Universidad de Concepción
Dirección de postgrado
Facultad de Ciencias Forestales- Programa de Magister en Ciencias Forestales

Identificación y validación de *Single Nucleotide Polymorphism* (SNPs) distribuidos en el genoma de *Eucalyptus globulus*

Tesis para optar al grado de Magister en Ciencias Forestales

NICOLE ANDREA MUNNIER GONZÁLEZ
CONCEPCIÓN – CHILE
2015

Profesora Guía: Sofía Valenzuela Águila
Dpto. de Silvicultura, Facultad de Ciencias Forestales
Universidad de Concepción

Identificación y validación de Single Nucleotide Polymorphism (SNPs) distribuidos en el genoma de *Eucalyptus globulus*

Comisión Evaluadora:

Sofía Valenzuela Águila (Profesor guía)

Bioquímico; Dr. _____

Regis Teixeira Mendonça (Profesor co-guía)

Ingeniero Químico; Dr _____

Claudio Balocchi Leonelli (Comisión evaluación)

Ingeniero Forestal, Dr. _____

David Neale (Comisión evaluación)

Forest Genetics; Ph.D. _____

Director de Postgrado:

Regis Teixeira Mendonça

Ingeniero Químico, Dr. _____

Decano Facultad de Ciencias Forestales:

Manuel Sánchez Olate.

Ingeniero Forestal, Dr. _____

Dedicado a mi familia,

porque siempre han estado ahí …

AGRADECIMIENTOS

Quisiera agradecer en primera instancia a mi familia, a quienes dedico este gran logro, por haberme apoyado en la decisión de estudiar este magister, por guiarme siempre, por acompañarme en el camino y por motivarme cada vez que salía un inconveniente a través del desarrollo de la tesis.

Agradecer a mi profesora guía, la Dra. Sofía Valenzuela por creer en mí y por darme la posibilidad de desarrollarme íntegramente a lo largo de todo el magister, en el cual además me otorgó la posibilidad de viajar y hacer una pasantía en UC Davis, CA. Esto me permitió ampliar mis conocimientos, desarrollándome como una profesional integral, conociendo otros laboratorios con distintas formas de trabajar y conociendo profesionales de otras culturas, lo que además me aportó de gran forma a un crecimiento personal, quisiera destacar a Pedro Martínez y John Lietchy quienes contaron con una gran paciencia para explicarme y ayudarme en el proceso incluso a distancia, a Irina Calic, una gran amiga del laboratorio y especialmente a David Neale, quien me recibió dentro de "Neale Lab" para poder desarrollar parte de mi tesis.

No puedo dejar de agradecer al personal profesional, estudiantil y académico del Centro de Biotecnología de la Universidad de Concepción, específicamente a los laboratorios de Biología Molecular y Secuenciación, con los que tuve la oportunidad de compartir además de experiencias laborales, personales, destacando a Carlos, Victoria, Catalina, Ricardo, Paula, Andrea, Pepe, Jany, Valentina, Mariela y Daniel, por su cariño, apoyo, ayuda y preocupación durante todo este tiempo.

Finalmente, quisiera agradecer al proyecto FONDEF d10i11221 EUCACHIP, que permitió el financiamiento de la investigación.

ABREVIATURAS

AFLP: Amplied Fragment Length Polymorphism, Amplificación de Fragmentos Polimórficos

*Ape*KI: *Aeropyrum pernix* K1

ADN: Ácido Desoxirribonucleico

CRoPS: Complexity Reduction of Polymorphic Sequences, Reducción de Complejidad de Secuencias Polimórficas

ER: Enzima de Restricción

EST: Expressed Sequence Tag, Secuencias Expresadas

GBS: Genotyping by Sequencing, Genotipificación por Secuenciación

InDels: Inserciones/Deleciones

NCBI: National Center for Biotechnology Information, Centro Nacional de Información Biotecnológica

NGS: Next Generation Sequencing, Secuenciación de Próxima Generación

PCR: Polymerase Chain Reaction, Reacción en Cadena de la Polimerasa

PE: Paired-end

RAD-Seq: Restriction-site Associated DNA Sequencing, Sitio de Restricción Asociado a marcadores de ADN

RAPD: Random Amplified Polymorphic DNA, Amplificacion Aleatoria de ADN

RFLP: Restriction Fragment Length Polymorphism, Polimorfismos de Fragmentos de Restricción

RRL: Reduced Representation Libraries, Representación de Librerías Reducidas

SE: Single-end

SNP: Single Nucleotide Polymorphism, Polimorfismos de Nucleótido Simple

SSR: Simple Sequence Repeat – Repeticiones de Secuencias Simples

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

I. RESUMEN

Single Nucleotide Polymorphisms (SNPs) son variaciones de un solo nucleótido que están presentes en al menos el 1% de la población. En este estudio se realizó la identificación de SNPs mediante *Genotyping by Sequencing* (GBS), esta es una metodología de secuenciación que consististe en la utilización de enzimas de restricción para reducir el genoma y adaptadores especiales que permiten la secuenciación múltiple en una reacción. La enzima *Ape*KI fue utilizada para la creación de librerías GBS mediante secuenciación Illumina, donde 1.115.163.456 lecturas fueron secuenciadas entre 506 genotipos de *E. globulus* estudiados. Las librerías fueron procesadas y mapeadas para la posterior identificación de SNPs utilizando un *pipeline* bioinformático desarrollado por el *Laboratorio de Bioinformática* del Centro de Biotecnología de la Universidad de Concepción. Debido a que *E. globulus* no cuenta con un genoma de referencia anotado, se realizó un ensamble *de novo* de la especie, mediante SOAPdenovo2 con 506 librerías SE y 2 librerías PE mediante un *pipeline* desarrollado en el *Neale Lab* de la Universidad de California, Davis. Se encontraron 80.851 SNPs putativos, los que fueron sometidos a diferentes filtros, quedando finalmente 1.044 SNPs polimórficos e informativos. Los SNPs identificados mediante GBS en una población clonal de *E. globulus* podrían ser utilizados en la correlación con características económicas relevantes para la especie, como por ejemplo la producción de celulosa.

II. ABSTRACT

Single Nucleotide Polymorphisms (SNPs) are one nucleotide variation that should be present at least in 1% of the population. In this study, SNPs calling was made by *Genotyping by Sequencing* (GBS). GBS is a method that uses specific restriction enzymes and adapters that allows multi-samples sequencing in a reaction. *Ape*KI RE was used for GBS library preparation with Illumina sequencing and 1.115.163.456 reads were sequenced from 506 *E. globulus* genotypes. The libraries were processing, mapped and SNP called using specific bioinformatic *pipeline* developed by *Laboratorio de Bioinformática* of Centro de Biotecnología of Universidad de Concepción. Due to a lack of *E. globulus* genome reference, a *de novo* assembly it was created by SOAPdenovo2 with 506 single-end libraries and 2 paired-end libraries through a bioinformatic *pipeline* developed by *Neal lab* of University of California, Davis. There were found 80.851 putatitve SNPs, these were filtered and finally we obtained 1.044 informative SNPs. SNPs calling by GBS within a clonal population of *E. globulus* can be useful to identify correlations of economically important traits of this specie, such as cellulose production with a specific genotype.

III. INTRODUCCIÓN

Eucalyptus globulus

E. globulus pertenece a la familia Myrtacea y es una de las especies más plantadas de las 700 que conforman el género Eucalyptus. Esta especie es ampliamente distribuida en España, Portugal, Italia, India, China, Australia y Chile. Su amplia distribución se debe principalmente a su rápido crecimiento, a que se distribuyen en lugares con climas templados y en una amplia gama de suelos, además de su corta edad de rotación (Nesbitt et al. 1994; Poke et al. 2005). La baja resistencia al frío de *E. globulus* es el principal factor que restringe su desarrollo y limita su crecimiento y área de cultivo (Gallino et al. 2007). A pesar de esto, *E. globulus* es una de las especies forestales de madera dura más importante del mundo en cuanto a la producción de pulpa y papel, debido a su alto rendimiento pulpable, y excelente calidad de fibra (Goulao et al. 2011). Debido a estas características *E. globulus* es una especie de alto interés económico, siendo la segunda especie forestal más plantada en el país (23% de las plantaciones forestales al año 2013), después de *Pinus radiata* (Infor 2014). Existen programas de mejoramiento genético de esta especie, utilizando marcadores moleculares (por ejemplo SSR) como herramientas para la evaluación de diversidad genética y genotipificación de las poblaciones (Jones et al. 2002).

Un marcador molecular corresponde a un fragmento de ADN de un organismo, que es fácilmente detectable y cuya herencia puede ser monitoreada (Kumar et al. 2009), basándose fundamentalmente en el análisis de los polimorfismos en secuencias de ADN entre individuos. Se caracterizan por poseer numerosas ventajas por sobre las alternativas convencionales basadas en selección fenotípica, ya que son estables y fáciles de detectar en cualquier tipo de tejido, no son alterados por el medio ambiente, además son independientes de crecimiento, estado de diferenciación y desarrollo (Agarwal et al. 2008). En los últimos años los marcadores moleculares se han utilizado en mapeo de genes de interés, construcción de mapas de ligamiento, selección asistida por marcadores moleculares, genética de poblaciones, estudios filogenéticos y establecer relaciones genéticas entre individuos (Kalia et al. 2011).

Hasta hace pocos años los marcadores más utilizados se clasificaban en dos categorías, aquellos basados en la técnica PCR y los basados en hibridización. Dentro de los primeros se encuentran los RAPDs (amplificación aleatoria del ADN polimórfico), AFLPs (polimorfismo en la longitud de los fragmentos amplificados), SSR (microsatélites o secuencias simples repetidas), entre otros, mientras que, de los marcadores basados en la técnica de hibridización, se utilizaba principalmente los RFLPs (polimorfismo de la longitud de los fragmentos de restricción) (Agarwal et al. 2008). A pesar que hasta hace un tiempo, estos marcadores habían sido útiles en estudios genéticos, presentaban ciertas desventajas en cuanto al uso de tiempo y recursos tanto para su desarrollo como para la evaluación de un gran número de individuos (Gupta et al. 2008). Existe otro tipo de marcador molecular que ha tomado importancia en los últimos tiempos que es el polimorfismo de un solo nucleótido, más conocido como SNP.

Polimorfismo de Nucleótido Simple (*Single Nucleotide Polymorphism*, SNP)

Como su nombre lo indica, los SNPs son variaciones en una secuencia que involucra 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 (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 and Salina 2006). Se caracterizan por ser estables genéticamente y poseer una baja tasa de mutación ($\sim 10^{-8}$) además requieren de una mínima cantidad de muestra para su identificación, sin embargo se necesita una alta densidad de SNPs para obtener un nivel de información adecuada, como la obtenida por SSR. 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 (Lindblad-Toh et al. 2000; Kennedy et al. 2003).

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) (Rao and Gu 2008). Está comprobado que los diferentes tipos de sustituciones no se distribuyen de forma equivalente en el genoma, existiendo un predominio de las transiciones sobre la transversiones en relación con el número esperado al azar (Santos 2011). En un estudio realizado en trigo por Lai et al. (2012) se analizó la frecuencia de cada tipo de sustitución (Tabla 1) siendo la más repetida la transición C – T.

Tabla 1. Frecuencia de los diferentes tipos de SNPs en *Triticum aestivum* (Modificado de Lai et al. 2012).

Cambio de base	Tipo	Frecuencia en el genoma de trigo
A-G	Transición	12.323
C-T	Transición	12.954
A-C	Transversión	3.324
A-T	Transversión	3.345
C-G	Transversión	3.625
G-T	Transversión	3.354

Como se mencionó anteriormente, los SNPs se distribuyen a lo largo de todo el genoma, incluyendo regiones codificantes (cSNP), regiones intrónicas (iSNP), regiones reguladoras (rSNP) y regiones intergenómicas (gSNP). Los SNPs se pueden clasificar en “no sinónimos” (nsSNP), 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 et al. 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 la actividad transcripcional del gen (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” (sSNP) o silenciosos, los cuales no alteran al aminoácido, sin embargo se ha descrito que algunos de

estos polimorfismos pueden tener consecuencias funcionales por algún tipo de mecanismo desconocido (Khlestina and Salina 2006).

Según Hendre et al. (2011) en *E. camaldulensis*, existen 1,3 veces más SNPs en zonas intrónicas que en exones, ocurriendo con una frecuencia de 1/65 pb en intrones y 1/108 pb en exones. Estos resultados coinciden con los obtenidos por Kulheim et al. (2009), en un estudio realizado con 23 genes secuenciados para 4 especies del género Eucalyptus *(globulus, nitens, camaldulensis, loxophleba)* (Tabla 2). Los autores reportaron que la distribución de los SNPs ocurre 1,5 veces más en intrones que exones. Además, observaron que la frecuencia de los SNPs en *E. nitens* es 1 cada 33 pb, en *E. globulus* 1 cada 31 pb y 1 cada 17 pb para *E. loxophleba*, siendo *E. camaldulensis* una de las especies con mayor frecuencia de SNPs en donde es posible encontrar 1 polimorfismo de un solo nucleótido cada 16 bases, estableciéndolo como una de las especies forestales con mayor cantidad de SNPs descritas hasta ese momento (2009). Considerando esto Thumma et al. (2012) posteriormente trabajaron en la correlación de los SNPs con la expresión diferencial de genes implicados en la respuesta al estrés hídrico en *E. camaldulensis*.

