Universidad de la República

Centro Universitario Regional Noreste

Sede Tacuarembó

Ingeniería Forestal

BÚSQUEDA DE GENES CANDIDATOS RELACIONADOS A LA PRECOCIDAD DEL CAMBIO DE FOLLAJE EN *Eucalyptus globulus*

Facundo Acuña Muñoz
Hugo Rodriguez Silveira

Tutor: Dr. Facundo Giorello

Tacuarembó, 17 de septiembre de 2024

PÁGINA DE APROBACIÓN

CENTRO UNIVERSITARIO DEL NORESTE – SEDE TACUAREMBÓ

El tribunal docente integrado por los abajo firmantes aprueba el Trabajo Final de Ingeniería Forestal:

Título:

BÚSQUEDA DE GENES CANDIDATOS RELACIONADOS A LA PRECOCIDAD DEL CAMBIO DE FOLLAJE EN *EUCALYPTUS GLOBULUS*

Autor/es:

Facundo Jesús Acuña Muñoz

Hugo Santiago Rodriguez Silveira

Tutor:

Dr. Facundo Giorello

Puntaje: _____

Tribunal:

…….(nombre y firma).

…….(nombre y firma).

…….(nombre y firma).

FECHA:

RESUMEN

La madera de *Eucalyptus globulus* tiene un alto valor y gran demanda en el mercado internacional. En Uruguay ésta fue una de las especies más utilizadas debido a su alta capacidad pulpable, sin embargo, una limitante es su gran susceptibilidad al ataque de enfermedades y patógenos. Uno de estos patógenos es *Teratosphaeria nubilosa* que produce la enfermedad conocida como "Mancha de Mycosphaerella". Esta enfermedad es una de las principales limitantes productivas en los sistemas de producción de madera de nuestro país. Trabajos recientes determinaron que el cambio precoz de follaje en *E. globulus* tiene el potencial de reducir la severidad de los efectos de la enfermedad debido a que el follaje adulto es más resistente a la infección que el follaje juvenil. Dichos trabajos, además, encontraron una considerable variación genética y alta heredabilidad para el inicio del cambio de follaje. Sabiendo de la preponderancia del factor genético en la característica, en este trabajo se estudió la base genética de la precocidad del cambio de follaje utilizando la aproximación XP-GWAS. Para este fin, se analizaron dos pooles de individuos con fenotipos extremos y opuestos de *E. globulus* en relación al inicio del cambio de follaje, y se estudiaron los loci en los que diferían sus frecuencias alélicas. A partir de estos estudios, se identificaron 86 SNPs significativos, de los cuales 45 mostraron una fuerte asociación con 9 genes ubicados a menos de 6.000 pares de bases. Dentro de este grupo de 9 genes, localizados en los cromosomas 3, 5, 7 y 8, se destacaron 4 genes asociados a resistencia y 2 que codificaban proteínas con funciones aún no determinadas. Además, se encontraron 3 genes no codificantes (ncRNA). Interesantemente, encontramos un SNP marginalmente significativo en las inmediaciones del microRNA 156. Este gen no codificante ha mostrado ser clave en el cambio de follaje en otros estudios.

ABSTRACT

The wood of *Eucalyptus globulus* is highly valued and in great demand on the international market. This species was one of the most widely used in Uruguay due to its high pulpability. However, one limitation is its high susceptibility to diseases and pathogens. One such pathogen is *Teratosphaeria nubilosa*, which causes a disease known as "*Mycosphaerella* leaf disease". This disease is a major constraint on the country's wood production systems. Recent studies have determined that early foliage change in *E. globulus* can potentially reduce the severity of the disease's effects, as mature foliage is more resistant to infection than juvenile foliage. These studies also found significant genetic variation and high heritability for the onset of foliage change. Given this trait's predominance of genetic factors, this work aimed to investigate the genetic basis of early foliage change using an XP-GWAS approach. To this end, two pools of *E. globulus* individuals with extreme and contrasting phenotypes regarding the onset of foliage change were analyzed, and loci with differing allele frequencies were studied. From these analyses, 86 significant SNPs were identified, 45 of which showed a strong association with 9 genes located within 6000 base pairs. Among these 9 genes, localized in chromosomes 3, 5, 7, and 8, 4 were notably associated with resistance genes, while 2 encoded proteins with functions yet to be determined. Additionally, 3 non-coding RNAs were identified. Interestingly, a marginally significant SNP was found near microRNA 156. In other studies, this non-coding RNA has been shown to be key in the onset of adult foliage.

Índice

1 Introducción

El género *Eucalyptus*, perteneciente a la familia de las Mirtáceas, constituye un conjunto de árboles altamente diversificados, abarcando más de 700 especies [1]. Además de su relevancia económica como árboles forestales, los *Eucalyptus* ejemplifican un fenómeno de radiación adaptativa [2]. Desde el período del Cenozoico medio, que se remonta a unos 25 a 10 millones de años atrás, las distintas especies pertenecientes a este género experimentaron una amplia diversificación aunque sus genomas son relativamente conservados (Recuadro 1). Durante este proceso, se adaptaron a una amplia gama de nichos ecológicos y a una diversidad de climas. Producto de su diversidad y numerosos usos, los *Eucalyptus* son árboles que tienen una gran importancia económica. En Uruguay, por ejemplo, se cultivan varias especies del género. La rapidez de crecimiento de los eucaliptos en entornos exóticos, junto con su eficiencia en el uso de nutrientes y agua, ha sido un factor clave para su éxito fuera de Australia [3].

En las últimas décadas, el área forestada destinada a la comercialización en el país ha aumentado de 45.000 ha en 1990 a 1.087.109 ha en el 2021, convirtiendo al rubro forestal en el segundo exportador más importante [4]. La especie *Eucalyptus globulus* fue una de las más utilizadas en el país [5] debido a las características de su madera, ideal para la producción de pulpa de celulosa y papel, con alta demanda en el mercado internacional. Sin embargo, según los datos de la última cartografía nacional forestal realizada por la Dirección General Forestal en 2021, se observa un cambio significativo en la distribución de las plantaciones forestales. En particular, se destaca una reducción en la superficie dedicada a las plantaciones de *E. globulus, Eucalyptus maidenii* y *Eucalyptus bicostata,* pasando de 149,329 ha en 2018 a 103,639 ha en 2021. Uno de los motivos que explica la reducción de las superficies plantadas de estas especies es su vulnerabilidad al ataque de patógenos.

El crecimiento del sector forestal ha traído consigo la introducción de más de 24 especies de patógenos que afectan las plantaciones de *E. globulus*. Entre ellas, la más importante debido a su impacto directo en la producción de biomasa, se encuentra la mancha foliar provocada por *Teratosphaeria nubilosa*, también conocida como “Mancha de Mycosphaerella” [6], [7]. *T. nubilosa* es un hongo patógeno oriundo de Australia que afecta el follaje juvenil de *E. globulus,* provocando manchas necróticas, defoliación y, en casos severos, muerte de ápices y ramas [7]

(Figura 1). Esta enfermedad tiene como consecuencia una reducción en la capacidad fotosintética y esto conlleva a un menor desarrollo de los árboles que, sumado al aumento de la mortalidad, determina una menor productividad de madera por unidad de superficie y/o tiempo [8].

Para abordar este problema, INIA ha implementado programas de mejoramiento genético desde 1990, con enfoque en la productividad y resistencia a *T. nubilosa* de *E. globulus* [9]. Estudios recientes han encontrado escasa variabilidad genética para la resistencia a *T. nubilosa*, pero una variabilidad genética importante para la precocidad en el cambio de follaje [10]. Esta característica es importante dado que muchos estudios sugieren que los *E. globulus* podrían escapar de las enfermedades foliares mediante el cambio ontogénico a un follaje adulto resistente [10], [11]. En varias especies de *Eucalyptus* las hojas juveniles son más susceptibles a patógenos y a insectos que las hojas adultas [7], [12]. Cabe destacar que el cambio de follaje es en gran medida un proceso genéticamente determinado aunque factores como la cantidad y calidad de los nutrientes, así como la intensidad de la luz, pueden influir en él [13]. Este proceso, sin embargo, parece ocurrir de manera independiente a los ataques de patógenos o la herbivoría [13]. El proceso de cambio de follaje, es parte del fenómeno conocido como heteroblastia, el cual describe el cambio de fase vegetativa. Las variaciones son particulares a cada especie y comprenden una gama de cambios, tales como la transición desde la fase juvenil de las hojas a la adulta, los patrones de ramificación, la distribución diferencial de la cera epicuticular, los patrones de producción de tricomas, la morfología celular, los patrones vasculares, la capacidad de generar raíces adventicias, la presencia o ausencia de compuestos fitoquímicos como la antocianina, y la resistencia frente a enfermedades o insectos [13], [14].

Recuadro 1. Genomas de *Eucalyptus* [15]

Los árboles del género *Eucalyptus* son organismos diploides cuyo tamaño genómico promedio es de 600 Mb y su inmensa mayoría cuentan con 11 cromosomas. Los genomas de *E. grandis* y *E. globulus*, están estrechamente relacionados y poseen una longitud de 640Mb y 530 Mb, respectivamente. El genoma de *E. grandis* contiene 36.376 genes codificantes, de los cuales un 84% se expresa en tejidos vegetativos y reproductivos. Además, un 34% de estos genes se encuentra en duplicaciones en tándem [15].

Figura 1. Hojas afectadas por Teratosphaeria nubilosa. Extraído de [16].

1.1 miRNAs y bases genéticas de la Heteroblastia

Los individuos pueden presentar diversas características fenotípicas (rasgos observables), las cuales pueden estar determinadas por un solo gen o por varios genes. Las características que están determinadas por un solo gen se denominan características monogénicas. Estas dependen únicamente de las variaciones en un gen específico, es decir, dependen de las variaciones alélicas de este. Por otro lado, las características que están determinadas por más de un gen se denominan características poligénicas. Estas características dependen de la interacción entre alelos de un mismo gen (dominancia), entre varios genes (epistasis) y son influidas por el ambiente. El efecto de los genes y el ambiente se estima en base al modelo infinitesimal de Fisher. Un ejemplo de una característica poligénica es el grado de resistencia a los hongos *Ceratocystis* y *Mycosphaerella cryptica* por parte de *Eucalyptus*. La resistencia a estos hongos está asociado a regiones genómicas que involucran múltiples genes [17], [18]. A su vez, factores ambientales como la humedad afectan la capacidad de contraer o no la infección.

Los genes que afectan las características se pueden dividir en genes codificantes y no codificantes. Los codificantes son aquellos que su expresión lleva a la generación de un proteína mientras que los no codificantes son genes que típicamente codifican una molécula de ARN (*e.g.,* ARN ribosómico, ARN mensajero, microARN). Ambos tipos de genes pueden influir las características fenotípica. Un gen codificante influye de manera directa de acuerdo a la proteína que codifique, mientras que los genes no codificante influyen en la características fenotípicas

regulando, en general. la expresión de genes codificantes. Un ejemplo de esto son los denominados miRNAs (microRNAs).

Los miRNAs son ARN endógenos no codificantes de aproximadamente 22 nucleótidos, que regulan negativamente la expresión génica al interferir con la traducción o degradación de ARNm. La síntesis de los miARNs es un proceso que ocurre en el núcleo y el citoplasma. Comienza con la transcripción del ADN a ARN, mediada por una ARN polimerasa, produciendo un microARN primario que posee una estructura de horquilla de ARN doble cadena. Este microARN primario es exportado al citoplasma y procesado por Dicer, generando un microARN doble cadena. Finalmente, ambas cadenas son separadas por acción de la proteína Argonauta. La cadena de miARN funcional (la otra se degrada) junto a Argonauta y la proteína RISC ya pueden inhibir la expresión de los genes blanco al reconocer sus mensajeros [19], [20]. Los microARNs ejercen, por tanto, su función reguladora a nivel post-transcripcional. Es posible que un único microARN puede dirigirse a cientos de ARN mensajeros específicos, lo que le brinda la capacidad de controlar la expresión de múltiples genes de manera coordinada. En el contexto de las plantas, es necesario que exista una complementariedad total entre el microARN y su ARN mensajero objetivo, para que se produzca el reconocimiento y la posterior degradación del mensajero. Por otra parte, en mamíferos, se requiere una complementariedad parcial entre el microARN y su blanco [19], [20].

