping of North American spring wheat breeding germplasm reveals loci conferring resistance to Ug99 and other African stem rust races. *BMC Plant Biology*, *15*(1), 1–19. https://doi.org/10.1186/s12870-015-0628-9

Bajgain, Prabin, Rouse, M. N., & Anderson, J. A. (2016). Comparing genotyping-by-sequencing and single nucleotide polymorphism chip genotyping for quantitative trait loci mapping in wheat. *Crop Science*, *56*(1), 232–248. https://doi.org/10.2135/cropsci2015.06.0389

Bak, S., Beisson, F., Bishop, G., Hamberger, B., Höfer, R., Paquette, S., & Werck-Reichhart, D. (2011). The Arabidopsis Book - Cytochromes P450. In *American Society of Plant Biologists* (Vol. 9). https://doi.org/10.1199/tab.0144

Ballesta, P., Bush, D., Silva, F. F., & Mora, F. (2020). Genomic Predictions Using Low-Density SNP Markers, Pedigree and GWAS Information: A Case Study with the Non-Model Species Eucalyptus cladocalyx. *Plants*, *9*(1), 99. https://doi.org/10.3390/plants9010099

Ballesta, P., Serra, N., Guerra, F. P., Hasbún, R., & Mora, F. (2018). Genomic prediction of growth and stem quality traits in Eucalyptus globulus Labill. at its southernmost distribution limit in Chile. *Forests*, *9*(12), 1–18. https://doi.org/10.3390/f9120779

Barchi, L., Lanteri, S., Portis, E., Acquadro, A., Valè, G., Toppino, L., & Rotino, G. L. (2011). Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *BMC Genomics*, *12*. https://doi.org/10.1186/1471-2164-12-304

Bartholomé, J., Van Heerwaarden, J., Isik, F., Boury, C., Vidal, M., Plomion, C., & Bouffier, L. (2016).

Performance of genomic prediction within and across generations in maritime pine. *BMC Genomics*, *17*(1), 604. https://doi.org/10.1186/s12864-016-2879-8

Beaulieu, J., Doerksen, T., Clément, S., Mackay, J., & Bousquet, J. (2014). Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. *Heredity*, *113*(4), 343–352. https://doi.org/10.1038/hdy.2014.36

Ben Langmead & Steven L Salzberg. (2012). *Fast gapped-read alignment with Bowtie 2*. https://doi.org/10.1038/nmeth.1923

Biyue Tan. (2018). *Genomic selection and genome- wide association studies to dissect quantitative traits in forest trees* (Dissertati). Retrieved from http://umu.diva-portal.org/

Botstein, D., & Risch, N. (2003, March). Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, Vol. 33, pp. 228–237. https://doi.org/10.1038/ng1090

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Braier, G., Esper, N., & Corinaldesi, L. (2004). *Informe nacional complementario. Tendencias y perspectivas del sector forestal al año 2020. Argentina.* Retrieved from http://www.fao.org/tempref/docrep/fao/009/j2053s/j2053s00.pdf

Brookes, A. J. (1999, July 8). The essence of SNPs. *Gene*, Vol. 234, pp. 177–186. https://doi.org/10.1016/S0378-1119(99)00219-X

Buckler IV, E. S., & Thornsberry, J. M. (2002). Plant molecular diversity and applications to genomics. *Current Opinion in Plant Biology*, Vol. 5, pp. 107–111. https://doi.org/10.1016/S1369-5266(02)00238-8

Buongiorno, J., & Zhu, S. (2014). Assessing the impact of planted forests on the global forest economy. *New Zealand Journal of Forestry Science*, *44*(Suppl 1), S2. https://doi.org/10.1186/1179-5395-44-S1-S2

Burghardt, L. T., Young, N. D., & Tiffin, P. (2017). A Guide to Genome-Wide Association Mapping in Plants. *Current Protocols in Plant Biology*, *2*(1), 22–38. https://doi.org/10.1002/cppb.20041

Campbell, E. O., Davis, C. S., Dupuis, J. R., Muirhead, K., & Sperling, F. A. H. (2017). Cross-platform compatibility of de novo-aligned SNPs in a nonmodel butterfly genus. *Molecular Ecology Resources*, *17*(6), e84–e93. https://doi.org/10.1111/1755-0998.12695

Cappa, E. P., de Lima, B. M., da Silva-Junior, O. B., Garcia, C. C., Mansfield, S. D., & Grattapaglia, D. (2019). Improving genomic prediction of growth and wood traits in Eucalyptus using phenotypes from non-genotyped trees by single-step GBLUP. *Plant Science*, *284*, 9–15. https://doi.org/10.1016/j.plantsci.2019.03.017

Cappa, E. P., El-Kassaby, Y. A., Garcia, M. N., Acuña, C., Borralho, N. M. G., Grattapaglia, D., & Marcucci Poltri, S. N. (2013). Impacts of population structure and analytical models in genome-wide association studies of complex traits in forest trees: A case study in Eucalyptus globulus. *PLoS ONE*, *8*(11). https://doi.org/10.1371/journal.pone.0081267

Cappa, E. P., El-Kassaby, Y. A., Muñoz, F., Garcia, M. N., Villalba, P. V., Klápště, J., & Marcucci Poltri, S. N. (2017). Improving accuracy of breeding values by incorporating genomic information in spatial-competition mixed models. *Molecular Breeding*, *37*(10). https://doi.org/10.1007/s11032-017-0725-6

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). *Stacks* : Building and Genotyping Loci *De Novo* From Short-Read Sequences. *G3: Genes|Genomes|Genetics*, *1*(3),

171–182. https://doi.org/10.1534/g3.111.000240

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/s13742-015-0047-8

Chang, M., He, L., & Cai, L. (2018). An overview of genome-wide association studies. In *Methods in Molecular Biology* (Vol. 1754, pp. 97–108). https://doi.org/10.1007/978-1-4939-7717-8_6

Chen, C., Mitchell, S. E., Elshire, R. J., Buckler, E. S., & El-Kassaby, Y. A. (2013). Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. *Tree Genetics and Genomes*, *9*(6), 1537–1544. https://doi.org/10.1007/s11295-013-0657-1

Christensen, O. F., Madsen, P., Nielsen, B., Ostersen, T., & Su, G. (2012). Single-step methods for genomic evaluation in pigs. *Animal*, *6*(10), 1565–1571. https://doi.org/10.1017/S1751731112000742

Christensen, Ole F, & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution*, *42*(1), 2. https://doi.org/10.1186/1297-9686-42-2

Clark, G. B., Dauwalder, M., & Roux, S. J. (1992). Purification and immunolocalization of an annexin-like protein in pea seedlings. *Planta*, *187*, 1–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11538119

Clarke, C. (n.d.). The profitable pulp mill. *Australian Forest Genetics Conference*.

Corder, E. H., Saunders, A. M., Risch, N. J., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., … Pericak-Vance, M. A. (1994). Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nature Genetics*, *7*(2), 180–184. https://doi.org/10.1038/ng0694-180

Croué, I., & Ducrocq, V. (2017). Genomic and single-step evaluations of carcass traits of young bulls in dual-purpose cattle. *Journal of Animal Breeding and Genetics*, *134*(4), 300–307. https://doi.org/10.1111/jbg.12261

Da Silva Perez, D., Guillemain, A., Alazard, P., Plomion, C., Rozenberg, P., Carlos Rodrigues, J., … Chantre, G. (2007). Improvement of Pinus pinaster Ait elite trees selection by combining near infrared spectroscopy and genetic tools. *Holzforschung*, *61*(6), 611–622. https://doi.org/10.1515/HF.2007.118

DaCosta, J. M., & Sorenson, M. D. (2014). Amplification Biases and Consistent Recovery of Loci in a Double-Digest RAD-seq Protocol. *PLoS ONE*, *9*(9), e106713. https://doi.org/10.1371/journal.pone.0106713

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Darrow, W. K. (1994). The effect of drought on eucalypt species growing on shallow soils in South Africa. *The Effect of Drought on Eucalypt Species Growing on Shallow Soils in South Africa.*, (No. 7/94).

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*(7), 499–510. https://doi.org/10.1038/nrg3012

Deeks, M. J., Hussey, P. J., & Davies, B. (2002, November 1). Formins: Intermediates in signal-transduction cascades that affect cytoskeletal reorganization. *Trends in Plant Science*, Vol. 7, pp. 492–498. https://doi.org/10.1016/S1360-1385(02)02341-5

Desta, Z. A., & Ortiz, R. (2014). Genomic selection: Genome-wide prediction in plant improvement. *Trends in Plant Science*, Vol. 19, pp. 592–601. https://doi.org/10.1016/j.tplants.2014.05.006

DeWan, A., Liu, M., Hartman, S., Zhang, S. S. M., Liu, D. T. L., Zhao, C., … Hoh, J. (2006). HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science*, *314*(5801), 989–992.

https://doi.org/10.1126/science.1133807

Doughty, R. W. (2000). *The eucalyptus : a natural and commercial history of the gum tree*. Johns Hopkins University Press.

Du, Q., Gong, C., Wang, Q., Zhou, D., Yang, H., Pan, W., … Zhang, D. (2016). Genetic architecture of growth traits in Populus revealed by integrated quantitative trait locus (QTL) analysis and association studies. *New Phytologist*, *209*(3), 1067–1082. https://doi.org/10.1111/nph.13695

Du, Q., Lu, W., Quan, M., Xiao, L., Song, F., Li, P., … Zhang, D. (2018). Genome-wide association studies to improve wood properties: Challenges and prospects. *Frontiers in Plant Science*, Vol. 871. https://doi.org/10.3389/fpls.2018.01912

Durán, R., Isik, F., Zapata-Valenzuela, J., Balocchi, C., & Valenzuela, S. (2017). Genomic predictions of breeding values in a cloned Eucalyptus globulus population in Chile. *Tree Genetics and Genomes*, *13*(4). https://doi.org/10.1007/s11295-017-1158-4

Durán, R., Zapata-Valenzuela, J., Balocchi, C., & Sofía Valenzuela, ·. (2018). Efficiency of EUChip60K pipeline in fingerprinting clonal population of Eucalyptus globulus. *Trees*, *32*, 663–669. https://doi.org/10.1007/s00468-017-1637-0

El-Dien, O. G., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., & El-Kassaby, Y. A. (2016). Implementation of the realized genomic relationship matrix to open-pollinated white spruce family testing for disentangling additive from nonadditive genetic effects. *G3: Genes, Genomes, Genetics*, *6*(3), 743–753. https://doi.org/10.1534/g3.115.025957

Elbasyoni, I. S., Lorenz, A. J., Guttieri, M., Frels, K., Baenziger, P. S., Poland, J., & Akhunov, E. (2018). A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. *Plant Science*, *270*(October 2017), 123–130. https://doi.org/10.1016/j.plantsci.2018.02.019

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*(5), 1–10. https://doi.org/10.1371/journal.pone.0019379

Escamilla-Treviño, L. L., Chen, W., Card, M. L., Shih, M. C., Cheng, C. L., & Poulton, J. E. (2006). Arabidopsis thaliana β-Glucosidases BGLU45 and BGLU46 hydrolyse monolignol glucosides. *Phytochemistry*, *67*(15), 1651–1660. https://doi.org/10.1016/j.phytochem.2006.05.022

Fahrenkrog, A. M., Neves, L. G., Resende, M. F. R., Vazquez, A. I., de los Campos, G., Dervinis, C., … Kirst, M. (2017). Genome-wide association study reveals putative regulators of bioenergy traits in *Populus deltoides*. *New Phytologist*, *213*(2), 799–811. https://doi.org/10.1111/nph.14154

FAO. (1981). *El eucalipto en la repoblación forestal* (O. D. L. N. U. P. L. A. Y. LA ALIMENTACION, Ed.). Retrieved from http://www.fao.org/3/AC459S/AC459S00.htm#TOC

FAO. (2012). *El estado de los recursos genéticos forestales en el mundo Informe Nacional- Argentina-Dirección de Producción Forestal del Ministerio de Agricultura, Ganadería y Pesca de la Nación para FAO*. Retrieved from http://www.fao.org/3/i3825e/i3825e1.pdf

FAO. (2015). *Recursos Forestales Mundiales 2015 ¿Cómo estan cambiando los bosques del mundo?* (O. D. L. N. U. P. L. A. Y. LA AGRICULTURA, Ed.). Retrieved from www.fao.org/forest-resources-assessment/es

FAO. (2018a). *El estado de los bosques del mundo - Las vías forestales hacia el desarrollo sostenible*. Roma.

FAO. (2018b). *Global Forest Products - Facts and Figures*.

Faria, D. A., Tanno, P., Reis, A., Martins, A., Ferreira, M. E., & Grattapaglia, D. (2012). Genotyping-by-Sequencing (GbS) the Highly Heterozygous Genome of Eucalyptus Provides Large Numbers of High Quality Genome-Wide SNPs. *International Plant and Animal Genome Conference PO521*. Retrieved

from https://pag.confex.com/pag/xx/webprogram/Paper4239.html

Fernando, R. L., Dekkers, J. C., & Garrick, D. J. (2014). A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genetics Selection Evolution*, *46*(1), 50. https://doi.org/10.1186/1297-9686-46-50

Flint-Garcia, S. A., Thornsberry, J. M., & Buckler, E. S. (2003). Structure of Linkage Disequilibrium in Plants. *Annual Review of Plant Biology*, *54*(1), 357–374. https://doi.org/10.1146/annurev.arplant.54.031902.134907

Freeman, J. S., Whittock, S. P., Potts, B. M., & Vaillancourt, R. E. (2009). QTL influencing growth and wood properties in Eucalyptus globulus. *Tree Genetics & Genomes*, *5*(4), 713–722. https://doi.org/10.1007/s11295-009-0222-0

Gamal El-Dien, O., Ratcliffe, B., Klápště, J., Chen, C., Porth, I., & El-Kassaby, Y. A. (2015). Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. *BMC Genomics*, *16*(1), 1–16. https://doi.org/10.1186/s12864-015-1597-y

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics*, *51*(4), 1440. https://doi.org/10.2307/2533274

Gion, J. M., Rech, P., Grima-Pettenati, J., Verhaegen, D., & Plomion, C. (2000). Mapping candidate genes in Eucalyptus with emphasis on lignification genes. *Molecular Breeding*, *6*(5), 441–449. https://doi.org/10.1023/A:1026552515218

Goddard, M. E., Hayes, B. J., & Meuwissen, T. H. E. (2010). Genomic selection in livestock populations. *Genetics Research*, *92*(5–6), 413–421. https://doi.org/10.1017/S0016672310000613

Goddard, M. E., Wray, N. R., Verbyla, K., & Visscher, P. M. (2009). Estimating Effects and Making Predictions from Genome-Wide Marker Data. *Statistical Science*, *24*(4), 517–529. https://doi.org/10.1214/09-STS306

Granato, I. S. C., Galli, G., de Oliveira Couto, E. G., e Souza, M. B., Mendonça, L. F., & Fritsche-Neto, R. (2018). snpReady: a tool to assist breeders in genomic analysis. *Molecular Breeding*, *38*(8). https://doi.org/10.1007/s11032-018-0844-8

Grattapaglia, D., & Bradshaw Jnr, H. D. (1994). Nuclear DNA content of commercially important Eucalyptus species and hybrids. *Canadian Journal of Forest Research*, *24*(5), 1074–1078. https://doi.org/10.1139/x94-142

Grattapaglia, D., & Bradshaw Jr., H. D. (1994). Nuclear DNA content of commercially important *Eucalyptus* species and hybrids. *Canadian Journal of Forest Research*, *24*(5), 1074–1078. https://doi.org/10.1139/x94-142

Grattapaglia, Dario. (2014). Breeding forest trees by genomic selection: Current progress and theway forward. In *Genomics of Plant Genetic Resources: Volume 1. Managing, Sequencing and Mining Genetic Resources* (pp. 651–682). https://doi.org/10.1007/978-94-007-7572-5_26

Grattapaglia, Dario, de Alencar, S., & Pappas, G. (2011). Genome-wide genotyping and SNP discovery by ultra-deep Restriction-Associated DNA (RAD) tag sequencing of pooled samples of E. grandis and E. globulus. *BMC Proceedings*, *5*(Suppl 7), P45. https://doi.org/10.1186/1753-6561-5-S7-P45

Grattapaglia, Dario, & Kirst, M. (2008). Eucalyptus applied genomics: From gene sequences to breeding tools. *New Phytologist*, *179*(4), 911–929. https://doi.org/10.1111/j.1469-8137.2008.02503.x

Grattapaglia, Dario, Mamani, E. M. C., Silva-Junior, O. B., & Faria, D. A. (2015). A novel genome-wide microsatellite resource for species of Eucalyptus with linkage-to-physical correspondence on the reference genome sequence. *Molecular Ecology Resources*, *15*(2), 437–448. https://doi.org/10.1111/1755-0998.12317