Estas variaciones de un nucleótido en el ADN no solo se han identificado en especies animales, vegetales y forestales, sino que también en el genoma humano, lo que tiene una gran importancia biológica, ya que determinan la mayor parte de la variabilidad genética de los individuos (Frazer et al. 2009). Esto representa una gran importancia en el avance de las nuevas tecnologías médicas, así como la identificación y etiología de diversas enfermedades, pues la susceptibilidad o resistencia individual a distintas enfermedades radica principalmente en los SNPs y en menor grado a inserciones, deleciones, secuencias repetitivas y/o rearreglos cromosómicos (Harismendy et al. 2009). Esto se debe a que el ADN del genoma humano está expuesto a múltiples modificaciones que pueden dar como resultado la aparición de enfermedades. En humanos se estima que estos polimorfismos ocurren cada 1.000 pares de bases, ya que el 90% de la diversidad fenotípica humana proviene de los SNPs. Estos corresponden a la mínima alteración que puede experimentar la secuencia de ADN de un individuo y es el más común de los polimorfismos (Spalvieri and Rotenberg 2004), identificándose más de 3,1 millones de SNPs en el genoma humano (Frazer et al. 2007).

Tabla 2. Número de SNPs encontrados en exones e intrones para 4 especies de Eucalyptus. Los "alelos comunes" están por sobre el 10% y los "alelos raros" son inferiores al 10% (Modificado de Kulheim et al. 2009).

Fragmento de ADN	**Tipo de alelo**	***Eucalyptus globulus***	***Eucalyptus nitens***	***Eucalyptus camaldulensis***	***Eucalyptus loxophleba***	***Total***
Exones	sinónimos comunes	96	63	142	136	437
	no sinónimos comunes	82	30	114	97	323
	sinónimos raros	133	175	358	316	982
	no sinónimos raros	174	203	273	238	888
	Total	485	471	887	787	2.630
Intrones	Común	367	344	634	720	2.065
	Raro	626	603	1.510	1.197	3.936
	Total	993	947	2.144	1.917	6.001
Exones + Intrones	Común	545	437	890	953	2.825
	Raro	933	981	2.141	1.751	5.806
	Total	1.478	1.418	3.031	2.704	8.631

Los SNPs son los principales marcadores moleculares utilizados en el mapeo genético y ensayos de diversidad en plantaciones de cultivo debido a su alta abundancia y distribución uniforme en el genoma (Chagne et al. 2008). Además entregan información específica de la historia y estructura de una población, especialmente cuando una cierta cantidad de SNPs, estrechamente separados, se combinan para definir un haplotipo (Rafalski 2002). En el área vegetal, los SNPs se han utilizado en reconstruir *pedigrees*, monitorear tasas de *out-crossing*, control de calidad de pureza genética desde huertos semilleros y en el establecimiento de líneas clonales. Además muchos estudios de genómica funcional se han desarrollado en base a los SNPs, como en genes regulatorios, transcritos y Expressed Sequence Tags (ESTs). Cabe destacar que han sido fundamentales en el análisis de genes y descubrimiento de la base genética molecular de importantes características agronómico-industriales, como por ejemplo en el arroz, se ha posibilitado la asociación de SNPs con la calidad de cocción, procesamiento y aroma del grano (Larkin and Park 2003). Este tipo de marcadores también ha sido

empleado en estudios evolutivos de conservación genética, genética de poblaciones y filogenia, identificación de variedad, análisis funcionales, análisis de asociación en el mejoramiento genético vegetal, entre otros (Morin et al. 2004). Las ventajas de usar estos marcadores, es que permiten realizar nuevas investigaciones, asociándose con ciertas características relevantes, lo que va reduciendo los tiempos de análisis y por ende, la disminución de los costos para el descubrimiento de genes de interés (Thavamanikumar et al. 2011; Kumar et al. 2012).

Los SNPs son identificados mediante el alineamiento de las secuencias contra un genoma de referencia. Esta se puede llevar a cabo fácilmente en genomas de plantas pequeñas con genomas de referencia disponibles, como en el caso de arroz (Yamamoto et al. 2010) o *Arabidopsis* (Ossowski et al. 2008), aunque la identificación de SNPs en genomas complejos sin un genoma de referencia como el trigo, cebada, avena y porotos, se puede lograr mediante las nuevas técnicas de secuenciación de nueva generación (Next Generation Sequencing, NGS) o incluso, usando datos transcriptómicos o representaciones de complejidad reducidas, las que son facilitadas por las técnicas de secuenciación masiva (Kumar et al. 2012). Además, los SNPs se pueden identificar mediante comparaciones de secuencias con genomas de referencia de especies con un alto porcentaje de identidad, como podría darse en el caso de *E. globulus* (que no cuenta con un genoma de referencia secuenciado) y que presenta un 75% de identidad y un 1.5% de divergencia con la especie del mismo género *E. grandis*, según Myburg et al. (2011).

La identificación de SNPs en especies que no cuentan con un genoma de referencia descrito, podría darse mediante la construcción de una referencia, esto es un ensamble *de novo*, utilizando secuencias cortas provenientes de tecnologías NGS. Un ensamble *de novo* es una estructura jerárquica de datos, en el cual las secuencias individuales se agrupan y fusionan para formar fragmentos contiguos denominados *contigs*, los que comparten la misma secuencia de nucleótidos. Posteriormente los *contigs* son ensamblados en *scaffolds* (o *supercontigs*) los que definen el orden y la orientación de los *contigs* y el tamaño de los *gaps* entre los *contigs* (Miller et al. 2010). Crear un ensamble *de novo* es uno de los grandes desafíos de las técnicas de secuenciación masiva (y del posterior análisis de datos), especialmente para aquellas especies con genomas complejos como los eucariontes (Zhang et

al.2011), lo que se debe principalmente a que estos presentan altos niveles de ploidía, presente en más de un 80% de las especies vegetales, además de la alta heterocigosidad y las familias de genes y pseudogenes derivadas de eventos como duplicación y transposones (Schatz et al. 2012). Es por esto que existe la posibilidad de que el polimorfismo detectado no sea un SNP propiamente tal, ya que elementos repetitivos y parálogos pueden generar ambigüedades en la calidad de la secuenciación y por ende, en la identificación de los SNPs (Treangen and Salzberg 2012).

Existe un gran número de especies vegetales en las cuales ya se han reportado un gran número de SNPs: más de 4.900 SNPs fueron validados en maíz (*Zea mays*) en más de 2.400 genes secuenciados mediante tecnología 454 (Barbazuk et al. 2007); por su parte, en arroz (*Oryza sativa*) se encontraron 384.341 posibles SNPs (Feltus et al. 2004); 9.448 SNPs candidatos en avena (*Avena sativa*) (Oliver et al. 2011). Por su parte, los estudios en búsqueda de SNPs en especies forestales están aumentando: en *Pinus taeda* se identificaron SNPs relacionados con características fisiológicas y propiedades de la madera (González-Martínez et al. 2007). Con respecto al género Eucalyptus, Thumma et al. (2005) reportaron la primera asociación entre SNPs y el ángulo microfibrilar en *E. globulus*. En *E. grandis* fueron detectados sobre 36.000 SNPs putativos, de los cuales 23.742 están relacionados con evolución genética y fue uno de los primeros estudios de *genome-wide* en diversidad nucleotídica en plantas no-modelo (Novaes et al. 2008).

Las nuevas técnicas de secuenciación masiva (NGS) han otorgado un rápido desarrollo en la detección de SNPs dentro de los genomas, aplicándose en varias especies vegetales, permitiendo además, la re-secuenciación completa de un genoma a un costo cada vez menor, esto sumado a la reducción de los costos de genotipificación, facilita la aplicación de la selección genómica, atrayendo la atención de mejoradores de especies agrícolas y plantas perennes (Wong and Bernardo 2008).

Secuenciación masiva

La secuenciación del ADN consiste en determinar el orden de las bases A, C, G y T en un fragmento de ADN (Jimenez-Escrig et al. 2012). Dentro de las tecnologías de secuenciación de

primera generación, se encuentra el método descrito por (Sanger et al. 1977), con el cual se logró la secuenciación del genoma humano, el que tenía una capacidad de lectura de aproximadamente 80 pb. Los avances de la tecnología, en conjunto con la gran necesidad de conocer el orden de las bases en el material genético hereditario y la búsqueda de abaratar costos, ha permitido el desarrollo de nuevas tecnologías de secuenciación masiva denominadas NGS, las que se basan en generar múltiples secuencias (más cortas que las generadas por Sanger). Por ejemplo Illumina genera secuencias de entre 35 - 150 pb aproximadamente (Desai and Jere 2012), generando millones de secuencias de forma paralela, lo que hace posible la secuenciación del genoma humano completo de un individuo en el mismo tiempo y a un costo mucho menor que la secuenciación de Sanger (Haas et al. 2011). Dentro de las NGS las que se caracterizan por su capacidad de entregar una gran cantidad de información a un menor costo, existiendo diferentes plataformas desarrolladas para la secuenciación (Metzker 2010). En el 2002 el costo de secuenciación era de US $9.000 por un millón de bases, pero gracias a la aparición de estas tecnologías, en el 2012 se redujo a US $0,1 (Hayden 2013). Tecnologías como HiSeq 2000 de Illumina, son capaces de generar hasta casi 100 Gpb por día e incluso secuenciar 3 Gpb de genoma por menos de US$ 10.000 (Schatz et al. 2012).

Dentro de las NGS, se encuentran Roche/454, Illumina/Solexa y Applied Biosystem/SOLID las que se diferencian de Sanger, por no requerir vectores para realizar la secuenciación, además las NGS generan un número mayor de secuencias pero de un tamaño mucho menor y a un menor costo. Roche se basa en pirosecuenciación e Illumina utiliza el método de terminador reversible, mientras que SOLID utiliza secuenciación por ligación (Nowrousian 2010), a su vez Roche utiliza el método de emulsión por PCR para amplificar las moléculas de ADN, Illumina, también conocida como secuenciación por puente, utiliza la amplificación en fase sólida diferenciándose ambas en las tecnologías de detección de fluorescencia (Metzker 2010).

Identificación de SNPs

Actualmente, dentro de las nuevas estrategias para identificación de SNPs, se encuentran la creación de bibliotecas de genoma reducido de ADN, para luego secuenciarlas y realizar los análisis bioinformáticos respectivos. Una biblioteca o librería genómica, es una compilación de secuencias de ADN, en la que se espera que esté incluido el mayor numero de fragmentos

posibles de un genoma (Alberts et al. 1994). Para realizar la digestión, se utilizan enzimas de restricción que generan cortes en regiones especificas del genoma, cuyo objetivo es eliminar regiones altamente repetidas y poco informativas, debido a que existen enzimas que son sensibles a las regiones metiladas y no son capaces de cortar en estos sitios (Davey et al. 2011). Los genomas complejos, presentan muchas secuencias repetidas, por ejemplo aproximadamente la mitad del genoma humano corresponde a repeticiones, estas surgen a partir de los mecanismos biológicos de replicación, donde surgen copias extras de un fragmento del genoma, sin embargo, algunas de estas repeticiones podrían ser funcionales, actuando incluso como marcadores moleculares (You et al. 2011).

Existen una serie de enfoques desarrollados en la creación de librerías genómicas que se basan en estrategias de reducción de complejidad, utilizando NGS con el objetivo de reducir el costo y simplificar la identificación de los SNPs, como lo son RNA-Seq, CRoPS (complexity reduction of polymorphic sequences), RAD-Seq (restriction-site-associated DNA sequencing) y GBS (*Genotyping by Sequencing*) (Kumar et al. 2012).

Genotipificación por Secuenciación (*Genotyping by Sequencing*, GBS)

GBS es un método de creación de librerías genómicas, que permite secuenciar un gran número de individuos de forma simultánea, esta técnica se basa en la reducción de la complejidad del genoma, en la que se utilizan enzima de restricción (ER) sensibles a metilación para digerir el genoma. Se diferencia de las otras metodologías previamente mencionadas (RAD-seq y RRL) básicamente en que GBS no realiza una selección de fragmentos por tamaño (los más grandes) antes de la secuenciación. Esta técnica presenta múltiples ventajas, pues al basarse en técnicas NGS, requiere una menor cantidad de ADN, además la secuenciación es bastante rápida, simple y reproducible. La creación de librerías genómicas mediante este método presenta un bajo costo de secuenciación por muestra (Davey et al. 2011; Lindner 2012).

Esta técnica requiere de dos tipos de adaptadores, los primeros corresponden a oligonucleótidos que son complementarios a la secuencia que se encuentra unido a la celda de flujo del secuenciador y los segundos son denominados *barcodes*, que corresponden a adaptadores que permiten la identificación de las muestras. Debido a que cada individuo presenta un *barcode*

diferente, es posible secuenciar hasta 384 muestras de forma simultánea en una corrida. La amplificación de los fragmentos con los adaptadores ocurre mediante PCR y se realiza una purificación para eliminar restos de adaptadores y reactivos, los que son evaluados mediante una electroforesis en gel de agarosa y un espectrofotómetro PicroGreen antes de la secuenciación. En la Figura 1, se observa un diagrama de la técnica GBS (Elshire et al. 2011).