Los microARN están ganando reconocimiento como reguladores cruciales de la actividad génica. Aunque pasaron desapercibidos hasta hace poco tiempo, los miRNAs se han revelado como una de las clases más abundantes de moléculas reguladoras en organismos multicelulares. Por eso, mediante el crecimiento de su importancia también fue creciendo el interés en estudiarlas [21].

Recientemente, mediante técnicas moleculares se ha logrado determinar que ciertos microARNs están asociados al cambio de follaje en *Arabidopsis* y *Eucalyptus*. Estos trabajos han apuntado a un controlador de la expresión génica, el microARN 156 (miR156)[22]. Sin embargo, hasta ahora existen pocos trabajos que relacionen dicho microRNA con el cambio de follaje en *Eucalyptus* y evidencia adicional ayudaría a consolidar el rol del miR156. Si bien son escasos los trabajos que relacionan al miR156 con el cambio de follaje en ellos se ha encontrado que una disminución en la actividad del miR156 lleva al cambio de fase vegetativa [20], [23].

Se ha observado que la señalización que maneja el cambio de fase proviene de los primordios foliares, ya que se ha notado un gran retraso en este cambio cuando estos tejidos están ausentes [20], [24]. Este retraso se ha asociado con un aumento en los niveles de miR156. Dichos niveles son significativamente elevados en brotes y tejidos juveniles, sin embargo, disminuyen significativamente durante la transición al estado adulto [20]. Se ha establecido que la expresión continua de miR156 prolonga la fase juvenil, mientras que la disminución de su actividad promueve el cambio de fase y una floración temprana [25].

El miR156 controla la transición de la fase vegetativa mediante la inhibición directa de la expresión de los factores de transcripción SQUAMOSA Promoter-Binding Protein-Like (SPL) [20]. Estos factores son cruciales en diversos procesos fisiológicos, como la transición de fase, la floración, el desarrollo de frutos, la estructura de la planta, la señalización de giberelinas, la esporogénesis, la homeostasis del cobre y la respuesta a toxinas fúngicas [26]. Los genes SPL muestran una expresión diferencial en las hojas durante las etapas juveniles y adultas en plantas tanto herbáceas como leñosas [20], [26]. Por ejemplo, SPL3 está relacionado con la inducción de la floración y la adquisición de características foliares adultas, mientras que SPL9 favorece el desarrollo de la morfología foliar adulta y coordina el desarrollo y la tolerancia al estrés [20], [26]. Los niveles de SPL3 y SPL9 son bajos al principio y aumentan con el envejecimiento de la planta, regulados a nivel post-transcripcional por miR156 [20], [26]

Este cambio también incluye al miR172, que interviene en la fase de floración [20]. MiR172 presenta un patrón de expresión que es inversamente proporcional al de miR156, incrementándose a medida que miR156 disminuye [20]. Se cree que las secuencias y funciones de miR156, SPLs y miR172 están evolutivamente conservadas en las plantas [27].

1.2 Herramientas para estudiar las bases genéticas

Típicamente las características poligénicas, se pueden estudiar mediante el mapeo de QTL (del inglés, “quantitative trait locus”) o mediante estudios de asociación del genoma completo (GWAS; del inglés, “genome wide association studies”)[28]. La primera aproximación, implica encontrar variantes en un locus que se correlacione con la variación de un rasgo cuantitativo mediante la construcción de poblaciones de mapeo. Por otro lado, en la aproximación GWAS se estudia si ciertas variantes alélicas están asociadas de manera preferente a los casos (*i.e.,* individuos que presentan la característica) en relación a los controles. Sin embargo, estas estrategias son costosas debido al genotipado de miles de individuos. Una alternativa es el método XP-GWAS o pool-GWAS, [29], [30] donde los individuos no se analizan por separado si no en pooles, permitiendo ahorrar considerablemente los costos del genotipado, ya que implica secuenciar solo dos bibliotecas en vez de una por individuo.

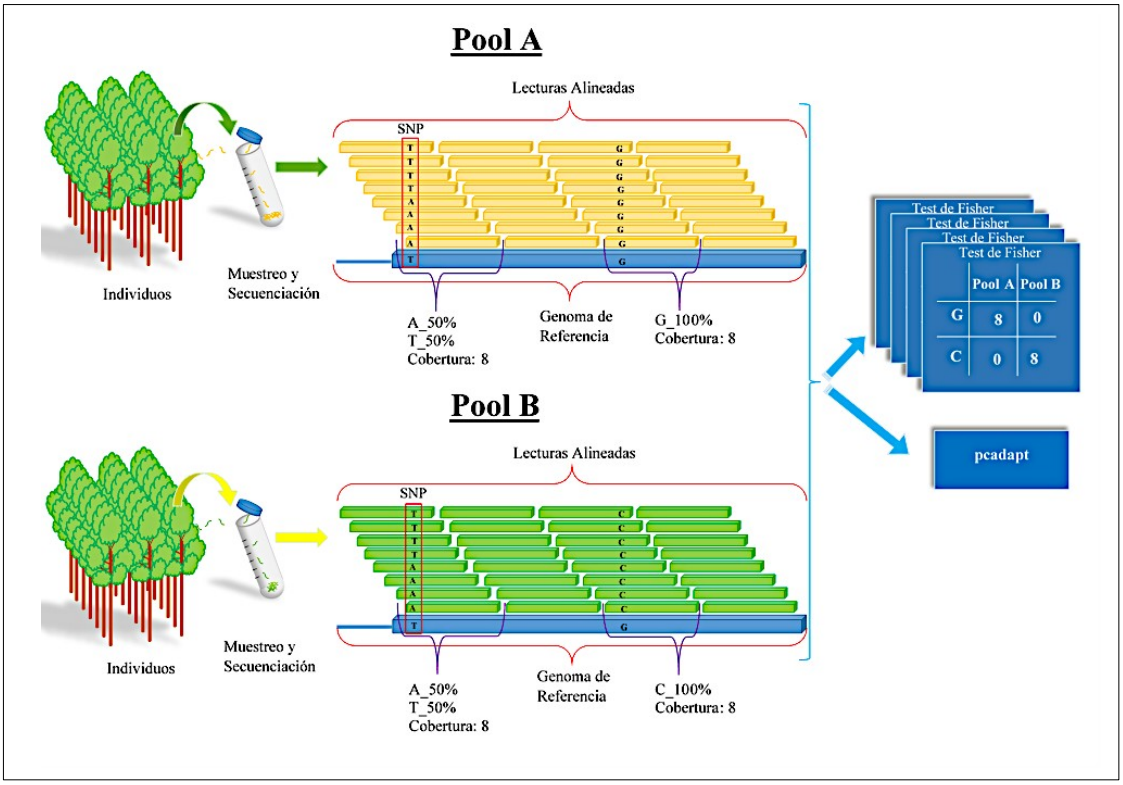

Figura 2. Proceso general del proyecto. Se seleccionaron dos grupos de individuos con fenotipos extremos y opuestos, a partir de los cuales se procedió a la secuenciación genómica de ambos conjuntos. Los fragmentos de ADN secuenciados fueron alineados contra el genoma de referencia de Eucalyptus globulus con el propósito de reconocer variantes alélicas y sus frecuencias entre los grupos. Posteriormente, se aplicó la prueba de Fisher e implementación de la librería pcadapt, que tiene en cuenta la estructura poblacional, para detectar aquellas diferencias en las frecuencias alélicas significativas entre grupos.

En el método del pool-GWAS se combinan muestras de individuos que exhiben fenotipos extremos y opuestos para una característica particular de interés [31]. Estos individuos se agrupan en dos pooles distintos, cada uno representando uno de los extremos fenotípicos. Por ejemplo, uno de los pooles puede agrupar individuos resistentes a la enfermedad, mientras que el otro pool agrupa únicamente individuos susceptibles.

El objetivo principal del pool-GWAS es identificar las variantes genéticas (como por ejemplo, polimorfismos de un solo nucleótido (SNPs) que se describirán más adelante) en las que los pooles difieren significativamente en sus frecuencias alélicas. Por ejemplo, para un nucleótido determinado uno de los pooles puede presentar dos variantes con la misma frecuencia, mientras que el otro pool puede presentar una sola de las variantes. Estas frecuencias se estiman a partir de las lecturas alienadas correctamente a la región genómica de interés (Figura 2). Por esto es importante considerar la profundidad de cobertura en cada pool, ya que una cobertura adecuada permite una mejor estimación de las frecuencias alélicas y también una mejor identificación de las variantes [31]. También es necesario corregir por la estructura poblacional para evitar falsos positivos (Recuadro 2). En definitiva, si el fenotipo de interés está desencadenado en gran medida por una de las variantes, es de esperar que esta variante esté en mucho mayor frecuencia en el pool que muestra el fenotipo.

Dado que se secuencia el genoma entero, las asociaciones significativas no tienen porqué depender de altos niveles de desequilibrio de ligamiento de largo alcance entre las variantes causales y los SNPs genotipados (Recuadro 3). Al identificarse potencialmente todos los SNPs presentes, se podrían identificar directamente los SNPs causales. Este método es ideal para estudiar las bases genéticas de características que no dependen de un gran número de genes, como la que presentaría la precocidad en el cambio de follaje [10].

Recuadro 2. Estructura Poblacional [32]

La estructura poblacional se refiere a la subdivisión de una población en subpoblaciones más pequeñas, conocidas como *demos*, que se diferencian genéticamente de la población general, por ejemplo en sus frecuencias alélicas. Estas subpoblaciones suelen estar geográficamente separadas y pueden interconectarse a través del flujo génico, es decir, el intercambio de individuos (y alelos) entre subpoblaciones. Estas subpoblaciones pueden formarse debido a diversos factores, tales como la distancia geográfica, barreras naturales (como montañas o ríos), migración, y la selección natural que favorece ciertas variantes genéticas en entornos específicos. Esta subdivisión poblacional puede llevar a diferencias genéticas significativas entre las subpoblaciones que se manifiestan y se pueden inferir por medio de sus frecuencia alélicas, lo que a su vez puede influir en la variabilidad genética total de la población. En cuanto a su influencia en estudios genéticos, como los GWAS, es importante considerar la estructura poblacional, ya que no tenerla en cuenta puede llevar a asociaciones falsas entre variantes genéticas y rasgos de interés. Esto puede suceder en los casos donde las poblaciones analizadas además de diferir para la característica de estudio y sus alelos causales, también difieran para muchas variantes genéticas a causa de que están estructuradas. Por lo cual realizar ajustes con la estructura poblacional ayuda a evitar estos sesgos y a identificar asociaciones genéticas verdaderas [32].

Recuadro 3. Desequilibrio de ligamiento [33]

El desequilibrio de ligamiento se refiere a la relación no aleatoria entre alelos que se encuentran en diferentes loci (diferentes posiciones) en un cromosoma, lo que hace que tiendan a heredarse juntos en mayor frecuencia de lo que se esperaría por azar, es decir, asumiendo que son independientes. La proximidad física entre los loci, por ejemplo, aumenta la probabilidad de que los alelos cercanos se hereden juntos con mayor frecuencia. El desequilibrio de ligamiento también se ve afectado por numerosos factores genético-poblacionales [33].

1.3 Tipo de Secuenciación

La secuenciación del ADN es el proceso mediante el cual se determina el orden exacto de las cuatro bases nucleotídicas (adenina, timina, citosina y guanina) que componen una molécula de ADN. El resultado de la secuenciación es un conjunto de datos en forma de “lecturas” o *reads*, que son fragmentos cortos de ADN secuenciados. La secuencia resultante proporciona información sobre la composición de una región específica del ADN. Dependiendo de la tecnología empleada en la secuenciación y el objetivo específico del estudio, los fragmentos de ADN pueden variar tanto en el número de lecturas obtenidas como en su longitud. Asimismo, dichos fragmentos pueden representar tanto el genoma completo como únicamente una región

específica de interés dentro del mismo. Este proceso es fundamental, ya que a través de estudios posteriores no solo permite identificar la ubicación de genes, sino que también posibilita inferir las variantes específicas. La detección de variantes es clave en genética de poblaciones y para los estudios de asociación genética [33]. Además, la secuenciación se utiliza para el análisis de la expresión génica cuando se secuencia ARN.

