UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias Biológicas y Ambientales

Desarrollo de un sistema de etiquetado para optimizar la secuenciación de amplicones y la determinación de alelos en el MinION

.

Mónica Becerra Wong

Ingeniería en Procesos Biotecnológicos

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniera en Procesos Biotecnológicos

Quito, 4 de mayo de 2020

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias Biológicas y Ambientales

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Desarrollo de un sistema de etiquetado para optimizar la secuenciación de amplicones y la determinación de alelos en el MinION

Mónica Becerra Wong

Nombre del profesor, Título académico **Andrés Torres, PhD**

Quito, 4 de mayo de 2020

DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Nombres y apellidos: Mónica Becerra Wong

Código: 00130529

Cédula de identidad: 1756573091

Lugar y fecha: Quito, mayo de 2020

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

La secuención nanoporo con el dispositivo MinION es una estrategia novedosa que permite el análisis de resultados en tiempo real. El MinION, al ser un equipo portable y una tecnología poco costosa, posee un amplio rango de aplicaciones. Una de estas aplicaciones es la secuenciación de amplicones y la determinación de alelos, lo cual se realizó en la presente investigación. Como modelo de estudio se empleó el gen de la S-RNasa del capulí, una especie de interés comercial en el Ecuador. La relevancia de estudiar este gen radica en que es uno de los determinantes de la autoincompatibilidad gametofítica en el capulí, por tanto, sus alelos delimitan qué cultivares son compatibles entre sí. El capulí es tetraploide y el gen de la S-RNasa posee alelos que tienen el mismo tamaño pero diferente secuencia, lo que determina incertidumbre en la discriminación de alelos utilizando marcadores moleculares tradicionales. Tomando en cuenta lo mencionado, se considera que el MinION, al generar lecturas de cadenas individuales, permitirá la distinción de los alelos S de forma efectiva. Durante la investigación, se amplificó el gen de la S-RNasa de 12 muestras de capulí, utilizando *primers* con etiquetas nucleotídicas únicas (códigos de barras), esto permitió optimizar la secuenciación al incluir varias muestras en una sola reacción. Posteriormente, el producto de amplificación fue purificado y secuenciado en el MinION. Durante el análisis, las lecturas fueron clasificadas de acuerdo con el código de barras de cada muestra. Luego las lecturas se filtraron y agruparon según su similitud de secuencia. Con las secuencias consenso de los grupos obtenidos, se identificaron los alelos representados en los individuos analizados. Los alelos identificados coinciden con los previamente reportados mediante la metodología CAPS. Se identificaron además dos nuevos alelos, sugiriendo mayor sensibilidad en la técnica utilizada. Las secuencias de alelos obtenidas son equiparables a las referencias, adquiridas mediante secuenciación Sanger. Esto demuestra la eficacia del análisis bioinformático realizado. El sistema de etiquetado y el análisis bioinformático diseñado son herramientas escalables a estudios similares. Esta investigación representa un avance en el campo de la Biotecnología en Ecuador, tomando en cuenta el uso novedoso del MinION.

Palabras clave: secuenciamiento nanoporo, alelos S, gen de la S-RNasa, bioinformática, código de barras, modificación de *primers*, etiquetado de muestras