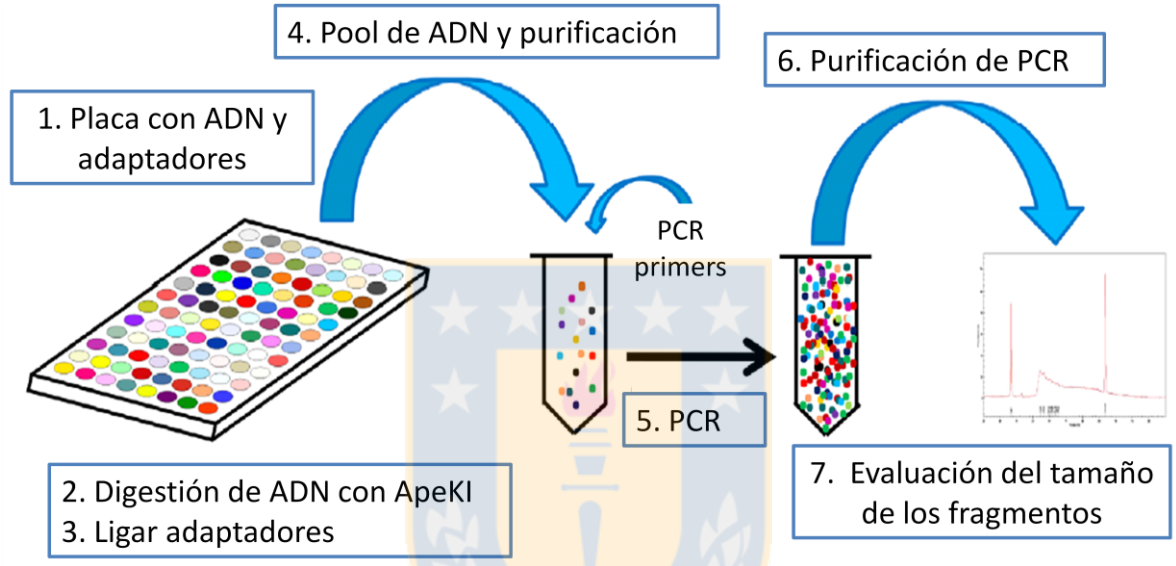

Figura 1. Esquema de la técnica Genotipificación por Secuenciación (Modificado de Elshire et al. 2011).

Es adecuada para estudios de población, filogenética, caracterización de germoplasma, selección genómica, mejoramiento y mapeo de diversos organismos, ya que el principal objetivo de la técnica es la identificación y genotipificación de SNPs. Esta puede ser generalizada a cualquier especie, ya ha sido descrita tanto especies de interés comercial como el maíz y la cebada (Elshire et al. 2011) como para especies forestales con genomas por sobre los 20.000 Mpb, como *Pinus contorta* y *P. glauca*, donde se encontraron sobre 17 mil SNPs para ambas especies (Chen et al. 2013).

Análisis bioinformático para la identificación de SNPs

Una vez obtenidos los datos de la secuenciación, la bioinformática es el área encargada de identificar las herramientas necesarias para identificar los SNPs. Si bien no existe una metodología modelo (*pipeline*) establecida con software específico para el análisis de las secuencias, lo que se debe a que cada análisis dependerá de las especies, el material utilizado, el método de construcción de librería, y el tipo de secuenciación utilizadas. La bioinformática utiliza una estrategia que se basa en tres pasos básicos: (1) alineamiento de las lecturas de diferentes genotipos contra un genoma de referencia (mapeo); (2) generación de una secuencia consenso para cada genotipo; y (3) la identificación de los SNPs bajo la comparación con un genoma de referencia (Azam et al. 2012).

Para el caso de especies que no cuentan con un genoma de referencia, como *E. globulus*, es necesario crear una referencia. Existen numerosos software para realizar ensambles *de novo*, tales como *Velvet* (Zerbino and Birney 2008), *MaSurCa* (Zimin et al. 2013), *SPAdes* (Bankevich et al. 2012), *SOAPdenovo* (Li 2009), entre otros. Estos difieren en el tipo de ácido nucleico (ADN o ARN) que son capaces de ensamblar, además de la estrategia con la que las muestras fueron secuenciadas, y el objetivo para el cual se está creando la referencia. De acuerdo a esto, cada ensamblador trabaja con diferentes algoritmos para ensamblar las secuencias. *SOAPdenovo* (Short Oligonuclotide Analysis Package) ha sido utilizado en la creación de varios genomas de referencia *de novo*, pero tiene algunas desventajas que fueron mejoradas en la siguiente versión: *SOAPdenovo2* (Luo et al. 2012). Este último utiliza un algoritmo diseñado para reducir el uso de memoria en la construcción gráfica, mejorar y disminuir los *gap* y en general, optimizar el ensamble para genomas complejos. *SOAPdenovo2* se basa en seis etapas: corrección de errores, construcción de gráfico de Bruijn, ensamble de *contigs*, mapeo de lecturas *paired-end* (PE), construcción de *scaffolds* y disminución de *gaps* (Luo et al. 2012).

Al igual que cuando se cuenta con un genoma de referencia, alinear las secuencias con la referencia es el primer paso, y el más lento, en la mayoría de los *pipelines* utilizados para genómica comparativa, expresión diferencial, entre otros. Muchos software de alineamiento utilizan un índice de genoma para disminuir la lista de ubicaciones de un posible alineamiento. *Bowtie2* extiende este índice de genoma para además permitir alinear los *gap*, y se comprobó

que es mejor y más rápido en comparación con otros software como *Bowtie2* (Langmead and Salzberg 2012), *BWA* (Li and Durbin 2009) y *SOAP2* (Li et al. 2009b).

Para identificar los SNPs luego del mapeo, *SAMtools* (Li et al. 2009a) es un software que soporta *reads* generados por diferentes plataformas de secuenciación. Es capaz de convertir diferentes formatos (*.sam/.bam*), ordenar datos (*sorted*), unir alineamientos, eliminar artefactos de PCR e identificar SNPs e InDels, entre otros.

El correcto análisis bioinformático de los datos, esto es una buena utilización de las estrategias de secuenciación, y la elección de los parámetros adecuados para los datos estudiados, permiten una identificación de SNPs confiable para potenciales asociaciones con características de interés económico, como podrían ser los genes involucrados en formación de madera de especies como *E. globulus*. Es por esto que uno de los objetivos de este estudio es la creación de un ensamble *de novo* de la especie, lo podría ser mejor estrategia para el llamado de SNPs, que la utilización de genomas de referencia de otras especies disponibles en bases de datos públicas.

IV. HIPÓTESIS

Existen suficientes SNPs (*Single Nucleotide Polymorphisms*) polimórficos, distribuidos uniformemente en el genoma de *Eucalyptus globulus* cuya identificación y validación tienen un uso potencial para selección genómica.

V. OBJETIVOS

Objetivo general

Identificar SNPs, su distribución y frecuencia en un genoma de referencia parcial de una población de *Eucalyptus globulus* a partir de bibliotecas génicas.

Objetivos específicos

- Establecer librerías génicas de *Eucalyptus globulus* a partir de datos obtenidos mediante la técnica de secuenciación *Genotyping by Sequencing* (GBS).
- Establecer un genoma de referencia parcial *de novo* de *Eucalyptus globulus* a partir de librerías génicas.
- Identificar SNPs y seleccionar aquellos que sean informativos en el genoma de *Eucalyptus globulus* con potencial uso en selección genómica.

VI. METODOLOGÍA

Material vegetal

Se utilizaron muestras de corteza de 600 genotipos diferentes de *Eucalyptus globulus*, provenientes de ensayos forestales y plantaciones operacionales ubicadas en la Región del Biobío, facilitadas por la empresa *Forestal Mininco S.A.* y *Bioforest S.A.* La estructura familiar de los individuos provenientes de *Bioforest S.A.* está constituida por 29 familias, provenientes de 18 padres diferentes. En cuanto a la estructura familiar de *Mininco S.A.* está constituida por 64 familias, provenientes de 34 padres diferentes.

Biblioteca genómica

Extracción de ADN

La extracción de ADN se realizó mediante el kit comercial DNeasy Plant Mini (Qiagen). Las muestras de ADN fueron cuantificadas en el equipo Agilent Bioanalyzer 2200 para determinar la concentración y verificar la calidad del ADN, de acuerdo al protocolo del fabricante.

Digestión enzimática

El protocolo de la técnica GBS exige un análisis de digestión enzimática de al menos un 10% de las muestras a enviar mediante alguna enzima de restricción, con el objetivo de asegurarse que la enzima escogida es capaz de digerir el ADN en fragmentos más pequeños para la posterior secuenciación y creación de las librerías genómicas. La digestión enzimática se realizó en 60 muestras de aproximadamente 10 ng/ul de ADN genómico, utilizando la enzima de restricción *Hind*III (New England Biolabs) de acuerdo al protocolo de la enzima.

Construcción de librerías genómica

Las librerías GBS fueron preparadas y analizadas en el Instituto de Diversidad Genómica (IGD) de la Universidad de Cornell (Ithaca, New York) de acuerdo al protocolo descrito por

Elshire et al. (2011), utilizando la enzima *Ape*KI para la digestión y la generación de las librerías utilizando 96 *barcodes* únicos para la secuenciación con la plataforma NGS Illumina/ HiSeq 2000/2500.

Pretratamiento de librerías genómicas

Previo al tratamiento bioinformático, es necesario realizar un pretratamiento para eliminar regiones que contengan bases con baja o mala calidad, bases altamente repetidas (elementos transponibles) y/o adaptadores propios de la secuenciación, ya que estas interfieren con el ensamblaje y/o mapeo.

Eliminación de Barcodes

Se utilizó el software *CutAdapt* (Martin 2011) para cortar entre 4 a 8 bases de adaptador (específico para cada librería) indicando que el *barcode* está ubicado en el extremo 5' de cada lectura y no entremedio de los fragmentos secuenciados.

Control de calidad

El control de calidad se realizó mediante el software *Sickle* (Joshi 2012), restringiendo el control de calidad con un largo mínimo de 55 pb y un *phred score* de 33 lo que se refiere a una aceptación de un 0,000631% de error.

Ensamble *de novo*

Antes de realizar el ensamble *de novo* de todos los genotipos, se realizó una prueba con un subgrupo de 20 librerías con cuatro valores de *kmer* diferentes para el software *SOAPdenovo2* (Luo et al. 2012): k=33, k=41, k=45, k=55 y cuatro valores *kmer*, k=33, k=41, k=45, k=55 para el software ABySS (Simpson et al. 2009).

Se descargaron 2 librerías SRA (*Sequence Read Archive*, archivo de lectura de secuencias) de un genotipo de *E.globulus* disponibles en la base de datos de NCBI (http://www.ncbi.nlm.nih.gov/sra/SRX116786) para incrementar la información y realizar una referencia más robusta. Estas librerías genómicas de tipo *paired-end* secuenciadas por la tecnología Illumina (Illumina Genome Analyzer II), con lecturas de 300 pb, fueron procesadas

por el mismo pretratamiento utilizado anteriormente, con los mismos parámetros para ambos software, antes de ser ensambladas por *SOAPdenovo2* y *ABySS*.

Scaffolding

El *scaffolding* se realizó con el software *Sspace* (Hunt et al. 2014) utilizando el input *scafSeq* obtenido a partir del ensamble realizado por *SOAPdenovo2* y *ABySS*. El software exige un archivo con las librerías *paired-end* a utilizar y ciertas características de ellas. Mediante el comando –u se indexó un archivo con todas las librerías *single-end* (SE) concatenadas previamente (-x 1 -m 32 -o 20 -t 0 -k 5 -a 0.70 -n 15 -p 0 -v 0).

Alineamiento del ensamble con la referencia de *E. grandis*

Con el objetivo de conocer la cobertura que tiene la referencia creada mediante el ensamble *de novo*, se realizó un alineamiento entre este y la referencia de *E. grandis* descargada de pythozome (http://www.phytozome.net/eucalyptus.php). Para esto se utilizó el algoritmo *bwasw* del software *BWA*, ya que es más conveniente en este caso, para realizar alineamientos locales con *reads* largos.

Mapeo contra la nueva referencia de *E. globulus*

Antes de realizar el mapeo contra el genoma de referencia recientemente creado, las muestras seleccionadas (aquellas que pasaron el filtro de calidad) fueron separadas de acuerdo a las poblaciones a las que pertenecen.

Para realizar el mapeo de cada librería con la nueva referencia de *E. globulus* ensamblada previamente mediante *SOAPdenovo2*, se utilizó *Bowtie2*, utilizando un *phred score* de 33 para el mapeo. Los datos del mapeo se observan utilizando *Qualimap*, pero previamente las librerías mapeadas deben ser transformadas a formato *sorted.bam*.

Llamado de SNPs

Una vez que cada genotipo fue mapeado con el genoma de referencia creado de *E. globulus*, se utilizaron diferentes herramientas de *SAMTools*, para identificar las variantes. Mediante la herramienta *–mpileup* se realizó el llamado de SNPs. El primer filtro realizado con *vcfutils.pl* se basó a los datos de profundidad de mapeo (*-D3*). Posteriormente mediante el software

vcftools (Danecek et al. 2011) se realizó un segundo filtro eliminando los InDels (--*indels*), y un tercer filtro, de aquellas variantes que no están presentes en el 1% de cada población (MAF >0,01) dejando solo aquellos SNPs informativos.