En los últimos años, la secuenciación de Illumina se ha consolidado como el método principal debido a su alta precisión y capacidad de paralelización, lo que permite una reducción significativa en los costos de secuenciación genómica. La secuenciación por síntesis de Illumina utiliza microscopios fluorescentes y microfluidos para secuenciar ADN en una celda de flujo con nanopocillos (Figura 3). Esta celda contiene canales por donde fluyen las bibliotecas, con una superficie de vidrio grabada con millones de nanopocillos los cuales tienen conectores sintéticos de ADN complementarios a los adaptadores. Luego de cargar la celda, en cada nanopocillo se hibrida un fragmento de la biblioteca y se amplifica la señal mediante un PCR especial, denominado PCR en puente. Posteriormente se agregan nucleótidos fluorescentes a la celda los cuales se incorporan a fragmentos agrupados base a base. La síntesis de ADN se detiene con la incorporación de nucleótidos lo que permite obtener imágenes de toda la celda bajo iluminación láser. Al finalizar un algoritmo traduce las imágenes tomadas en los ciclos de secuenciación por síntesis de bases en una secuencia [34]. Las plataformas de Illumina tienen la capacidad de generar *reads* con una longitud que varía entre 36 pb y 300 pb [35]. Dependiendo del tipo de plataforma utilizada, la cantidad de reads por corrida puede variar desde 4 millones hasta 52 mil millones de *reads* [36].

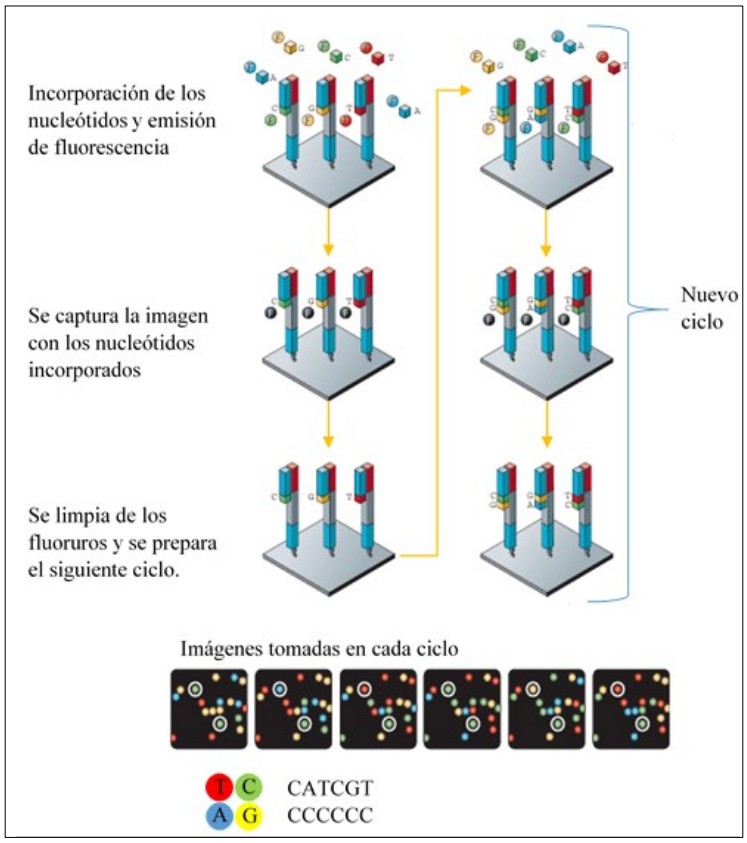

Figura 3. Secuenciación por síntesis de la plataforma Illumina. Antes de iniciar este proceso, se realizó la amplificación en puente, donde las moléculas de ADN se unieron a una superficie sólida y se replicaron en clústeres. Posteriormente, se da inicio a los ciclos de secuenciación, en los cuales se incorporan nucleótidos marcados con fluoróforos. En el primer paso, ocurre la incorporación del nucleótido y se emite una fluorescencia. En el segundo paso, se captura una imagen que permite identificar los nucleótidos agregados. Después, se eliminan los fluoróforos para preparar la muestra para el siguiente ciclo. Este proceso se repite hasta completar la secuenciación de los fragmentos de ADN, permitiendo la lectura de millones de secuencias en paralelo. Como se puede apreciar en la parte inferior de la figura, se ilustra una secuencia de fragmentos después de seis ciclos. Adaptado de [37].

Para secuenciar una muestra de ADN primero es necesario realizar la preparación de una biblioteca o librería. Típicamente la preparación de una biblioteca involucra varios pasos cruciales [34]. En primer lugar, el proceso de *Shearing* se utiliza para fragmentar el ADN en fragmentos de longitud específica, dependiendo de la tecnología de secuenciación utilizada. Estos fragmentos suelen tener un tamaño específico que generalmente oscila entre 200 y 500 pb, [38]. Luego, se realiza la reparación del ADN, que implica fosforilar los extremos y agregar

colas para hacerlos compatibles con adaptadores. La ligación de adaptadores es esencial, ya que estos contienen secuencias necesarias para la amplificación y secuenciación. Recientemente, se han desarrollado protocolos basados en transposasas que simplifican la preparación de librerías. Para esto las transposasas son agregadas a la muestra de ADN; éstas pueden realizar el corte y ligamiento de adaptadores precargados en un solo paso [34]. La tecnología de secuenciación Illumina puede requerir la amplificación por PCR dependiendo de la cantidad de ADN disponible. Si se cuenta con suficiente ADN se pueden utilizar bibliotecas PCR-free, que no se amplifican. El uso de *barcoding* permite agrupar múltiples bibliotecas para reducir costos y mantener la identificación de las muestras originales.

Cabe mencionar que existen otros métodos de secuenciación, tales como *"Oxford Nanopore Technology"* (ONT), que utiliza nanoporos de proteínas y un circuito integrado de aplicación específica (ASIC). Este chip detecta cambios en la corriente iónica cuando la molécula de ADN pasa a través del nanoporo. Este tipo de secuenciación tiene la capacidad de generar lecturas ultralargas [34]. Otro método de secuenciación es el de PacBio, que es similar a Illumina en cuanto a la microscopía de fluorescencia, pero difiere en que permite la secuenciación de una única molécula en tiempo real. Aunque ambos métodos pueden procesar lecturas largas, no cuentan con la capacidad de paralelización que ofrece Illumina. A pesar de la existencia de estos métodos, Illumina sigue siendo ampliamente utilizada para identificar variantes debido a su fiabilidad y costo [34].

1.4 Polimorfismo de nucleótido único y llamado de variantes

Los polimorfismos de un nucleótido único o SNP (del inglés, "Single Nucleotide Polymorphism") son una variación en la secuencia del ADN que ocurre cuando una sola base (nucleótido) en el genoma se altera que se utilizan como marcadores bialélicos. Este tipo de polimorfismo se produce cuando hay una sustitución de un solo nucleótido por otro en una posición específica en la secuencia del ADN. Los SNPs son uno de los tipos de variaciones genéticas más comunes en los genomas de los organismos y pueden tener un impacto en el desarrollo de enfermedades, la respuesta a factores externos y otras características.

El proceso llamada de variantes (en inglés: *SNP calling*), consta de siete etapas generales [39], las cuales se describirán a continuación.

Llamada de bases

Como mencionamos previamente, un dispositivo captura la señal luminosa emitida por la hebra recién generada. Una vez adquirida esta señal, se debe llevar a cabo la conversión de las imágenes en bases nucleotídicas. Diversas plataformas encargadas de esta tarea utilizan diferentes modelos estadísticos que buscan proporcionar una medida de la certeza de cada llamada de base. Estos modelos estiman el error potencial mediante varios parámetros, como la intensidad de la señal de la imagen registrada, el número de ciclos de secuenciación y la distancia con respecto a otras *clusters* de secuencias. Esta certeza puede expresarse en forma de puntuaciones de calidad, como las puntuaciones de Phred, que se calculan mediante el logaritmo decimal de la probabilidad del error esperado para esa base: $Q_{\text{Phred}} = -10 * \log(P, \text{error})$. Utilizando esta fórmula, un error con una probabilidad del 5% se traduce en una puntuación de Phred de 13.

Control de calidad

La distribución de las puntuaciones de calidad en cada secuencia es uno de los parámetros más relevantes para evaluar la calidad general. Por lo general, el software de llamada de bases proporciona una visión inicial de la calidad de los datos. Sin embargo, para obtener una visión más completa y detallada, es necesario emplear otras herramientas especializadas, como FastQC [40].

Alineamiento

El siguiente paso crucial es la alineación de las lecturas contra un genoma de referencia. Dado que es necesario alinear millones de lecturas frente a una referencia, se han desarrollado diversos algoritmos eficientes para esta tarea. BWA [41] y HISAT [42] son algunos ejemplos de alineadores. La elección tanto de la herramienta de alineación como de los ajustes correspondientes tiene un impacto significativo en el resultado final del análisis. Es importante encontrar un equilibrio: permitir solo alineaciones perfectas en relación con la referencia impedirá la detección de SNPs, mientras que permitir un alto número de diferencias entre la lectura y la referencia generará múltiples alineamientos erróneos y una abundancia de SNPs falsos positivos en los análisis subsiguientes.

Pos-procesamiento del alineamiento

El paso siguiente, que precede a la etapa de llamada de variantes, implica organizar los alineamientos en función de su ubicación cromosómica. Esta tarea se puede llevar a cabo utilizando herramientas como SAMtools [43] o Picard [44].

Luego, debido a que la PCR utilizada para ampliar la biblioteca y añadir adaptadores puede introducir artefactos, es decir, lecturas o pares de lecturas que comienzan exactamente en la misma posición y tienen la misma longitud de inserción, es una práctica común eliminar o marcar estos artefactos/duplicados de PCR. Una vez más, SAMtools y Picard proporcionan las capacidades necesarias para abordar esta tarea. El siguiente paso en el proceso de post-procesamiento implica la eliminación de todos los alineamientos no únicos, es decir, lecturas que tienen más de una alineación óptima.

Recalibración de la puntuación de calidad

Investigaciones previas han revelado que las puntuaciones de calidad tipo Phred generadas por las primeras plataformas de secuenciación a menudo pueden desviarse de manera sistemática desde la verdadera tasa de error. La precisión en estas puntuaciones de calidad es fundamental para los algoritmos modernos de llamada de SNP, dado que estas puntuaciones Phred se consideran al realizar el llamado de bases. Sin embargo el impacto de la re-calibración en mejorar la calidad de los SNPs no es universal, depende del alineador y del programa utilizado para hacer la llamada de bases [45].

Llamada de variantes y genotipos

En esta etapa, la principal tarea es detectar todos los SNPs que fueron encontrados a partir del alineamiento. Los primeros enfoques para la llamada de variantes se basaban en la simple cuenta de la abundancia de nucleótidos de alta calidad en un solo sitio. Sin embargo, recientemente se han adoptado enfoques probabilísticos que integran diversas fuentes de información. Esta técnica resulta especialmente beneficiosa en situaciones de baja o media cobertura, donde solo unas pocas lecturas se alinean en la posición del posible SNP.

Filtrado de SNPs Candidatos

Tiene como objetivo principal reducir el número de falsos positivos en las llamadas de variantes. Los filtros aplicados en general buscan detectar desviaciones en el equilibrio de Hardy-Weinberg (HWE), establecer valores mínimos y máximos para la profundidad de lectura, considerar la proximidad a indels y evaluar posibles sesgos de hebra.

1.5 Objetivos

1.5.1 General

Explorar las bases genéticas que intervienen en el cambio de follaje en la especie *E. globulus* utilizando la estrategia de XP-GWAS.