ABSTRACT

Nanopore sequencing with the MinION device is a novel strategy that allows the real-time analysis of results. The MinION, as a portable equipment and as an inexpensive technology, has a wide range of applications. One of these applications is the sequencing of amplicons and the determination of alleles, what was done in the present investigation. As a study model, the S-RNase gene from Black Cherry was used. This species has a commercial interest in Ecuador. The relevance of studying this gene lies in the fact that it is one of the determinants of gametophytic self-incompatibility in Black Cherry; therefore, its alleles define which cultivars are compatible. Black Cherry is a tetraploid species, and the S-RNase gene has alleles that have the same size but different sequence, this fact determines uncertainty in allele discrimination using traditional molecular markers. Therefore, it is considered that the MinION device has the potential to discriminate the S alleles effectively, because of its capacity for generating single-chain readings. During the investigation, the S-RNase gene was amplified from 12 samples of Black Cherry, using primers modified with barcodes, this allowed to optimize sequencing by including several samples in a single reaction. Subsequently, the PCR products were purified and sequenced in the MinION. As part of the analysis, the readings were classified according to the barcode, then, the readings were filtered and grouped conforming to their sequence similarity. With the consensus sequences of those groups, it was possible to identify the alleles represented in each individual. The identified alleles coincide with the ones previously reported using CAPS as molecular marker. Two new alleles were also identified, suggesting a greater sensitivity of the used technique. The allele sequences obtained are equivalent to the references acquired by Sanger sequencing. This demonstrates the effectiveness of the bioinformatic pipeline designed. The multiplex sequencing system and the bioinformatic analysis are tools that can be applied to similar studies. This research represents an advance in the Biotechnology field in Ecuador, considering the innovative use of the MinION.

Key words: nanopore sequencing, S alleles, S-RNase gene, bioinformatics, barcode, primer modification, multiplexing

TABLA DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

La secuenciación de ácidos nucleicos por nanoporo representa un avance importante en los estudios moleculares y genómicos. Esta tecnología se basa en la lectura directa de moléculas de ADN de cadena simple, sin la necesidad de amplificación por PCR o etiquetado químico-colorimétrico de la muestra (Metzker, 2010). Esta tecnología ha sido una de las más populares para secuenciación desde su lanzamiento en el 2014 por parte de la compañía Oxford Nanopore Technologies (ONT) (Heather & Chain, 2016).

Un nanoporo consiste en un agujero de pocos nanómetros de diámetro, formado por proteínas celulares transmembranarias. Durante la secuenciación, dicho nanoporo está sumergido en un fluido conductor al que se le aplica un voltaje, por tanto, existe constantemente un flujo de iones a través del nanoporo. Este flujo iónico es interrumpido por el paso de los nucleótidos, que disminuyen la intensidad de la corriente en un período de tiempo proporcional a la longitud y naturaleza del ácido nucleico. Mediante su paso por el nanoporo, cada nucleótido altera la corriente de manera muy específica, lo cual puede detectarse y cuantificarse (van Dijk et. al., 2018).

La secuenciación por nanoporo permite generar datos en tiempo real, que se pueden visualizar en términos de número de lecturas y distribución de tamaños (Benítez et. al., 2016). Una de las principales ventajas de esta secuenciación, a comparación de otras tecnologías populares como Sanger o Illumina, es su alto rendimiento por la longitud de lecturas que genera (Laszlo et. al., 2014). Las plataformas de ONT para secuenciación de ácidos nucleicos y proteínas incluyen GridION, PromethION y MinION (Varshney et. al., 2018). Estas plataformas incluso permiten la detección de cambios epigenéticos en las secuencias de ADN, mediante el uso de herramientas computacionales como NanoMod (Liu et.al., 2019).

El MinION™ es una de las plataformas de secuenciación nanoporo más utilizadas debido a su versatilidad. Este es un dispositivo que permite la generación de datos de forma

mucho más sencilla y rápida, debido a la propia naturaleza del equipo (Heather & Chain, 2016). El MinION posee un tamaño similar al de una memoria flash, y es operado desde una computadora a través de un puerto USB 3.0 (Benítez et. al., 2016). Comercialmente, el MinION representa una tecnología poco costosa y accesible. Para ejecutar el dispositivo y obtener las secuencias, se necesita del software MinKNOW y de un *basecaller*, un programa que traduce el cambio de corriente en nucleótidos (ONT, 2020). La naturaleza compacta y portable del MinION ofrece la oportunidad de descentralizar la secuenciación, pudiendo incluso implementarse en el campo. Por tanto, esta tecnología puede revolucionar no solo la naturaleza de datos que se producen, sino también dónde, cuándo y por quién se producen (Heather & Chain, 2016).