VII. RESULTADOS

Extracción de ADN y digestión enzimática

De las 600 extracciones de ADN de los diferentes genotipos provenientes de *Forestal Mininco* S.A.y *Bioforest* S.A., solo fueron seleccionadas aquellas que presentaban una concentración mayor a los 10 ng/ul. En la Figura 2 se muestra un subgrupo de 7 muestras aleatorias cuantificadas mediante Bioanalyzer, donde se observa una baja cantidad de ADN degradado, los fragmentos de ADN está íntegros de acuerdo al estándar de peso molecular y presentan aproximadamente entre 15.000 y 48.500 pb. De las 600 extracciones de ADN, 30 se descartaron debido a que no superaban los 10 ng/ul de ADN exigidos por protocolo para la creación de librerías genómicas GBS. Una gran cantidad de muestras no presentaban o tenían muy poco felógeno visible, por lo que se solicitó un nuevo muestreo de estos genotipos, aun así, en algunos casos no se logró la concentración requerida. Para algunos casos, la extracción de ADN se realizó a partir de tejido floemático y para otros, los genotipos fueron eliminados.

La digestión enzimática realizada con la enzima de restricción *Hind*III, generó un *smear* indicando que los fragmentos de ADN presentan entre 23.000 y 500 pb.

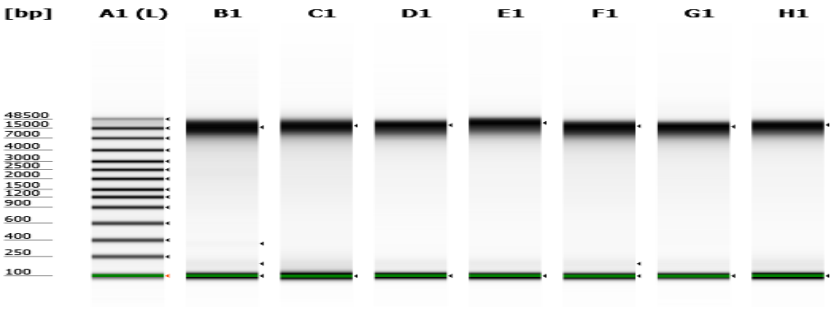

Figura 2. Resultado de cuantificación de algunas de las muestras obtenido mediante equipo Bioanalyzer.

Análisis bioinformático

Como producto de la secuenciación de las muestras, se recibieron 6 archivos diferentes (95 muestras cada archivo, correspondientes a cada placa de secuenciación), por lo que antes de realizar el análisis de las secuencias, se procedió a separar las bibliotecas de cada genotipo de acuerdo a los *barcodes* que marcan cada *read* secuenciado, hasta obtener los 570 archivos en formato *fastq* correspondientes a cada librería/genotipo. Las librerías son producto de la secuenciación Illumina, son de tipo SE y de un tamaño de 101 pb.

Los 570 genotipos fueron clasificados entre *fail* y *pass* de acuerdo a la calidad de secuenciación. El 11% de ellas fueron calificadas como *fail*, ya que presentaban una cantidad de *reads* menor al 10% del promedio de la línea de secuenciación, debido a esto fueron eliminadas puesto que no son útiles para crear un ensamble *de novo*, mapeo y llamado de SNPs, lo que dejó un total de 506 librerías.

La secuenciación de estas 506 muestras generó librerías génicas de entre 143.157 y 15.761.239 *reads*, sumando un total de 1.115.163.456 secuencias, dentro de las cuales *Bioforest* suma 541.591.133 *reads* y *Mininco* 573.572.323 *reads*.

Pretratamiento

Una vez que los barcodes fueron removidos, se calculó el largo promedio de todas las librerías analizadas, el que corresponde a 95 pb, eliminando en promedio 6 pb de las secuencias originales. Posteriormente, se realizó el control de calidad con *Sickle*, donde el largo promedio de las librerías disminuyó a 88 pb (Tabla 3). Luego del pretratamiento, los *reads* de las librerías disminuyeron su largo en promedio 13 pb. Por su parte, el número de lecturas disminuyó de 1.115.163.456 a 1.036.297.707 *reads*, por lo que durante el control de calidad se eliminó un 8% de *reads*.

Tabla 3. Resumen comparativo del promedio del número de *reads* antes y después del pretratamiento de las secuencias.

	Librerías en bruto		**Librerías post-tratamiento**	
Mínimo y máximo de secuencias	Min: 143.157	Max: 15.761.239	Min: 133,705	Max: 13.843.208
Número total de secuencias	1.115.163.456		1.036.297.707	
Promedio de secuencias	2.205.991		1.672.246	
Promedio largo secuencias	101 pb		88 pb	

Ensamble *de novo* de *E. globulus*

Se descargaron dos librerías *paired-end* de un genotipo de *E. globulus* de la base de datos de *National Center for Biotechnology Information* (NCBI) y fueron pretratados con *Sickle* utilizando los mismos parámetros aplicados anteriormente en las librerías de *E. globulus* secuenciadas. La prueba del ensamble y posterior *scaffolding* de estas dos librerías PE y un subgrupo de 10 librerías SE con *SOAPdenovo2* se observa en la Tabla 4 y con *ABySS* en la Tabla 5. En la Figura 3 y 4 además se observa un histograma del número de *contigs* y el valor N50 para cada una de las pruebas utilizando *SOAPdenovo2* y *ABySS* respectivamente.

Tabla 4. Ensamble *de novo* con *SOAPdenovo2* y posterior *scaffolding* utilizando 10 librerías SE y 2 librerías PE de *E. globulus* descargadas de NCBI comparando cuatro valores *kmer*s diferentes.

Software	**Parámetros**	**K33**	**K41**	**K45**	**K55**
Ensamble con *SOAPdenovo2*	*Contigs*	2.065.391	2.137.924	2.109.396	2.254.664
	N50	1.127	1.174	1.181	938
Scaffolding* con *Sspace	*Scaffolds*	712.774	700.599	682.170	959.440
	N50	1.978	2.152	2.152	1.508

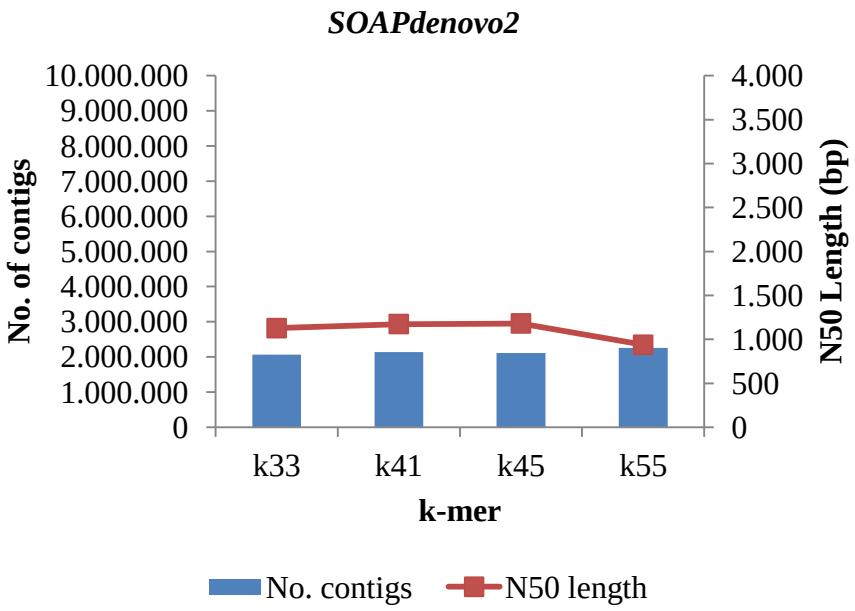

Figura 3. Histograma representativo del número de *contigs* y valor N50 para cada una de las pruebas utilizando *SOAPdenovo2* como ensamblador. Las barras indican el número de *contigs* obtenidos en cada ensamble (eje izquierdo) y las líneas muestran el valor N50 (eje derecho).

Tabla 5. Ensamble *de novo* con *ABySS* y posterior *scaffolding* utilizando 10 librerías SE y 2 librerías PE de *E.globulus* descargadas de NCBI comparando cuatro valores *kmer*s diferentes.

Software	**Parámetros**	**K33**	**K41**	**K45**	**K55**
Ensamble	*Contigs*	9.170.250	5.974.387	4.983.012	3.350.942
con *ABySS*	N50	3.615	3.244	2.982	2.085
Scaffolding	*Scaffolds*	6.724.372	5.233.865	4.221.109	2.527.428
con *Sspace*	N50	512	523	775	1.220

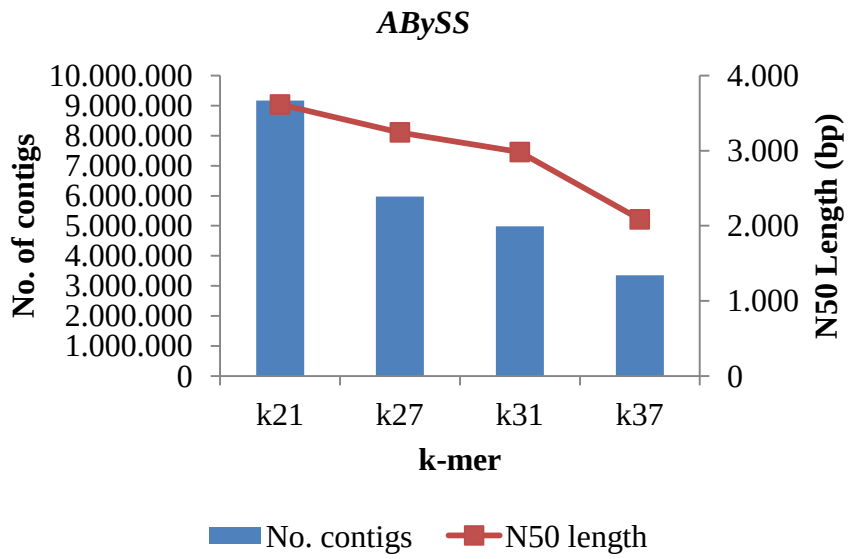

Figura 4. Histograma representativo del número de *contigs* y valor N50 para cada una de las pruebas utilizando *ABySS* como ensamblador. Las barras indican el número de *contigs* obtenidos en cada ensamble (eje izquierdo) y las líneas muestran el valor N50 (eje derecho).

Estas pruebas fueron realizadas para seleccionar el valor de *kmer* que mejor se ajusta a los datos de *E.globulus* estudiados, buscando un menor número de *contigs* más largos y con un mayor N50 (valor estadístico de un grupo de *contigs*/*scaffolds*). En las Tablas 4 y 5 se puede observar que además, el software *SOAPdenovo2* genera un menor número de *contigs* que *ABySS*, y a pesar que los valores de N50 obtenidos por *ABySS* son más altos, la relación entre número de *contigs*-N50 es mejor para *SOAPdenovo2*.

Es por esta razón que se realizó el ensamble de todas las muestras con *SOAPdenovo2*: dos librerías PE provenientes de NCBI sumadas a las 506 librerías *single-end*, ocupando los mismos parámetros previamente utilizados y seleccionando el valor *k*=45, ya que este es el que entregaba un mejor ensamble.

El ensamble final con *SOAPdenovo2* generó 1.544.370 *scaffolds* y *singleton*, los detalles del ensamble se observan en la Tabla 6.

Tabla 6. Resumen de datos del ensamble *de novo* obtenidos con *SOAPdenovo2*.

Estadísticas	Ensamble *de novo* con *SOAPdenovo2*
Contigs* y *singleton	1.544.370
Suma (pb)	643.110.452
***Gaps* (pb)**	22.656.321
***Contig* mas largo (pb)**	42.278
***Contig* mas pequeño (pb)**	100
N50	816
promedio de tamaño del *contig* (pb)	416

Scaffolding

Utilizando el archivo proveniente desde *SOAPdenovo2* con 1.544.370 *contigs*, luego del *scaffolding* con *Sspace*, se obtuvo una referencia con un total de 1.060.306 *scaffolds* (Tabla 7). Este archivo es utilizado posteriormente en el mapeo y búsqueda de SNPs.

Tabla 7. Resumen de estadísticas de *scaffolding* obtenidos con *Sspace*.

	Scaffolding* con *Sspace
Scaffolds	1.060.306
Suma (pb)	651.735.217
***Gaps* (pb)**	20.675.536
***Scaffold* mas largo (pb)**	58.000
***Scaffold* mas pequeño (pb)**	100
N50	1.723
Promedio del tamaño de los *scaffolds* (pb)	614

Alineamiento con *E. grandis*

Un 82 % del ensamble *de novo* de *E. globulus* fue alineado con el genoma de referencia de *E. grandis*, utilizando el algoritmo *bwasw* de *BWA*. El alineamiento presentó una cobertura de 0,86 y la calidad de mapeo de 40,26.

Se mapearon un total de 1.336.641 *scaffolds* del ensamble *de novo* a los 4.942 scaffolds de la referencia de *E. grandis*, siendo el scaffold 8, el que presenta una mayor cantidad de reads mapeados, sin necesariamente, ser el más grande. En la Tabla 8 se muestra la cantidad de *scaffolds* del ensamble *de novo* mapeados a cada uno de los 11 cromosomas más grandes de *E. grandis*.

Tabla 8. Número de *scaffolds* del ensamble *de novo* de *E. globulus* mapeados contra 11 de los 4.952 scaffolds del genoma de *E. grandis*.