1.5.2 Específicos

Analizar datos de secuenciación masiva e identificación de SNPs.
Analizar las frecuencias alélicas entre ambos pooles.
Adecuar la aproximación estadística en base a la eventual presencia de estructura poblacional.
Identificar los genes potencialmente asociados a la precocidad en el cambio de follaje.
Evaluar los genes candidatos mediante revisión de bibliografía especializada.

2 Materiales y Métodos

2.1 Individuos seleccionados y secuenciación

En este trabajo se utilizaron datos de secuenciación previamente obtenidos como parte del proyecto Vaz Ferreira 2017, titulado "Caracterización de los genes involucrados en el cambio de follaje de *Eucalyptus globulus*" (Responsable: Cecilia Da Silva). Estos datos se obtuvieron a partir de individuos con fenotipos extremos seleccionados dentro del Plan de Mejoramiento Genético del Programa Forestal del INIA. Los individuos fueron clasificados como de cambio foliar precoz si el proceso comenzaba antes de los 14 meses, y como tardíos si comenzaba después de los 2 años. Los individuos provenían de un ensayo de INIA en el departamento de Lavalleja, Uruguay (Lat. 34° 11' S; Long. 54° 54' O; Alt. 206 m), el cual contiene 194 familias de polinización abierta, compuestas por un total de 4601 individuos [46]. Es necesario destacar que, al cumplir aproximadamente un año, los individuos contrajeron la infección de *Teratosphaeria nubilosa*. A partir de esta selección, se formaron dos grupos o *pooles* de 50 individuos cada uno. Para crear cada *pool*, se extrajo el ADN de cada individuo utilizando el protocolo CTAB y se combinaron cantidades equimolares de ADN de cada uno [47]. Una vez conformados los pooles de ADN, estos fueron secuenciados en Macrogen (Macrogen Inc., Seúl, Corea del Sur) empleando la plataforma Illumina NovaSeq, con lecturas pareadas de 151 pares de bases y un inserto de 350 pares de bases.

2.2 Identificación de SNPs y estudio de asociación.

Antes de la etapa de alineamiento, se realizó un control de calidad de las lecturas empleando la herramienta FastQC. Para alinear las lecturas de ambos pooles secuenciados, se utilizó el programa BWA, tomando como referencia el genoma de *E. globulus* de NCBI (ASM1418254v1). La calidad de este genoma es excelente, ya que cuenta con el ensamblado completo de los 11 cromosomas. Una vez finalizado el alineamiento, el siguiente paso fue convertir los archivos en formato SAM a BAM utilizando la herramienta Samtools. Además, con esta misma herramienta, se aplicó el comando "sort" para ordenar las lecturas según su posición cromosómica en el formato BAM, lo que facilita el posterior análisis de estos archivos y la función *depth* para estimar la cobertura de ambos pooles. La identificación y eliminación de

las lecturas duplicadas se realizó mediante el lenguaje de programación Java utilizando el programa Picard, con los comandos MarkDuplicates y REMOVE_DUPLICATES respectivamente. Para la identificación de SNPs, se utilizó el software STRELKA, el cual se ejecutó con sus parámetros por defecto. Luego se utilizó un *script* de Python para transformar el archivo VCF resultante de la llamada de variante a un archivo de formato SYNC.

El análisis de la variación en las frecuencias alélicas se llevó a cabo para cada SNP identificado. En una primera instancia, para calcular estas frecuencias, se realizó un *script* (Anexo 7.1) en Python que incorporó el paquete fast-fisher. Este paquete permite realizar la prueba de Fisher de manera rápida y eficiente, además de manejar grandes volúmenes de datos [48]. El Test de Fisher es una prueba estadística utilizada para evaluar si existe una asociación significativa entre dos variables categóricas. Los datos se organizan en tablas de contingencia 2x2, donde cada celda contiene los valores de una combinación de ambas variables. Este test considera una hipótesis nula (H_0), que asume que las variables son independientes, y una hipótesis alternativa (H_a), que propone que las variables no son independientes, es decir, que existe una asociación entre ellas. La decisión entre estas hipótesis se basa en un nivel de significancia (por ejemplo, 0,05). Si el p-valor es menor que este nivel, se rechaza H_0 en favor de H_a, lo que indica que hay una asociación significativa entre las variables [49]. La implementación del *script* de Python con fast-fisher permitió detectar los SNPs que son significativamente diferentes. La presencia de estructura poblacional se evaluó mediante la distribución de los p-valores en un gráfico de cuantiles (gráfico Q-Q). La estructura poblacional puede distorsionar la significancia de los SNPs, generando falsos positivos, ya que introduce diferencias en las frecuencias alélicas entre los grupos que no se deben directamente a la característica de interés.

Al detectar estructura poblacional, empleamos, en una segunda instancia, el software R junto con la librería pcadapt [50]. Esta librería utiliza un enfoque de análisis de componentes principales para corregir la estructura poblacional, partiendo del supuesto de que la variante relacionada con la característica presenta una estructura más pronunciada que la de fondo. Para emplear la librería pcadapt, se requieren varios pasos para lograr el resultado (script de R, Anexo 7.2). Para esto, se utilizaron las siguientes librerías “poolfstat”[51], [52], “MASS”[53], “data.table”[54], [55], “qvalue”[56], [57] y “qqman”[58]. Estas librerías permitieron generar archivos intermedios necesarios para alimentar a la librería “pcadapt” con nuestros datos y así generar la salida esperada. En estos pasos intermedios, también se utilizaron *scripts* de Python que permitieron ordenar y transformar archivos. Tanto para los análisis con fast-fisher como

para pcadapt, consideramos únicamente los SNPs que presenten una cobertura mínima de 30x (es decir, deberán estar cubiertos por al menos 30 lecturas) y una cobertura máxima de hasta 150x. Esto nos permitirá evitar la inclusión de SNPs falsos causados por errores de secuenciación, los cuales generalmente se asocian con una cobertura muy baja, así como SNPs resultantes de un alineamiento incorrecto de las lecturas, que pueden producir coberturas excepcionalmente altas.

Posteriormente, mediante Python, en las proximidades del miR156, se realizó una búsqueda con un filtro de un q-value menor a 0,1 para encontrar aquellos SNPs marginalmente significativos, que pueden no haber sido detectados debido a la falta de poder estadístico de la herramienta o al acotado desequilibrio de ligamiento con el que se trabajó (6kb, ver más adelante).

2.3 Genes asociados a los SNPs codificantes.

Debido a que el genoma de *E. globulus* no está anotado; es decir, no disponíamos de coordenadas que indiquen la ubicación de cada gen en el genoma, proyectamos las anotaciones del genoma de *Eucalyptus grandis* (Ensamblado: ASM1654582v1; Anotación RefSeq: GCF_016545825.1) al de *E. globulus*. Para lograr esto, se utilizó el programa LIFTOFF [59]. Todos los genes que se encuentren a 6 kb cascada arriba y abajo de cada SNP significativo fueron reportados y analizados mediante una revisión bibliográfica adecuada. Se utiliza este filtrado de 6 kb ya que estudios han demostrado que el desequilibrio de ligamiento en *Eucalyptus* se encuentra, en promedio, entre 4 y 6 kb [60].

3 Resultados

Las muestras que se obtuvieron para ambos pooles fueron de excelente calidad (Figura 4) y por lo tanto no fueron recortadas y se procedió a su alineamiento contra el genoma de *E. globulus*. En la Tabla 1 se muestra el número de lecturas alineadas por cada grupo de individuos (*pool*), así como la cobertura utilizada en cada uno de los alineamientos. En el grupo de individuos con cambio rápido de follaje, se obtuvo un total de 505.056.755 lecturas alineadas, mientras que en el grupo de individuos con cambio lento de follaje se registraron 513.971.051 lecturas alineadas. Con respecto a la cobertura promedio para el pool de cambio rápido se obtuvo una cobertura de 120x, y para el de cambio tardío una de 124x. En cuanto al porcentaje del genoma cubierto, observamos que se logró un valor próximo al 100% para ambos pooles.

Pooles	**Lecturas alineadas**	**Cobertura promedio**	**Genoma cubierto > 15x**
Cambio rápido	505.056.755	120x	0.92%
Cambio lento	513.971.051	124x	0.93%

Tabla 1. Datos de alineación de lecturas.

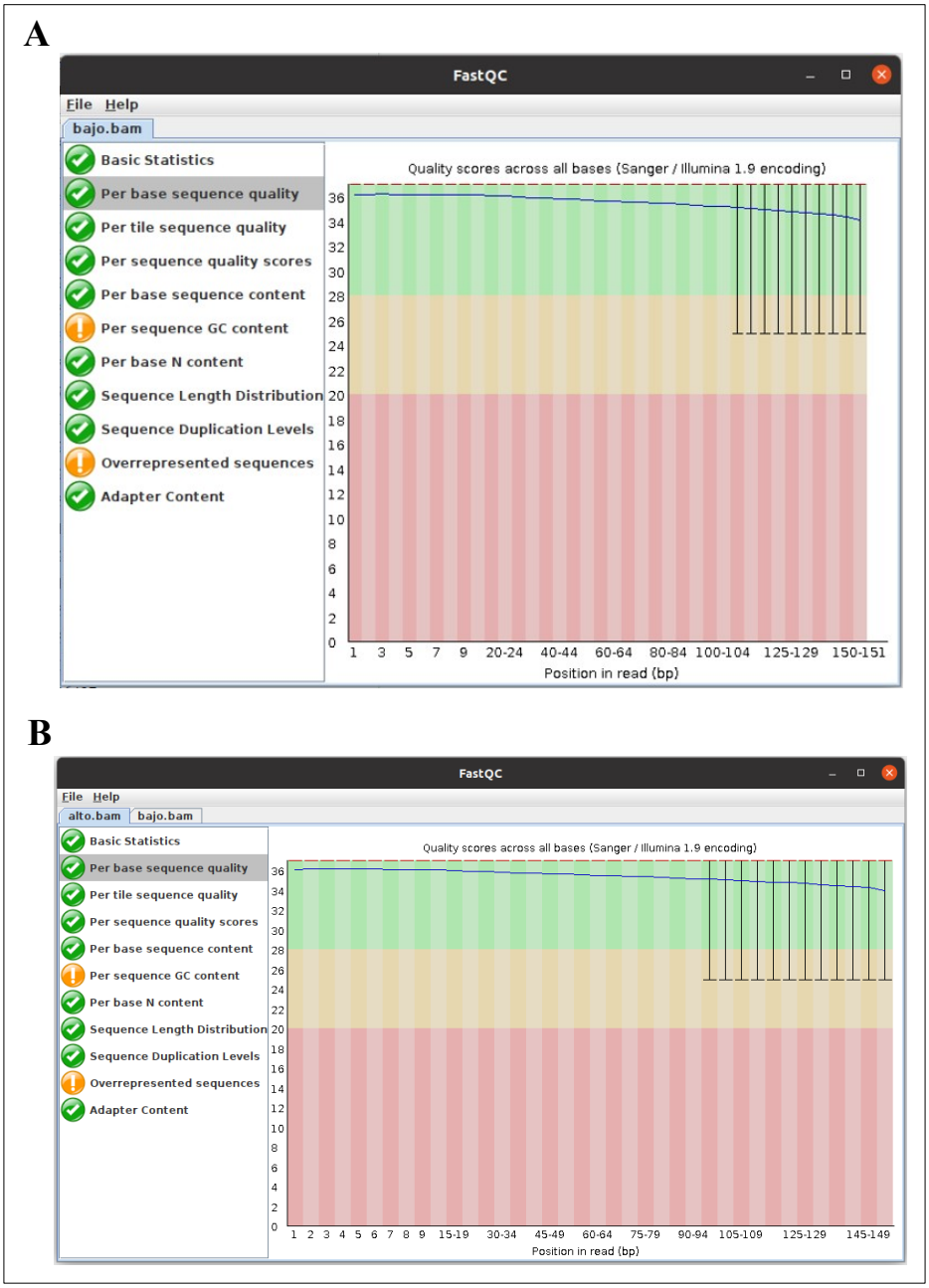

Figura 4. Valor Phred de las lecturas. A. Calidad de las lecturas para el pool "cambio lento". B. Calidad de las lecturas para el pool "cambio rápido". En la figura se muestra el valor Phred promedio para cada posición de la lectura. Todas las posiciones de la lectura tienen un valor de calidad Phred mayor o igual a 28 (línea azul). Solo en las últimas 40 bases la calidad decae como se deduce a partir de su desviación estándar.