Grattapaglia, Dario, & Resende, M. D. V. (2011). Genomic selection in forest tree breeding. *Tree Genetics and Genomes*, *7*(2), 241–255. https://doi.org/10.1007/s11295-010-0328-4

Grattapaglia, Dario, Vaillancourt, R. E., Shepherd, M., Thumma, B. R., Foley, W., Külheim, C., … Myburg, A. A. (2012). Progress in Myrtaceae genetics and genomics: Eucalyptus as the pivotal genus. *Tree Genetics and Genomes*, *8*(3), 463–508. https://doi.org/10.1007/s11295-012-0491-x

Grattapaglia, Dario, Vilela Resende, M., Resende, M., Sansaloni, C., Petroli, C., Missiaggia, A., … Kilian, A. (2011a). Genomic Selection for growth traits in Eucalyptus: accuracy within and across breeding populations. *BMC Proceedings*, *5*(Suppl 7), O16. https://doi.org/10.1186/1753-6561-5-s7-o16

Grattapaglia, Dario, Vilela Resende, M., Resende, M., Sansaloni, C., Petroli, C., Missiaggia, A., … Kilian, A. (2011b). Genomic Selection for growth traits in Eucalyptus: accuracy within and across breeding populations. *BMC Proceedings*, *5*(Suppl 7), O16. https://doi.org/10.1186/1753-6561-5-S7-O16

Groover, A. T. (2007). Will genomics guide a greener forest biotech? *Trends in Plant Science*, *12*(6), 234–238. https://doi.org/10.1016/j.tplants.2007.04.005

Guo, X., Christensen, O. F., Ostersen, T., Wang, Y., Lund, M. S., & Su, G. (2015). Improving genetic evaluation of litter size and piglet mortality for both genotyped and nongenotyped individuals using a single-step method. *Journal of Animal Science*, *93*(2), 503–512. https://doi.org/10.2527/jas.2014-8331

Guo, Y., Yang, G. Q., Chen, Y., Li, D., & Guo, Z. (2018). A comparison of different methods for preserving plant molecular materials and the effect of degraded DNA on ddRAD sequencing. *Plant Diversity*, *40*(3), 106–116. https://doi.org/10.1016/j.pld.2018.04.001

Hayes, B., & Goddard, M. (2010). Genome-wide association and genomic selection in animal breedingThis article is one of a selection of papers from the conference "Exploiting Genome-wide Association in Oilseed Brassicas: a model for genetic improvement of major OECD crops for sustainable farming". *Genome*, *53*(11), 876–883. https://doi.org/10.1139/G10-076

Heavens, D., Accinelli, G. G., Clavijo, B., & Clark, M. D. (2015). A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost. *BioTechniques*, *59*(1). https://doi.org/10.2144/000114310

Henderson, C. R. (1984). Applications of linear models in animal breeding. *Applications of Linear Models in Animal Breeding.*

Hendre, P. S., Kamalakannan, R., Rajkumar, R., & Varghese, M. (2011). High-throughput targeted SNP discovery using Next Generation Sequencing (NGS) in few selected candidate genes in Eucalyptus camaldulensis. *BMC Proceedings*, *5*(S7). https://doi.org/10.1186/1753-6561-5-s7-o17

Hirakawa, H., Nakamura, Y., Kaneko, T., Isobe, S., Sakai, H., Kato, T., … Sato, S. (2011). Survey of the genetic information carried in the genome of Eucalyptus camaldulensis. *Plant Biotechnology*, *28*(5), 471–480. https://doi.org/10.5511/plantbiotechnology.11.1027b

Hirschhorn, J. N., & Daly, M. J. (2005, February). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, Vol. 6, pp. 95–108. https://doi.org/10.1038/nrg1521

Hodel, R. G. J., Segovia-Salcedo, M. C., Landis, J. B., Crowl, A. A., Sun, M., Liu, X., … Soltis, P. S. (2016). The Report of My Death was an Exaggeration: A Review for Researchers Using Microsatellites in the 21st Century. *Applications in Plant Sciences*, *4*(6), 1600025. https://doi.org/10.3732/apps.1600025

Hoisington, D., González de León, D., & Khairallah, M. (1994). *Laboratory protocols: CIMMYT applied molecular genetics laboratory protocols* (2da edició). México, CIMMYT, 1994.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data, R package version 1.5., 2012. *Journal of Statistical Software*, *45*(7), 1–3.

Inglis, P. W., Pappas, M. de C. R., Resende, L. V., & Grattapaglia, D. (2018). Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for

high-throughput SNP genotyping and sequencing applications. *PLOS ONE*, *13*(10), e0206085. https://doi.org/10.1371/journal.pone.0206085

Isik, F. (2014). Genomic selection in forest tree breeding: The concept and an outlook to the future. *New Forests*, Vol. 45, pp. 379–401. https://doi.org/10.1007/s11056-014-9422-z

Isik, F., Kumar, S., Martínez-García, P. J., Iwata, H., & Yamamoto, T. (2015). Acceleration of Forest and Fruit Tree Domestication by Genomic Selection. In *Advances in Botanical Research* (Vol. 74). https://doi.org/10.1016/bs.abr.2015.05.002

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Jombart, T., & Collins, C. (2015). *A tutorial for Discriminant Analysis of Principal Components (DAPC) using adegenet 2.0.0*.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1), 94. https://doi.org/10.1186/1471-2156-11-94

Jones, N., Naidoo, S., Kanzlera, A., & Myburg, A. (2019). Genomic Prediction by Combining Data Across Eucalyptus dunnii Populations. *IUFRO*.

Kainer, D., Padovan, A., Degenhardt, J., Krause, S., Mondal, P., Foley, W. J., & Külheim, C. (2019). High marker density <scp>GWAS</scp> provides novel insights into the genomic architecture of terpene oil yield in *Eucalyptus*. *New Phytologist*, *223*(3), 1489–1504. https://doi.org/10.1111/nph.15887

Kainer, D., Stone, E. A., Padovan, A., Foley, W. J., & Külheim, C. (2018). Accuracy of genomic prediction for foliar terpene traits in Eucalyptus polybractea. *G3: Genes, Genomes, Genetics*, *8*(8), 2573–2583. https://doi.org/10.1534/g3.118.200443

Kang, M. H., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, *178*(3), 1709–1723. https://doi.org/10.1534/genetics.107.080101

Kerem, B. S., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., … Tsui, L. C. (1989). Identification of the cystic fibrosis gene: Genetic analysis. *Science*, *245*(4922), 1073–1080. https://doi.org/10.1126/science.2570460

Kess, T., Gross, J., Harper, F., & Boulding, E. G. (2015). Low-cost ddRAD method of SNP discovery and genotyping applied to the periwinkle *Littorina saxatilis*. *Journal of Molluscan Studies*, eyv042. https://doi.org/10.1093/mollus/eyv042

Khlestkina, E. K., & Salina, E. A. (2006, June). SNP markers: Methods of analysis, ways of development, and comparison on an example of common wheat. *Russian Journal of Genetics*, Vol. 42, pp. 585–594. https://doi.org/10.1134/S1022795406060019

Klahre, U., Friederich, E., Kost, B., Louvard, D., & Chua, N. H. (2000). Villin-like actin-binding proteins are expressed ubiquitously in Arabidopsis. *Plant Physiology*, *122*(1), 35–47. https://doi.org/10.1104/pp.122.1.35

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., … Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, *308*(5720), 385–389. https://doi.org/10.1126/science.1109557

Komar, A. A. (2016). The Yin and Yang of codon usage. *Human Molecular Genetics*, *25*(R2), R77–R85. https://doi.org/10.1093/hmg/ddw207

Külheim, C., Hui Yeoh, S., Maintz, J., Foley, W. J., & Moran, G. F. (2009). Comparative SNP diversity among four Eucalyptus species for genes from secondary metabolite biosynthetic pathways. *BMC Genomics*, *10*, 452. https://doi.org/10.1186/1471-2164-10-452

Kullan, A. R. K., van Dyk, M. M., Hefer, C. A., Jones, N., Kanzler, A., & Myburg, A. A. (2012). Genetic dissection of growth, wood basic density and gene expression in interspecific backcrosses of Eucalyptus grandis and E. urophylla. *BMC Genetics*, *13*, 60. https://doi.org/10.1186/1471-2156-13-60

Ladiges, P. Y., Udovicic, F., & Nelson, G. (2003). Australian biogeographical connections and the phylogeny of large genera in the plant family Myrtaceae. *Journal of Biogeography*, *30*(7), 989–998. https://doi.org/10.1046/j.1365-2699.2003.00881.x

Lamara, M., Raherison, E., Lenz, P., Beaulieu, J., Bousquet, J., & MacKay, J. (2016). Genetic architecture of wood properties based on association analysis and co-expression networks in white spruce. *New Phytologist*, *210*(1), 240–255. https://doi.org/10.1111/nph.13762

Lange, V., Böhme, I., Hofmann, J., Lang, K., Sauter, J., Schöne, B., … Schmidt, A. H. (2014). Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genomics*, *15*(1). https://doi.org/10.1186/1471-2164-15-63

Lee, S. J., Ban, S. H., Kim, G. H., Kwon, S. Il, Kim, J. H., & Choi, C. (2017). Identification of potential gene-associated major traits using GBS-GWAS for Korean apple germplasm collections. *Plant Breeding*, *136*(6), 977–986. https://doi.org/10.1111/pbr.12544

Legarra, A., Aguilar, I., & Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*, *92*(9), 4656–4663. https://doi.org/10.3168/jds.2009-2061

Legarra, A., Calenge, F., Mariani, P., Velge, P., & Beaumont, C. (2011). Use of a reduced set of single nucleotide polymorphisms for genetic evaluation of resistance to Salmonella carrier state in laying hens. *Poultry Science*, *90*(4), 731–736. https://doi.org/10.3382/ps.2010-01260

Legarra, Andrés, Robert-Granié, C., Manfredi, E., & Elsen, J. M. (2008). Performance of genomic selection in mice. *Genetics*, *180*(1), 611–618. https://doi.org/10.1534/genetics.108.088575

Lenz, P. R. N., Beaulieu, J., Mansfield, S. D., Clément, S., Desponts, M., & Bousquet, J. (2017). Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (Picea mariana). *BMC Genomics*, *18*(1), 335. https://doi.org/10.1186/s12864-017-3715-5

Lepais, O., & Weir, J. T. (2014). SimRAD: An R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. *Molecular Ecology Resources*, *14*(6), 1314–1321. https://doi.org/10.1111/1755-0998.12273

Li, B., & Kimmel, M. (2013a). Factors influencing ascertainment bias of microsatellite allele sizes: Impact on estimates of mutation rates. *Genetics*, *195*(2), 563–572. https://doi.org/10.1534/genetics.113.154161

Li, B., & Kimmel, M. (2013b). Factors influencing ascertainment bias of microsatellite allele sizes: Impact on estimates of mutation rates. *Genetics*, *195*(2), 563–572. https://doi.org/10.1534/genetics.113.154161

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Subgroup, D. P. (2009). The Sequence Alignment/Map format and SAMtools. *BIOINFORMATICS APPLICATIONS NOTE*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, X. Y., Lin, E. P., Huang, H. H., Niu, M. Y., Tong, Z. K., & Zhang, J. H. (2018). Molecular characterization of SQUAMOSA PROMOTER BINDING PROTEIN-LIKE (SPL) gene family in Betula luminifera. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.00608

Li, Y., Levran, O., Kim, J. J., Zhang, T., Chen, X., & Suo, C. (2019). Extreme sampling design in genetic association mapping of quantitative trait loci using balanced and unbalanced case-control samples. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-51790-w

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., … Zhang, Z. (2012). GAPIT: Genome association and prediction integrated tool. *Bioinformatics*, *28*(18), 2397–2399. https://doi.org/10.1093/bioinformatics/bts444

Lopez-Casado, G., Urbanowicz, B. R., Damasceno, C. M., & Rose, J. K. (2008, June). Plant glycosyl hydrolases and biofuels: a natural marriage. *Current Opinion in Plant Biology*, Vol. 11, pp. 329–337. https://doi.org/10.1016/j.pbi.2008.02.010

López, J. A., Borralho, N. M., López, A. J., Marcó, M. A., & Harrand, L. (2012). Variación Genética del Índice de Rajado Rollizos en Eucalyptus dunnii. *Simposio IUFRO- Eucaliptos Geneticamente Mejorados Para Aumentar La Competitividad Del Sector Forestal En America Latina*. Pucón, Chile.

Mackay, I., & Powell, W. (2007, February). Methods for linkage disequilibrium mapping in crops. *Trends in Plant Science*, Vol. 12, pp. 57–63. https://doi.org/10.1016/j.tplants.2006.12.001

Mackay, T. F. C. (2001). The Genetic Architecture of Quantitative Traits. *Annual Review of Genetics*, *35*(1), 303–339. https://doi.org/10.1146/annurev.genet.35.102401.090633

Mandrou, E., Hein, P. R. G., Villar, E., Vigneron, P., Plomion, C., & Gion, J. M. (2012). A candidate gene for lignin composition in Eucalyptus: Cinnamoyl-CoA reductase (CCR). *Tree Genetics and Genomes*, *8*(2), 353–364. https://doi.org/10.1007/s11295-011-0446-7

Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, *27*, 209–220.

Marcó, M. A. (2005). Eucalyptus de Rápido Crecimiento para Usos Sólidos. *IDIA XXI. Forestal Ed., INTA*, *5*(8), 178–179. Retrieved from http://www.biblioteca.org.ar/libros/210585.pdf

Marcó, M., & White, T. L. (2002). Genetic Parameter Estimates and Genetic Gains for Eucalyptus Grandis and E. Dunnii in Argentina. *Forest Genetics*, *9*(3), 205–215. Retrieved from http://lf.tuzvo.sk/files/fg/volumes/2002/FG09-3_205-215.pdf

Marcucci Poltri S. N. , Zelener N., Rodriguez Traverso J., G. P. and H. H. E. (2003). *Selection of a seed orchard of Eucalyptus dunnii based on genetic diversity criteria calculated using molecular markers*. https://doi.org/https://doi.org/10.1093/treephys/23.9.625

McCann, M. C., & Carpita, N. C. (2008, June 1). Designing the deconstruction of plant cell walls. *Current Opinion in Plant Biology*, Vol. 11, pp. 314–320. https://doi.org/10.1016/j.pbi.2008.04.001

Mckown, A. D., Klápště, J., Guy, R. D., Geraldes, A., Porth, I., Hannemann, J., … Douglas, C. J. (2014). Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of Populus trichocarpa. *New Phytologist*, *203*(2), 535–553. https://doi.org/10.1111/nph.12815

Merino, G. (2018). *Imputación de Genotipos Faltantes en Datos de Secuencación Masiva*. Universidad Nacional de Córdoba, Córdoba, Argentina.