Otra de las ventajas de la secuenciación con MinION es la posibilidad de realizar secuenciaciones con múltiples muestras, donde cada una esté etiquetada de forma específica (Stubbs et. al., 2020). El proceso de etiquetado funciona añadiendo un código de barras de naturaleza nucleotídica específico para cada una de las muestras (Karamitros & Magiorkinis, 2017). Posteriormente, durante la etapa de análisis post-secuenciación, las lecturas generadas podrán ser separadas y agrupadas de acuerdo con los códigos de barras añadidos, en un procedimiento que se conoce como resolución por códigos de barras. Una reacción con múltiples muestras permite hacer un uso más eficiente de las celdas de secuenciación, reduciendo de esta forma el costo de secuenciación por muestra analizada. Esto es útil principalmente cuando la cantidad de datos requeridos por muestra es menor que la cantidad total de datos que se pueden generar a partir de una sola celda de secuenciación (Currin et. al., 2019).

Los kits ofrecidos por ONT varían en la forma de añadir los códigos de barras, algunos incluyen el uso de una transposasa, que corta las cadenas de ADN y añade los códigos de barras, como el *Rapid Barcoding Kit* (SQK-RBK004). Otros, requieren la necesidad de PCR

como el *Low Input by PCR Barcoding Kit* (SQK-LWB001) (ONT, 2020). En el presente estudio se evalúa la efectividad de añadir códigos de barras empleando un método alternativo, que consiste en la modificación de *primers*, sin la necesidad de utilizar kits comerciales para ello.

De forma específica, en la investigación se utiliza el MinION para la secuenciación de amplicones y la determinación de alelos. Para ello, se empleará como modelo de estudio el sistema de alelos S del gen de la S-RNasa del capulí (*Prunus serotina* subsp. *capuli*). El interés de estudiar este locus radica en que es uno de los determinantes de la autoincompatibilidad gametofítica en el capulí, por tanto, sus alelos determinan qué cultivares son compatibles entre sí (Broothaerts et. al., 1995). El capulí es una especie de interés comercial en el Ecuador, y para garantizar un rendimiento adecuado en la producción de frutos, se deben sembrar individuos molecularmente compatibles (Gordillo et. al., 2013), es aquí donde radica la importancia de dilucidar los alelos S.

Este sistema de alelos S presenta varios desafíos que determinan que la secuenciación nanoporo sea la metodología de haplotipado más viable y efectiva. El primer desafío se relaciona a la tetraploidía del capulí, que determina que cada individuo pueda presentar hasta 4 alelos diferentes en un locus (Marquis, 2019). En estudios previos, se detectó que existen alelos con un mismo tamaño de amplicón, pero diferente secuencia (Gordillo et. al., 2013). Esta característica dificulta la distinción de algunos alelos mediante electroforesis en geles de agarosa. Un segundo desafío es la presencia de regiones altamente conservadas a lo largo del gen (Wu et. al., 2013), lo que complica el ensamblado y reconstrucción de la secuencia completa de cada alelo, en caso de utilizar métodos de secuenciación que fragmenten los amplicones. Estructuralmente, el gen de la S-RNasa se compone de cinco regiones conservadas, denominadas secuencialmente como C1-C5, posee además dos intrones con amplio polimorfismo y una región hipervariable (RHV) (Wu et. al., 2013) (Anexo A). Otro

desafío se asocia al tamaño relativamente grande de este gen, que se halla entre 1 kb y 2 kb (Wu et. al., 2013). La tecnología de secuenciación con nanoporo permite secuenciar productos largos de PCR (> 2 kb) (Laver et. al., 2015), siendo por tanto viable para la investigación.

El hecho de que el gen de la S-RNasa presente alelos con un mismo tamaño, pero diferente secuencia, conllevó al desarrolló de la metodología de haplotipado CAPS (Cleaved Amplified Polymorphic Sequence) (Correa, 2018). Este enfoque utiliza enzimas de restricción para digerir los amplicones de la región hipervariable del gen, y de esta forma diferenciar alelos que tengan un mismo tamaño, pero diferente secuencia (Correa, 2018). Sin embargo, esta diferenciación no siempre es posible. Además, el método CAPS necesita que la secuencia de interés tenga sitios de reconocimiento para enzimas de restricción comerciales, y que en dichos sitios existan polimorfismos. También, CAPS es un sistema difícil de automatizar, lo que se refleja sobre todo durante el análisis (Konieczny & Ausubel, 1993). Lo mencionado, limita el uso de esta metodología de haplotipado, a la vez que supone incertidumbre en la determinación de algunos alelos.