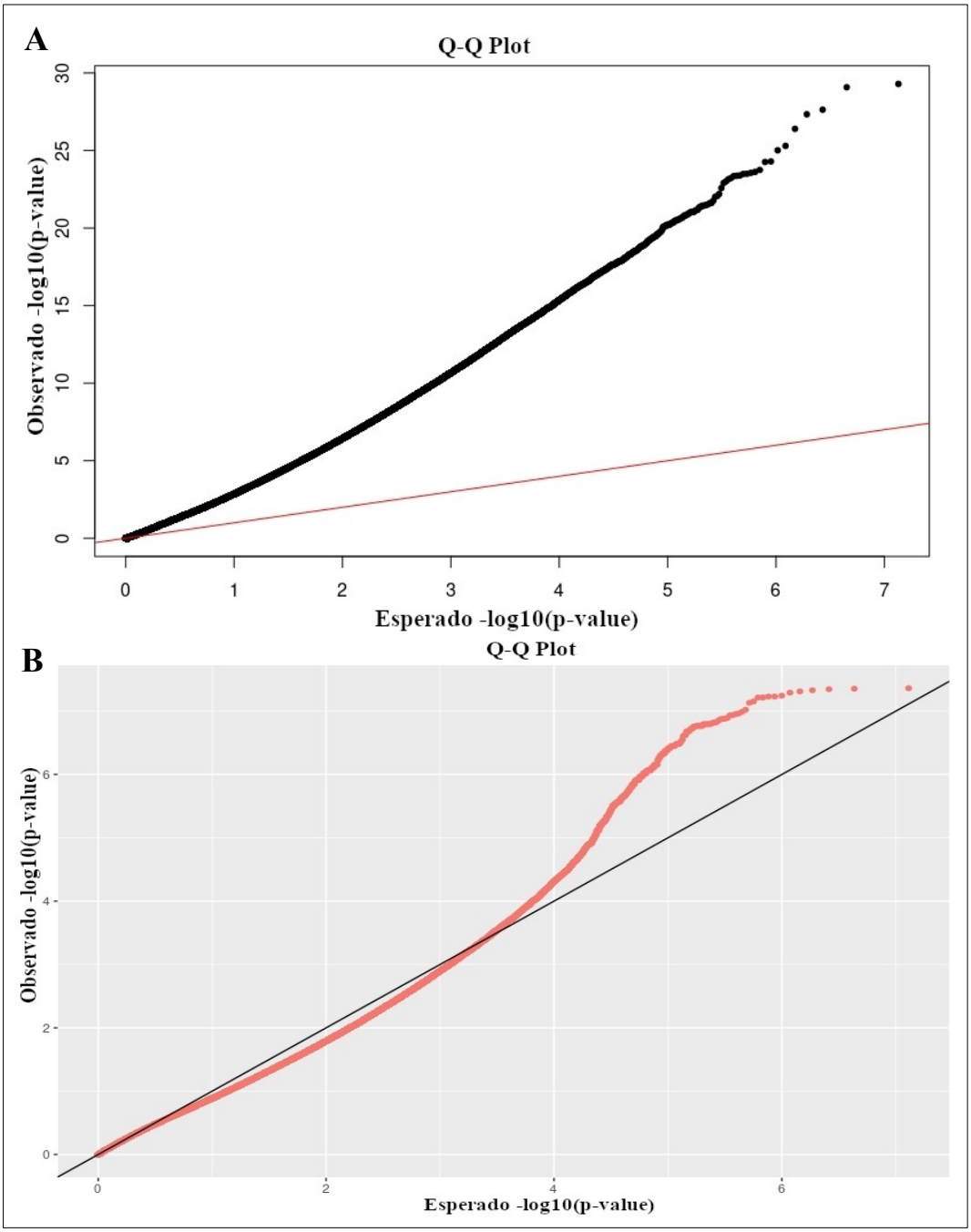

Figura 5. A- Q-Q Plot de los p-valores de test de fisher sin corregir por estructura poblacional. B- Q-Q Plot de los p-valores corregidos por estructura poblacional utilizando el programa pcadapt.

Los *pooles* muestran evidencia de estructuración genética (Figura 5A). Cómo se puede observar en la Figura 5, los p-valores observados (puntos) difieren significativamente de lo esperado (línea roja). Si la distribución de los p-valores no estaría sesgado por la estructura poblacional,

es de esperar que estos p-valores tengan una distribución uniforme y se ajusten a la línea roja. Luego de utilizar el pcadapt para corregir la estructura genética se puede observar un mejor ajuste (Figura 5B). Luego de corregir por estructura poblacional en los 9 millones de SNPs iniciales, se identificaron 86 SNPs significativos con un q-valor menor a 0.05 (Figura 6B. *Manhattan Plot* y Tabla 2). A modo de ejemplo, si los p-valores no son corregidos por estructura poblacional y se procede a estimar sus q-valores, obtenemos millones de SNPs menores a 0.05. Esto indica que el pcadapt elimina satisfactoriamente falsos positivos producto de la estructura genética.

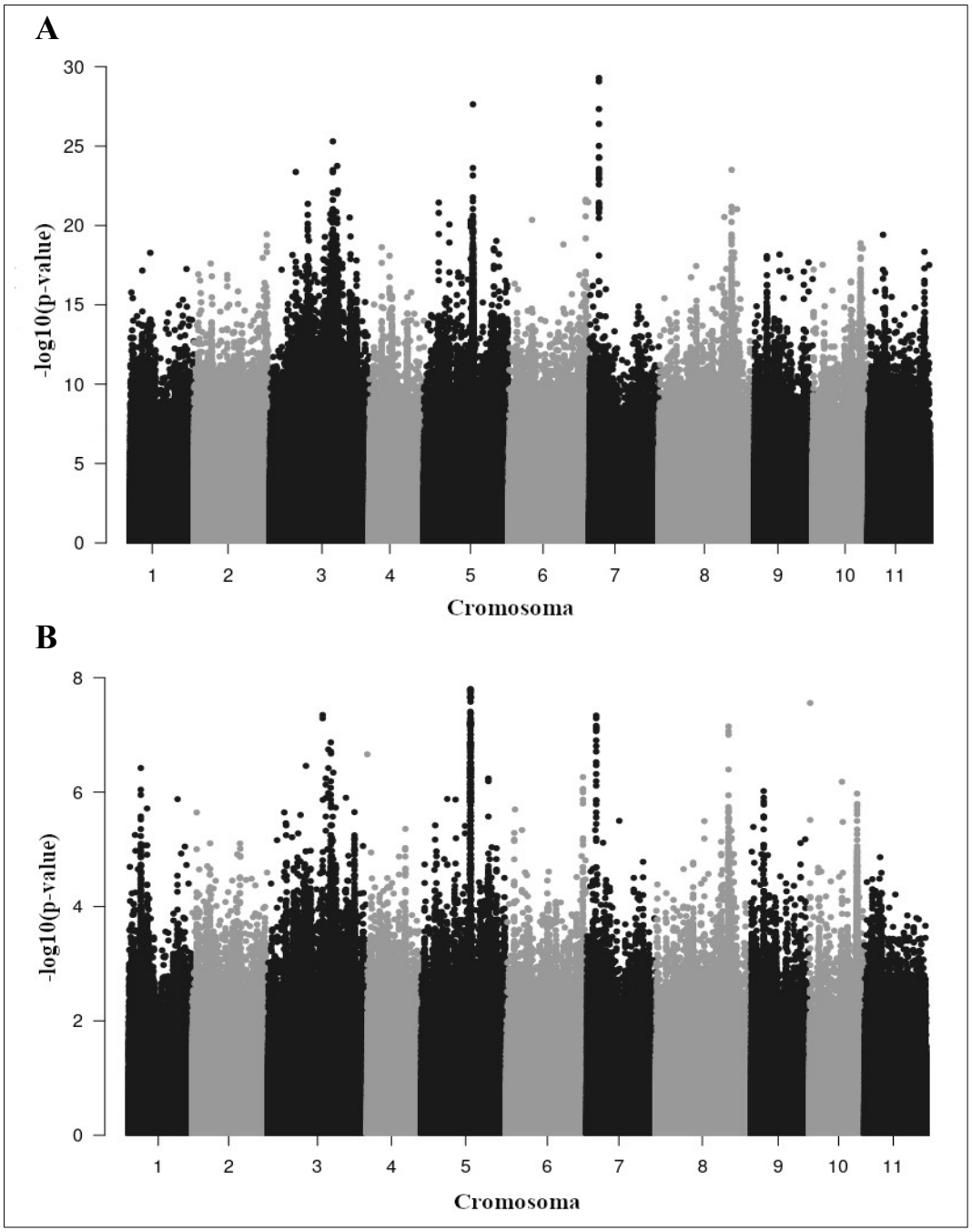

Figura 6. Manhattan Plot de los p-valores sin corregir (A) y de los p-valores corregidos (B). En ambos Manhattan plot se observan los p-valores para cada SNP de cada uno de los 11 cromosomas de Eucalyptus globulus. Estos p-valores se encuentran transformados de tal manera que aquellos con un mayor valor indican que las frecuencias alélicas en el SNP analizado son muy dispares entre los pooles.

A grandes rasgos se puede observar que el *Manhattan plot* de los p-valores sin corregir (Figura 6A) y corregidos por estructura poblacional (Figura 6B) señalan las mismas regiones y SNPs

como las más significativas. A partir de la Figura 6B, se puede observar 7 grandes grupos de SNP significativos dentro de los cromosomas 1, 3, 5, 7, 8, 9 y 10, de los cuales 4 resultaron significativos (Tabla 2). Estos grupos de SNPs identificados en los cromosomas antes mencionados se diferencian de los demás debido a que se pueden notar columnas de SNPs, producto del desequilibrio de ligamiento, que sobresalen, mientras que en los demás cromosomas no se observan estos grupos diferenciados de SNPs.

Gen	**Nombre**	**Cromosoma**	**Inicio / Final (bp)**	**SNPs-Asociados**
LOC104445514	Calcium-dependent protein kinase 24-like	3	36973037 / 36973822	2
LOC120294978	ncRNA	3	36974559 / 36976450	2
LOC108958191	TMV resistance protein N-like	3	42538763 / 42541412	2
LOC120293477	ncRNA	5	33039211 / 33042517	2
LOC120293716	Ankyrin repeat-containing protein BDA1-like	5	33109540 / 33110170	33
LOC120286457	Disease resistance protein RUN1-like	7	6364500 / 6368171	1
LOC104415333	Protein coding	8	49165739 / 49169401	1
LOC120287401	Protein coding	8	49170233 / 49170618	1
LOC104415334	ncRNA	8	49175304 / 49178588	1

Tabla 2. Genes asociados a los SNPs significativos.

De los 86 SNPs significativos, 45 de ellos se encuentran asociados a 9 genes a menos de 6000 pares de bases (Tabla 2). Estos genes están presentes en los cromosomas 3, 5, 7 y 8. Para el resto de los SNP significativos no se encontraron genes asociados. Dentro de los 9 genes, se pueden observar que 4 de ellos están asociados a genes de resistencia. Esto son: calcium-

dependent protein kinase 24-like (CDPK), TMV resistance protein N-like (TMV), ankyrin repeat-containing protein BDA1-like (BDA1) y disease resistance protein RUN1-like (RUN1-like). Entre los 9, se encuentran además dos genes codificantes para proteínas que tienen una función desconocida y 3 genes no codificantes (ncRNA).

De los tres ncRNA, luego de ser interrogados mediante BLAST, solo uno de ellos (LOC104415334), mostró homología con un gen conocido, Ankyrin-1-Like.