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

Meuwissen, T. H. E., Luan, T., & Woolliams, J. A. (2011). The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. *Journal of Animal Breeding and Genetics*, *128*(6), 429–439. https://doi.org/10.1111/j.1439-0388.2011.00966.x

Mewalal, R., Mizrachi, E., Mansfield, S. D., & Myburg, A. A. (2014). Cell Wall-Related Proteins of Unknown Function: Missing Links in Plant Cell Wall Development. *Plant and Cell Physiology*, *55*(6), 1031–1043. https://doi.org/10.1093/pcp/pcu050

Misztal, I. (2016). Is genomic selection now a mature technology? *Journal of Animal Breeding and Genetics*, *133*(2), 81–82. https://doi.org/10.1111/jbg.12209

Misztal, I., Legarra, A., & Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of Dairy Science*, *92*(9), 4648–4655. https://doi.org/10.3168/jds.2009-2064

Mitchell-Olds, T. (2010). Complex-trait analysis in plants. *Genome Biology*, *11*(4), 113. https://doi.org/10.1186/gb-2010-11-4-113

Mizrachi, E., & Myburg, A. A. (2016, April 1). Systems genetics of wood formation. *Current Opinion in Plant Biology*, Vol. 30, pp. 94–100. https://doi.org/10.1016/j.pbi.2016.02.007

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G. Y., & Myles, S. (2015). LinkImpute: Fast and accurate genotype imputation for nonmodel organisms. *G3: Genes, Genomes, Genetics*, *5*(11), 2383–2390. https://doi.org/10.1534/g3.115.021667

Müller, B. S. F., de Almeida Filho, J. E., Lima, B. M., Garcia, C. C., Missiaggia, A., Aguiar, A. M., … Grattapaglia, D. (2019). Independent and Joint-GWAS for growth traits in Eucalyptus by assembling genome-wide data for 3373 individuals across four breeding populations. *New Phytologist*, *221*(2), 818–833. https://doi.org/10.1111/nph.15449

Müller, B. S. F., de Almeida Filho, J. E., Lima, B. M., Garcia, C. C., Missiaggia, A., Aguiar, A. M., … Grattapaglia, D. (2019). Independent and Joint- <scp>GWAS</scp> for growth traits in *Eucalyptus* by assembling genome-wide data for 3373 individuals across four breeding populations. *New Phytologist*, *221*(2), 818–833. https://doi.org/10.1111/nph.15449

Müller, B. S. F., Neves, L. G., de Almeida Filho, J. E., Resende, M. F. R., Muñoz, P. R., dos Santos, P. E. T., … Grattapaglia, D. (2017). Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of Eucalyptus. *BMC Genomics*, *18*(1). https://doi.org/10.1186/s12864-017-3920-2

Muñoz, F., & Sánchez, L. (2014). breedR: Statistical Methods for Forest Genetic Resources Analysts. Retrieved from https://github.com/famuvie/breedR

Muñoz, P. R., Resende, M. F. R., Gezan, S. A., Resende, M. D. V., de los Campos, G., Kirst, M., … Peter, G. F. (2014). Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics*, *198*(4), 1759–1768. https://doi.org/10.1534/genetics.114.171322

Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., … Schmutz, J. (2014). The genome of Eucalyptus grandis. *Nature*, *510*(7505), 356–362. https://doi.org/10.1038/nature13308

Naidoo, R.; Jones, N.; Kanzler, A.; Myburg, A. (2018). *Genomic selection modelling of growth and wood properties in Eucalyptus dunnii*. Retrieved from IUFRO

Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology*, *15*(6), 1044–1045. https://doi.org/10.1093/beheco/arh107

Nasholm, T., Palmroth, S., Ganeteg, U., Moshelion, M., Hurry, V., & Franklin, O. (2014). Genetics of superior growth traits in trees are being mapped but will the faster-growing risk-takers make it in the wild? *Tree Physiology*, *34*(11), 1141–1148. https://doi.org/10.1093/treephys/tpu112

Nazareno, A. G., Bemmels, J. B., Dick, C. W., & Lohmann, L. G. (2017). Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Molecular Ecology Resources*, *17*(6), 1136–1147. https://doi.org/10.1111/1755-0998.12654

Neale, D. B., & Kremer, A. (2011). Forest tree genomics: Growing resources and applications. *Nature Reviews Genetics*, *12*(2), 111–122. https://doi.org/10.1038/nrg2931

Negro, S. S., Millet, E. J., Madur, D., Bauland, C., Combes, V., Welcker, C., … Nicolas, S. D. (2019). Genotyping-by-sequencing and SNP-arrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies. *BMC Plant Biology*, *19*(1). https://doi.org/10.1186/s12870-019-1926-4

Nejati-Javaremi, A., Smith, C., & Gibson, J. P. (1997). Effect of Total Allelic Relationship on Accuracy of Evaluation and Response to Selection. *Journal of Animal Science*, *75*(7), 1738–1745.

https://doi.org/10.2527/1997.7571738x

Núñez, C. (2004). Microestructura de la madera. *PROCYP, Universida*, 5.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y. C., Scofield, D. G., … Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, *497*(7451), 579–584. https://doi.org/10.1038/nature12211

Pan, J., Wang, B., Pei, Z. Y., Zhao, W., Gao, J., Mao, J. F., & Wang, X. R. (2015). Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. *Molecular Ecology Resources*, *15*(4), 711–722. https://doi.org/10.1111/1755-0998.12342

Parchman, T. L., Gompert, Z., Mudge, J., Schilkey, F. D., Benkman, C. W., & Buerkle, C. A. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, *21*(12), 2991–3005. https://doi.org/10.1111/j.1365-294X.2012.05513.x

Parchman, T. L., Jahner, J. P., Uckele, K. A., Galland, L. M., & Eckert, A. J. (2018). RADseq approaches and applications for forest tree genetics. *Tree Genetics and Genomes*, *14*(3). https://doi.org/10.1007/s11295-018-1251-3

PATTERSON, H. D., & THOMPSON, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554. https://doi.org/10.1093/biomet/58.3.545

Peña, D. la, Ramos, & Flores. (2012). Las heladas del 7 8 y 9 de Junio en la costa del río Uruguay y sus efectos sobre las plantaciones de eucalipto. Retrieved from https://www.researchgate.net/publication/320347397_Las_heladas_del_7_8_y_9_de_Junio_en_la_costa_del_rio_Uruguay_y_sus_efectos_sobre_las_plantaciones_de_eucalipto/references

Pérez-Rodríguez, P., Crossa, J., Rutkoski, J., Poland, J., Singh, R., Legarra, A., … Dreisigacker, S. (2017). Single-Step Genomic and Pedigree Genotype × Environment Interaction Models for Predicting Wheat Lines in International Environments. *The Plant Genome*, *10*(2), plantgenome2016.09.0089. https://doi.org/10.3835/plantgenome2016.09.0089

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*(5). https://doi.org/10.1371/journal.pone.0037135

Peterson, G. W., Dong, Y., Horbach, C., & Fu, Y. B. (2014). Genotyping-by-sequencing for plant genetic diversity analysis: A lab guide for SNP genotyping. *Diversity*, *6*(4), 665–680. https://doi.org/10.3390/d6040665

Petzold, H. E., Rigoulot, S. B., Zhao, C., Chanda, B., Sheng, X., Zhao, M., … Brunner, A. M. (2018). Identification of new protein–protein and protein–DNA interactions linked with wood formation in Populus trichocarpa. *Tree Physiology*, *38*(3), 362–377. https://doi.org/10.1093/TREEPHYS/TPX121

Plomion, C., Leprovost, G., & Stokes, A. (2001, December 1). Wood formation in trees. *Plant Physiology*, Vol. 127, pp. 1513–1523. https://doi.org/10.1104/pp.010816

Poke, F. S., Vaillancourt, R. E., Potts, B. M., & Reid, J. B. (2005, September). Genomic research in Eucalyptus. *Genetica*, Vol. 125, pp. 79–101. https://doi.org/10.1007/s10709-005-5082-4

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE*, *7*(2). https://doi.org/10.1371/journal.pone.0032253

Poland, J. A., & Rife, T. W. (2012). Genotyping-by-Sequencing for Plant Breeding and Genetics. *The Plant Genome Journal*, *5*(3), 92. https://doi.org/10.3835/plantgenome2012.05.0005

Porth, I., Klapšte, J., Skyba, O., Hannemann, J., McKown, A. D., Guy, R. D., … Mansfield, S. D. (2013). Genome-wide association mapping for wood characteristics in *Populus* identifies an array of candidate single nucleotide polymorphisms. *New Phytologist*, *200*(3), 710–726. https://doi.org/10.1111/nph.12422

Porth, I., Ranjan, P., Klapšte, J., Guy, R. D., Tuskan, G. A., Ehlting, J., … Hannemann, J. (2013). Genome-wide association mapping for wood characteristics in Populus identifies an array of candidate single nucleotide polymorphisms . *New Phytologist*, 710–726. https://doi.org/10.1111/nph.12422

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Pritchard, J. K., Stephens, M., Rosenberg, N. A., & Donnelly, P. (2000). Association mapping in structured populations. *American Journal of Human Genetics*, *67*(1), 170–181. https://doi.org/10.1086/302959

Putz, A. M., Tiezzi, F., Maltecca, C., Gray, K. A., & Knauer, M. T. (2018). A comparison of accuracy validation methods for genomic and pedigree-based predictions of swine litter size traits using Large White and simulated data. *Journal of Animal Breeding and Genetics*, *135*(1), 5–13. https://doi.org/10.1111/jbg.12302

Pyne, R., Honig, J., Vaiciunas, J., Koroch, A., Wyenandt, C., Bonos, S., & Simon, J. (2017). *A first linkage map and downy mildew resistance QTL discovery for sweet basil (Ocimum basilicum) facilitated by double digestion restriction site associated DNA sequencing (ddRADseq)*. https://doi.org/10.1371/journal.pone.0184319

Qin, H., Yang, G., Provan, J., Liu, J., & Gao, L. (2017). Using MiddRAD-seq data to develop polymorphic microsatellite markers for an endangered yew species. *Plant Diversity*, *39*(5), 294–299. https://doi.org/10.1016/j.pld.2017.05.008

Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., … Turner, D. J. (2008). A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, *5*(12), 1005–1010. https://doi.org/10.1038/nmeth.1270

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

R Core Team. (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, (Vienna, Austria). Retrieved from http://www.r-project.org/

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology*, Vol. 5, pp. 94–100. https://doi.org/10.1016/S1369-5266(02)00240-6

Rafalski, J. A. (2010, April). Association genetics in crop improvement. *Current Opinion in Plant Biology*, Vol. 13, pp. 174–180. https://doi.org/10.1016/j.pbi.2009.12.004

Ragauskas, A. J., Nagy, M., Kim, D. H., Eckert, C. A., Hallett, J. P., & Liotta, C. L. (2006, March). From wood to fuels. *Industrial Biotechnology*, Vol. 2, pp. 55–65. https://doi.org/10.1089/ind.2006.2.55

Ramensky, V. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, *30*(17), 3894–3900. https://doi.org/10.1093/nar/gkf493

Ratcliffe, B., El-Dien, O. G., Klápště, J., Porth, I., Chen, C., Jaquish, B., & El-Kassaby, Y. A. (2015). A comparison of genomic selection models across time in interior spruce (Picea engelmannii × glauca) using unordered SNP imputation methods. *Heredity*, *115*(6), 547–555. https://doi.org/10.1038/hdy.2015.57

Ratcliffe, Blaise, El-Dien, O. G., Cappa, E. P., Porth, I., Klápště, J., Chen, C., & El-Kassaby, Y. A. (2017a). Single-step BLUP with varying genotyping effort in open-pollinated Picea glauca. *G3: Genes, Genomes, Genetics*, *7*(3), 935–942. https://doi.org/10.1534/g3.116.037895

Ratcliffe, Blaise, El-Dien, O. G., Cappa, E. P., Porth, I., Klápště, J., Chen, C., & El-Kassaby, Y. A. (2017b). Single-Step BLUP with Varying Genotyping Effort in Open-Pollinated *Picea glauca*. *G3: Genes|Genomes|Genetics*, *7*(3), 935–942. https://doi.org/10.1534/g3.116.037895

Rencoret, J., Gutié Rrez, A., & Del Río, J. C. (2007). Lipid and lignin composition of woods from different

eucalypt species. *Holzforschung, 61*, 165–174. https://doi.org/10.1515/HF.2007.030

Resende, J. F. R., Muñoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., … Kirst, M. (2012). Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). *Genetics*, *190*(4), 1503–1510. https://doi.org/10.1534/genetics.111.137026

Resende, M. D. V., Resende, M. F. R., Sansaloni, C. P., Petroli, C. D., Missiaggia, A. A., Aguiar, A. M., … Grattapaglia, D. (2012a). Genomic selection for growth and wood quality in Eucalyptus: Capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytologist*, *194*(1), 116–128. https://doi.org/10.1111/j.1469-8137.2011.04038.x

Resende, M. F. R., Muñoz, P., Acosta, J. J., Peter, G. F., Davis, J. M., Grattapaglia, D., … Kirst, M. (2012). Accelerating the domestication of trees using genomic selection: Accuracy of prediction models across ages and environments. *New Phytologist*, *193*(3), 617–624. https://doi.org/10.1111/j.1469-8137.2011.03895.x

Resende, R. T., Resende, M. D. V., Silva, F. F., Azevedo, C. F., Takahashi, E. K., Silva-Junior, O. B., & Grattapaglia, D. (2017). Assessing the expected response to genomic selection of individuals and families in Eucalyptus breeding with an additive-dominant model. *Heredity*, *119*(4), 245–255. https://doi.org/10.1038/hdy.2017.37

Resende, Rafael Tassinari, Resende, M. D. V., Silva, F. F., Azevedo, C. F., Takahashi, E. K., Silva-Junior, O. B., & Grattapaglia, D. (2017a). Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in *Eucalyptus*. *New Phytologist*, *213*(3), 1287–1300. https://doi.org/10.1111/nph.14266

Resende, Rafael Tassinari, Resende, M. D. V., Silva, F. F., Azevedo, C. F., Takahashi, E. K., Silva-Junior, O. B., & Grattapaglia, D. (2017b). Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in Eucalyptus. *New Phytologist*, *213*(3), 1287–1300. https://doi.org/10.1111/nph.14266

Resende, M. D. V., Resende, M. F. R., Sansaloni, C. P., Petroli, C. D., Missiaggia, A. A., Aguiar, A. M., … Grattapaglia, D. (2012b). Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytologist*, *194*(1), 116–128. https://doi.org/10.1111/j.1469-8137.2011.04038.x

Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Retrieved from https://cran.r-project.org/package=psych

Rezende, G. D. S. P., de Resende, M. D. V., & de Assis, T. F. (2014). *Eucalyptus Breeding for Clonal Forestry*. https://doi.org/10.1007/978-94-007-7076-8_16

Rochette, N. C., & Catchen, J. M. (2017). Deriving genotypes from RAD-seq short-read data using Stacks. *Nature Protocols*, *12*(12), 2640–2659. https://doi.org/10.1038/nprot.2017.123

Rodrigues, J., Faix, O., & Pereira, H. (1998). Determination of lignin content of Eucalyptus globulus wood using FTIR spectroscopy. *Holzforschung*, *52*(1), 46–50. https://doi.org/10.1515/hfsg.1998.52.1.46

Rollins, J. A., Drosse, B., Mulki, M. A., Grando, S., Baum, M., Singh, M., … von Korff, M. (2013). Variation at the vernalisation genes Vrn-H1 and Vrn-H2 determines growth and yield stability in barley (Hordeum vulgare) grown under dryland conditions in Syria. *Theoretical and Applied Genetics*, *126*(11), 2803–2824. https://doi.org/10.1007/s00122-013-2173-y

Roy, S. C., Moitra, K., & De Sarker, D. (2017). Assessment of genetic diversity among four orchids based on ddRAD sequencing data for conservation purposes. *Physiology and Molecular Biology of Plants*, *23*(1), 169–183. https://doi.org/10.1007/s12298-016-0401-z

Rutkoski, J. E., Poland, J., Jannink, J. L., & Sorrells, M. E. (2013). Imputation of unordered markers and the impact on genomic selection accuracy. *G3: Genes, Genomes, Genetics*, *3*(3), 427–439. https://doi.org/10.1534/g3.112.005363

Sansaloni, C. P., Petroli, C. D., Carling, J., Hudson, C. J., Steane, D. A., Myburg, A. A., … Kilian, A. (2010). A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in Eucalyptus. *Plant Methods*, *6*(1), 1–11. https://doi.org/10.1186/1746-4811-6-16

Savolainen, O., & Kärkkäinen, K. (1992). Effect of forest management on gene pools. *New Forests*, *6*(1–4), 329–345. https://doi.org/10.1007/BF00120651

Scaglione, D., Fornasiero, A., Pinto, C., Cattonaro, F., Spadotto, A., Infante, R., … Testolin, R. (2015). A RAD-based linkage map of kiwifruit (Actinidia chinensis Pl.) as a tool to improve the genome assembly and to scan the genomic region of the gender determinant for the marker-assisted breeding. *Tree Genetics and Genomes*, *11*(6). https://doi.org/10.1007/s11295-015-0941-3

Scheben, A., Batley, J., & Edwards, D. (2017). Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnology Journal*, *15*(2), 149–161. https://doi.org/10.1111/pbi.12645

Schimleck, L., Raymond, C., Beadle, C., Downes, G., Kube, P., & French, J. (2000). Applications of NIR spectroscopy to forest research. *Appita Journal: Journal of the Technical Association of the Australian and New Zealand Pulp and Paper Industry*. Retrieved from https://epubs.scu.edu.au/cpcg_pubs/591

Schwanninger, M., & Hinterstoisser, B. (2002). Klason lignin: Modifications to improve the precision of the standardized determination. *Holzforschung*, *56*(2), 161–166. https://doi.org/10.1515/HF.2002.027

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Secretaría de Agroindustria. (2018). *Sector Forestal 2017*. Retrieved from https://www.magyp.gob.ar/sitio/areas/ss_desarrollo_foresto_industrial/estadisticas/_archivos//000000_Sector Forestal/000000_Informes/170000_2017 - Sector Forestal.pdf

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Sharmaa, A., Lee, J. S., Dang, C. G., Sudrajad, P., Kim, H. C., Yeon, S. H., … Lee, S. H. (2015, October 1). Stories and challenges of genome wide association studies in livestock - a review. *Asian-Australasian Journal of Animal Sciences*, Vol. 28, pp. 1371–1379. https://doi.org/10.5713/ajas.14.0715