Por estas razones, se considera que la secuenciación es la alternativa más viable para determinar los alelos del gen de la S-RNasa del capulí. Previamente, en la investigación de Correa (2018), los alelos del gen también fueron secuenciados mediante el método Sanger. No obstante, en la secuenciación de amplicones con Sanger, no es posible distinguir varios alelos presentes en una misma banda, ya que el resultado es una secuencia consenso, generalmente ruidosa (Heather & Chain, 2016). En este aspecto, ya que la secuenciación nanoporo genera lecturas de cadenas individuales (Laver et. al., 2015), se pueden obtener los diferentes alelos representados en una muestra, independientemente si los alelos tienen el mismo tamaño o no.

De acuerdo con lo planteado, se considera que existe evidencia suficiente para afirmar que la secuenciación en el MinION es un método viable para detectar los alelos el gen de la S-RNasa del capulí. No obstante, a pesar de las múltiples ventajas que trae consigo el uso del

dispositivo portable MinION, es importante recalcar que los perfiles observados de calidad de las lecturas son relativamente bajos (Heather & Chain, 2016). Esto determina la necesidad de análisis post-secuenciación estrictos, que permitan filtrar y estudiar lecturas fiables. Para ello, se encuentran disponibles varios programas de análisis bioinformático, escritos específicamente para secuenciación nanoporo. Sin embargo, no existe un programa que integre todas las funciones necesarias para el análisis que se quiere realizar. Es decir, un programa que permita el manejo de datos crudos de secuenciación y la determinación de alelos. Por ello, uno de los retos del presente estudio fue minimizar el uso de programas bioinformáticos y aún así garantizar resultados viables, que validen el uso del MinION en la determinación de alelos de una especie tetraploide.

De esta forma, el objetivo principal de la investigación es utilizar el dispositivo MinION para secuenciar amplicones y determinar alelos utilizando un modelo tetraploide. Además, se determinará la viabilidad de la modificación de *primers* como alternativa para añadir códigos de barras a los amplicones, de forma que se pueda optimizar la secuenciación. Por último, se diseñará una metodología de análisis bioinformático que permita estudiar lecturas fiables provenientes del MinION y determinar alelos en el modelo utilizado.

MÉTODOS

Selección de muestras

La presente investigación se enfocó en el análisis de 12 accesiones de capulí, las cuales fueron caracterizadas previamente mediante la metodología CAPS y secuenciadas utilizando tecnología Sanger (Correa, 2018). Dichas accesiones son representativas de una amplia variabilidad de alelos S del capulí de la Sierra Ecuatoriana. Utilizar estas muestras permitió la comparación directa de los resultados obtenidos con los reportados previamente utilizando la metodología de secuenciamiento Sanger y el método CAPS.

Las muestras de ADN fueron obtenidas de la colección de ejemplares de *P. serotina* almacenados en el Laboratorio de Biotecnología Vegetal de la Universidad San Francisco de Quito. Los nombres específicos y el origen de las muestras utilizadas pueden encontrarse en el Anexo B.

Diseño de *primers*

Para secuenciar el gen de la S-RNasa en su totalidad se diseñaron *primers* para las regiones extremas del gen (SPR y C5; Anexo C) utilizando el programa Primer3 (Primer3web, 2019). En el programa se comprobó la Tm (Temperatura de fusión), el contenido de GC, y la probabilidad de formación de estructuras secundarias entre los *primers*. Se garantizó que los *primers* cumplieran con todos los parámetros adecuados de diseño. El *primer forward* corresponde a la región SPR y el *reverse* a la región C5.

Para el etiquetado específico de los amplicones de las 12 muestras, el *primer forward* se modificó en el extremo 5´ con una cola de 8 nucleótidos, específica para cada una de las muestras. Estos *primers* modificados también cumplieron con los parámetros evaluados en Primer3 (Primer3web, 2019).