SNP	Cromosoma	Posición	Q-Value	Distancia al miR156 (pb)
3_44170060	3	44.170.060	0,06	135.541

Tabla 3. SNP marginalmente significativo cercano al miR156.

Luego de buscar SNPs con q-valores menores a 0.1 alrededor del mirRNA156, encontramos un SNP en la posición 44.170.060 del cromosoma 3 a 135.541 pares de bases de distancia de dicho miRNA.

4 Discusión

La tecnología XP-GWAS representa una alternativa excelente para explorar las bases genéticas debido a su costo reducido en comparación con otras tecnologías. En este trabajo exploramos las bases genéticas de la heteroblastia en dos pooles de *E. globulus* con fenotipos extremos para la característica estudiada.

Ambos pooles mostraron una cobertura promedio alrededor de 120x, suficiente para realizar la llamada de variantes a partir de pooles [31]. La cobertura y extensión obtenida en este trabajo es mayor que la reportada por Giorello et al., 2023. En en este trabajo utilizamos el genoma de *E. globulus* como referencia, a diferencia de Giorello et al., 2023 que utilizó el genoma de *E. grandis*, y utilizamos el alineador BWA en vez de HISAT2. Estas diferencias, y principalmente el genoma utilizado como referencia, podría explicar las diferencias en la cobertura y extensión reportadas. Al obtener una alta a muy alta cobertura (mayor 100x) se obtiene una mejor estimación de la frecuencias alélicas de los pooles, mejorando la posible identificación de SNPs significativos. Las muestras mostraron una estructuración genética que fue posteriormente corregida con pcadapt. Sin embargo, el sesgo en los p-valores causado por la estructura genética de las muestras no pudo erradicarse del todo (Figura 5). Luego de ajustar por la tasa de descubrimiento falso (q-value menor a 0.05) encontramos 86 SNP, de los cuales 45 se asocian a 9 genes que se describen a continuación.

4.1 SNPs asociados al miR156

En términos estrictos, no encontramos ningún SNP significativo asociado al miR156. Sin embargo, encontramos un SNP marginalmente significativo (q-value 0.06) en las inmediaciones del miR156. Cabe destacar que Giorello et al., 2023 tampoco encontró SNP significativos asociados al miR156, aunque sí marginalmente significativos, mientras que Quezada et al., 2023, reportó un SNP significativo asociado al miR156. Quezada et al., 2023, analizó genotipos individuales significativos en el contexto de un estudio de asociación del genoma completo (GWAS), el cual permite utilizar técnicas estadísticas más complejas y potentes.

Si bien el SNP en cuestión no se encuentra cerca del miRNA de interés, no se puede descartar la posibilidad de que ambos estén en desequilibrio de ligamiento. Nuestro SNP marginalmente significativo se encuentra a 135.541 pares de bases de distancia del miR156 y estudios han demostrado que en eucaliptos es posible encontrar un haplotipo, es decir, un grupo de variantes genómicas que tienden a heredarse juntas, de hasta 175.5 kb [61].

Por otro lado, al estudiar las frecuencias alélicas en las inmediaciones del miR156 (menor a 6kb de distancia) observamos que decenas de SNPs cercanos al miR156 muestran una diferencia importante en las frecuencias alélicas pero con una particularidad. SNPs con más de 100 de cobertura se muestran para un pool como homocigotas mientras que para el otro pool como heterocigota (Anexo 7.3). Aunque este tipo de distribución alélica no es la que está asociada con la menor probabilidad (existen distribuciones alélicas aún mas extremas), sería interesante evaluar en un futuro si se trata de una señal causal o casual.

4.2 Genes asociados a SNPs significantes

En relación a los genes detectados, el RUN1-like, codifica para una proteína de resistencia en plantas para ciertas enfermedades, especialmente al hongo *Erysiphe necator*, que causa el oídio en la vid. Originalmente identificado en *Muscadinia rotundifolia*, este gen ha sido transferido con éxito a *Vitis vinifera*, proporcionando a estas plantas una notable resistencia contra este patógeno [62]. La importancia del RUN1-like reside en su capacidad para activar una respuesta inmune robusta y efectiva en la planta. Esta respuesta incluye la generación de especies reactivas de oxígeno (ROS), la acumulación de callosa, la inducción de la muerte celular programada (PCD) y la activación de genes específicos de defensa como VvSTS36 y VvPEN1 [63]. Estos mecanismos de defensa trabajan en conjunto para prevenir la penetración y el desarrollo del hongo, reduciendo considerablemente la tasa de infección y el crecimiento de las hifas del patógeno [62], [63], [64].

También encontramos otros genes relacionados a la resistencia, tales como CDPK y BDA1 y TMV. El gen CDPK, desempeña un papel importante en la regulación del crecimiento de las plantas, así como en su desarrollo y tolerancia al estrés abiótico y de patógenos [65], [66], [67]. Los genes CDPKs son capaces de detectar cambios en la concentración intracelular de Ca^{2+} y convertir estas variaciones en eventos de fosforilación, los cuales activan procesos de

señalización a niveles posteriores. Numerosos estudios funcionales y de expresión sobre diversos CDPKs y sus genes codificadores han corroborado su papel multifacético en la respuesta al estrés [68]. BDA1 es un gen que juega un papel en la regulación de la inmunidad en plantas [69]. Codifica una proteína con repeticiones de anquirina en el extremo N-terminal y dominios transmembrana en el extremo C-terminal [70]. Además, las plantas con mutaciones en BDA1 muestran una mayor susceptibilidad a los patógenos, lo que refuerza su papel como componente clave en la señalización de defensa [69]. Por otro lado, una variante de ganancia de función de BDA1 provoca la activación constante de la muerte celular y las respuestas de defensa, resaltando su importancia en la regulación de la inmunidad de las plantas [69]. TMV resistance protein N-like es otro de los genes que también se ha encontrado relacionado a resistencia en plantas [71].

De los genes mencionados anteriormente TMV y RUN1-like coinciden con lo reportado por Giorello et al., 2023, donde se evaluó el potencial de XP-GWAS para *Eucalyptus*. Además, cabe mencionar que Quezada et al., 2023, también reportó la presencia de TMV. Sin embargo el gen BDA1-Like, encontrado en este trabajo y asociado a 33 SNPs significativos no fue reportado en Giorello et al., 2023 y tampoco en Quezada et al., 2023. Esto en parte se podría deber a los distintos genoma de referencia utilizados, a las distintas herramientas de alineación utilizadas, así como a los programas utilizados para realizar la llamada de variantes. En el presente trabajo se utilizó STRELKA mientras que Giorello et al., 2023 utilizó GATK.

Cabe mencionar en relación al gen BDA1-Like, que dos SNPs encontrados se localizaron en un exón, es decir, en una región codificante. El resto de los SNPs reportados no se encontraron dentro de ningún gen. Los SNPs en la región codificante podría estar modificando el producto proteico.

La presencia de genes asociados a la resistencia es llamativa y se podría explicar por la presencia de *T. nubilosa* durante los ensayos. El grupo de individuos con cambio tardío de follaje fue seleccionado a partir de árboles que fueron capaces de sobrevivir ante la presencia de *T. nubilosa* durante más tiempo, con el agregado que lo hicieron con follaje juvenil, el cual es más susceptible a la mancha foliar. Esto pudo haber traído consigo una selección indirecta de árboles más resistentes. Los genes aquí reportados y sus variantes podrían tener que ver con esta

sobrevivencia diferencial por parte de estos árboles. En línea con esto, Giorello et al., 2023, encontró que el pool con cambio de follaje tardío posee un mayor número de copias de genes de resistencia.

5 Conclusiones

Mediante una estrategia de XP-GWAS detectamos un SNP asociado al miR156 aunque marginalmente significativo. Estos resultados se suman a los ya reportados en otros trabajos que señalan al miR156 como regulador de la heteroblastia. Por otro lado, también identificamos varios genes potencialmente relacionados a la resistencia que pueden ser de particular interés tanto para *E. globulus* como para otras especies de eucaliptos. Aunque la tecnología utilizada no ofrece un gran poder estadístico, permite obtener una visión general de los cromosomas donde se localizan los SNPs significativos y los genes asociados. XP-GWAS surge como una alternativa considerablemente más económica y útil para una primera aproximación en estudios de este tipo. Esto es particularmente valioso en casos donde no se dispone de información previa sobre la característica a estudiar, generando inicialmente datos que orientan hacia una búsqueda más profunda. Así, se sientan las bases para futuros estudios más costosos, como los análisis de asociación del genoma completo.

5.1 Perspectivas

Del presente trabajo se derivan dos grandes lineas que se podrían continuar estudiando. Una de ellas, es cómo afecta los alineadores y los programas de llamado de variantes a la identificación de SNPs significativos. En nuestro análisis, aunque encontramos genes en común con Giorello et al., 2023, también encontramos algunas diferencias importantes. Por otro lado, sería interesante analizar otros programas para detectar SNP significativos mediante pooles, en particular GWAlpha [72]. Este programa utiliza 3 pooles, de los cuales dos de ellos está compuesto por individuos con fenotipos extremos y opuestos mientras que el tercer es obtenido muestreando al azar la población. Este tercer pool, permite cuantificar el efecto del alelo y a su vez podría aumenta la potencia estadística. Para estudiar su potencial, en primera instancia se podrían realizar simulaciones de muestreo de alelos a partir de los datos genotípicos reportados en Quezada et al., 2023.

6 Bibliografía

[1] M. Bayly, «Phylogenetic studies of eucalypts: fossils, morphology and genomes», *Proceedings of the Royal Society of Victoria*, vol. 128, p. 12, ene. 2016, doi: 10.1071/RS16002.

[2] A. R. Kersting, E. Mizrachi, E. Bornberg-Bauer, y A. A. Myburg, «Protein domain evolution is associated with reproductive diversification and adaptive radiation in the genus Eucalyptus», *New Phytologist*, vol. 206, n.º 4, pp. 1328-1336, jun. 2015, doi: 10.1111/nph.13211.

[3] R. G. Florence, *Ecology and silviculture of eucalypt forests*. CSIRO publishing, 2004.

[4] SPF, «Uruguay Forestal», Sociedad de Productores Forestales del Uruguay. Accedido: 26 de mayo de 2023. [En línea]. Disponible en: https://spf.com.uy/uruguay-forestal/

[5] División Evaluación & Información DGF-MGAP, «Superficie Forestal Del Uruguay (Bosques Plantados) Período 1975 -2021», 2022, [En línea]. Disponible en: https://www.gub.uy/ministerio-ganaderia-agricultura-pesca/datos-y-estadisticas/estadisticas/superficie-forestal-bosques-plantados-del-uruguay

[6] G. Balmelli *et al.*, «Impact of Teratosphaeria nubilosa over tree growth and survival of Eucalyptus globulus and Eucalyptus maidenii in Uruguay», *New Forests*, vol. 47, n.º 6, pp. 829-843, nov. 2016, doi: 10.1007/s11056-016-9547-3.

[7] G. C. Hunter, P. W. Crous, A. J. Carnegie, y M. J. Wingfield, «Teratosphaeria nubilosa, a serious leaf disease pathogen of Eucalyptus spp. in native and introduced areas», *Molecular plant pathology*, vol. 10, n.º 1, pp. 1-14, ene. 2009, doi: 10.1111/j.1364-3703.2008.00516.x.

[8] E. A. Pinkard y C. L. Mohammed, «Photosynthesis of Eucalyptus globulus with Mycosphaerella leaf disease», *New Phytologist*, vol. 170, n.º 1, pp. 119-127, mar. 2006, doi: 10.1111/j.1469-8137.2006.01645.x.