Scaffold* genoma de *E. grandis	**Tamaño del *scaffold* de *E. grandis***	**N° de *reads* de *E. globulus* mapeados**
***scaffold*_1**	40.297.282	84.544
***scaffold*_2**	64.237.462	125.259
***scaffold*_3**	80.088.348	146.716
***scaffold*_4**	41.978.404	86.174
***scaffold*_5**	74.731.017	149.030
***scaffold*_6**	53.893.726	97.666
***scaffold*_7**	52.447.651	117.514
***scaffold*_8**	74.330.457	149.579
***scaffold*_9**	39.019.482	77.542
***scaffold*_10**	39.359.118	79.791
***scaffold*_11**	45.510.589	90.049

Mapeo

Para poder realizar el llamado de SNPs, es necesario hacer un mapeo de cada una de las secuencias contra el genoma de referencia, que en este caso, es el ensamble *de novo* creado. Como resultado de este alineamiento, se obtiene un archivo (*.sam*) para cada una de las muestras, que contiene la información del mapeo contra el genoma de referencia. Para observar las estadísticas del mapeo con *Qualimap* (Tabla 9), los archivos con extensión *.sam* fueron transformados a *.sorted.bam*.

Tabla 9. Tabla resumen de las estadísticas de mapeo utilizando la referencia *de novo* de *E. globulus.*

	Ensamble *de novo* *E. globulus*
Referencia (*scaffolds*)	1.060.306
Promedio de *reads* a mapear	1.672.246
Número de *reads* mapeados	862.049
Porcentaje *reads* mapeados	51,4
Duplicación	42,29
Cobertura	0,21
Calidad de mapeo	2,35

Llamado de SNPs

En la identificación de SNP en toda la población de *E. globulus* estudiada (Tabla 10), se obtuvo un archivo con formato *.bcf* mediante la utilización de *SAMTools*. Este archivo, contiene el número de SNPs putativos, por lo que tuvo que ser utilizado el paquete de herramientas *vcftools* para realizar el filtrado de frecuencia alélica y de InDels. Si consideramos que la referencia de *E. globulus* utilizada tiene aproximadamente 651 Mpb, la distribución de SNPs es 1 cada 2,6 Mb. Para la población de *Mininco* se encontró 1 SNP cada 1 Mb, y para de *Bioforest* la distribución de las variantes fue de 1 cada 1,4 Mb.

Tabla 10. Número total de SNPs y los respectivos filtros realizados para los 506 genotipos estudiados.

	Número de SNPs
SNPs totales	2.123.426
Filtro por calidad de mapeo	55.028
Filtro de frecuencia alélica	47.482
Filtro de InDels	243

Este estudio se realizó con dos poblaciones diferentes de *E. globulus* (una de cada empresa forestal), por lo que en la Tabla 11 se encuentra el número de SNPs identificados para cada

una de ellas y los filtros en los cuales inicialmente se filtró por calidad de mapeo, donde se removió una gran cantidad de variantes que no fueron considerados como SNPs, además se realizó la remoción de InDels (otro tipo de polimorfismo común en el ADN) y aquellos SNPs que no estaban presentes en el 1% de la población (~3 individuos para ambas poblaciones).

Para la empresa *Bioforest*, los 438 SNPs están distribuidos en 328 *scaffolds*, para *Mininco*, las 606 variantes están en 411 *scaffolds*, en la Tabla 12 se muestra el número de SNPs y su frecuencia de su distribución, debido a que no coinciden de acuerdo a la ubicación dentro de los scaffolds, se consideraron los once primeros scaffolds que contienen SNPs. En la Figura 5, se observa un histograma representativo de la distribución de las variantes en los *scaffolds*, en la que hay un bajo número máximo de SNPs por *scaffold*: mientras que *Bioforest* cuenta con 4 *scaffolds* que presentan 5 SNPs (máximo de SNPs por *scaffold*), *Mininco* tiene 8 *scaffolds* con 5 SNPs. Para estos casos, en los que es posible encontrar más de un SNP en un mismo scaffold, la densidad y la distancia que hay entre ellos es heterogénea, existiendo SNPs más cercanas que otros, donde se pueden encontrar a una distancia aproximada entre 1 y 500 pb como máximo.

Mininco presenta 603 SNPs únicos para su población, en tanto que *Bioforest* cuenta con 435. Ambas poblaciones presentan 3 SNPs en común, como se observa en el diagrama de Venn de la Figura 6.

Tabla 11. Número total de SNPs y los respectivos filtros realizados para cada población.

	***Mininco* S.A.**	***Bioforest* S.A.**
SNPs totales	1.999.427	2.015.144
Filtro calidad mapeo	39.121	41.730
Filtro InDels	34.332	37.598
Filtro frecuencia alélica	606	438

De acuerdo a los SNPs polimórficos e informativos para cada población, las transiciones se encuentran con una mayor frecuencia en el genoma, siento la sustitución C-T la más común,

representado por un 32-34%, lo que es posible observar en la Figura 7 para *Bioforest* y Figura 8 para *Mininco*.

Tabla 12. Distribución de SNPs en los once primeros *scaffolds* en los que se encontraron SNPs para cada población.

	Bioforest			*Mininco*		
Scaffold	**Tamaño (pb)**	**N° SNPs**	**Frecuencia (pb)**	**Tamaño (pb)**	**N° SNPs**	**Frecuencia (pb)**
*Scaffold*_1	29.625	1	1/29.625	27.185	1	1/27.185
*Scaffold*_2	17.120	1	1/17.120	17.188	1	1/17.188
*Scaffold*_3	16.068	4	1/4.017	16.922	1	1/16.922
*Scaffold*_4	14.015	3	1/4.672	16.568	2	1/8.284
*Scaffold*_5	13.530	1	1/13.530	15.831	1	1/15.831
*Scaffold*_6	13.210	1	1/13.210	14.507	1	1/14.507
*Scaffold*_7	11.985	1	1/11.985	13.400	1	1/13.400
*Scaffold*_8	11.964	1	1/11.964	13.329	1	1/13.329
*Scaffold*_9	11.875	1	1/11.875	12.905	2	1/6.453
*Scaffold*_10	11.683	1	1/11.683	12.238	3	1/4.079
*Scaffold*_11	11.646	2	1/5.823	11.939	1	1/11.939
Otros	11.624-103	421		11.913-109	591	

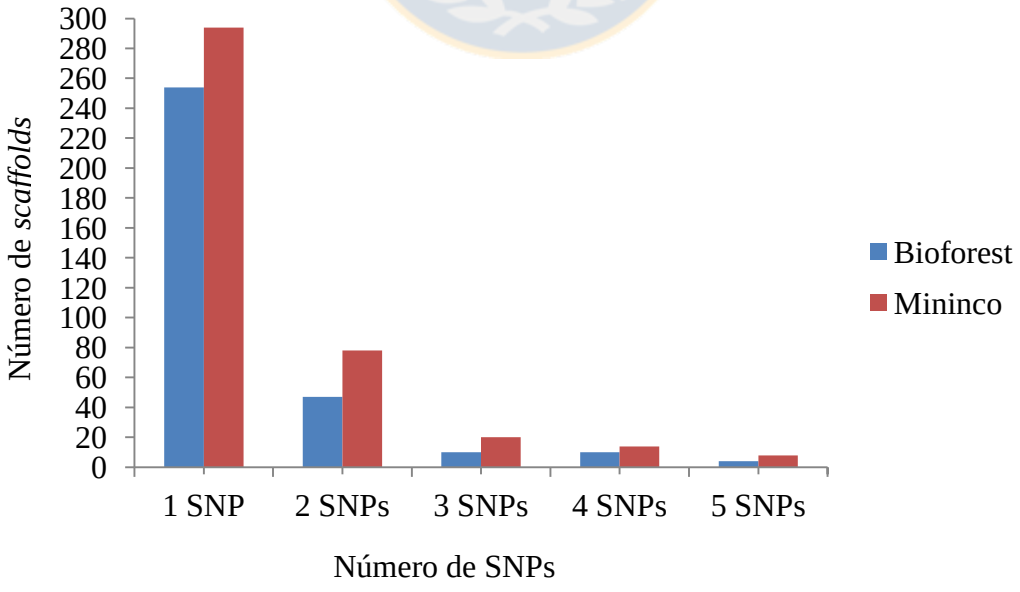

Figura 5. Histograma que representa el N° de SNPs por *scaffold* para cada población.

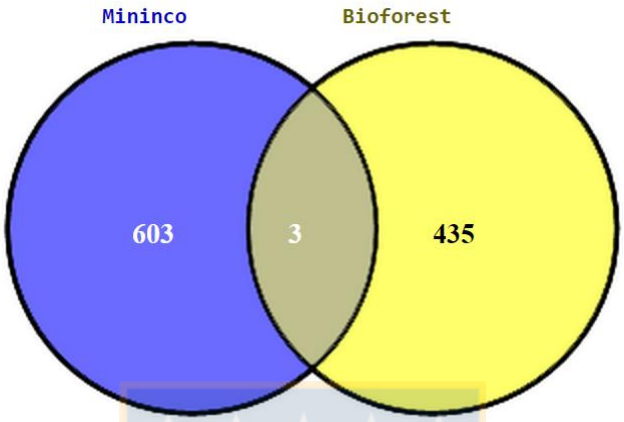

Figura 6. Diagrama de Venn de los SNPs comunes entre ambas poblaciones de *E. globulus.*

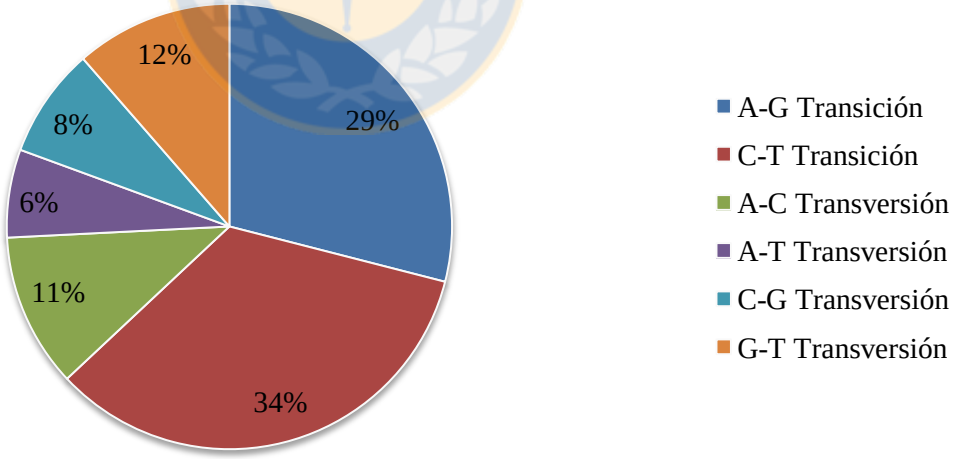

Figura 7. Gráfico representativo de los tipos de SNPs identificados en la población de *Bioforest*.

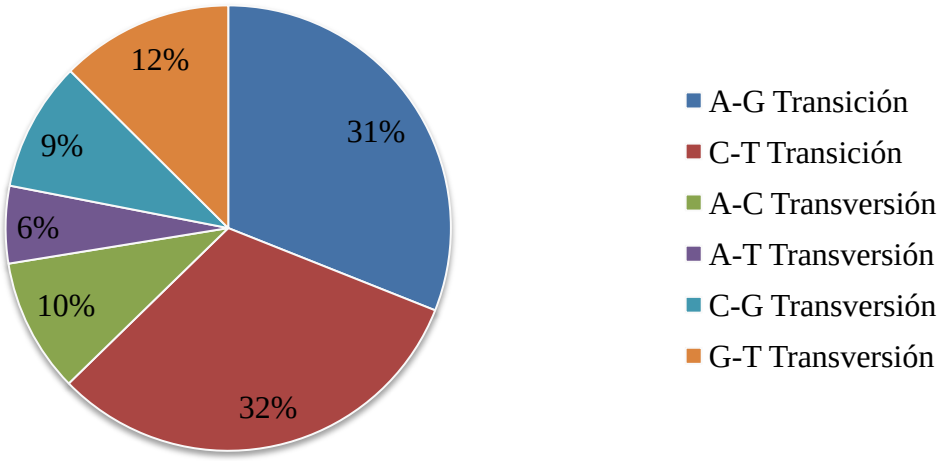

Figura 8. Gráfico representativo de los tipos de SNPs identificados en la población de *Mininco*.

VIII. DISCUSIÓN

Calidad de las librerías

Para obtener un buen ensamble, es necesario que ocurra un buen *overlap* entre los *reads*, es decir que, mientras mayor número de secuencias y mayor similitud tenga el *overlap* mejor es el ensamblaje (Surget-Groba and Montoya-Burgos 2010). Sin embargo, pueden ocurrir ciertos errores debido a genes duplicados o regiones repetitivas del genoma, los que se pueden encontrar tanto en zonas codificantes como no codificantes. Esta es una de las razones por las cuales, los métodos de creación de librerías más utilizados actualmente, como GBS, se basan en el uso de enzimas de restricción (una o más), eliminando regiones repetitivas, para así obtener una representación reducida del genoma (Davey et al. 2011). Además la correcta elección de ésta, determina el número de fragmentos y la cobertura de secuenciación de los fragmentos (Gardner et al. 2014). Para el caso de *E. globulus*, se realizó previamente una digestión *in silico* evaluando la presencia de sitios de restricción de la enzima *Ape*KI, la que tiene un sitio de reconocimiento de 5 bases (GCWGC, donde W puede ser una A o una T). Se encontraron 5 sitios de corte en el gen f5h, separados aproximadamente por 500 pb (datos no mostrados). Además los *barcodes* utilizados para la identificación de los *reads* secuenciados, están diseñados para ser ligados a los sitios de corte de *Ape*KI (Elshire et al. 2011), sin embargo, existen estudios en los que se han utilizado combinaciones de 2 o más ER de corte poco frecuente (Poland and Rife 2012). El uso de más de una ER para la digestión del genoma de *E. globulus*, o el uso de una enzima de restricción con mayor numero de sitios de restricción en la especie, podría generar fragmentos más cortos, lo que hubiera mejorado la calidad de la secuenciación de librerías Illumina, y de esta forma también una mayor cobertura y profundidad de ésta en conjunto con una menor perdida de datos (Peterson et al. 2014).