Los *primers* se sintetizaron a través de Midland Certified Reagent Co. (Estados Unidos) y se estandarizaron para determinar sus condiciones óptimas de amplificación.

Amplificación y purificación del gen de la S-RNasa

El gen de la S-RNasa de las 12 muestras seleccionadas fue amplificado utilizando las siguientes condiciones de PCR: Buffer 1X, Cloruro de Magnesio 1.5 mM, dNTP´s 0.2 mM, *Primer forward* modificado 0.5 µM, *Primer reverse* 0.5 µM, Taq Polimerasa Platinum (Invitrogen) 1U y ADN 40 ng. El volumen final de la reacción en cada caso fue de 25 µl.

El programa de termociclado utilizado consistió en: Denaturación inicial a 94°C por 2 minutos; seguido de 35 ciclos de: Denaturación a 94 °C por 1 minuto, Annealing a 59 °C por 1 minuto, Elongación a 68 °C por 4 minutos; y una Elongación final a 68 °C por 10 minutos. Los amplicones fueron visualizados mediante electroforesis en geles de agarosa al 1.5%, para comprobar la integridad de la amplificación, y observar si el tamaño y número de bandas coincidía con el esperado para cada una de las muestras. Las condiciones de electroforesis fueron de 80 voltios durante 30 minutos y se utilizó TBE 1X como buffer de corrida. Los amplicones fueron visualizados utilizando SYBR Safe (Invitrogen) y su tamaño fue estimado utilizando como referencia el Ladder 100 pb (Promega).

Los productos de PCR de las 12 muestras fueron purificados utilizando el kit *Wizard® SV Gel and PCR Clean-Up System* (Promega). Se utilizó la metodología de purificación por centrifugación, empleando columnas. Las muestras fueron resuspendidas en un volumen final de 25 µl de agua de PCR y la concentración de ADN recuperado se determinó utilizando MultiSkan (Thermo Scientific).

Preparación de librerías

Para lograr obtener una mayor profundidad de secuenciamiento de los amplicones de cada una de las 12 muestras, se decidió realizar 2 ensayos de secuenciación independientes, con seis muestras cada uno.

En cada evento se mezclaron los seis productos de amplificación para obtener un pool de ADN, en un volumen total de 55 µl. Posteriormente, se utilizó el *Ligation Sequencing Kit*

SQK-LSK109 (ONT). El primer elemento de preparación de librerías consistió en la reparación de los extremos de los amplicones, en donde se utilizó el kit *Next Companion Module for ONT* (New England Biolabs). Para la purificación se utilizó el kit *DNA Purification SPRI Magnetic Beads* (abm).

Los adaptadores de secuenciación se ligaron y el producto se purificó nuevamente, utilizando *SPRI Magnetic Beads* (abm) y el *Short Fragment Buffer* (ONT) como buffer de lavado. Finalmente, el producto purificado fue resuspendido en 15 µl de buffer de elución y la concentración de ADN fue cuantificada mendiante fluorometría utilizando Qubit (Thermo Scientific).

Secuenciación

Como primer paso, se verificó que en la celda de secuenciación a utilizar existiera al menos el número mínimo de poros viables (800). Una vez comprobado esto, se cargó el buffer y la librería en la celda de secuenciación, según las indicaciones del protocolo *Priming and loading the SpotON flow cell* (ONT).

La adquisición de datos y el *basecalling* (traducción de cambio de corriente en nucleótidos) se realizó con el software MinKNOW (ONT). Sin embargo, durante la primera secuenciación (muestras 1-6) este proceso se detuvo; por tanto, para continuar el *basecalling* se utilizó el programa Guppy (ONT). Para la segunda secuenciación se utilizó la misma celda de secuenciación, previamente lavada según el protocolo *Washing flow cells* (ONT).