[9] G. Balmelli, V. Marroni, N. Altier, y R. Garcia, *Potencial Del Mejoramiento Genético Para El Manejo De Enfermedades En Eucalyptus globulus*. en Tecnica, no. 143. Andes 1365, Piso 12. Montevideo - Uruguay: Unidad de Agronegocios y Difusión del INIA, 2004. [En línea]. Disponible en: http://www.ainfo.inia.uy/digital/bitstream/item/2888/1/15630021107132338.pdf

[10] G. Balmelli *et al.*, «Susceptibility to Teratosphaeria nubilosa and precocity of vegetative phase change in Eucalyptus globulus and E. maidenii (Myrtaceae)», *Australian Journal of Botany*, vol. 61, n.º 8, pp. 583-591, mar. 2014, doi: 10.1071/BT13225.

[11] A. W. Milgate, B. M. Potts, K. Joyce, C. Mohammed, y R. E. Vaillancourt, «Genetic variation inEucalyptus globulus for susceptibility to Mycosphaerella nubilosa and its association with tree growth», *Australasian Plant Pathology*, vol. 34, n.º 1, pp. 11-18, mar. 2005, doi: 10.1071/AP04073.

[12] M. J. Steinbauer, «Oviposition preference and neonate performance of Mnesampela privata in relation to heterophylly in Eucalyptus dunnii and E. globulus», *Agricultural and Forest Entomology*, vol. 4, n.º 4, pp. 245-253, nov. 2002, doi: 10.1046/j.1461-9563.2002.00151.x.

[13] G. Zotz, K. Wilhelm, y A. Becker, «Heteroblasty--a review», *The Botanical Review*, vol. 77, n.º 2, p. 109+, jun. 2011.

[14] R. S. Poethig, «The Past, Present, and Future of Vegetative Phase Change», *Plant Physiology*, vol. 154, n.º 2, pp. 541-544, oct. 2010, doi: 10.1104/pp.110.161620.

[15] A. A. Myburg *et al.*, «The genome of Eucalyptus grandis», *Nature*, vol. 510, n.º 7505, pp. 356-362, jun. 2014, doi: 10.1038/nature13308.

[16] G. D. Balmelli Hernández, «Impacto de Mycosphaerella en Uruguay variabilidad genética para la resistencia a la enfermedad en Eucalyptus globulus Y Eucalyptus maidenii», info:eu-repo/semantics/doctoralThesis, 2014. doi: 10.35376/10324/7220.
[17] J. S. Freeman, B. M. Potts, y R. E. Vaillancourt, «Few Mendelian Genes Underlie the Quantitative Response of a Forest Tree, Eucalyptus globulus, to a Natural Fungal Epidemic», *Genetics*, vol. 178, n.º 1, pp. 563-571, ene. 2008, doi: 10.1534/genetics.107.081414.
[18] C. C. G. Rosado *et al.*, «QTL mapping for resistance to Ceratocystis wilt in Eucalyptus», *Tree Genetics & Genomes*, vol. 12, n.º 4, p. 72, ago. 2016, doi: 10.1007/s11295-016-1029-4.
[19] C. Catalanotto, C. Cogoni, y G. Zardo, «MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions», *Int J Mol Sci*, vol. 17, n.º 10, p. 1712, oct. 2016, doi: 10.3390/ijms17101712.
[20] D. Manuela y M. Xu, «Juvenile Leaves or Adult Leaves: Determinants for Vegetative Phase Change in Flowering Plants», *International Journal of Molecular Sciences*, vol. 21, n.º 24, 2020, doi: 10.3390/ijms21249753.
[21] N. Zhang, G. Hu, T. G. Myers, y P. R. Williamson, «Protocols for the Analysis of microRNA Expression, Biogenesis, and Function in Immune Cells», *Current Protocols in Immunology*, vol. 126, n.º 1, p. e78, sep. 2019, doi: 10.1002/cpim.78.
[22] C. J. Hudson *et al.*, «Genetic Control of Heterochrony in Eucalyptus globulus», *G3 Genes|Genomes|Genetics*, vol. 4, n.º 7, pp. 1235-1245, jul. 2014, doi: 10.1534/g3.114.011916.
[23] G. Wu, M. Y. Park, S. R. Conway, J.-W. Wang, D. Weigel, y R. S. Poethig, «The Sequential Action of miR156 and miR172 Regulates Developmental Timing in Arabidopsis», *Cell*, vol. 138, n.º 4, pp. 750-759, ago. 2009, doi: 10.1016/j.cell.2009.06.031.
[24] L. Yang, S. R. Conway, y R. S. Poethig, «Vegetative phase change is mediated by a leaf-derived signal that represses the transcription of miR156», *Development*, vol. 138, n.º 2, pp. 245-249, ene. 2011, doi: 10.1242/dev.058578.
[25] G. Wu y R. S. Poethig, «Temporal regulation of shoot development in Arabidopsis thalianaby miR156 and its target SPL3», *Development*, vol. 133, n.º 18, pp. 3539-3547, sep. 2006, doi: 10.1242/dev.02521.
[26] X. Chen, Z. Zhang, D. Liu, K. Zhang, A. Li, y L. Mao, «SQUAMOSA Promoter-Binding Protein-Like Transcription Factors: Star Players for Plant Growth and Development», *Journal of integrative plant biology*, vol. 52, pp. 946-51, nov. 2010, doi: 10.1111/j.1744-7909.2010.00987.x.
[27] Y. Luo, Z. Guo, y L. Li, «Evolutionary conservation of microRNA regulatory programs in plant flower development», *Developmental Biology*, vol. 380, n.º 2, pp. 133-144, ago. 2013, doi: 10.1016/j.ydbio.2013.05.009.
[28] T. F. C. Mackay, E. A. Stone, y J. F. Ayroles, «The genetics of quantitative traits: challenges and prospects», *Nature Reviews Genetics*, vol. 10, n.º 8, pp. 565-577, ago. 2009, doi: 10.1038/nrg2612.
[29] H. Bastide *et al.*, «A Genome-Wide, Fine-Scale Map of Natural Pigmentation Variation in Drosophila melanogaster», *PLOS Genetics*, vol. 9, n.º 6, p. e1003534, jun. 2013, doi: 10.1371/journal.pgen.1003534.
[30] J. Yang *et al.*, «Extreme-phenotype genome-wide association study (XP-GWAS): a method for identifying trait-associated variants by sequencing pools of individuals selected from a diversity panel», *The Plant Journal*, vol. 84, n.º 3, pp. 587-596, 2015.
[31] C. Schlötterer, R. Tobler, R. Kofler, y V. Nolte, «Sequencing pools of individuals — mining genome-wide polymorphism data without big funding», *Nat Rev Genet*, vol. 15, n.º 11, pp. 749-763, nov. 2014, doi: 10.1038/nrg3803.
[32] A. Fontdevila y A. Moya, *Introducción a la genética de poblaciones*. 1999.

[33] Benjamin. A. Pierce, *Genética Un enfoque conceptual Edición 5ª*, 5ª edición. Madrid, España: Editorial Médica Panamericana, S.A., 2015.
[34] L. Aigrain, «Beginner's guide to next-generation sequencing», *The Biochemist*, vol. 43, n.º 6, pp. 58-64, dic. 2021, doi: 10.1042/bio_2021_135.
[35] H. Satam *et al.*, «Next-Generation Sequencing Technology: Current Trends and Advancements», *Biology*, vol. 12, n.º 7, p. 997, jul. 2023, doi: 10.3390/biology12070997.
[36] «Sequencing Read Length | How to calculate NGS read length». Accedido: 16 de junio de 2024. [En línea]. Disponible en: https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/read-length.html
[37] M. L. Metzker, «Sequencing technologies — the next generation», *Nat Rev Genet*, vol. 11, n.º 1, pp. 31-46, ene. 2010, doi: 10.1038/nrg2626.
[38] «An Introduction to Next-Generation Sequencing Technology».
[39] A. Altmann, P. Weber, D. Bader, M. Preuß, E. B. Binder, y B. Müller-Myhsok, «A beginners guide to SNP calling from high-throughput DNA-sequencing data», *Hum Genet*, vol. 131, n.º 10, pp. 1541-1554, oct. 2012, doi: 10.1007/s00439-012-1213-z.
[40] «Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data». Accedido: 28 de agosto de 2024. [En línea]. Disponible en: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[41] H. Li y R. Durbin, «Fast and accurate short read alignment with Burrows–Wheeler transform», *bioinformatics*, vol. 25, n.º 14, pp. 1754-1760, 2009.
[42] D. Kim, J. M. Paggi, C. Park, C. Bennett, y S. L. Salzberg, «Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype», *Nat Biotechnol*, vol. 37, n.º 8, pp. 907-915, ago. 2019, doi: 10.1038/s41587-019-0201-4.
[43] H. Li *et al.*, «The Sequence Alignment/Map format and SAMtools», *Bioinformatics*, vol. 25, n.º 16, pp. 2078-2079, ago. 2009, doi: 10.1093/bioinformatics/btp352.
[44] P. Toolkit, «Broad institute», *(No Title)*, 2019.
[45] S. Tian, H. Yan, M. Kalmbach, y S. L. Slager, «Impact of post-alignment processing in variant discovery from whole exome data», *BMC Bioinformatics*, vol. 17, p. 403, oct. 2016, doi: 10.1186/s12859-016-1279-z.
[46] M. Quezada, F. M. Giorello, C. C. Da Silva, I. Aguilar, y G. Balmelli, «Single-step genome-wide association study for susceptibility to Teratosphaeria nubilosa and precocity of vegetative phase change in Eucalyptus globulus», *Front. Plant Sci.*, vol. 14, p. 1124768, jul. 2023, doi: 10.3389/fpls.2023.1124768.
[47] F. M. Giorello, J. Farias, P. Basile, G. Balmelli, y C. C. Da Silva, «Evaluating the potential of XP-GWAS in Eucalyptus: Leaf heteroblasty as a case study», *Plant Gene*, vol. 36, p. 100430, dic. 2023, doi: 10.1016/j.plgene.2023.100430.
[48] *fast-fisher: Calculate Fisher's exact test very quickly.* Python. Accedido: 4 de agosto de 2024. [En línea]. Disponible en: https://github.com/mrtomrod/fast_fisher
[49] A. Agresti, «Introduction to Categorical Data Analysis».
[50] F. Privé, K. Luu, B. J. Vilhjálmsson, y M. G. B. Blum, «Performing Highly Efficient Genome Scans for Local Adaptation with R Package pcadapt Version 4», *Molecular Biology and Evolution*, vol. 37, n.º 7, pp. 2153-2154, jul. 2020, doi: 10.1093/molbev/msaa053.
[51] M. Gautier, R. Vitalis, L. Flori, y A. Estoup, «f-Statistics estimation and admixture graph construction with Pool-Seq or allele count data using the R package poolfstat», *Molecular Ecology Resources*, vol. 22, n.º 4, pp. 1394-1416, may 2022, doi: 10.1111/1755-0998.13557.
[52] V. Hivert, R. Leblois, E. J. Petit, M. Gautier, y R. Vitalis, «Measuring Genetic Differentiation from Pool-seq Data», *Genetics*, vol. 210, n.º 1, pp. 315-330, sep. 2018, doi: 10.1534/genetics.118.300900.