Una vez secuenciadas, el 11% de las muestras fueron calificadas como *fail*, lo que podría atribuirse a que el ADN se encuentre con restos de reactivos, ARN u otros polisacáridos que puedan inhibir o interferir en la reacción de amplificación previa a la secuenciación, ya que la cuantificación visual mediante electroforesis en gel de agarosa (utilizando marcador lambda) era superior a la concentración medida en espectrofotómetros (no mostrado), por lo que podría

estar dada por restos de ARN o proteínas que no fueron removidos durante la extracción de ADN.

Las tecnologías de NGS generan una gran cantidad de secuencias cortas, donde aproximadamente la mitad de estas no son utilizables, en promedio el 55% de las secuencias Illumina pasan los filtros de calidad (Harismendy et al. 2009). Esto se debe a que existen artefactos propios de la secuenciación como restos de adaptadores y primers que no fueron ligados, además de errores de lectura o en el llamado de las bases que provocan baja calidad de secuenciación y que podrían interferir en el análisis posterior de los datos y generar resultados erróneos (Patel and Jain 2012). Es por esto que al 89% de librerías de *E. globulus* secuenciadas (que fueron calificadas como *pass*) también fue necesario realizarles un control de calidad, donde el ~8% de los *reads* fue eliminado. En *Ipomoea batatas* (Wang et al. 2010) de un total de 59 millones de secuencias Illumina de tipo PE de 75 pb de largo, se redujeron a aproximadamente 51 millones de secuencias (13% de *reads* removidos), para algunas secuencias de tipo Illumina del genoma humano, se eliminó el 45% de los *reads* (Harismendy et al. 2009), mientras que para especies forestales como *E. camaldulensis* (Hendre et al. 2012) y *Pinus contorta* (Parchman et al. 2012), ambas secuenciadas con Illumina, se removieron 38% y 16%, respectivamente. La diferencia entre el porcentaje de *reads* removidos al realizar el control de calidad entre *E. globulus* con los estudios mencionados, se debe principalmente a la calidad, profundidad y cobertura de secuenciación, además a que corresponden a diferentes géneros y especies, por lo que los tamaños de sus genomas y la ploidía de cada uno de ellos son diferentes, incluso la cantidad de regiones repetitivas influye en la cantidad de *reads* obtenidos y filtrados, por lo que es primordial la elección o combinación de las enzimas de restricción utilizadas para digerir el genoma y generar librerías.

Sofatware *ABySS* v/s *SOAPdenovo2*

En un análisis comparativo de ensambladores para secuencias NGS (Lin et al. 2011), se clasificaron los datos entre el contenido de GC, el largo de las secuencias y el tipo de read (SE o PE) para obtener un mejor ensamble. Según Lin, los datos SE de *E. globulus* utilizados en el presente estudio son calificados con “un alto contenido de GC” (53%) y “*reads* largos” (aprox

88 pb), por lo que el orden de software más apropiado para el ensamble es el siguiente: $N50_{SOAPdenovo}$>$N50_{Edena}$≥$N50_{Velvet}$≈$N50_{ABySS}$>$N50_{SSAKE}$>$N50_{EULER\text{-}sr}$>$N50_{VCAKE}$, donde el mejor (mayor) N50 es obtenido por *SOAPdenovo*, el que según Lin et al. (2011) en conjunto con *ABySS*, también es considerado uno de los programas ensambladores más eficientes en cuanto a tiempo de ejecución y uso de memoria. Según Miller et al. (2010), las estadísticas N50 provenientes de diferentes programas de ensamblaje no son comparables a menos que cada uno sea calculado utilizando los mismos parámetros y valores de longitud, es por esta razón que la elección del mejor *kmer* (de los 4 probados para cada programa) no fue solo en base al ensamble realizado (*SOAPdenovo2* y *ABySS*), sino que también a las estadísticas entregadas por los *scaffolding* (*Sspace*) de cada uno de ellos.

Al realizar las pruebas de ensamblaje y *scaffolding* con ambos programas, utilizando los mismos parámetros, los resultados indicaron que el ensamble realizado con *SOAPdenovo2* fue mejor que el obtenido mediante *ABySS*, lo que coincide con otro estudios de ensamblaje realizados, donde a pesar de ser especies completamente diferentes se demostró que *SOAPdenovo* genera mejores resultados basados en la relación N° de *contigs*-N50: para el genoma humano *SOAPdenovo* produjo ensambles más contiguos y completos que *ABySS* (Li et al. 2009b), además generó mejores ensambles que *ABySS*, *Velvet*, *EULER*-SR, *SSAKE* y *Edena* en el genoma de *E. coli* (Paszkiewicz and Studholme 2010).

El valor *kmer*, se refiere a la fragmentación de una secuencia en pequeños subgrupos de largo *k* para luego ser ensambladas (Birol et al. 2009). Para conocer el valor *kmer* se ajusta a los datos de *E. globulus* de este estudio, y así obtener un mejor ensamble, se toman como referencia varias estadísticas que entregan los programas ensambladores, como el promedio del largo de los *gaps* y el promedio del número de *gaps* por *scaffold*, sin embargo las más importantes son el número de *scaffolds* y el valor N50 (Yandell and Ence 2012). Considerando estos dos parámetros, el mejor ensamble fue realizado con el valor k45, ya que tiene la mejor relación N50-número de *contigs*.

Ensamble *de novo* - *scaffolding*

El pequeño tamaño de las lecturas Illumina, constituye una de las limitaciones principales para realizar ensambles *de novo* para especies que no cuentan con un genoma de referencia, además de poseer gran cantidad de regiones repetitivas, ya que las secuencias cortas podrían generar falsos solapamientos al momento del ensamblaje (Schatz et al. 2012). Cuanto más largo es el fragmento, con mayor exactitud se produce el alineamiento entre secuencias, asignando la dirección correcta en el caso de alinear con una referencia (Harismendy et al. 2009). Es por esto que para el caso de no contar con un genoma de referencia, como sucede con *E. globulus*, se sugiere la utilización de *reads* más largos, por lo que mediante secuencias obtenidas por 454 (~450 pb) (Luo et al. 2012) se podría obtener un mejor solapamiento entre los *reads*, además de la combinación de secuencias SE con PE y un nivel medio de cobertura de secuenciación (Ahmad et al. 2011) lo que facilitaría el ensamblaje del cromosoma en pocos *contigs*.

El uso de estas secuencias cortas, crean un mayor número de contigs y de menor tamaño lo que genera una gran cantidad de *gaps* (Li and Homer 2010). Un *gap* es una región no secuenciada entre pares de *contigs*/*scaffolds*, y están representados como *N* (Yandell and Ence 2012). Según Myburg et al. (2011), de 642 Mpb distribuidos en 32.762 *contigs* de *E. grandis*, aproximadamente el 7,3% corresponden a *gaps*. Por su parte, para el caso de *E. globulus*, de los ~651 Mpb del largo de la referencia creada *de novo*, un 3,6% corresponden a *gaps*, sin embargo en otras especies secuenciadas mediante metodologías NGS, los ensambles *de novo* contienen un número mucho menor de *gaps*, ya que según Paszkiewicz and Studholme (2010) la mayoría de los ensambles tienen un *gap* cada 10.000 pb, ya que no superan unas pocas megabases de largo y contienen cientos de estos (0,01%).

Con la información que proporcionan las lecturas *paired-end* durante el *scaffolding*, los *gaps* pueden ser eliminados o reducidos, ya que una secuencia PE que abarca dos *contigs* sirve como evidencia de la yuxtaposición de esos dos *contigs* dentro del genoma (Paszkiewicz and Studhome 2010). Al eliminar un *gap*, dos *contigs* se unen, por lo que esto genera la disminución en el número de *contigs*/*scaffolds.* Esta es la razón por la cual aumentó el largo de la referencia (2%), además del *contig*/*scaffold* más largo (27%) y también el valor N50 (53%), al realizar el *scaffolding*.

En un estudio realizado por (Boetzer et al. 2011) en *Ailuropoda melanoleuca*, el ensamble *de novo*, mediante *SOAPdenovo* generó 4.585 *contigs* los que después del *scaffolding* (con *Sspace*) de la referencia se redujo en más de un 50% a 2.041 *scaffolds* (37 librerías PE utilizadas). La razón por la cual la diferencia en el tamaño (número de *contigs*/*scaffolds*) entre el ensamble *de novo* y el *scaffolding* realizado para *E. globulus* es menor (~30%) que lo observado en *A. melanoleuca*, se debe a la poca cantidad de secuencias de tipo *paired-end* (37 librerías PE en *A. melanoleuca* vs 2 en *E. glubulus*), ya que son estas las que permiten la unión de los *contigs* en el proceso de *scaffolding*. Este tipo de secuencias (PE) no solo disminuyen el numero de gaps y contigs/scaffols, pues también aumentan el valor N50, esto es una medida estadística de la calidad del ensamble y se refiere a la longitud del último contig/scaffold que es añadido para conseguir el 50% del ensamble (Schatz et al. 2010). En el genoma humano (Li and Homer 2010) el valor N50 fue mejorado en la etapa de *scaffolding* al haber adicionado lecturas de tipo PE; cuando se agregaron 2,6 kb PE el N50 aumentó de 17,3 kb a 103,5 kb (83%), en cuanto se adicionaron 6kb de lecturas PE, el N50 aumentó a 203,5 kb, es decir un 50% más y cuando se adicionaron 9,6 kb de PE el valor N50 aumentó a 446,2 kb (54.3 %). Al sumar más cantidad de secuencias (PE), el *contig* más largo también aumenta, por lo que el valor N50 mejora. Para el caso de *E .globulus*, el N50 aumentó un 52% en la etapa de *scaffolding* de 816 a 1.723, debido a las secuencias PE provenientes de NCBI, que se encargan de eliminar los *gaps* y alargar los *contigs*. Idealmente la etapa de *scaffolding* debería realizarse usando una mayor cantidad de librerías de tipo PE, sin embargo, la generación de este tipo de librerías dobla el costo de la secuenciación (Surget-Groba and Montoya-Burgos 2010).

El ensamble *de novo* creado con las 506 muestras de *E. globulus*, generó una referencia de ~650 Mpb, sin embargo, en un estudio realizado por (Grattapaglia and Bradshaw 1994) se estimó el contenido de ADN nuclear en diferentes especies del género Eucalyptus mediante citometría de flujo. De acuerdo a este análisis, el genoma de *E. globulus* está compuesto por 530 Mpb. Una de las razones de la diferencia en el tamaño de la especie mediante ambas metodologías se debe a la sobrerrepresentación en la secuenciación de una región especifica del genoma, considerándolas así, el ensamblador, como regiones repetitivas y de este modo alargar la secuencia esperada.

Mapeo ensamble *de novo* de *E. globulus* con referencia *E. grandis*

E. globulus y *E. grandis* pertenecen al subgénero Symphyomyrtus y tienen aproximadamente un 1.5% de divergencia genética entre ellos (Myburg et al. 2011), por lo que el alto porcentaje de alineamiento del ensamble *de novo* de *E. globulus* contra el genoma de referencia de *E. grandis* indica que este representa gran parte del genoma de *E. grandi*s. Ambas especies sólo se diferencian en un 18% de secuencias, pues a pesar de pertenecer al mismo género y además de tener un alto porcentaje de identidad, son especies diferentes, y la gran cantidad de diferencias entre genes ortólogos con el aumento de la distancia evolutiva, conduce a una baja calidad de alineamiento entre ambas (Surget-Groba and Montoya-Burgos 2010). Parte del 18% del ensamble que no se alineó, también puede estar dada por regiones repetitivas y/o por regiones que pueden ser exclusivas de *E. globulus* y no existir en el genoma de *E. grandis*, como *gaps* o incluso por la presencia de variantes estructurales (SV) (Metzker 2010). Por otra parte, la baja cobertura de alineamiento se debe a que es un alineamiento de genomas y no de librerías genómicas, las cuales están compuestas de numerosos *reads*.