Análisis

Se analizó la calidad promedio de todas las lecturas obtenidas y la distribución de tamaños, utilizando el programa FastQC (Andrews & Krueger, 2019). En el siguiente paso, las lecturas fueron separadas de acuerdo con el código de barras colocado, utilizando el programa Porechop (PoreCamp Australia 2017). En este programa se probaron dos metodologías de separación. Una consistió en colocar en la base de datos del programa toda la región del *primer*

con los 8 nucleótidos diferentes para cada muestra, y en usar el valor de 75% en el parámetro Barcode_Threshold. La segunda metodología de separación empleó solamente 5 nucleótidos del *primer* y los 8 nucleótidos específicos de cada muestra, en este caso el valor del parámetro Barcode_Threshold se colocó en 100%.

Cuando se tuvieron las lecturas separadas de acuerdo con el código de barras, estas se filtraron por calidad y tamaño, utilizando el programa NanoFilt (De Coster et. al., 2018). Se eliminaron todas las lecturas con calidad menor a Q=12 y con un tamaño menor a 800 pb. Se eligió Q=12 como umbral de calidad ya que es un valor promedio para secuenciación con MinION (Zanchetta & Manno, 2017). El umbral de tamaño se escogió en 800 pb para garantizar en su mayoría la integridad del gen en las lecturas a analizar. Complementario a esto, se obtuvo la cantidad de información remanente luego de aplicar los filtros. Esto se realizó con el programa SeqKit (Shen et. al., 2016).

Después del filtrado, las lecturas se reordenaron según su tamaño de mayor a menor, utilizando el paquete BBMap, de BBtools (Bushnell, 2015). Este paso es esencial para que el siguiente programa (Meshclust) reconozca los archivos. En MeshClust (James et. al., 2018) las lecturas se agruparon utilizando un porcentaje de identidad de 90%. Esto implica que, al alinearlas, las lecturas deben tener este porcentaje de similitud entre sí para colocarse en el mismo grupo.

Como resultado del agrupamiento se obtuvo un archivo con el nombre de las lecturas asociadas a cada uno de los grupos, estas lecturas se seleccionaron y colocaron en un nuevo archivo, lo cual fue realizado con la función -filterbyname de BBMap (Bushnell, 2015). Para maximizar la probabilidad de determinar los alelos correctos en cada muestra se eligieron aquellos grupos representativos, es decir, los que tuvieran un número significativo de lecturas asociadas, donde se estableció un umbral de 100 lecturas para cada grupo. Las lecturas de los grupos representativos se etiquetaron según el código de barra y el grupo al que pertenecían,

lo cual se realizó con SeqKit (Shen et. al., 2016). La manipulación de lecturas mencionada hasta el momento se realizó a nivel de terminal bajo el sistema operativo Linux.

Las lecturas de cada uno de los grupos se alinearon en el programa UGENE (Okonechnikov et. al., 2012), utilizando el algoritmo MUSCLE. Después del alineamiento, se eliminaron las columnas con 60% de *gaps* utilizando la opción *Edit.* Esto permitió la obtención de una secuencia consenso de cada uno de los grupos, la cual fue colocada en un nuevo archivo ".fasta."

La integridad y el sentido de las secuencias consenso obtenidas se analizó en MEGA (Kumar et. al., 2018). El sentido se analizó mediante el uso de *primers* en disposición 5´- 3´. A las secuencias identificadas como *reverse* se les cambió el sentido mediante la función *Reverse Complement* de MEGA. La naturaleza de las secuencias consenso se confirmó en la base de datos del NCBI, utilizando la herramienta BLAST (nucleótido-nucleótido) (NCBI, 2019), donde todos los parámetros de búsqueda se mantuvieron en modo *default*.

Las secuencias consenso validadas y las referencias reportadas por Correa (2018) se alinearon en MEGA con el algoritmo ClustalW. Con las secuencias alineadas se realizó un dendrograma, utilizando el método de *Maximum Likelihood*. Así mismo, se delimitaron las regiones del Intrón I y C2-C3 en todas las secuencias, con esto también se realizaron alineamientos y los dendrogramas correspondientes, utilizando la misma metodología empleada para el todo el gen (Anexo D). La información obtenida en los dendrogramas mostró qué alelos estaban representados en cada una de las muestras. Esto fue posteriormente comprobado mediante alineamientos en MEGA y con CAPS *in silico*, utilizando el programa Genome Compiler (Genome Compiler Corporation, 2015) con las enzimas: RsaI, MboI y HinfI. Los patrones CAPS obtenidos se compararon con los reportados previamente por Correa (2018) (Anexo E). Los detalles y comandos del análisis bioinformático realizado se encuentran en el Anexo G.