[53] «Modern Applied Statistics with S, 4th ed». Accedido: 4 de agosto de 2024. [En línea]. Disponible en: https://www.stats.ox.ac.uk/pub/MASS4/
[54] T. Barrett, M. Dowle, A. Srinivasan, J. Gorecki, M. Chirico, y T. Hocking, «data.table: Extension of `data.frame`». p. 1.15.4, 15 de abril de 2006. doi: 10.32614/CRAN.package.data.table.
[55] M. Dowle y A. Srinivasan, «data. table: Extension of 'data. frame'», *R package version*, vol. 1, n.º 8, 2019.
[56] «qvalue», Bioconductor. Accedido: 4 de agosto de 2024. [En línea]. Disponible en: http://bioconductor.org/packages/qvalue/
[57] J. D. Storey y R. Tibshirani, «Statistical significance for genomewide studies», *Proceedings of the National Academy of Sciences*, vol. 100, n.º 16, pp. 9440-9445, 2003.
[58] S. D. Turner, «qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots», *JOSS*, vol. 3, n.º 25, p. 731, may 2018, doi: 10.21105/joss.00731.
[59] A. Shumate y S. L. Salzberg, «Liftoff: accurate mapping of gene annotations», *Bioinformatics*, vol. 37, n.º 12, pp. 1639-1643, jul. 2021, doi: 10.1093/bioinformatics/btaa1016.
[60] O. B. Silva-Junior y D. Grattapaglia, «Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of ucalyptus grandis», *New Phytologist*, vol. 208, n.º 3, pp. 830-845, 2015, doi: 10.1111/nph.13505.
[61] C. E. Valenzuela Pinto y F. (Profesor guía) Mora Poblete, «Determinantes genéticos del crecimiento y calidad de la madera de Eucalyptus cladocalyx: un enfoque de haplotipo y asociación de genoma completo», Thesis, Universidad de Talca. Instituto de Ciencias Biológicas, 2021. Accedido: 5 de agosto de 2024. [En línea]. Disponible en: http://dspace.utalca.cl/handle/1950/12807
[62] A. Feechan *et al.*, «Genetic dissection of a TIR-NB-LRR locus from the wild North American grapevine species Muscadinia rotundifolia identifies paralogous genes conferring resistance to major fungal and oomycete pathogens in cultivated grapevine», *The Plant Journal*, vol. 76, n.º 4, pp. 661-674, 2013, doi: 10.1111/tpj.12327.
[63] M. Agurto *et al.*, «RUN1 and REN1 Pyramiding in Grapevine (Vitis vinifera cv. Crimson Seedless) Displays an Improved Defense Response Leading to Enhanced Resistance to Powdery Mildew (Erysiphe necator)», *Front Plant Sci*, vol. 8, p. 758, may 2017, doi: 10.3389/fpls.2017.00758.
[64] I. b. Dry *et al.*, «Molecular strategies to enhance the genetic resistance of grapevines to powdery mildew», *Australian Journal of Grape and Wine Research*, vol. 16, n.º s1, pp. 94-105, 2010, doi: 10.1111/j.1755-0238.2009.00076.x.
[65] S. Ray, P. Agarwal, R. Arora, S. Kapoor, y A. K. Tyagi, «Expression analysis of calcium-dependent protein kinase gene family during reproductive development and abiotic stress conditions in rice (Oryza sativa L. ssp. indica)», *Mol Genet Genomics*, vol. 278, n.º 5, pp. 493-505, nov. 2007, doi: 10.1007/s00438-007-0267-4.
[66] T. Romeis, A. A. Ludwig, R. Martin, y J. D. G. Jones, «Calcium-dependent protein kinases play an essential role in a plant defence response», *The EMBO Journal*, vol. 20, n.º 20, pp. 5556-5567, oct. 2001, doi: 10.1093/emboj/20.20.5556.
[67] B. Wan, Y. Lin, y T. Mou, «Expression of rice Ca2+-dependent protein kinases (CDPKs) genes under different environmental stresses», *FEBS Letters*, vol. 581, n.º 6, pp. 1179-1189, mar. 2007, doi: 10.1016/j.febslet.2007.02.030.
[68] S. D. Dekomah *et al.*, «The role of CDPKs in plant development, nutrient and stress signaling», *Front. Genet.*, vol. 13, sep. 2022, doi: 10.3389/fgene.2022.996203.
[69] Y. Yang, Y. Zhang, P. Ding, K. Johnson, X. Li, y Y. Zhang, «The Ankyrin-Repeat Transmembrane Protein BDA1 Functions Downstream of the Receptor-Like Protein

SNC2 to Regulate Plant Immunity1[C][OA]», *Plant Physiol*, vol. 159, n.º 4, pp. 1857-1865, ago. 2012, doi: 10.1104/pp.112.197152.

[70] C. Becerra, T. Jahrmann, P. Puigdomènech, y C. M. Vicient, «Ankyrin repeat-containing proteins in *Arabidopsis*: characterization of a novel and abundant group of genes coding ankyrin-transmembrane proteins», *Gene*, vol. 340, n.º 1, pp. 111-121, sep. 2004, doi: 10.1016/j.gene.2004.06.006.

[71] R. Marathe, R. Anandalakshmi, Y. Liu, y S. P. Dinesh-Kumar, «The tobacco mosaic virus resistance gene, N», *Molecular Plant Pathology*, vol. 3, n.º 3, pp. 167-172, 2002, doi: 10.1046/j.1364-3703.2002.00110.x.

[72] A. Fournier-Level, C. Robin, y D. Balding, «GWAlpha: Genome-Wide estimation of additive effects (Alpha) based on trait quantile distribution from pool-sequencing experiments», *Bioinformatics (Oxford, England)*, vol. 33, dic. 2016, doi: 10.1093/bioinformatics/btw805.

7 Anexo

7.1 Script implementado en python

```
from fast_fisher import fast_fisher_exact, odds_ratio

file = open("file.sync", "r")
x = 150
y = 30

for linea in file:
    lento = linea.strip().split("\t")
    cromosoma = lento[0]
    numposicion = lento[1]
    nombre_columna1 = lento[3]
    nombre_columna2 = lento[4]
    column_elements = nombre_columna1.split(":")
    column_elements2 = nombre_columna2.split(":")
    column_elements_int = [int(elemento) for elemento in column_elements]
    column_elements_int2 = [int(elemento) for elemento in column_elements2]

    posicion1 = [i for i, val in enumerate(column_elements_int) if val != 0]
    posicion2 = [i for i, val in enumerate(column_elements_int2) if val != 0]

    cobertura1 = 0
    for pos in posicion1:
        cobertura1 = cobertura1 + column_elements_int[pos]
    cobertura2 = 0
    for pos in posicion2:
        cobertura2 = cobertura2 + column_elements_int2[pos]


    if y <= cobertura1 <= x and y <= cobertura2 <= x:
        if len(posicion1) == 2:
            valor1 = column_elements_int[posicion1[0]]
            valor2 = column_elements_int[posicion1[1]]
            valor3 = column_elements_int2[posicion1[0]]
            valor4 = column_elements_int2[posicion1[1]]
            p_value = fast_fisher_exact(valor1, valor2, valor3, valor4)
            print(cromosoma+"_"+numposicion,cromosoma,numposicion,p_value)

        elif len(posicion2) == 2:
            valor1 = column_elements_int[posicion2[0]]
            valor2 = column_elements_int[posicion2[1]]
            valor3 = column_elements_int2[posicion2[0]]
            valor4 = column_elements_int2[posicion2[1]]
            p_value = fast_fisher_exact(valor1, valor2, valor3, valor4)
            print(cromosoma+"_"+numposicion,cromosoma,numposicion,p_value)
```

```
    elif len(posicion1) == 1 and len(posicion2) == 1:
        if posicion1[0] == posicion2[0]:
            valor1 = column_elements_int[posicion1[0]]
            valor2 = 0
            valor3 = column_elements_int2[posicion2[0]]
            valor4 = 0
            p_value = fast_fisher_exact(valor1, valor2, valor3, valor4)
            print(cromosoma+"_"+numposicion,cromosoma,numposicion,p_value)

        else:
            valor1 = column_elements_int[posicion1[0]]
            valor2 = column_elements_int2[posicion2[0]]
            valor3 = 0
            valor4 = 0
            p_value = fast_fisher_exact(valor1, valor2, valor3, valor4)
            print(cromosoma+"_"+numposicion,cromosoma,numposicion,p_value)

file.close()
```

7.2 Script implementado en R

```
setwd("/media/vegetal/f83ce719-d341-4907-aca6-2898c3ea32f6/Documents/STRELKA/
strelka-2.9.10.centos6_x86_64/bin/definitivo_strelka/results/variants")

library(poolfstat)

pooldata=vcf2pooldata(vcf.file="onlyPASS_SNPs.vcf.recode.vcf",min.cov.per.pool=30,poolsiz
es=rep(50,2),poolnames=c("SAMPLE1","SAMPLE2"), max.cov.per.pool = 150)

pooldata2genobaypass(pooldata)

res.fst=computeFST(
        pooldata,
        method = "Anova",
        nsnp.per.bjack.block = 0,
        sliding.window.size = 0,
        verbose = TRUE)

pcadapt_traspuesta<-read.delim("/media/vegetal/f83ce719-d341-4907-aca6-2898c3ea32f6/
Documents/STRELKA/strelka-2.9.10.centos6_x86_64/bin/definitivo_strelka/results/variants/
STRELKA/def_cov_30_150/De nuevo Definitivo 35_150_alpha_0,05/pcadapt_HF_Definitivo",
encoding="858", header=FALSE)

pcadapt_input←t(pcadapt_traspuesta)

library(MASS)

write.matrix(pcadapt_input, file = "pcadapt_HF_trasp_definitivo", sep = " ")

library(pcadapt)
library(data.table)
```

```
pool.data <- fread("pcadapt_HF_trasp2")

filename <- read.pcadapt(pool.data, type = "pool")

res <- pcadapt(filename, min.maf=0.01)
summary(res)
pvalues <- res$pvalues


plot (res, option="manhattan")
plot (res, option="qqplot")

library(qvalue)
qval <- qvalue(res$pvalues)$qvalues
alpha <- 0.05
outliers <- which(qval < alpha)
length(outliers)


write.matrix(outliers, file = "index_cov_30_150_m0.05_a0.05", sep = "")
write.matrix(qval, file = "QVALUES_cov_30_150_m0.05_a0.05", sep = "")
write.matrix(pvalues, file = "PVALUES_cov_30_150_m0.05_a0.05", sep = "")

write.matrix(res$pvalues, file = "PVALUES_2", sep = "")

library(qqman)
manhattan(Manhattanpostapvalues)

manhattan(gwasResults,suggestiveline=FALSE, genomewideline=FALSE)


gwasResults2 <- read.delim("ARCHIVO_PA_MANHATTAN_PVALUES")
gwasResults2 <- read.delim("qvalues_35_150_HF")

manhattan(gwasResults2,suggestiveline=FALSE,genomewideline=1,ylab="-log10(q-value)")
```

7.3 Frecuencias alélicas en las inmediaciones del miR156.

CM024719.1	44026417	52 :53 :0 :0 :0 :0	0 :134 :0 :0 :0 :0
CM024719.1	44026422	0 :50 :52 :0 :0 :0	0 :0 :135 :0 :0 :0
CM024719.1	44026450	26 :65 :0 :0 :0 :0	32 :103 :0 :0 :0 :0
CM024719.1	44026451	87 :6 :0 :0 :0 :0	:40 :0 :0 :0 :0
CM024719.1	44026490	0 :40 :66 :0 :0 :0	0 :0 :135 :0 :0 :0
CM024719.1	44026515	22 :0 :0 :99 :0 :0	35 :0 :0 :97 :0 :0
CM024719.1	44026549	0 :81 :29 :0 :0 :0	0 :32 :78 :0 :0 :0
CM024719.1	44026748	26 :0 :0 :32 :0 :0	78 :0 :0 :28 :0 :0
CM024719.1	44026756	0 :52 :47 :0 :0 :0	0 :34 :71 :0 :0 :0
CM024719.1	44026772	46 :0 :0 :62 :0 :0	4 :0 :0 :110 :0 :0
CM024719.1	44026811	0 :37 :43 :0 :0 :0	0 :104 :0 :0 :0 :0
CM024719.1	44026841	0 :19 :83 :0 :0 :0	0 :37 :78 :0 :0 :0
CM024719.1	44026883	0 :46 :62 :0 :0 :0	0 :0 :124 :0 :0 :0
CM024719.1	44026885	3 :0 :0 :109 :0 :0	38 :0 :0 :88 :0 :0
CM024719.1	44026889	47 :0 :0 :66 :0 :0	0 :0 :0 :129 :0 :0
CM024719.1	44026892	0 :47 :66 :0 :0 :0	0 :0 :129 :0 :0 :0

Figura 7: En esta figura se observa la distribución de frecuencias alélicas en dos pools de individuos. Uno de los pools muestra una distribución equilibrada, con un 50% de los reads asignados a cada una de las variantes alélicas. En cambio, el otro pool presenta una distribución con el 100% de los reads correspondientes a una sola variante alélica.