Silva-Junior, O. B., Faria, D. A., & Grattapaglia, D. (2015). A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 Eucalyptus tree genomes across 12 species. *New Phytologist*, *206*(4), 1527–1540. https://doi.org/10.1111/nph.13322

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., … Belzile, F. (2013). An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. *PLoS ONE*, *8*(1), 1–9. https://doi.org/10.1371/journal.pone.0054603

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(16), 9440–9445. https://doi.org/10.1073/pnas.1530509100

Strandén, I., & Garrick, D. J. (2009). Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of Dairy Science*, *92*, 2971–2975. https://doi.org/10.3168/jds.2008-1929

Sun, G., Zhu, C., Kramer, M. H., Yang, S. S., Song, W., Piepho, H. P., & Yu, J. (2010). Variation explained in mixed-model association mapping. *Heredity*, *105*(4), 333–340. https://doi.org/10.1038/hdy.2010.11

Suontama, M., Klápště, J., Telfer, E., Graham, N., Stovold, T., Low, C., … Dungey, H. (2019). Efficiency of genomic prediction across two Eucalyptus nitens seed orchards with different selection histories. *Heredity*, *122*(3), 370–379. https://doi.org/10.1038/s41437-018-0119-5

Tan, B., Grattapaglia, D., Martins, G. S., Ferreira, K. Z., Sundberg, B., & Ingvarsson, P. K. (2017). Evaluating

the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. *BMC Plant Biology*, *17*(1), 110. https://doi.org/10.1186/s12870-017-1059-6

Tan, G., Opitz, L., Schlapbach, R., & Rehrauer, H. (2019). Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-39076-7

Thavamanikumar, S., McManus, L. J., Tibbits, J. F. G., & Bossinger, G. (2011). The significance of single nucleotide polymorphisms (SNPs) in Eucalyptus globulus breeding programs. *Australian Forestry*, *74*(1), 23–29. https://doi.org/10.1080/00049158.2011.10676342

Thiel, T., Michalek, W., Varshney, R. K., & Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theoretical and Applied Genetics*, *106*(3), 411–422. https://doi.org/10.1007/s00122-002-1031-0

Thomas, D., Henson, M., Joe, B., Boyton, S., & Dickson, R. (2009). Review of growth and wood quality of plantation-grown Eucalyptus dunnii Maiden. *Australian Forestry*, *72*(1), 3–11. https://doi.org/10.1080/00049158.2009.10676283

Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D., & Buckler, E. S. (2001). Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics*, *28*(3), 286–289. https://doi.org/10.1038/90135

Thumma, B. R., Southerton, S. G., Bell, J. C., Owen, J. V., Henery, M. L., & Moran, G. F. (2010). *Quantitative trait locus (QTL) analysis of wood quality traits in Eucalyptus nitens*.

Timm, H., Weigand, H., Weiss, M., Leese, F., & Rahmann, S. (2018). ddrage: A data set generator to evaluate ddRADseq analysis software. *Molecular Ecology Resources*, *18*(3), 681–690. https://doi.org/10.1111/1755-0998.12743

Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R., & Bird, C. E. (2013). ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, *1*, e203. https://doi.org/10.7717/peerj.203

Torales, S. L., Rivarola, M., Gonzalez, S., Inza, M. V., Pomponio, M. F., Fernández, P., … Marcucci Poltri, S. N. (2018). De novo transcriptome sequencing and SSR markers development for Cedrela balansae C. DC., a native tree species of northwest Argentina. *PLoS ONE*, *13*(12). https://doi.org/10.1371/journal.pone.0203768

Torkamaneh, D., & Belzile, F. (2015). Scanning and filling: Ultra-dense SNP genotyping combining genotyping-by-sequencing, SNP array and whole-genome resequencing data. *PLoS ONE*, *10*(7), 1–16. https://doi.org/10.1371/journal.pone.0131533

Torkamaneh, D., Boyle, B., & Belzile, F. (2018). Efficient genome-wide genotyping strategies and data integration in crop plants. *Theoretical and Applied Genetics*, *131*(3), 499–511. https://doi.org/10.1007/s00122-018-3056-z

Torkamaneh, D., Laroche, J., & Belzile, F. (2016). Genome-wide SNP calling from genotyping by sequencing (GBS) data: A comparison of seven pipelines and two sequencing technologies. *PLoS ONE*, *11*(8), 1–14. https://doi.org/10.1371/journal.pone.0161333

Truong, H. T., Ramos, A. M., Yalcin, F., de Ruiter, M., van der Poel, H. J. A., Huvenaars, K. H. J., … van Eijk, M. J. T. (2012). Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLoS ONE*, *7*(5). https://doi.org/10.1371/journal.pone.0037565

Turner, S. R., & Somerville, C. R. (1997). Collapsed xylem phenotype of arabidopsis identifies mutants deficient in cellulose deposition in the secondary cell wall. *Plant Cell*, *9*(5), 689–701. https://doi.org/10.1105/tpc.9.5.689

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Vargas, O. M., Ortiz, E. M., & Simpson, B. B. (2017). Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: *Diplostephium*). *New Phytologist*, *214*(4), 1736–1750. https://doi.org/10.1111/nph.14530

Varshney, R. K., Roorkiwal, M., & Sorrells, M. E. (2017). Genomic selection for crop improvement: New molecular breeding strategies for crop improvement. In *Genomic Selection for Crop Improvement: New Molecular Breeding Strategies for Crop Improvement* (pp. 1–258). https://doi.org/10.1007/978-3-319-63170-7

Vilella, F. (2019). Bioeconomía 2030 – Forestación en Argentina. Retrieved from http://portallarroque.com.ar/bioeconomia-2030-forestacion-en-argentina/

Villanueva, B., Pong-Wong, R., Fernández, J., & Toro, M. A. (2005). Benefits from marker-assisted selection under an additive polygenic genetic model1. *Journal of Animal Science*, *83*(8), 1747–1752. https://doi.org/10.2527/2005.8381747x

Walker, J. C. F. (2006). Basic wood chemistry and cell wall ultrastructure. In *Primary Wood Processing: Principles and Practice* (Vol. 9781402043932, pp. 23–67). https://doi.org/10.1007/1-4020-4393-7_2

Wang, J., Song, J., Clark, G., & Roux, S. J. (2018). ANN1 and ANN2 function in post-phloem sugar transport in root tips to affect primary root growth. *Plant Physiology*, *178*(1), 390–401. https://doi.org/10.1104/pp.18.00713

Wang, M., Yan, J., Zhao, J., Song, W., Zhang, X., Xiao, Y., & Zheng, Y. (2012). Genome-wide association study (GWAS) of resistance to head smut in maize. *Plant Science*, *196*, 125–131. https://doi.org/10.1016/j.plantsci.2012.08.004

Wang, Shenhao, Yang, X., Xu, M., Lin, X., Lin, T., Qi, J., … Huang, S. (2015). A Rare SNP Identified a TCP Transcription Factor Essential for Tendril Development in Cucumber. *Molecular Plant*, *8*(12), 1795–1808. https://doi.org/10.1016/j.molp.2015.10.005

Wang, Shi, Meyer, E., Mckay, J. K., & Matz, M. V. (2012). 2b-RAD: A simple and flexible method for genome-wide genotyping. *Nature Methods*, *9*(8), 808–810. https://doi.org/10.1038/nmeth.2023

Wang, Y., Cao, X., Zhao, Y., Fei, J., Hu, X., & Li, N. (2017). Optimized double-digest genotyping by sequencing (ddGBS) method with high-density SNP markers and high genotyping accuracy for chickens. *PLOS ONE*, *12*(6), e0179073. https://doi.org/10.1371/journal.pone.0179073

Weir, B. S. (2008). Linkage Disequilibrium and Association Mapping. *Annual Review of Genomics and Human Genetics*, *9*(1), 129–142. https://doi.org/10.1146/annurev.genom.9.081307.164347

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Retrieved December 27, 2019, from Springer; New York; NY; USA website: https://www.springer.com/gp/book/9780387981413

Wickland, D. P., Battu, G., Hudson, K. A., Diers, B. W., & Hudson, M. E. (2017). A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. *BMC Bioinformatics*. https://doi.org/10.1186/s12859-017-2000-6

Wimmer, V., Albrecht, T., Auinger, H.-J., & Schön, C.-C. (2012). synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics*, *28*(15), 2086–2087. https://doi.org/10.1093/bioinformatics/bts335

Wingfield, M. J., Brockerhoff, E. G., Wingfield, B. D., & Slippers, B. (2015, August 21). Planted forest health: The need for a global strategy. *Science*, Vol. 349, pp. 832–836. https://doi.org/10.1126/science.aac6674

Wu, J., Feng, F., Lian, X., Teng, X., Wei, H., Yu, H., … Mei, H. (2015). Genome-wide Association Study (GWAS) of mesocotyl elongation based on re-sequencing approach in rice. *BMC Plant Biology*, *15*(1), 218. https://doi.org/10.1186/s12870-015-0608-0

Xiao, Y., Liu, H., Wu, L., Warburton, M., & Yan, J. (2017, March 6). Genome-wide Association Studies in Maize: Praise and Stargaze. *Molecular Plant*, Vol. 10, pp. 359–374.

https://doi.org/10.1016/j.molp.2016.12.008

Xing, A., Gao, Y., Ye, L., Zhang, W., Cai, L., Ching, A., … Li, J. (2015). A rare SNP mutation in Brachytic2 moderately reduces plant height and increases yield potential in maize. *Journal of Experimental Botany*, *66*(13), 3791–3802. https://doi.org/10.1093/jxb/erv182

Xing, C., & Xing, G. (2009). Power of selective genotyping in genome-wide association studies of quantitative traits. *BMC Proceedings*, *3*(S7), 1–5. https://doi.org/10.1186/1753-6561-3-s7-s23

Yan, L., Hofmann, N., Li, S., Ferreira, M. E., Song, B., Jiang, G., … Song, Q. (2017). Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. *BMC Genomics*, *18*(1), 529. https://doi.org/10.1186/s12864-017-3922-0

Yan, Y. Y., Burbridge, C., Shi, J., Liu, J., & Kusalik, A. (2019). Effects of input data quantity on genome-wide association studies (GWAS). *International Journal of Data Mining and Bioinformatics*, *22*(1), 19–43. https://doi.org/10.1504/IJDMB.2019.099286

Yang, G. Q., Chen, Y. M., Wang, J. P., Guo, C., Zhao, L., Wang, X. Y., … Guo, Z. H. (2016). Development of a universal and simplified ddRAD library preparation approach for SNP discovery and genotyping in angiosperm plants. *Plant Methods*, *12*(1), 1–17. https://doi.org/10.1186/s13007-016-0139-1

Yang, Z., Li, Z., & Bickel, D. R. (2013). Empirical Bayes estimation of posterior probabilities of enrichment: A comparative study of five estimators of the local false discovery rate. *BMC Bioinformatics*, *14*, 1–12. https://doi.org/10.1186/1471-2105-14-87

Yin, T., & König, S. (2019). Genome-wide associations and detection of potential candidate genes for direct genetic and maternal genetic effects influencing dairy cattle body weight at different ages. *Genetics Selection Evolution*, *51*(1), 4. https://doi.org/10.1186/s12711-018-0444-4

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., … Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203–208. https://doi.org/10.1038/ng1702

Zapata-Valenzuela, J., Isik, F., Maltecca, C., Wegrzyn, J., Neale, D., McKeand, S., & Whetten, R. (2012). SNP markers trace familial linkages in a cloned population of Pinus taeda-prospects for genomic selection. *Tree Genetics and Genomes*, *8*(6), 1307–1318. https://doi.org/10.1007/s11295-012-0516-5

Zapata-Valenzuela, J., Whetten, R. W., Neale, D., McKeand, S., & Isik, F. (2013). Genomic estimated breeding values using genomic relationship matrices in a cloned population of loblolly pine. *G3: Genes, Genomes, Genetics*, *3*(5), 909–916. https://doi.org/10.1534/g3.113.005975

Zelener, N., Poltri, S. N. M., Bartoloni, N., López, C. R., & Hopp, H. E. (2005). Selection strategy for a seedling seed orchard design based on trait selection index and genomic analysis by molecular markers: a case study for Eucalyptus dunnii. *Tree Physiology*, *25*(11), 1457–1467. https://doi.org/10.1093/TREEPHYS/25.11.1457

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., … Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, *42*(4), 355–360. https://doi.org/10.1038/ng.546

Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., … Nordborg, M. (2007). An Arabidopsis Example of Association Mapping in Structured Samples. *PLoS Genetics*, *3*(1), e4. https://doi.org/10.1371/journal.pgen.0030004

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

Zhong, R., Cui, D., & Ye, Z. H. (2019, March 1). Secondary cell wall biosynthesis. *New Phytologist*, Vol. 221, pp. 1703–1723. https://doi.org/10.1111/nph.15537

Zhou, X., Xia, Y., Ren, X., Chen, Y., Huang, L., Huang, S., … Jiang, H. (2014). Construction of a SNP-based genetic linkage map in cultivated peanut based on large scale marker development using next-generation double-digest restriction-site-associated DNA sequencing (ddRADseq). *BMC Genomics*, *15*(1), 1–14. https://doi.org/10.1186/1471-2164-15-351

Zhu, C., Gore, M., Buckler, E. S., & Yu, J. (2008). Status and Prospects of Association Mapping in Plants. *The Plant Genome*, *1*(1), 5–20. https://doi.org/10.3835/plantgenome2008.02.0089

7 ANEXO

1. Desarrollo de una Metodología de Genotipificación Masiva para *E. dunnii*

7.1.1 Individuos utilizados para la puesta a punto de los protocolos 1 y 2 de ddRADseq para E. dunnii.

Tabla 7.1.1. *Individuos utilizados para la puesta a punto de los protocolos 1 y 2 de ddRADseq para E. dunnii*. Ind.: letra (A o B) y número (1 al 24) que identifica al individuo de *E. dunnii*; Flia.: Número de familia; Origen: Origen australiano o precedencia local, SD: procedencia Local (Oliveros Santa Fe, Argentina; origen Australiano, Moleton), SY: South Yabra, DHAC: Dearth Horse Track Region y Acacia Creek, OC: Oaky Creek, BCUN: Boomi Creek y Unumungar State Forest; Resultados de secuenciación en equipos de Illumina MiSeq o NextSeq: Genotecas secuenciadas: nombres de genotecas secuenciadas; N° Lecturas PE: número de lecturas pareadas (PE) obtenidas; % Mapeado: porcentaje de lecturas mapeadas en el genoma de referencia de *E. grandis*; Cobertura: número de lecturas promedio para cada *locus*; N° *loci*: número de *loci* definidos por genoteca. Resultados del análisis con genoma de referencia con Stacks (Catchen et al 2013).