RESULTADOS

Amplificación y purificación de amplicones

Se modificó el extremo 5´ del *primer forward* con ocho nucleótidos específicos para cada una de las 12 muestras amplificadas, información expuesta en la **Tabla 1**. Los *primers* modificados demostraron ser efectivos ya que todas las muestras amplificaron y generaron el tamaño y número de bandas esperado (**Figura 1**). No obstante, sí se observó la presencia de dímeros en la amplificación.

Los parámetros de calidad 260/280 y 260/230 para los productos amplificados, promediaron respectivamente 2.24 (rango: 1.97-2.34) y 2.40 (rango: 1.40-2.89). Los dos pools de amplicones obtenidos para cada evento de secuenciación tuvieron una cantidad final de ADN de 217 femtomoles, adecuada para la secuenciación de amplicones en el MinION.

Secuenciación de amplicones

La secuenciación se realizó en dos eventos independientes utilizando una misma celda de secuenciación; en cada evento se analizaron 6 genotipos simultáneamente. La primera secuenciación se realizó con 1346 poros activos, mientras que la segunda secuenciación se realizó con 734 poros activos. Respectivamente, el primer y segundo evento de secuenciación generaron 1.6 y 5 GB de información. En ambos eventos, el tamaño promedio de lectura fue de 700 pb, con un mínimo de 1 pb y un máximo de 2.5 kb. La calidad promedio obtenida en las lecturas fue de Q=12.

Separación de las muestras por códigos de barras

En esta fase de manejo de lecturas se utilizó el programa Porechop. Se emplearon dos metodologías de separación, una de las cuales no fue lo suficiente robusta y efectiva. Se identificó que la resolución por códigos de barras era inespecífica al utilizar el valor *default* de 75% en el parámetro Barcode_Threshold, y emplear todo el *primer* con los 8 nucleótidos específicos para cada muestra. En este caso, no se encontraron lecturas asociadas a los códigos

de barras 7 y 8, y, luego de las diferentes fases de análisis, no se detectaron los alelos esperados para cada una de las muestras. En la **Tabla 2** se detallan los alelos obtenidos al utilizar la metodología de separación no robusta.

La segunda metodología de separación, que consistió en el uso de los 8 nucleótidos del código de barras con 5 nucleótidos del *primer*, y el valor de 100% en el parámetro Barcode_Threshold, sí fue robusta. En este caso, se encontraron lecturas asociadas a cada código de barras y los alelos obtenidos luego de los diferentes pasos de análisis sí fueron consecuentes con lo esperado (**Tabla 3**). No obstante, es relevante mencionar que esta separación exigente determinó una pérdida de información considerable, ya que de los 6.6 GB que se tenían originalmente, solo se lograron recuperar 700 MB. Así mismo, tras el proceso de resolución por códigos de barras, se detectó una distribución heterogénea de la cantidad de lecturas asociadas a cada código de barras (**Figura 2**).

Identificación y validación de alelos

Posterior al proceso de separación por códigos de barras, se empleó el programa NanoFilt para seleccionar únicamente secuencias fiables y de alta calidad. El proceso de filtración eliminó un 61% del total de lecturas obtenidas tras el proceso de separación (**Figura 2**). Por otro lado, en el paso de agrupamiento, donde se utilizó el programa MeshClust y una identidad de 90%, se esperaban un máximo de 4 grupos por muestra, correspondientes a los 4 posibles alelos. Sin embargo, se obtuvieron en todos los casos un mayor número de grupos, con un promedio de 75 por muestra.