Se mapearon un total de 1.336.641 *scaffolds* del ensamble de *E. globulus* al genoma de referencia de *E. grandis.* Si consideramos que el ensamble *de novo* creado está compuesto por 1.060.306 *scaffolds*, se puede deducir que un *scaffold* se mapeó más de una vez en el genoma de *E. grandis*, esto se debe al algoritmo del software del mapeo (*BWA*) que permite la realización de alineamientos múltiples. En el *scaffold* 1 (40.297.282 pb de largo) y el *scaffold* 2 (64.237.462 pb de largo) de *E. grandis* se mapearon 84.544 y 125.259 *scaffolds* del ensamble *de novo* respectivamente, de los cuales comparten 4.158 *reads*. En el *scaffold* 3 (80.088.348 pb de largo) se mapearon 146.716 *reads* y comparte 5.481 *reads* con el *scaffold* 1 y 7.604 *reads* con el *scaffold* 2; entre los 3 primeros *scaffolds* de *E. grandis* se repiten 1.048 *reads* del ensamble *de novo* (datos no mostrados). Esto podría generar un mal llamado de SNPs, exclusivamente para los SNPs presentes en aquellos *scaffolds* que están alineados más de una vez en el genoma de *E. grandis.*

Mapeo librerías y llamado de SNPs

Las estadísticas de mapeo fueron similares entre todas las muestras analizadas, donde la cobertura y la calidad de mapeo de cada librería contra la referencia *de novo* son bastante bajas considerando que fueron digeridas con *Ape*KI, una enzima de restricción de corte frecuente, que ya se ha utilizado en estudios en *E. grandis* (comunicación personal Instituto de Diversidad Genómica de la Universidad de Cornell) por lo que se esperaban resultados similares para *E. globulus.*

Aproximadamente un 51% de los *reads* filtrados fueron mapeados contra el genoma de *E. globulus*. Debido a la calidad de mapeo obtenida, la probabilidad de que el alineamiento sea fiable es baja, esto podría deberse principalmente al pequeño tamaño de los reads, pues tienen mayor facilidad para alinearse de forma equivocada o incluso alinearse a múltiples regiones del genoma de forma simultánea, debido al algoritmo utilizado para el mapeo. La baja cobertura, indica que las librerías secuenciadas están abarcando solo 21% del genoma de *E. globulus*, los bajos índices de cobertura y calidad de mapeo son más bajos de lo aceptado; Thumma et al. (2012) estableció un límite mínimo de cobertura 8, para el descubrimiento de SNPs en el caso de *E. camaldulensis* y una calidad de mapeo sobre 20 para establecer un 99% de confianza. Los altos niveles de duplicación podrían ser menores si es que inicialmente las librerías se hubieran sometido a un filtro de duplicación, ya que este es un resultado directo de la secuenciación, donde ciertas partes del genoma están sobrerrepresentados y enriquecidos de reads secuenciados. Además la gran cantidad de elementos repetitivos en genomas complejos generan un problema debido a que consumen la capacidad de secuenciación, lo que es consistente con las altas tasas de duplicación obtenidos tanto en la secuenciación como en el mapeo.

La frecuencia con la que se producen los SNPs varía drásticamente, sobre todo en las especies arbóreas, donde según Thavamanikumar et al. (2011), *E. globulus* corresponde a una de las especies forestales estudiadas con mayor frecuencia de SNPs (entre *Pseudotsuga menziesii*, *Pinus taeda* y *Populus trichocarpa*), en un estudio de 20 genes relacionados a la formación de madera, en el cual *E. globulus* presenta 1 SNP cada 45 pb. En el presente estudio, la identificación de SNPs se realizó en el genoma completo, donde para la población de la empresa *Bioforest*, de un total de aproximadamente 41 mil variantes, solo un 1% de ellas

fueron consideradas como SNPs informativos mientras que para *Mininco*, de los más de 39 mil SNPs putativos, el 1,5% de ellos pasaron todos los filtros, determinando una frecuencia promedio de 1 SNP cada 1.2 Mpb, para ambas poblaciones. Además, se identificó que el tipo de SNP más común para este estudio son las transiciones, entre las cuales la transición C-T tiene una leve predominancia por sobre la transición A-G, lo que coincide con otros trabajos mencionados anteriormente (Lai et al. 2012).

El bajo número de SNPs encontrados en las librerías de *E. globulus*, se debe a la calidad de secuenciación de las librerías, el llamado de SNPs basado en secuencias cortas NGS genera dos tipos de problemas: la mala calidad de los *reads*, especialmente en el extremo 3', lo que afecta directamente con la exactitud del llamado de cada variante (SNP), sobretodo en regiones de baja cobertura y por otra parte la precisión en el mapeo de los *reads*, más aún cuando estos además de cortos son de tipo SE (Azam et al. 2012), incluso una mala elección de la ER en el momento de la digestión para la reducción de librerías genómicas, influye directamente con el número de variantes identificadas y la heterocigozidad (Lu et al. 2013). Por otra parte, el bajo número de SNPs podría atribuirse a una referencia poco representativa del genoma, o incluso elementos parálogos que contienen un alto nivel de similitud podrían ser ensamblados en el mismo contig, dado que el ensamblador no puede distinguir si corresponden a diferentes regiones del genoma, debido al pequeño tamaño de las secuencias (Novaes et al. 2008), lo que al igual que el alto nivel de duplicación de las secuencias, podría inducir en un llamado de SNPs erróneo.

Para un estudio realizado en *Prunus persica*, se identificaron 165 mil SNPs de los cuales se eliminó un 96% de ellos (Ahmad et al. 2011), sin embargo en *Malus x domestica* (Chagné et al. 2012), se identificaron casi 11 millones de SNPs, de los cuales el 19,5 % pasaron los filtros, considerando esto, la frecuencia de distribución fue de un SNP cada 288 pb. En cuanto a los estudios realizados en especies leñosas como *E. grandis*, se encontraron ~23 mil SNPs solo en secuencias EST, de las cuales un 83% fue validado (Novaes et al. 2008) mientras que para *E. camaldulensis* se identificaron más de 12 mil SNPs putativos en 41 genes candidatos implicados en crecimiento y finalmente solo se seleccionaron 1.191 SNPs (10%) luego del filtro. Como se puede observar, existe gran variación entre el porcentaje de SNPs finales para las diferentes especies mencionadas, lo que podría estar relacionado directamente con los tipos

de filtros y a los parámetros a los que son sometidos cada uno de ellos, que dependen exclusivamente de las características del genoma y del conjunto de datos (Kumar et al. 2012).

Dentro de los filtros más utilizados en el llamado de SNPs, está eliminar aquellas variantes que estén cerca de *gaps*, donde se define un mínimo de bases de distancia, o se establece un mínimo de *reads* en los que debe estar presente la variante, por ejemplo en un estudio en *P. persica* (Ahmad et al. 2011), se definió un mínimo de 60 pb de distancia a un *gap*, y que la variante esté presente en al menos 10 *reads*, evitando así falsos SNPs provocados por homopolímeros. Para el caso de *E. globulus*, este filtro por profundidad de *reads*, no se realizó debido a la baja profundidad de secuenciación que tenían las muestras. Otro de los filtros más utilizados es la eliminación de haplotipos, es decir grupos de variantes en que la distancia entre cada uno sea al menos 2 kb (Chagné et al. 2012). En este caso, si se hubiera aplicado el filtro de haplotipos probablemente el número de SNPs encontrados en cada población de *E. globulus* estudiada, sería mucho menor al encontrado, debido a que la distancia entre cada uno de ellos dentro de un *scaffold*, va entre 1 y 500 pb como máximo, por lo que incluso, algunos de ellos que se separan por pocas bases, podrían corresponder a la misma variante, y no a diferentes SNPs, provocado por la baja calidad de secuenciación. Para ambas poblaciones analizadas, se observa que el filtro más selectivo (después del filtro por calidad de mapeo, en el cual las variantes no son considerados SNPs) en cuanto al número de polimorfismos es el de MAF, seguido por la eliminación de InDels, ya que estos son uno de los tipos de polimorfismo más abundantes en el ADN (Lijavetzky et al. 2007), después de los SNPs.

A pesar que cada estudio utiliza sus propios parámetros y filtros para reducir el número de SNPs, el filtro por mínima frecuencia alélica (MAF) es uno de los filtros comunes en varios trabajos, donde se define generalmente entre un 1 y un 5% de frecuencia alélica, esto va a depender de la categoría funcional de los SNPs (Xu and Taylor 2009). Para *E. globulus*, se eliminaron todas aquellas variantes que no estuvieran presenten en al menos el 1% de la población, lo que corresponde a aproximadamente 3 individuos (para cada empresa), sin embargo, debido al pequeño tamaño de ambas poblaciones, es probable que algunos de los SNPs con una frecuencia menor al 1%, también sean informativos. Se esperaba que los SNPs encontrados para ambas poblaciones fueran comunes, sin embargo solo 3 de ellos se compartían. Esto puede deberse a que posiblemente se secuenciaron zonas diferentes del

genoma para cada especie, pudiendo encontrar una cantidad de SNPs diferentes en cada población, además del número de individuos estudiados, pues si este aumentara, aumenta el 1% de la población y por ende podría aumentar el número de SNPs comunes entre ambas poblaciones.

Para especies que cuentan con un genoma de referencia, la identificación de SNPs se simplifica, al no tener que crear una referencia *de novo*. A pesar de esto, la alineación con el genoma de referencia, no es fácil, debido a regiones con altos niveles de diversidad entre el genoma de referencia y el genoma secuenciado. Luego del alineamiento, el llamado de SNPs putativos y el filtro de ellos es lo que sigue. Por otra parte, para aquellas especies que no cuentan con un genoma de referencia, además de la creación de la referencia *de novo*, existe la posibilidad de utilizar un genoma de referencia de otras especies que esté relacionada con la estudiada. En un estudio paralelo (datos no mostrados) se utilizaron los mismos genotipos (506 muestras) utilizadas en este trabajo para la identificación de SNPs, utilizando el genoma de referencia de *E. grandis*, un genoma de 640 Mpb que cuenta con 4.952 *scaffolds*. En este caso, el número de SNPs putativos fue mucho mayor que el encontrado con el ensamble *de novo* de *E. globulus*, por lo que un gran número de diferencias podría deberse a fragmentos de ADN propios de cada especie, y no a SNPs propiamente tal: para la población de *Bioforest* se encontraron cerca de 62 mil variantes, mientras que para *Mininco* aproximadamente 35 mil. Para ambas poblaciones, fueron aplicados los mismos filtros que los mencionados en este estudio (ver materiales y métodos), además de un filtro para eliminar aquellos falsos SNPs que solo se deben a las diferencias entre especies. Finalmente el número de variantes disminuyó a cerca de mil SNPs para cada población, aproximadamente el doble de los SNPs encontrados utilizando el ensamble *de novo*. Esta diferencia podría estar dada por un ensamble *de novo* poco representativo de la especie, con regiones sobrerrepresentadas y otras que no fueron cubiertas debido a la baja cobertura de la secuenciación, por lo que algunos de los SNPs encontrados en el estudio con *E. grandis*, podrían no estar presentes en el ensamble *de novo* de *E. globulus* recientemente creado. De las variantes encontradas para Bioforest, utilizando ambas metodologías, existe un SNP en común, al igual que para Mininco, considerando solo aquellos SNPs que se encuentran ubicados en algún gen (anotados en base a *E. grandis*). De acuerdo a las características de estos marcadores, se espera que las variantes encontradas utilizando ambas metodologías correspondan a los mismos SNPs, sin embargo al utilizar una

referencia creada bajo un ensamble *de novo* de la especie, sin validación previa, y comparar con los SNPs encontrados utilizando un genoma de otra especie, es bastante probable que estos no coincidan, a pesar de ser provenientes de poblaciones genéticamente iguales. Como se mencionó anteriormente, al contar con un genoma anotado de la especie, se evitan las ambigüedades en cuanto al número de variantes compartidas entre ambas metodologías y ambas poblaciones, en las que se deberían encontrar al menos, un número mayor de SNPs en común, extrapolables a cualquier población de *E. globulus*.

IX. CONCLUSIÓN

Es posible generar librerías génicas a partir de ADN de *E. globulus* mediante la metodología *Genotyping by Sequencing*, utilizando la enzima de restricción *Ape*KI para la reducción del genoma.

Es posible establecer un ensamble *de novo* representativo del genoma de la especie, a partir de librerías génicas obtenidas mediante técnicas GBS.

Un alto porcentaje de mapeo entre la referencia *de novo* de *E. globulus* contra el genoma de referencia de *E .grandis*, no asegura la fiabilidad ensamble, ya que en este caso, los pequeños s*caffolds* del ensamble *de novo*, son capaces de alinearse fácilmente a múltiples *scaffolds* a lo largo del genoma de referencia (*E. grandis*).

Existen 80,851 SNPs putativos distribuidos uniformemente a lo largo de todo genoma de *E. globulus*, de los cuales sólo un 1.3% de ellos fueron polimórficos e informativos para *E. globulus*. Este número podría ser mayor si la especie contara con un genoma de referencia anotado.

Considerando como referencia el ensamble *de novo* de *E. globulus* creado en este estudio, existe una frecuencia promedio de 1 SNP cada 1.2 Mpb, para ambas poblaciones.

Debido a que para selección genómica son requeridos miles de marcadores, el bajo número de SNPs polimórficos identificados en *E. globulus*, no presentan un uso potencial en esta área, a menos que se aumentara el tamaño efectivo de la población, y aumentaran de esta forma, el numero de SNPs identificados.

X. BIBLIOGRAFÍA

Agarwal M, Shrivastava N, Padh H (2008) Advances in molecular marker techniques and their applications in plant sciences. Plant cell reports 27:617-631. doi:10.1007/s00299-008-0507-z.

Ahmad R, Parfitt DE, Fass J, et al. (2011) Whole genome sequencing of peach (Prunus persica L.) for SNP identification and selection. BMC genomics 12:569. doi:10.1186/1471-2164-12-569.

Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD (1994) Molecular biology of the cell. Garland Publising. Nueva York, USA. 1220p.

Azam S, Thakur V, Ruperao P, et al. (2012) Coverage-based consensus calling (CbCC) of short sequence reads and comparison of CbCC results to identify SNPs in chickpea (Cicer arietinum; Fabaceae), a crop species without a reference genome. American journal of botany 99:186-192. doi:10.3732/ajb.1100419.

Bankevich A, Nurk S, Antipov D, et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19:455-477. doi:10.1089/cmb.2012.0021.

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. The Plant journal: for cell and molecular biology 51:910-918 doi:10.1111/j.1365-313X.2007.03193.x.