Ind.	Flia.	Origen	Genotecas secuenciadas		N° Lecturas PE		% Mapeado		Cobertura		Nº *loci*	
			MiSeq	NextSeq	MiSeq	NextSeq	MiSeq	NextSeq	MiSeq	NextSeq	MiSeq	NextSeq
A	237	DHAC	A	/	976759	/	78,44	/	19,26	/	77885	/
B	259	OC	B	/	1118618	/	86,07	/	29,05	/	71395	/
1	202	SD	MiSeq.1.202.SD	NextSeq.1.202.SD	153740	404742	85,38	82,86	4,33	8,25	12503	31733
2	215	SY	MiSeq.2.215.SY	NextSeq.2.215.SY	129755	1557495	84,18	82,07	4,86	13,22	17846	83532
3	219	SY	MiSeq.3.219.SY	NextSeq.3.219.SY	103217	1722699	84,64	81,88	4,7	13,59	16199	90587
4	222	DHAC	MiSeq.4.222.DH	NextSeq.4.222.DH	154290	870702	85,86	82,61	4,36	10,15	13319	63444
5	222	DHAC	MiSeq.5.222.DH	NextSeq.5.222.DH	137777	786091	84,53	82,16	4,76	9,64	16781	56951
6	228	DHAC	MiSeq.6.228.DH	NextSeq.6.228.DH	167245	942753	83,04	82,79	4,24	11,63	16058	58792
7	229	DHAC	MiSeq.7.229.DH	NextSeq.7.229.DH	92914	764896	85,61	83,11	4,28	9,31	18392	57578
8	231	DHAC	MiSeq.8.231.DH	NextSeq.8.231.DH	154509	1176239	85,17	81,81	4,34	10,87	19317	76898
9	233	DHAC	MiSeq.9.233.DH	NextSeq.9.233.DH	148645	760072	84,55	81,27	4,33	8,32	19358	59665
10	234	DHAC	MiSeq.10.234.DH	NextSeq.10.234.DH	171116	1385383	85,5	82,76	5,07	12,69	19701	78873

Ind.	Flia.	Origen	Genotecas secuenciadas		N° Lecturas PE		% Mapeado		Cobertura		Nº *loci*	
			Miseq	NextSeq	MiSeq	NextSeq	MiSeq	NextSeq	MiSeq	NextSeq	MiSeq	NextSeq
11	234	DHAC	MiSeq.11.234.DH	NextSeq.11.234.DH	117062	737492	85,18	83,09	4,17	9,23	14957	56017
12	239	DHAC	MiSeq.12.239.DH	NextSeq.12.239.DH	70699	840898	85	81,93	4,33	9,12	18684	61044
13	240	DHAC	MiSeq.13.240.DH	NextSeq.13.240.DH	141608	1370570	85,59	82,41	4,86	15,92	17707	62204
14	242	DHAC	MiSeq.14.242.DH	NextSeq.14.242.DH	142974	1480508	84,98	81,57	4,48	11,85	21233	88292
15	244	DHAC	MiSeq.15.244.DH	NextSeq.15.244.DH	161496	1026524	84,65	82,86	4,12	11,09	10151	66427
16	247	DHAC	MiSeq.16.247.DH	NextSeq.16.247.DH	146026	1777519	85,16	82,47	5	13,83	19639	94857
17	247	DHAC	MiSeq.17.247.DH	NextSeq.17.247.DH	138475	1011053	85,37	83,48	4,33	11,95	18358	63455
18	249	DHAC	MiSeq.18.249.DH	NextSeq.18.249.DH	129145	2280731	85,23	82,08	5,32	15,38	21438	110951
19	259	OC	MiSeq.19.259.OC	NextSeq.19.259.OC	99820	815768	85,3	83,01	4,56	12,22	15241	47306
20	262	BCUN	MiSeq.20.262.BC	NextSeq.20.262.BC	136232	1143559	84,78	82,52	3,94	11,84	7521	70083
21	262	BCUN	MiSeq.21.262.BC	NextSeq.21.262.BC	147082	598662	85,21	81,47	4,18	7,21	16640	52549
22	267	BCUN	MiSeq.22.267.BC	NextSeq.22.267.BC	151728	1151626	85,14	82,76	4,41	14,34	19943	56905
23	273	DHAC	MiSeq.23.273.AC	NextSeq.23.273.AC	158938	1134681	84,84	81,36	4,23	10,08	16543	77045
24	273	DHAC	/	NextSeq.24.273.AC	/	1659639	/	83,14	/	15,64	15854	81749
				Media	**137151,9**	**1141679,3**	**84,99**	**82,39**	**4,49**	**11,56**	**16807,6**	**68622,37**
				Desviación estándar	**25435,29**	**440543**	**0,58**	**0,62**	**0,35**	**2,44**	**3358,56**	**17386,07**

7.1.2 Protocolo de extracción de ADN con CTAB para E. dunnii

La extracción de ADN se realizó a partir del protocolo modificado, descrito por Marcucci Poltri y col. (2003). Los materiales utilizados fueron hojas jóvenes previamente liofilizadas mediante un liofilizador Labconco (EE.UU.) durante 48-72 horas (a -50ºC y 5-10 atmósferas); las hojas de cada muestra fueron trituradas por separado en un homogeneizador de tejidos denominado Tissue lyser (Qiagen, EE.UU.) hasta que se obtuvo un polvo fino que fue utilizado para la extracción del material genético.

Pasos:

1) Colocar 50 mg de polvo de cada muestra en tubos de 2 ml.
2) Agregar un 1 ml de buffer de extracción CTAB (ver Tabla I). Agitar con cuidado para evitar la formación de cúmulos de material.
3) Incubar en el baño a 65º C durante 60-90 minutos, sacándolos y volteándolos cada 20 minutos para una mejor homogenización.
4) Dejar enfriar a temperatura ambiente durante 4-5 minutos.
5) Agregar 500 µl de Cloroformo:Octanol (24:1) y agitar suavemente los tubos por 5-10 minutos.
6) Centrifugar 15 minutos a 13.000 rpm (conviene repetir el centrifugado para conseguir que precipiten los restos del material lo mejor posible).
7) Trasvasar el sobrenadante a un nuevo tubo de 2 ml (aprox. se obtienen unos 850 µl).
8) Agregar 1/10 de volumen de acetato de sodio 3M (85 µl) y 1 volumen de isopropanol (850 µl). Mezclar con cuidado hasta que comience a verse la madeja de ADN (es posible que la madeja no aparezca o que aparezca muy fragmentada).
9) Centrifugar 10 minutos a 13.000 rpm y desechar con cuidado el sobrenadante.
10) Agregar 1 ml de etanol 70% y agitar con cuidado.
11) Volver a centrifugar 10 minutos a 13.000 rpm, descartar el sobrenadante y dejar secar bajo campana hasta que se evapore el etanol. Si es necesario, dar un spin para desechar todo el sobrenadante.
12) Resuspender en 100 µl de TE 1X (10mM Tris-HCl, pH8; 1mM EDTA, pH8) o en 100 µl de agua ultra pura.
13) Incubar con 1 µl de ARNasa A (10mg/ml) durante 30 minutos a temperatura ambiente para eliminar el ARN de la muestra.
14) Guardar el ADN a -20ºC.

Tabla I: Composición del Buffer de Extracción CTAB

Buffer de extracción	Cc Final	10ml	100ml
H_2O		6,5ml	65ml
1M Tris-HCL pH 8	100mM	1ml	10ml
0,5M EDTA pH 8	50mM	1ml	10ml
5M NaCl	700mM	1,4ml	14ml
Mercaptoetanol	140mM	0,1ml	1ml
Bromuro de cetometilamonio (CTAB)	2%	0,2gr	2gr
Polivinilpolipirrolidona (PVP)	1%	0,1gr	1gr

- El CTAB está incorporado a la preparación del buffer de extracción y, tanto el PVP como el mercaptoetanol, se agregan en el momento de uso.
- Calentar primero el CTAB a 65 ºC.
- Cloroformo:Octanol => 24:1 (vol. total=25).

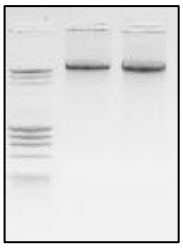
7.1.3 Verificación de la integridad del ADN genómico

-NanoDrop: Observar la calidad a través de la relación de absorbancia 260/280 y 260/230, según los siguientes valores óptimos:

- 260/280: entre 1,8 y 2,0 (>1,6 = aceptable),
- 260/230: mayor a 2 (> 1,8 = aceptable).

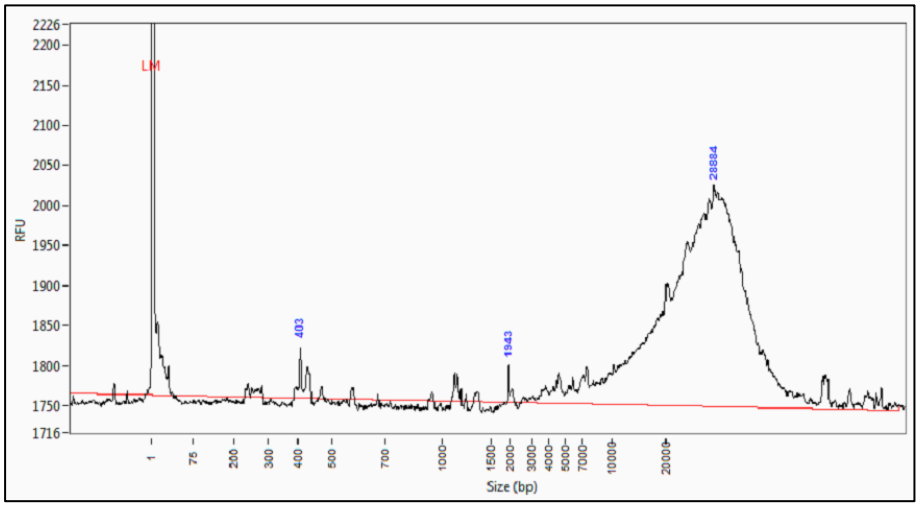
-Gel de Agarosa: Correr las muestras en geles de agarosa 1% teñidos con Bromuro de Etidio, utilizando un marcador de peso molecular como patrón de comparación (preferentemente con una banda = o > a 10 Kpb). Se debe observar una banda de ADN íntegro, sin degradación (Ver ANEXO Figura 1).

Figura 1: Gel de agarosa para verificación de Calidad de ADN genómico. De izquierda a derecha: Marker 23 Kpb (banda más grande) y 2 ADN genómicos de *Eucalyptus dunnii* de buena calidad.



-Fragment Analyzer (FA) o Bioanalyzer: Correr algunas muestras si es la primera vez que se verifica el método de extracción de ADN para la especie (Ver ANEXO Figura 2).

Figura 2: Corrida en Fragment Analyzer de ADN genómico de *Eucalyptus dunnii*, PM muy bueno de 28884pb.



7.1.4 Cuantificación de ADN genómico

Qubit: Cuantificar ADN genómico con kit dsDNA Broad Range:

-Utilizar tubos de pared delgada aptos para esta medición, y estándares 1 y 2 (0 y 100 ng/µl, respectivamente)

-Diluir Intercalante en Buffer 1/200, calculando 200µl de solución por cada tubo a medir + 2 extras: 200*(24 muestras+2 estándares + 2 extras) µl. Ejemplo para 24 muestras:

- ✓ Volumen total: 200*28 µl = 5600 µl.
- ✓ Intercalante: 28 µl.
- ✓ Buffer: 5600-28 = 5572 µl.

-Colocar 2 µl por muestra + 198 µl de Intercalante en Buffer 1/200 (rango de lectura: de 1 a 500ng/ µl).

-Mezclar con vórtex, dar un spin, y dejar incubar por 2 min a °t ambiente y protegidos de la luz.

-Para comenzar la medición en el equipo, elegir las opciones DNA, dsDNA broad Range, Read New Standars? YES (es preferible siempre preparar estándares nuevos y medirlos, para que tengan la misma solución de Intercalante en Buffer que las muestras).

-Medir Estándares y luego las muestras (Insert Assay tube – READ, Read Next Sample). Se le puede pedir al equipo que calcule la concentración final, indicando que el volumen de muestra utilizado fue de 2 µl (Opciones: Calculate Stock Conc., Volumen of Sample Used 2 µl, Choose Units: ng/µl), y guardar la tabla de salida en un pendrive.

7.1.5 Adaptadores y primers para ddRADSeq

Secuencias de oligonucleótidos para adaptadores universales (modificado de Peterson et al. 2014) **para PROTOCOLO ddRADseq IABiMo-INTA 1.0**

Oligo name	Barcode Sequence	Oligo sequence	5' Modification	Purification
SphIA	/	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCATG	/	/
SphIB	/	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
MboIA	/	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	/	/
MboIB	/	GATCAGATCGGAAGAGCGAGAACAA	/5Phos/	HPSF

Secuencias de oligonucleótidos para adaptadores con barcodes (modificado de Peterson *et al.* 2014 con 24 barcodes de los 48 de Poland *et al.* 2012) **para PROTOCOLO ddRADseq IABiMo-INTA 2.0**

Oligo name	Barcode Sequence	Oligo sequence	5' Modification	Purification
SphIA1	AAGTGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGTGACATG	/	/
SphIA3	ACAGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACAGACATG	/	/
SphIA4	ACCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACCACATG	/	/
SphIA5	AGAATGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGAATGACATG	/	/
SphIA8	AGTGTTAA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTGTTAACATG	/	/
SphIA10	ATCATACCT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATCATACCTCATG	/	/
SphIA14	CACGACCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCACGACCACATG	/	/
SphIA18	CCACTGG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCACTGGCATG	/	/
SphIA19	CCATCCACT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCATCCACTCATG	/	/
SphIA20	CCTG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCTGCATG	/	/
SphIA21	CGACG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGACGCATG	/	/
SphIA24	CTCACT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTCACTCATG	/	/
SphIA29	GACTCGG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGACTCGGCATG	/	/
SphIA31	GATGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGATGACATG	/	/
SphIA34	GGAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGAGCATG	/	/
SphIA35	GGCCGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGCCGACATG	/	/
SphIA36	GTATA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTATACATG	/	/
SphIA37	GTGCACCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTGCACCACATG	/	/
SphIA39	TATTCCACT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTATTCCACTCATG	/	/
SphIA42	TCTCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCTCACATG	/	/
SphIA43	TGCGAGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCGAGACATG	/	/
SphIA45	TGGC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGCCATG	/	/
SphIA46	TTGACCAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTGACCAGCATG	/	/
SphIA48	TTGTAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTGTAGCATG	/	/
SphIB1	TCACTT	TCACTTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB3	TCTGT	TCTGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB4	TGGT	TGGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB5	TCATTCT	TCATTCTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB8	TTAACACT	TTAACACTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB10	AGGTATGAT	AGGTATGATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB14	TGGTCGTG	TGGTCGTGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB18	CCAGTGG	CCAGTGGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB19	AGTGGATGG	AGTGGATGGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB20	CAGG	CAGGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB21	CGTCG	CGTCGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB24	AGTGAG	AGTGAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB29	CCGAGTC	CCGAGTCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB31	TCATC	TCATCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB34	CTCC	CTCCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB35	TCGGCC	TCGGCCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB36	TATAC	TATACAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB37	TGGTGCAC	TGGTGCACAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB39	AGTGGAATA	AGTGGAATAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB42	TGAGA	TGAGAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB43	TCTCGCA	TCTCGCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB45	GCCA	GCCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB46	CTGGTCAA	CTGGTCAAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
SphIB48	CTACAA	CTACAAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	/5Phos/	HPSF
MboIA1	GTGAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTGAC	/	/
MboIA2	CACGTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCACGTA	/	/
MboIB1	GTCAC	GATCGTCACAGATCGGAAGAGCGAGAACAA	/5Phos/	HPSF
MboIB2	TATCGTG	GATCTACGTGAGATCGGAAGAGCGAGAACAA	/5Phos/	HPSF

Secuencias de primers (Lange *et al* 2014)

Index2, Forward	Index Sequence	Outer Primer as Ordered
F1_01	TAGATCGC	AATGATACGGCGACCACCGAGATCTACACTAGATCGCACACTCTTTCCCTACACGA
F1_02	CTCTCTAT	AATGATACGGCGACCACCGAGATCTACACCTCTCTATACACTCTTTCCCTACACGA
F1_03	TATCCTCT	AATGATACGGCGACCACCGAGATCTACACTATCCTCTACACTCTTTCCCTACACGA
F1_04	AGAGTAGA	AATGATACGGCGACCACCGAGATCTACACAGAGTAGAACACTCTTTCCCTACACGA
F1_05	GTAAGGAG	AATGATACGGCGACCACCGAGATCTACACGTAAGGAGACACTCTTTCCCTACACGA
F1_06	ACTGCATA	AATGATACGGCGACCACCGAGATCTACACACTGCATAACACTCTTTCCCTACACGA
F1_07	AAGGAGTA	AATGATACGGCGACCACCGAGATCTACACAAGGAGTAACACTCTTTCCCTACACGA
F1_08	CTAAGCCT	AATGATACGGCGACCACCGAGATCTACACCTAAGCCTACACTCTTTCCCTACACGA
F1_09	TGAACCTT	AATGATACGGCGACCACCGAGATCTACACTGAACCTTACACTCTTTCCCTACACGA
F1_10	TGCTAAGT	AATGATACGGCGACCACCGAGATCTACACTGCTAAGTACACTCTTTCCCTACACGA
F1_11	TAAGTTCC	AATGATACGGCGACCACCGAGATCTACACTAAGTTCCACACTCTTTCCCTACACGA
F1_12	ATAGAGGC	AATGATACGGCGACCACCGAGATCTACACATAGAGGCACACTCTTTCCCTACACGA
F1_13	GGCTCTGA	AATGATACGGCGACCACCGAGATCTACACGGCTCTGAACACTCTTTCCCTACACGA
F1_14	AGGCGAAG	AATGATACGGCGACCACCGAGATCTACACAGGCGAAGACACTCTTTCCCTACACGA
F1_15	TAATCTTA	AATGATACGGCGACCACCGAGATCTACACTAATCTTAACACTCTTTCCCTACACGA
F1_16	CAGGACGT	AATGATACGGCGACCACCGAGATCTACACCAGGACGTACACTCTTTCCCTACACGA
Index1, Reverse (some of the 96)	**Index Sequence**	**Outer Primer as Ordered**
R_8	TACTCTAC	CAAGCAGAAGACGGCATACGAGATTACTCTACGTGACTGGAGTTCAGACGTG
R_9	ATAAGCTA	CAAGCAGAAGACGGCATACGAGATATAAGCTAGTGACTGGAGTTCAGACGTG
R_20	TACCGGTG	CAAGCAGAAGACGGCATACGAGATTACCGGTGGTGACTGGAGTTCAGACGTG
R_21	ATCGAAAC	CAAGCAGAAGACGGCATACGAGATATCGAAACGTGACTGGAGTTCAGACGTG
R_31	ATAGGAAT	CAAGCAGAAGACGGCATACGAGATATAGGAATGTGACTGGAGTTCAGACGTG
R_32	ATTAGTTG	CAAGCAGAAGACGGCATACGAGATATTAGTTGGTGACTGGAGTTCAGACGTG
R_43	TAAAGCTA	CAAGCAGAAGACGGCATACGAGATTAAAGCTAGTGACTGGAGTTCAGACGTG
R_44	ATCGATTA	CAAGCAGAAGACGGCATACGAGATATCGATTAGTGACTGGAGTTCAGACGTG
R_55	ATGGAACT	CAAGCAGAAGACGGCATACGAGATATGGAACTGTGACTGGAGTTCAGACGTG
R_56	ATTGACAT	CAAGCAGAAGACGGCATACGAGATATTGACATGTGACTGGAGTTCAGACGTG
R_67	ATATTATA	CAAGCAGAAGACGGCATACGAGATATATTATAGTGACTGGAGTTCAGACGTG
R_68	ATTCGGGA	CAAGCAGAAGACGGCATACGAGATATTCGGGAGTGACTGGAGTTCAGACGTG
R_79	TATGACAT	CAAGCAGAAGACGGCATACGAGATTATGACATGTGACTGGAGTTCAGACGTG
R_80	TAGGACGG	CAAGCAGAAGACGGCATACGAGATTAGGACGGGTGACTGGAGTTCAGACGTG
R_91	TAGTCGTC	CAAGCAGAAGACGGCATACGAGATTAGTCGTCGTGACTGGAGTTCAGACGTG
R_92	TAGCTGTA	CAAGCAGAAGACGGCATACGAGATTAGCTGTAGTGACTGGAGTTCAGACGTG

Combinaciones de primers utilizadas:

Forward	Reverse
F1_01	R1_80
F1_02	R1_92
F1_03	R1_09
F1_04	R1_21
F1_05	R1_31
F1_06	R1_43
F1_07	R1_55
F1_08	R1_67
F1_09	R1_79
F1_10	R1_91
F1_11	R1_08
F1_12	R1_20
F1_13	R1_32
F1_14	R1_44
F1_15	R1_56
F1_16	R1_68

Annealing de oligonucleótidos –Armado de adaptadores doble cadena

-Resuspender o diluir oligos a 100 µM.

-Preparar Annealing buffer 10X con 100mM Tris HCl (pH 8), 500 mM NaCl y 10mM EDTA.

-En tubos de 0,2 ml **colocar** por cada adaptador a obtener:

Reactivo	**Volumen (µl)**
H_2O	50
Oligo. Enzima Frecuente o rara A	20
Oligo. Enzima Frecuente o rara B	20
Annealing Buffer	10
Volumen Final (20 µM)	**100**

-En un termociclador, incubar a 94°C por 2,5 minutos, y luego bajar no más de 1°C por minuto hasta alcanzar 21°C.

-Preparar diluciones de uso:

-Adaptador enzima corte raro a 1 µM: 5 µl + 95 µl de agua.

-Adaptador enzima corte frecuente a 2 µM: 10 µl + 90 µl de agua.

-Guardar a -20°C.

7.1.6 ddRADseq optimizado: PROTOCOLO 1

(Puesta a punto en nuevas especies con bajo número de muestras)

Digestión

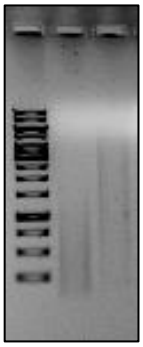
- **Nota:** Elegir las enzimas adecuadas para cada especie, haciendo una prueba de digestiones con distintos pares enzimáticos como, por ejemplo, SphI-MboI o PstI-MspI (Ver ANEXO Figura 3, patrones ejemplo en *Eucalyptus dunnii*). Además, es recomendable realizar digestiones *in silico*, tanto si la especie cuenta con genoma de referencia o con uno de una especie cercana, como si no lo posee, pero si existen estimaciones del tamaño del genoma y el porcentaje de GC esperado. Entre los softwares a utilizar se encuentra SimRAD (Lepais and Weir, 2014).
- Templado: 15 µl de ADN genómico en concentración 10 ng/µl (cuantificado con Qubit – dsDNA Broad Range).
- Preparar mix de digestión, **Nota:** Vortexear bien el Buffer antes de usarlo:

	x1 reacción (µl)
H_2O	11,4
SphI-HF (20U/µl)	0,12
MboI (5U/µl)	0,48
Buffer Cutsmart	3
vf	15

- En hielo, mezclar y dispensar 15 µl de mix en cada muestra de ADN (Volumen final = 30 µl).
- Mezclar e incubar:
  - 1.30 hs a 37 ºC
  - 20 minutos a 65 ºC (para inactivación de enzima)
- Sembrar 2 µl de algunas muestras en Fragment Analyzer (Figura 3 de resultados) para ver *smear* de fragmentos < 3 Kpb (Figura 3 anexo).
- Purificar con Ampure XP. **Nota:** Llevar las Beads a temperatura ambiente 30 minutos antes de usarlas.

Figura 3: Gel de agarosa con patrones ejemplo de digestiones enzimáticas de ADN de *E. dunnii*. De izquierda a derecha: Marker 1 Kpb, digestión con SphI-MboI y digestión con PstI-MspI.

Para *E. dunnii* se eligió la combinación enzimática SphI-MboI, porque presenta la totalidad de sus fragmentos menores a 2-3 Kpb.



Purificación con AmpureXP

- Añadir 1.5 volúmenes de *beads* (45µl) a cada muestra y mezclar (vórtex);
- Incubar a temperatura ambiente 5 minutos;
- Precipitar las *beads* en cama magnética por 5 minutos;
- Remover sobrenadante;
- Lavar 2 veces con 200 µl de Etanol 80% (preparado fresco) - 30 segundos por vez;
- Dejar secar por 5 minutos;
- Agregar 20 µl de buffer EB (Kit QIAGEN o 10mM TrisHCl, pH 8,5)
- Mezclar e incubar 5 minutos a temperatura ambiente;
- Poner las muestras nuevamente en cama magnética por 2 minutos;
- Transferir 18 µl de muestra purificada a nuevo tubo.

Nota: el protocolo puede detenerse acá y guardar las muestras a 4ºC.

Ligación

- Descongelar adaptadores en hielo (ver Annealing de adaptadores, diluciones y secuencias en ANEXO).
- Preparar mix de ligación:

	x1 reacción (µl)
T4 DNA ligasa Invitrogen (1U/µl)	2,4
Buffer T4 DNA ligasa (5X)	6
H_2O	2,1
Adaptador SphI (1µM)	2
Adaptador MboI (2µM)	2,5
vf	15

- A temperatura ambiente, mezclar y dispensar 15 µl de mix a 15 µl de digestión purificada (Volumen final = 30 µl).
- Mezclar e incubar:
    - 1 hora a 23 ºC;
    - 1 hora a 20 ºC;
    - 20 minutos a 65 ºC.

Nota: el protocolo puede detenerse acá y guardar las muestras a 4ºC.

Purificación con AmpureXP

- Añadir 1.5 volúmenes de *beads* a cada muestra y mezclar (vortex);
- Incubar a temperatura ambiente 5 minutos;
- Precipitar las *beads* en cama magnética por 5 minutos;
- Remover sobrenadante;
- Lavar 2 veces con 200 µl de Etanol 80% (preparado fresco) - 30 segundos por vez;
- Dejar secar por 5 minutos;
- Agregar 20 µl de buffer EB (Kit QIAGEN)
- Mezclar e incubar 5 minutos a temperatura ambiente;
- Poner las muestras nuevamente en cama magnética por 2 minutos;
- Transferir 18 µl de muestra purificada a nuevo tubo.

Cuantificación (con QUBIT) de las ligaciones

- Cuantificar las muestras en QUBIT con dsDNA Broad Range (2 µl de muestra + 198 µl de buffer con intercalante).

PCR

	x1 reacción (µl)
Buffer HF NEB (5X)	10
dNTPS 10 uM	1
H_2O	20,5
Phusion HF DNA polymerase 2U/µl NEB	0,5
Primer F 12,5 µM (con índice, Lange *et al* 2014)	1
Primer R 12,5 µM (con índice, Lange *et al* 2014)	1
vf	34

- En hielo, mezclar y dispensar 34 µl de mix a 16 µl de ligación purificada (Volumen final = 50 µl).
- Ciclado:

1 ciclo	95 ºC	3 min
10 ciclos	95 ºC	30 seg
	60 ºC	30 seg
	72 ºC	45 seq
1 ciclo	72 ºC	2 min
Hold	8 ºC	

Purificación con AmpureXP

- Añadir 1.2 volúmenes de *beads* a cada muestra y mezclar (vortex);
- Incubar a temperatura ambiente 5 minutos;
- Precipitar las *beads* en cama magnética por 5 minutos;
- Remover sobrenadante;
- Lavar 2 veces con 200 µl de Etanol 80% (preparado fresco) - 30 segundos por vez;
- Dejar secar por 5 minutos;
- Agregar 20 µl de buffer EB (Kit QIAGEN)
- Mezclar e incubar 5 minutos a temperatura ambiente;
- Poner las muestras nuevamente en cama magnética por 2 minutos;
- Transferir 20 µl (TODO) de muestra purificada a nuevo tubo.

Cuantificación (con QUBIT) de las PCRs

- Cuantificar las muestras en QUBIT con dsDNA Broad Range (2 µl de muestra + 198 µl de buffer con intercalante).
- Comparar las concentraciones de cada PCR con las correspondientes a sus ligaciones, verificando si hubo amplificación.

Pool de muestras

- *Poolear* las muestras en cantidades equimolares (tomando como referencia la muestra de menor concentración, usar todo el volumen).
- Evaporar en SpeedVac hasta llegar a volumen de 10 µl (Ej.: partiendo de 85µl, dejar 45 minutos a 30 ºC).

Siembra en gel de agarosa para *Size-selection*

- Armar gel de agarosa 1,5 % en TAE 1 X autoclavado. Previamente lavar bien la cama con lavandina. Opcional, poner a UV.
- Sembrar la *library*: 10 µl de muestra + 2 µl de *xylene-cyanol*, rodeado de 5 µl de *ladder* 100 pb (5 µl de *ladder* a la izquierda, una calle vacía, 12 µl de *library*, calle vacía, 5 µl *ladder* a la derecha).
- Correr 1 hora a 90 V.

Size selection (450 a 550 pb – corresponde a fragmentos de 310 a 410 pb + 140 pb de adaptadores)

- Cubrir la base del transiluminador UV con film, poner el gel, prender el UV, y con tips marcar en ambos *ladder*s las regiones de 450 pb y 550 pb.
- Apagar el UV y cortar con bisturí la sección delimitada por los tips.
- Poner en tubo Eppendorf 1.5 ml (previamente pesado).

Purificación con columnas de QIAGEN (seguir protocolo del KIT con algunos pequeños cambios)

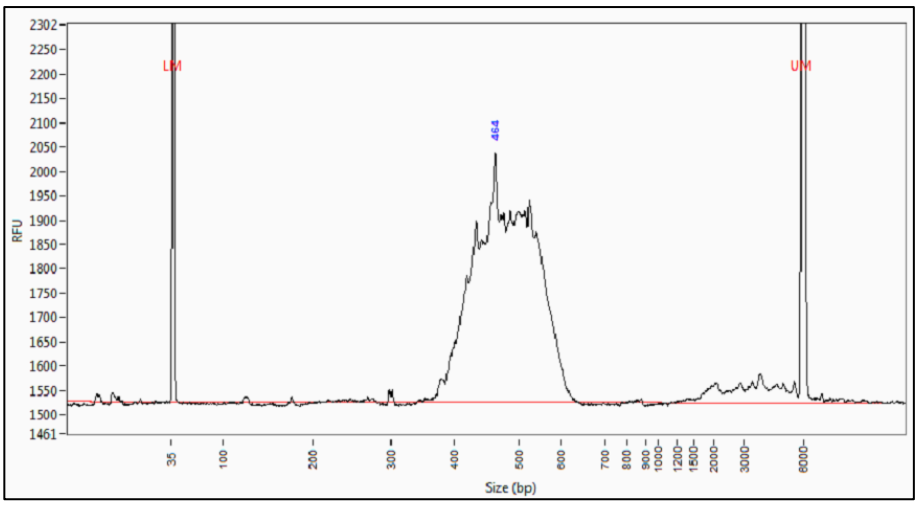
- Para disolver el gel de agarosa, agregar 3 volúmenes de buffer QG e incubar en termobloque a 37 ºC (Ej.: si el taco de agarosa pesa 200 mg, agregar 600 µl de buffer QG).
- 1era y 2da centrifugación se hacen a 8000 rpm; la 3era (para secado) a 13000 rpm.
- Eluir en 17 µl de buffer EB.
- Nota: las bibliotecas se pueden guardar a -20°C hasta un año antes de secuenciar, chequear con qPCR antes de usar y volver a cuantificar.

Purificar nuevamente con Ampure Beads

Pasos previos a secuenciación

- Cuantificar con QUBIT la *library* (dsDNA Broad Range) – Se esperan alrededor de 10 ng/µl para 8 muestras/*library*.
- Correr en Fragment Analyzer (Kit NGS) para chequear la calidad de la *library* (ver que todos los fragmentos tienen el tamaño esperado, como en ejemplo ANEXO Figura 4).

Figura 4: Corrida en Fragment Analyzer de biblioteca final de *Eucalyptus dunnii*.



- Calcular la molaridad de la muestra:

$$\frac{concentración\left(\frac{ng}{uL}\right) * 10^6}{660\frac{g}{mol} * promedio\_pb\_fragmentos} = Concentración\ en\ nM$$

- Conociendo la molaridad, diluir a 4 nM.
- Preparar *Library* 20 pM:

Library 4 nM	5 µl
NaOH 0,2N (fresco)	5 µl
HT1	990 µl
vf (*library* 20 pM)	1000 µl

- Diluir *library* a 15 pM (450µl *library* 20 pM + 150 µl HT1).
- Preparar: 540 µl de *library* 15 pM + 60 µl de PhiX 20 pM (queda PhiX 10%).
- Meter en *cartridge* de MiSeq

Armar *sample-sheet*

- Abrir el programa "Illumina experiment manager";
- Ir a la opción "create sample sheet". Seleccionar "miseq" – "other" – "Fasta only" (chequear "Cycles read 1" y "Cycles read 2");
- En el "Sample sheet wizard": seleccionar "add blank row" (1 por cada muestra de la *library* – poner el nombre de cada muestra) y combinar diferentes "index1" con distintos "index2";
- Abrir el Excel y editar "secuencia de índices" (todo en .csv).

Nota: finalmente queda la *sample sheet* con el nombre de los índices de Nextera pero con la secuencia de nuestros índices.

7.1.7 ddRADSeq optimizado: PROTOCOLO 2

(Escalado a población de muestras)

Digestión para 24 muestras:

Notas:

- ✓ Vortexear bien el Buffer antes de usarlo.
- ✓ Mantener refrigerados los reactivos y la mix durante todo el proceso.
- ✓ Este protocolo toma como ejemplo la digestión doble con las enzimas SphI-HF y MboI, si se utiliza otra combinación de enzimas:
  - Utilizar el Buffer de NEB que asegure el 100% de actividad en ambas.
  - Chequear las temperaturas de actividad e inactivación correspondientes.