Birol I, Jackman SD, Cydney BN, et al. (2009) De novo transcriptome assembly with ABySS. Bioinformatics 25:2872-2877. doi:10.1093/bioinformatics/btp367.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578-579. doi:10.1093/bioinformatics/btq683.

Brookes AJ (1999) The essence of SNPs. Gene 234:177-186.

Chagné D, Gasic K, Crowhurst RN, et al. (2008) Development of a set of SNP markers present in expressed genes of the apple. Genomics 92:353-358. doi:10.1016/j.ygeno.2008.07.008.

Chagné D, Crowhurst RN, Troggio M, et al. (2012) Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. PloS one 7:e31745. doi:10.1371/journal.pone.0031745.

Chen C, Mitchell SE, Elshire RJ, Buckler ES, El-Kassaby YA (2013) Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. Tree Genetics & Genomes 9:1537-1544. doi:10.1007/s11295-013-0657-1.

Danecek P, Auton A, Abecasis G, et al. (2011) The variant call format and VCFtools. Bioinformatics 27:2156-2158. doi:10.1093/bioinformatics/btr330.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature reviews Genetics 12:499-510. doi:10.1038/nrg3012.

Desai AN, Jere A (2012) Next-generation sequencing: ready for the clinics?. Clinical genetics 81:503-510. doi:10.1111/j.1399-0004.2012.01865.x.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one 6:e19379. doi:10.1371/journal.pone.0019379.

Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH (2004) An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. Genome research 14:1812-1819. doi:10.1101/gr.2479404.

Frazer KA, Ballinger DG, Cox DR (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861. doi:10.1038/nature06258.

Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nature Reviews Genetics 10:241-251. doi:10.1038/nrg2554.

Gallino JP, Fernández M, Tapias R, Alcuña MM, Cañas I (2007) Aclimatación al frío en diferentes clones de Eucalyptus globulus Labill durante el régimen natural de endurecimiento. Boletín del CIDEU 4:77-83.

Gardner KM, Brown P, Cooke T, et al. (2014) Fast and cost-effective genetic mapping in apple using next-generation sequencing. G3: Genes| Genomes| Genetics 4:1681-1687. doi:10.1534/g3.114.011023.

Gonzalez-Martinez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB (2007) Association genetics in Pinus taeda L. I. Wood property traits. Genetics 175:399-409. doi:10.1534/genetics.106.061127.

Goulao LF, Vieira-Silva S, Jackson PA (2011) Association of hemicellulose and pectin modifying gene expression with Eucalyptus globulus secondary growth. Plant Physiology and Biochemistry 49:873-881. doi:10.1016/j.plaphy.2011.02.020.

Grattapaglia D, Bradshaw Jr H (1994) Nuclear DNA content of commercially important Eucalyptus species and hybrids. Canadian Journal of Forest Research 24:1074-1078.

Gupta P, Rustgi S, Mir R (2008) Array-based high-throughput DNA markers for crop improvement. Heredity 101:5-18. doi:10.1038/hdy.2008.35.

Haas J, Katus HA, Meder B (2011) Next-generation sequencing entering the clinical arena. Molecular and cellular probes 25:206-211. doi:10.1016/j.mcp.2011.08.005.

Harismendy O, Ng PC, Strausberg RL, et al. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10:R32. doi:10.1186/gb-2009-10-3-r32.

Hayden EC (2013) Tepid showing for genomics X prize. Nature 497:546-547. doi:10.1038/497546a.

Hendre PS, 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:O17. doi:10.1186/1753-6561-5-S7-O17.

Hendre PS, Kamalakannan R, Varghese M (2012) High-throughput and parallel SNP discovery in selected candidate genes in Eucalyptus camaldulensis using Illumina NGS platform. Plant biotechnology journal 10:646-656. doi:10.1111/j.1467-7652.2012.00699.x

Hunt M, Newbold C, Berriman M, Otto TD (2014) A comprehensive evaluation of assembly scaffolding tools. Genome Biol 15: 42. doi:10.1186/gb-2014-15-3-r42.

Instituto Forestal. (2014). Boletín Estadístico. Anuario Forestal 2008. INFOR. Disponible en: http://www.infor.cl/index.php/quienes-somos/destacados/597-anuario-forestal-2014-boletin-estadistico-n-144.

Jimenez-Escrig A, Gobernado I, Sanchez-Herranz A (2012) Whole genome sequencing: a qualitative leap forward in genetic studies. Revista de neurologia 54:692-698.

Jones RC, Steane DA, Potts BM, Vaillancourt RE (2002) Microsatellite and morphological analysis of Eucalyptus globulus population. Canadian Journal of Forest Research 32:59-66. doi: 10.1139/X01-172.

Joshi NA, Fass JN (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle.

Kalia RK, Rai MK, Kalia S, Singh R, Dhawan A (2011) Microsatellite markers: an overview of the recent progress in plants. Euphytica 177:309-334. doi:10.1007/s10681-010-0286-9.

Kennedy GC , Matsuzaki H, Dong S, et al. (2003) Large-scale genotyping of complex DNA. Nature biotechnology 21:1233-1237. doi:10.1038/nbt869.

Khlestkina EK, Salina EA (2006) SNP markers: Methods of analysis, ways of development, and comparison on an example of common wheat. Russian Journal of Genetics 42:585-594. doi:10.1134/s1022795406060019.

Kulheim C, Yeoh SH, Maintz J, Foley WJ, Moran GF (2009) Comparative SNP diversity among four Eucalyptus species for genes from secondary metabolite biosynthetic pathways. BMC genomics 10:452. doi:10.1186/1471-2164-10-452.

Kumar P, Gupta V, Misra A, Modi D, Pandey B (2009) Potential of molecular markers in plant biotechnology. Plant Omics Journal 2:141-162.

Kumar S, Banks TW, Cloutier S (2012) SNP Discovery through Next-Generation Sequencing and Its Applications. International journal of plant genomics 2012:831460. doi:10.1155/2012/831460.

Lai K, Duran C, Berkman PJ, et al. (2012) Single nucleotide polymorphism discovery from wheat next-generation sequence data. Plant biotechnology journal 10:743-749. doi:10.1111/j.1467-7652.2012.00718.x.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357-359. doi:10.1038/nmeth.1923.

Larkin PD, Park WD (2003) Association of waxy gene single nucleotide polymorphisms with starch characteristics in rice (Oryza sativa L.). Molecular Breeding 12:335-339. doi:10.1023/B:MOLB.0000006797.51786.92.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009a) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079. doi:10.1093/bioinformatics/btp352.

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

Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Briefings in bioinformatics 11:473-483. doi:10.1093/bib/bbq015.

Li R (2009) Short Oligonucleotide Analysis Package: SOAPdenovo 1.03 Beijing Genomics Institute, Beijing.

Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J (2009b) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966-1967. doi:10.1093/bioinformatics/btp336.

Lijavetzky D, Cabezas JA, Ibáñez A, Rodríguez V, Martínez-Zapater JM (2007) High throughput SNP discovery and genotyping in grapevine (Vitis vinifera L.) by combining a re-sequencing approach and SNPlex technology. BMC genomics 8:424. doi:10.1186/1471-2164-8-424.

Lindblad-Toh K, Winchester E, Daly MK, et al. (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. Nature genetics 24:381-386. doi:10.1038/74215.

Luo R, Liu B, Xie Y, et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18. doi:10.1186/2047-217X-1-18.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal 17:10-12. doi:http://dx.doi.org/10.14806/ej.17.1.200.

Metzker ML (2010) Sequencing technologies—the next generation. Nature Reviews Genetics 11:31-46. doi:10.1038/nrg2626.

Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95:315-327. doi:10.1016/j.ygeno.2010.03.001.

Morin PA, Luikart G, Wayne RK (2004) SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution 19:208-216. doi:10.1016/j.tree.2004.01.009.

Myburg A, Grattapaglia D, Tuskan G, et al. (2011) The Eucalyptus grandis Genome Project: Genome and transcriptome resources for comparative analysis of woody plant biology. BMC Proceedings 5:I20. doi:10.1186/1753-6561-5-s7-i20.

Nesbitt KA, Potts BM, Vaillancourt RE, West AK, Reid JB (1995) Partitioning and distribution of RAPD variation in a forest tree species, Eucalyptus globulus (Myrtaceae). Heredity 74: 628-637. doi:10.1038/hdy.1995.86.

Novaes E, Drost DR, Farmerie WG, Pappas GJ, Jr., Grattapaglia D, Sederoff RR, Kirst M (2008) High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC genomics 9:312. doi:10.1186/1471-2164-9-312.

Nowrousian M (2010) Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. Eukaryotic cell 9:1300-1310. doi:10.1128/EC.00123-10.

Oliver RE, Lazo GR, Lutz JD, et al. (2011) Model SNP development for complex genomes based on hexaploid oat using high-throughput 454 sequencing technology. BMC genomics 12:77. doi:10.1186/1471-2164-12-77.

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome research 18:2024-2033. doi:10.1101/gr.080200.108.

Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle CA (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. Molecular ecology 21:2991-3005. doi:10.1111/j.1365-294X.2012.05513.x.

Paszkiewicz K, Studholme DJ (2010) De novo assembly of short sequence reads. Briefings in bioinformatics 11:457-472. doi:10.1093/bib/bbq020.

Patel RK, Jain M (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PloS one 7:e30619. doi:10.1371/journal.pone.0030619.

Peterson G, Dong Y, Horbach C, Fu Y-B (2014) Genotyping-By-Sequencing for Plant Genetic Diversity Analysis: A Lab Guide for SNP Genotyping. Diversity 6:665-680. doi:10.3390/d6040665.

Poke F, Vaillancourt R, Potts B, Reid J (2005). Genomic research in Eucalyptus. Genetica. 125:79–101. doi:10.1007/s10709-005-5082-4.

Poland JA, Rife TW (2012) Genotyping-by-Sequencing for Plant Breeding and Genetics. The Plant Genome Journal 5:92. doi:10.3835/plantgenome2012.05.0005.

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Current opinion in plant biology 5:94-100. doi:10.1016/S1369-5266(02)00240-6.

Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic acids research 30:3894-3900. doi:10.1093/nar/gkf493.

Rao DC, Gu CC (2008) Genetic dissection of complex traits. Academic Press. London, UK. 760p.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74:5463-5467.

Santos J (2011). Epidemiologia Genética. Editorial Mediterráneo Ltda. Santiago, Chile. 233p

Schatz MC, Delcher AL, Salzberg SL (2010) Assembly of large genomes using second-generation sequencing. *Genome research* 20:1165-1173. doi:10.1101/gr.101360.109

Schatz MC, Witkowski J, McCombie WR (2012) Current challenges in de novo plant genome sequencing and assembly. Genome Biology 13:243. doi:10.1186/gb-2012-13-4-243.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome research 19:1117-1123. doi:10.1101/gr.089532.108.

Spalvieri MP, Rotenberg RG (2004) Medicina genómica: Aplicaciones del polimorfismo de un nucleótido y micromatrices de ADN. Medicina (Buenos Aires) 64:533-542.

Surget-Groba Y, Montoya-Burgos JI (2010) Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome research 20:1432-1440. oi:10.1101/gr.103846.109.

Thavamanikumar S, McManus LJ, Tibbits JFG, Bossinger G (2011) The significance of single nucleotide polymorphisms (SNPs) inEucalyptus globulusbreeding programs. Australian Forestry 74:23-29. doi:10.1080/00049158.2011.10676342.

Thumma BR, Nolan MF, Evans R, Moran GF (2005) Polymorphisms in cinnamoyl CoA reductase (CCR) are associated with variation in microfibril angle in Eucalyptus spp. Genetics 171:1257-1265. doi:10.1534/genetics.105.042028.

Thumma BR, Sharma N, Southerton SG (2012) Transcriptome sequencing of Eucalyptus camaldulensis seedlings subjected to water stress reveals functional single nucleotide polymorphisms and genes under selection. BMC genomics 13:364. doi:10.1186/1471-2164-13-364.

Todd & Steven 2012

Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature reviews Genetics 13:36-46. doi:10.1038/nrg3117.

Wang Z, Fang B, Chen J, Zhang X, Luo Z, Huang L, Chen X (2010) De novo assembly and characterization of root transcriptome using Illumina *paired-end* sequencing and development of cSSR markers in sweetpotato (Ipomoea batatas). BMC genomics 11:726. doi:10.1186/1471-2164-11-726.

Wong CK, Bernardo R (2008) Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations TAG Theoretical and applied genetics. Theoretische und angewandte Genetik 116:815-824. doi:10.1007/s00122-008-0715-5.

Xu Z, Taylor JA (2009) SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. Nucleic acids research 37:W600-W605. doi:10.1093/nar/gkp290.

Yamamoto T, Nagasaki H, Yonemaru J-i, Ebana K, Nakajima M, Shibaya T, Yano M (2010) Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. BMC genomics 11:267. doi:10.1186/1471-2164-11-267.

Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. Nature reviews Genetics 13:329-342. doi:10.1038/nrg3174.

You FM, Hou N, Deal KR, Gu YQ, Luo MC, McGuire PE, Dvorak J, Anderson OD (2011) Annotation-based genome-wide SNP discovery in the large and complex Aegilops tauschii genome using next-generation sequencing without a reference genome sequence. BMC genomics 12:59. doi:10.1186/1471-2164-12-59.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research 18:821-829. doi:10.1101/gr.074492.107.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., & Shen, B. (2011). A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PloS one, 6, e17915. doi:10.1371/journal.pone.0017915

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA (2013) The MaSuRCA genome assembler. Bioinformatics 29:2669-2677. doi:10.1093/bioinformatics/btt476.