Pasos:

- Realizar 15 µl dilución de ADN genómico a una concentración de 10 ng/µl en 3 columnas de una placa de 96 pocillos.
- Preparar mix de digestión para SphI-HF/MboI (NEB):

	1 X(µl)	**25 X**	**50X**
H_2O	11,4	285	570
SphI-HF (20U/µl)	0,12 (2,4 U)	3	6
MboI (5U/µl)	0,48 (2,4 U)	12	24
Buffer Cutsmart	3	75	150
vf	15	375	750

- O preparar mix de digestión para PstI-HF/MspI (NEB):

	1 X(µl)	**25 X**	**50X**
H_2O	11,76	294	588
PstI-HF (20U/µl)	0,12 (2,4 U)	3	6
MspI (20U/µl)	0,12 (2,4 U)	3	6
Buffer Cutsmart	3	75	150
vf	15	375	750

- Mezclar y dispensar 15 µl de mix en cada muestra de ADN (Volumen final = 30 µl).
- Mezclar e incubar:
  - ✓ 1.30 hs a 37 ºC
  - ✓ 20 minutos a 65 ºC

Purificación con Beads Agencourt AMPure XP (Beckman Coulter)

- Notas:

  - ✓ Preparar el etanol al 80% en el momento (para 24 muestras: 5ml totales, 4ml EtOH y 1 ml de agua de cartucho).
  - ✓ En los lavados, agregar el etanol, dejarlo actuar 30 segundos, y retirarlo (contando los 30 segundos desde la incorporación del mismo a la primera columna de muestras).
  - ✓ Retirar el etanol con pipeta multicanal o monocanal, según comodidad. Es importante que no queden restos de etanol, si es necesario utilizar una pipeta de 10 µl para retirar restos del fondo de los pocillos;
  - ✓ Para una resuspensión efectiva y más fácil, **No secar de más** (secar menos tiempo si se demoró mucho en retirar el etanol del último lavado);
  - ✓ Con los microlitros restantes: Guardarlos en otros tubos y correr en el FA 2 µl de algunas muestras (ej.: 3) para ver el patrón o la población de fragmentos generados. Se espera una digestión completa del ADN genómico evidenciada en la obtención de una población de fragmentos homogéneamente distribuidos homogéneamente a partir de aproximadamente 3 Kpb hacia fragmentos de menor tamaño (Ver ANEXO Figura 3).
  - ✓ El protocolo puede detenerse una vez finalizada la purificación y guardar las muestras a 4ºC.

Pasos:

- Llevar las Beads a temperatura ambiente 30 minutos antes de usarlas. Resuspenderlas bien.
- Añadir 1.5 volúmenes de beads (45 µl) a cada digestión y mezclar (Sugerencia: usar vórtex de placas, cerrar placas con tapas de silicona o adhesivas);
- Incubar a temperatura ambiente 5 minutos, y dar spin;
- Precipitar las beads en una cama magnética para placas por 5 minutos;
- Remover el sobrenadante (ingresar al pocillo con el tip levemente inclinado, cuidando de no arrastrar beads);
- Hacer 2 lavados (30 segundos por vez) con 100 µl de Etanol 80%, y dejar secar por 5 minutos.
- Agregar 20 µl de buffer EB (Kit QIAGEN)
- Mezclar (en vórtex de placas), e incubar 5 minutos a temperatura ambiente;
- Dar un spin a baja velocidad (250g) y poner las muestras nuevamente en la cama magnética por 2 minutos;
- Transferir 15 µl de digestión purificada a un nuevo tubo (Sugerencia: utilizar las columnas 4, 5 y 6 de la placa).

Ligación

- Descongelar los adaptadores en hielo y mantener los reactivos refrigerados (ver Annealing de adaptadores, diluciones y secuencias en ANEXO).
- Preparar mix de ligación, elegir protocolo según ligasa de Invitrogen o NEB:

Ligasa de Invitrogen:

	1X (µl)	**25X**	**50X**
T4 DNA ligasa Invitrogen (1 Weiss Unit/µl)	2,4	60	120
Buffer T4 DNA ligasa (5X)	6	150	300
H_2O	2,1	52,5	105
Vf	10,5	262,5	525

Ligasa de NEB:

	1X (µl)	**25X**	**50X**
T4 DNA ligasa NEB (1U/µl)	0,4	10	20
Buffer T4 DNA ligasa (5X)	3	75	150
H_2O	7,1	177,5	355
Vf	10,5	262,5	525

- Mezclar la mix y agregar 10,5 µl de la misma a los 15 µl de cada digestión purificada.
- Agregar:
  - ✓ 2 µl de adaptadores SphI o PstI 1 µM con *barcodes* (códigos de barra) o 2 pM por reacción. Utilizar una pipeta multicanal y dispensar uno por cada una de las 24 muestras (24 adaptadores en 3 columnas de la placa de uso, utilizar la misma disposición en la placa de muestras).
  - ✓ 2,5 µl de adaptadores MboI o MspI 2 µM con *barcodes* o 5 pM por reacción. En este caso se diseñaron 2 *barcodes*, preferentemente usar uno para cada pool de muestras, el N° 1 en la mitad de los pooles y el N° 2 en la otra mitad.
- Cada reacción tiene un volumen final de 30 µl.
- Mezclar e incubar:
  - ✓ 1 hora a 23 ºC;
  - ✓ 1 hora a 20 ºC;
  - ✓ 20 minutos a 65 ºC (para inactivación de enzima).

Nota: el protocolo puede detenerse acá y guardar las muestras a 4ºC.

Cuantificación de digestiones:

Cuantificar las digestiones con un fluorómetro de placas FluorStar Óptima (BMG u otro fluorómetro):

- Utilizar una placa adecuada para mediciones en fluorómetros (ejemplo: GREINER 6).

- Preparar la curva de calibración con ADN de Timo de ternero (Sigma-Aldrich), utilizar por ejemplo 6 puntos de diluciones seriadas a ½ en agua partiendo de 50ng/µl de concentración (50, 25, 12.5, 6.25, 3.12 y 1.56 ng/µl).
- Diluir 1/200 el intercalante fluorescente Picogreen (protegido de la luz con papel aluminio, ThermoFisher Scientific) en buffer TE 1X, calculando 100µl de solución por cada pocillo a medir + curva de calibración + 1 control positivo y un blanco + 2 extras: 100*(24 muestras+ 6 para curva de calibración + 2 controles + 2 extras) µl. Ejemplo para 24 muestras:
    - ✓ Volumen total: 100*34 µl = 3400 µl.
    - ✓ Intercalante: 17 µl.
    - ✓ Buffer: 3400-17 = 3383 µl.
- Pipetear 2 µl de las muestras, controles y curva de calibración en la placa en los pocillos correspondientes.
- Agregar 98 µl de Picogreen en buffer a cada pocillo.
- Cerrar placa con una tapa adhesiva o silicona, vortexear y dar un spin.
- Incubar por dos minutos y al resguardo de la luz.
- Encender computadora y el equipo FluorStar Optima, seleccionar un usuario, completar la contraseña y seleccionar RUN.
- Abrir un protocolo de trabajo y setear cada parámetro deseado, entre ellas:
    - ✓ Seleccionar la placa que se utilizará (ejemplo: GREINER 6).
    - ✓ Elegir la opción Optic TOP.
    - ✓ Filtro de excitación 485nm, Filtro de emisión 520nm.
    - ✓ Well scanning: 3x3
    - ✓ Layout: Indicar los pocillos BLANCO, ESTANDAR y MUESTRA.
    - ✓ Seleccionar Measure de la barra de herramientas y dentro de la misma:
        - “Plate out” para colocar la placa en la bandeja del equipo, “Plate in” para cerrar.
        - “Measure”, “OK” al protocolo a utilizar y “Start Mesure”.
        - Plate ID: poner nombre a la placa.
- Acceder a los resultados desde la barra de herramientas, Results, Open, Open MARS Data Analysis software, seleccionar por Plate ID, open. El formato de *Table view* se ve mejor y todo puede ser EXPORTADO a un formato Excel.
- Calcular concentraciones en Excel:
    - ✓ Abrir el archivo en Excel.
    - ✓ Restar el blanco a todos los pocillos medidos.
    - ✓ Calcular la ecuación de la recta según los estándares, utilizando la concentración conocida y la fluorescencia de cada uno.
    - ✓ Calcular la concentración de cada muestra.

Pool de muestras

- *Poolear* las ligaciones en cantidades equimolares (tomando como referencia la digestión de menor concentración, usar la mitad del volumen total, 15 µl), utilizando un tubo de 1,5 ml o uno de 2ml para mayor superficie de evaporación (o dividir en 2 o 3 tubos cada pool).
- Evaporar en Speedvac durante 1 hora a 45°C o hasta llegar a un volumen menor a 100µl totales por pool.

Purificación con AmpureXP

- Añadir 1 volumen de *beads* a cada pool y mezclar (vórtex);
- Incubar a temperatura ambiente 5 minutos;
- Precipitar las *beads* en cama magnética por 5 minutos;
- Remover sobrenadante;
- Lavar 2 veces con 200 µl de Etanol 80% (preparado fresco) - 30 segundos por vez;
- Dejar secar por 5 minutos;
- Agregar 20 µl de buffer EB (Kit QIAGEN)
- Mezclar e incubar 5 minutos a temperatura ambiente;
- Poner las muestras nuevamente en cama magnética por 2 minutos;
- Transferir 65 µl de muestra purificada a nuevo tubo.

Repetir Purificación con AmpureXP

Cuantificación (con QUBIT) del pool de ligaciones

- Cuantificar el pool en QUBIT con dsDNA Broad Range (2 µl de muestra + 198 µl de buffer con intercalante).

Size Selection automatizado con SAGE ELF (Sage Science)

- Colocar en un tubo (0,5ml) 30 µl de pool purificado y agregarle 10 µl de DNA marker solution (a temperatura ambiente). Mezclar y dar spin.
- Realizar la calibración Óptica del equipo según manual.
- Crear o seleccionar el protocolo de fraccionamiento Ejemplo Protocolo GBS:
    - ✓ Cassette definition: 2% 100pb – 2300pb,
    - ✓ Separation Mode: size-based,
    - ✓ Target value: 450pb y
    - ✓ Mover el slider para setear el pocillo número 9, en el cual vamos a recuperar los fragmentos de un promedio de 450pb.
    - ✓ Save as: guardar el protocolo de fraccionamiento.
- Preparar el cassette de gel de agarosa al 2%.
    - ✓ Eliminar burbujas según el manual.
    - ✓ Retirar cobertura adhesiva, sosteniendo firmemente el cassette para no derramar el buffer.
    - ✓ Cambiar el buffer de los pocillos 13 de elusión por 30 µl de Buffer nuevo en cada uno.
    - ✓ Sellar nuevamente los pocillos de elusión (desde el 13 al 1) con una tapa adhesiva provista por el kit.
    - ✓ Colocar el cassette en el instrumento y rellenarlo con el buffer a nivel como índica el manual (llenar hasta ver un meñisco de buffer por sobre el cassette, colocar la vista a nivel de cassette para observar mejor, y retirar 1ml con una pipeta).
- Realizar el Electrophoresis current Test como indica el manual.
- Corroborar que el pocillo de siembra de muestra este lleno al ras, si no es así completar con Buffer.
- Retirar 40 µl de buffer del pocillo de siembra y sembrar los 40 µl de pool de ligaciones más el DNA marker.
- Asegurarse de que el pocillo de siembra quede lleno al ras, si no es así completar con Buffer.

- Cerrar tapa con cuidado y seleccionar "Run Protocol" (tener en cuenta que la corrida dura 3hs aproximadamente).
- Retirar todas las fracciones eluidas desde pocillo 1 al 12 (guardarlas en strips). Prestar especial atención en retirar el pocillo (ejemplo N°9) con el rango seteado y guardarlo en un tubo de 0,5ml.
- Enjuagar los electrodos según el manual.

Purificación Size Selection con AmpureXP

- Seguir los pasos del punto 8 pero con las siguientes modificaciones:
  - ✓ Añadir 0.8 volúmenes de *beads* a la elusión del pocillo 9 y mezclar (vortex); está proporción descarta fragmentos menores a 300pb (dímero de adaptadores).
  - ✓ Al finalizar la purificación, eluir en 32 µl de buffer EB (Kit QIAGEN) y transferir 30 µl del size selection purificado a un nuevo tubo.

Cuantificación (con QUBIT) del Size Selection

- Cuantificar la salida del SAGE purificada en QUBIT con dsDNA Broad Range (2 µl de muestra + 198 µl de buffer con intercalante).

PCR

- Realizar dos PCRs a partir de la Size Selection purificado (colocar 13 µl en cada uno de los dos tubos de 0,2ml).

	x1 reacción (µl)
H_2O	23,5
Buffer HF NEB (5X)	10
dNTPs 10 µM	1
Phusion HF DNA polymerase 2U/µl NEB	0,5
Primer F 12,5 µM (con índice, Lange *et al* 2014)	1
Primer R 12,5 µM (con índice, Lange *et al* 2014)	1
vf	37

- En hielo, mezclar y dispensar 37 µl de mix a cada tubo con 13 µl del pool salida del SAGE purificado (Volumen final = 50 µl).

- Ciclado:

1 ciclo	95 ºC	3 min
10 ciclos	95 ºC 60 ºC 72 ºC	30 seg 30 seg 45 seg
1 ciclo	72 ºC	2 min
Hold	8 ºC	

Purificación de la PCR con AmpureXP

- Seguir los pasos del punto 8 pero con las siguientes modificaciones:
  - ✓ Juntar ambos tubos de PCR y añadir 1 volumen de *beads* y mezclar (vortex).
  - ✓ Al finalizar la purificación, eluir en 50 µl de buffer EB (Kit QIAGEN) y transferir la PCR purificada a un nuevo tubo.
  - ✓ Repetir la purificación añadiendo nuevamente 1 volumen de *beads* y mezclar (vortex) y eluyendo finalmente en 20 µl de buffer EB.

Cuantificación de la PCR con QUBIT

- Cuantificar la PCR en QUBIT con dsDNA Broad Range (2 µl de muestra + 198 µl de buffer con intercalante).

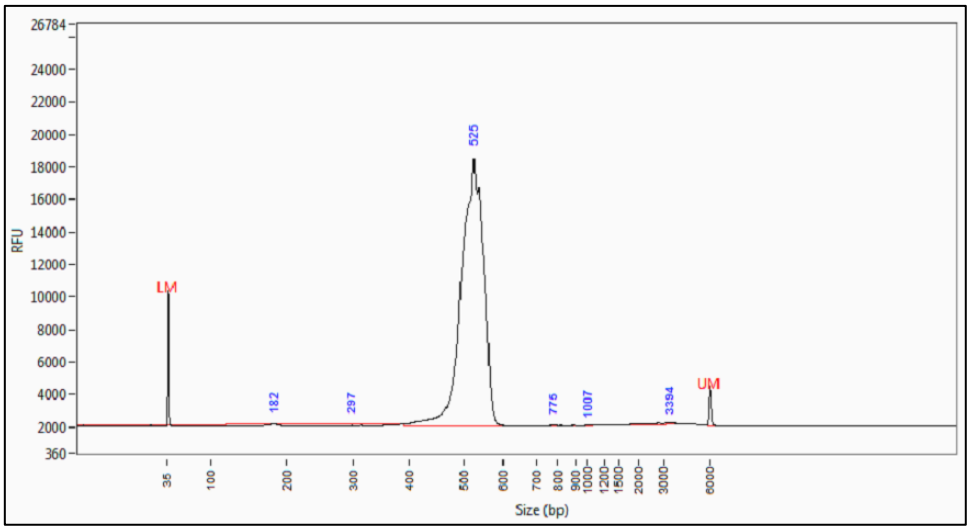
- Notas:

  - ✓ Comparar las concentraciones de entrada y salida a la PCR. Se esperan alrededor de <1 ng/µl para el Size Selection purificado y 5 ng/µl para una PCR.

Secuenciación NGS:

- Correr en Fragment Analyzer para chequear la calidad de la *library* (pico con una media aproximadamente de 520pb y un ancho de 70pb) – KIT NGS (ANEXO Figura 5).

- **Figura 5:** Corrida en Fragment Analyzer de biblioteca final de pool de 24 muestras de *Eucalyptus dunnii*.



- Calcular la molaridad de la muestra:

$$\frac{concentración\left(\frac{ng}{uL}\right) * 10^6}{660\frac{g}{mol} * promedio\_pb\_fragmentos} = Concentración\ en\ nM$$

- Conociendo la molaridad, diluir a 4 nM

- Preparar *Library* 20 pM:

Library 4 nM	5 µl
NaOH 0,2N (fresco)	5 µl
HT1	990 µl
vf (*library* 20 pM)	1000 µl

- Diluir *library* a 15 pM (450µl *library* 20 pM + 150 µl HT1).
- Preparar: 540 µl de *library* 15 pM + 60 µl de PhiX 20 pM (queda PhiX 10%).
- Meter en *cartridge* de *NextSeq*

2. Aplicación de metodologías genómicas para el mejoramiento molecular de *E. dunnii* mediante Mapeo por Asociación y Selección Genómica

7.2.1 SNPs asociados según el umbral ad hoc de -log(1E-03)

Se reportaron para la matriz conjunta 256 SNPs asociados según el umbral *ad hoc* de -log(1E-03) y 42 de ellos compartidos entre características fenotípicas. Como resultado general, se evidenciaron dos grupos de características que presentaron algún SNP asociado compartido entre ellas y estos grupos concordaron con los grupos que mostraron correlaciones fenotípicas altas y significativas. Uno de los grupos fue compuesto por los caracteres de calidad de madera, que fueron aquellos siete estimados mediante NIR y el otro gran grupo fue el de los caracteres de crecimiento. Sin embargo, las características de ir20 y for6 no presentaron SNPs compartidos con ningún otro fenotipo.

Así, observando los SNPs asociados compartidos entre lig20, cel20 y klas20 en el diagrama de Venn de la figura 3.3.6 a, se puede ver que se evidenciaron 27 SNPs para lig20, 23 para klas20 y 24 para cel20. De este conjunto de marcadores, 11 fueron comunes a lig20 y klas20, siete entre lig20 y cel20, y tres compartidos entre estas tres características. Estas coincidencias, como ya se mencionó, son de esperar debido a que las tres características fenotípicas mostraron fuertes correlaciones entre sí (ver apartado 3.2.1; klas20 y cel20: -0,76; klas20 y lig20: 0,90; lig20 y cel20: -0,82) y todas muy significativas (p-valor<0,001).

De estos SNPs compartidos, dos fueron comunes a lig20, klas20, cel20 y extet20, lo que se puede observar en más detalle en la tabla 7.2.1, en donde se enumeran aquellos marcadores que presentaron asociación en al menos dos características fenotípicas (umbral *ad hoc* de -log(1E-03)). Uno de ellos fue el ya mencionado SNP de GBS 13753_15 (cromosoma 3, punto azul para lig20, verde para cel20 y naranja para klas20 en gráficos Manhattan de Figura 7.2.1 b, c y d, respectivamente) y el otro fue EuBR09s427144, el de mayor asociación para klas20 (p-valor: 4,45E-05, cromosoma 9, punto verde para klas20, naranja para lig20 y cel20 en gráficos de Manhattan de figura 7.2.1 d, b y c, respectivamente).

Por otro lado, el SNP 13763_65 de GBS (cromosoma 3), a sólo 59,5 Kpb del SNP 13753_15, también presentó p-valores por debajo de 1E-03 para lig20, klas20 y cel20, como se observa en la tabla 7.2.1. Por último, la característica db20 compartió el SNP EuBR01s6731755 (p-valor: 4,41 E-04, Cromosoma 1) con sg20 (p-valor: 3,57 E-04), y el SNP EuBR01s34250726 (p-valor: 7,59 E-04, Cromosoma 1) con cel20, que fue el SNP que presentó menor valor de p-valor para este último carácter (2,10 E-05, Tablas 7.2.1 y 7.2.1).

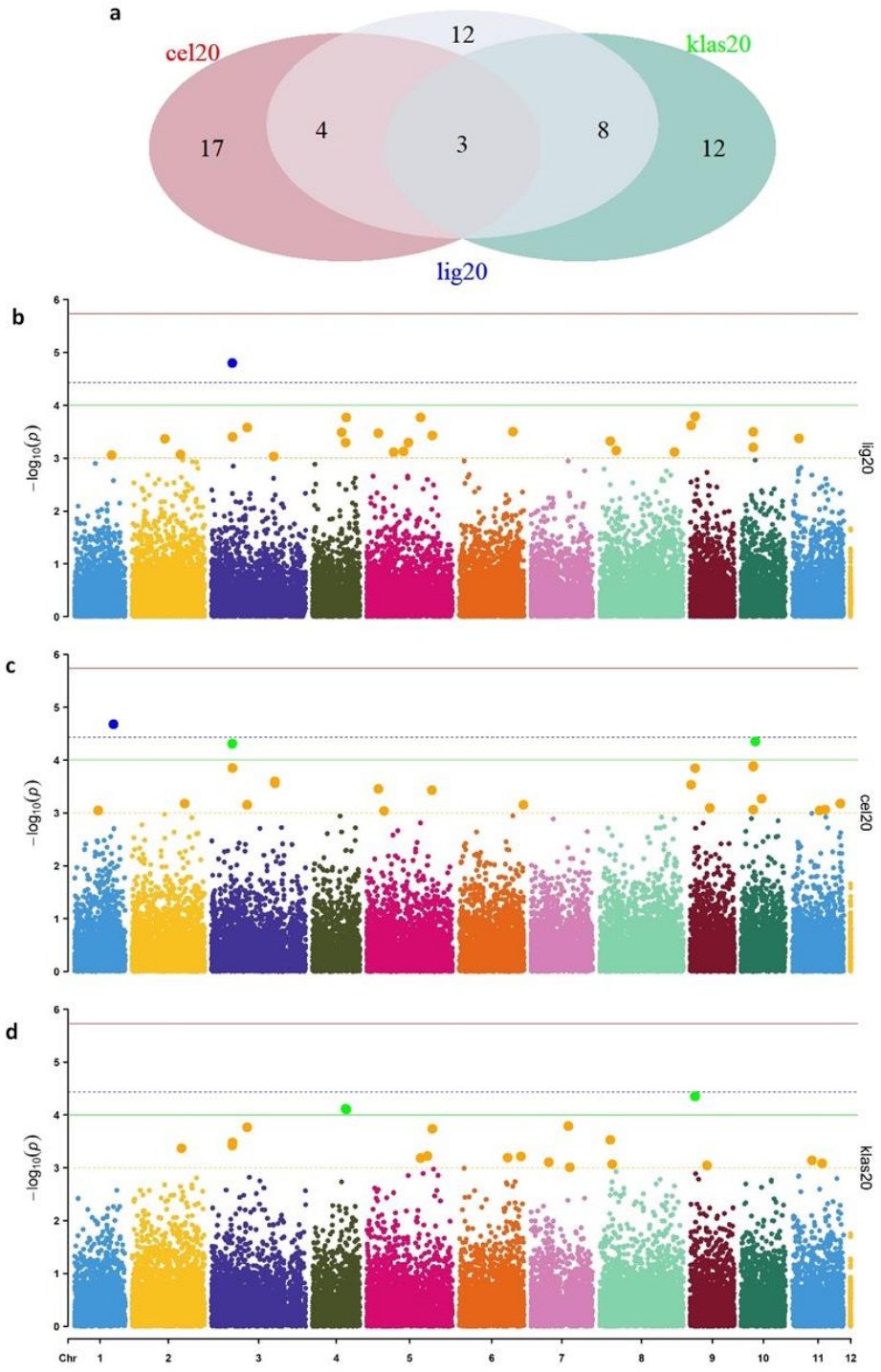


Figura 7.2.1: *SNPs asociados (umbral ad hoc de -log(1E-03)) compartidos entre lignina, celulosa y lignina Klason:* SNPs provenientes de GWAS con la matriz de Chip-GBS: a) Diagrama de Venn entre SNPs que superaron umbral *ad hoc* de -log(1E-03) para lig20, klas20 y cel20; Gráficos de Manhattan para: a) lig20, b) cel20, c) klas20. Eje de abscisas: Número de cada cromosoma y Posición en pares de bases de cada SNP (Crom 12: *Scaffolds*); Eje de ordenadas: -log10(p): logaritmo negativo en base 10 de los p-valores de cada SNP. Umbral naranja punteado: *ad hoc* de -log(1E-03); verde: *ad hoc* de -log(1E-04); azul punteado: -log(1/n); y rojo: Bonferroni. lig20: contenido total de Lignina; cel20: contenido total de Celulosa; klas20: Contenido de Lignina Klason.

Tabla 7.2.1. *SNPs que presentaron asociación en al menos dos caracteres fenotípicos en el análisis de GWAS utilizando SNPs provenientes de la matriz de Chip-GBS.* SNP: SNPs que superaron umbral *ad hoc* de -log(1E-03) y que son compartidos por al menos dos características fenotípicas; Crom: Cromosoma; Posición: Posición dentro del cromosoma en pb; Dap: diámetro a la altura de pecho a los 6, 11 y 20 años; at: altura total a los 6 y 20 años; for6: forma de fuste a los 6 años; ir20: índice de rajado a los 20 años; extet20: Extractivos etanólicos; exttot20: Extractivos totales; klas20: lignina klason; lig20: lignina; sg20: siringilo/guayacilo; cel20: celulosa; db20: densidad básica

SNP	Crom.	Posición	dap6	at6	for6	dap11	dap20	at20	ir20	extet20	exttot20	klas20	lig20	sg20	cel20	db20
EuBR01s6731755	1	6731755												**X**		**X**
EuBR01s34250726	1	34250726													**X**	**X**
EuBR02s1634702	2	1634702								**X**	**X**					
EuBR02s29023731	2	29023731								**X**	**X**		**X**	**X**		
EuBR02s57010174	2	57010174	**X**			**X**										
EuBR03s2637451	3	2637451				**X**	**X**									
13753_15	3	18501528								**X**		**X**	**X**		**X**	
13763_65	3	18561096										**X**	**X**		**X**	
EuBR03s31586288	3	31586288										**X**	**X**			
EuBR03s31586288rep1	3	31586288										**X**	**X**			
16495_56	3	49681560								**X**	**X**					
EuBR03s55795675	3	55795675				**X**		**X**								
21975_49	4	3122855				**X**	**X**									
EuBR04s29197264rep1	4	29197264										**X**	**X**			
EuBR04s29197264	4	29197264										**X**	**X**			
EuBR04s29699828	4	29699828										**X**	**X**			
EuBR04s32002058	4	32002058					**X**	**X**								
EuBR05s10051221	5	10051221											**X**		**X**	
EuBR05s23794379	5	23794379								**X**			**X**			
EuBR05s47611317	5	47611317										**X**	**X**			
28672_21	5	53723552								**X**	**X**	**X**				

SNP	Crom.	Posición	dap6	at6	for6	dap11	dap20	at20	ir20	extet20	exttot20	klas20	lig20	sg20	cel20	db20
EuBR05s58012795	5	58012795										**X**	**X**			
EuBR05s70539257	5	70539257	**X**			**X**										
EuBR05s70539257rep1	5	70539257	**X**			**X**										
EuBR06s52971844	6	52971844								**X**	**X**					
EuBR07s10247657	7	10247657				**X**	**X**									
EuBR08s9463598	8	9463598										**X**	**X**			
EuBR08s18155320	8	18155320				**X**	**X**									
48478_24	8	43190754								**X**	**X**					
49352_37	8	49447479								**X**	**X**					
EuBR08s56214808	8	56214808								**X**	**X**					
EuBR08s74030967	8	74030967								**X**	**X**					
EuBR09s686199	9	686199											**X**		**X**	
EuBR09s4271448	9	4271448								**X**		**X**	**X**		**X**	
53340_56	9	14374451								**X**	**X**					
53441_57	9	14876854								**X**	**X**					
62304_25	10	3752310								**X**	**X**					
63306_29	10	9125105				**X**	**X**									
EuBR10s10603441	10	10603441											**X**		**X**	
EuBR10s10649445	10	10649445											**X**		**X**	
EuBR10s30027250	10	30027250								**X**	**X**					
EuBR11s17351865	11	17351865		**X**						**X**	**X**					

Con respecto a extet20 y exttot20, se evidenciaron 28 SNPs con p-valor menor a 1 E-03 para extet20 y 23 para exttot20, de los 14 SNPs fueron comunes entre ambos, lo que coincide con la alta correlación observada entre los fenotipos (r^2 = 0,99, p-valor < 0,001), dado que la medición de extractivos totales incluye a los extractivos etanólicos. Dichos SNPs compartidos se pueden observar en los gráficos de Manhattan de la Figura 7.2.2. Por otro lado, exttot20 presentó asociado el SNP de GBS 28672_21 (cromosoma 5), que además de haberse encontrando entre los SNPs asociados a extet20, estuvo presente entre los 23 SNPs de kla20. Asimismo, el SNP EuBR02s29023731 del Chip (Cromosoma 2) estuvo asociado con extet20, exttot20, lig20 y sg20 y el SNP EuBR05s23794379 (cromosoma 5) es común a extet20 y lig20 (Tabla 7.2.1).

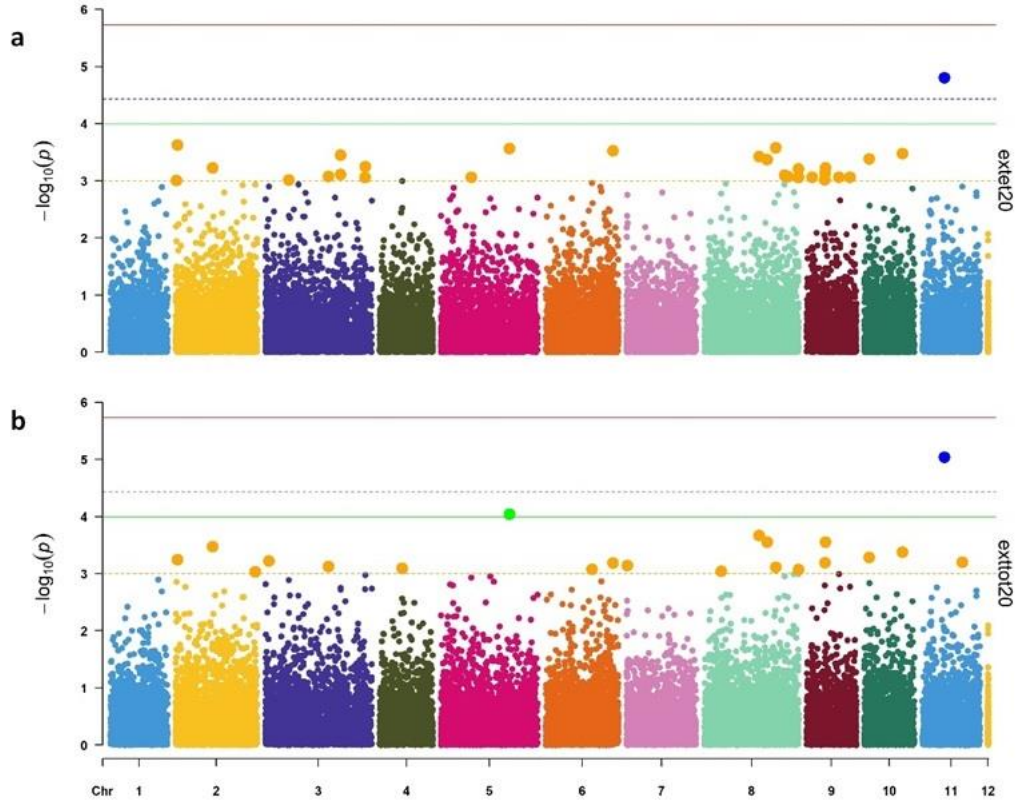


Figura 7.2.2: *Gráficos de Manhattan de GWAS utilizando SNPs provenientes de la matriz de Chip-GBS para: a) Extractivos Etanólicos (extet20), b) Extractivos Totales (exttot20).* Eje de abscisas: Número de cada cromosoma y Posición en pares de bases de cada SNP (Crom 12: scaffolds); Eje de ordenadas: -log10(p): logaritmo negativo en base 10 de los p-valores de cada SNP. Umbral naranja punteado: *ad hoc* de -log(1E-03); verde: *ad hoc* de -log(1E-04); azul punteado: -log(1/n); y rojo: Bonferroni.

Por otro lado, con respecto a los SNPs compartidos entre las características fenotípicas de crecimiento, se observó que dap11 y dap20 compartieron cinco SNPs asociados, de los cuales el segundo de menor p-valor para dap11 (1,88 E-04) fue el marcador de menor p-valor para dap20 (7,03E-05; EuBR03s2637451, en cromosoma 3, tabla 7.2.1). A su vez, dap11 compartió dos SNPs con dap6 (EuBR02s57010174 y EuBR05s70539257) y uno con at20 (EuBR03s55795675), y dap20 presentó un SNP en común con at20 (EuBR04s32002058). Como se vio anteriormente, los caracteres de dap y at a distintas edades presentaron

correlaciones altas y significativas entre la mayoría de ellos, por lo que encontrar SNPs asociados a las mismas características medidas a distintas edades permitiría, por ejemplo, aplicar MAS a distintas edades con los mismos marcadores. At6 presentó un p-valor de 6,42E-04 para el SNP EuBR11s17351865 del Chip asociado a extet20 y exttot20, aunque no hubo correlación entre dichos caracteres fenotípicos.

Una forma similar de análisis fue aplicada en abeto blanco (*Picea glauca*) por Lamara et al. (2016), quienes consideraron SNPs con asociaciones con rasgos de madera nominalmente significativas en $P < 0,05$, omitiendo la corrección para las pruebas múltiples, para maximizar el descubrimiento y obtener información sobre la arquitectura genómica y los procesos biológicos subyacentes a los rasgos cuantitativos. Como ejemplo, fueron compartidos un 22% del total de los SNPs asociados tanto para lignina total como para lignina Klason, presentando entre ellos una correlación positiva. Entre otros ejemplos se encontraron la correlación negativa observada entre celulosa y lignina, que mostró un 13,2% del total de los SNPs asociados comunes para ambos caracteres, siendo todos los efectos de los marcadores opuestos entre sí, y la alta correlación positiva entre los extractivos, que presentaron en común el 27,5% de los SNPs asociados.