El análisis del sentido de las secuencias consenso reveló que existen SNPs en la región de los *primers* de las diferentes partes del gen. Así mismo, con la herramienta BLAST se comprobó de forma efectiva que todas las secuencias consenso obtenidas correspondían al gen de la S-RNasa.

Los dendrogramas realizados en MEGA permitieron identificar de forma inequívoca los alelos representados en cada una de las muestras. Los tres dendrogramas realizados (con todo el gen, con la región C2-C3, con el intrón I) fueron consecuentes entre sí (Anexo D). En la **Tabla 3** se presentan los alelos S encontrados para cada genotipo analizado. Estos coinciden con los alelos previamente reportados por Correa (2018). La secuenciación con MinION y el análisis bioinformático obtenido también permitieron detectar dos nuevos alelos (SX, SY) (Anexo F) y obtener la secuencia completa del alelo S6, del que previamente solo se disponía la secuencia del intrón I (Anexo F). La **Figura 3** muestra un ejemplo de los alineamientos llevados a cabo para validar los alelos S de un genotipo determinado. Los alineamientos realizados validaron la efectividad del análisis, que permitió obtener secuencias equiparables a las referencias, adquiridas mediante secuenciación Sanger (Correa, 2018). La restricción enzimática *in silico* de la región C2-C3 también arrojó patrones consecuentes con lo esperado según la investigación de Correa (2018) (Anexo E).

Delimitar las regiones del gen permitió determinar su tamaño en los alelos identificados. Los tamaños obtenidos en las diferentes regiones son consecuentes con lo reportado para la secuencia de la S-RNasa en otras especies del género *Prunus* sp. Cabe destacar el tamaño de la región hipervariable, que se encontró en el rango de 357 pb a 1264 pb. El alelo más largo corresponde al S8 (1926 pb), y el más corto al S14 (940 pb).

DISCUSIÓN

El MinION es una plataforma nanoporo que permite de forma efectiva la secuenciación amplicones y la determinación de alelos. Esto se evidenció durante la investigación, ya que, tras amplificar y secuenciar el gen de la S-RNasa del capulí, se lograron determinar de forma inequívoca los alelos representados en 12 muestras (Tabla 3). Los alelos identificados corresponden a los previamente detectados mediante CAPS (Correa, 2018). Además, la herramienta utilizada, en conjunto con el análisis realizado, permitió la identificación de dos nuevos alelos en el pool de muestras empleado (Anexo F), no detectados en estudios previos. Esto sugiere una mayor sensibilidad del método utilizado en la determinación de alelos a comparación de las metodologías de haplotipado empleadas previamente. Así mismo, se debe destacar la obtención de la secuencia completa de alelo S6 (Anexo F), de la que previamente solo se disponía el intrón I. Estos resultados confirman la hipótesis planteada al inicio de la investigación, ya que se demostró que el MinION es una plataforma que permite la distinción de alelos en un modelo tetraploide. Esta distinción fue posible independientemente de la existencia de diferentes alelos con igual tamaño en una misma muestra. Así mismo, se obtuvieron secuencias completas del gen de la S-RNasa, de tamaño relativamente grande. Esto confirmó la característica de la secuenciación nanoporo para generar de lecturas de cadena larga y completa (Heather & Chain, 2016). No obstante, como se verá a lo largo de la discusión, para lograr obtener los alelos adecuados en cada una de las muestras, fue necesario realizar análisis bioinformáticos estrictos, que determinaron una pérdida considerable de información.

El éxito de la secuenciación con el MinION comienza desde el ADN utilizado como material de entrada. Es importante que este ADN cumpla con los requisitos de calidad y cantidad. Utilizar poco o demasiado ADN, o de baja calidad (altamente fragmentado o con contaminantes remanentes), puede afectar la preparación de librerías y la secuenciación, ya que los nanoporos pueden bloquearse por la presencia de contaminantes con un tamaño mayor a la

molécula de ADN (Zhang et. al., 2018). En el estudio, los valores de calidad y cantidad obtenidos luego de la purificación de los amplicones, revelaron que el material de entrada cumplía con las características necesarias. Por tanto, se esperaron resultados de secuenciación adecuados.

Sin embargo, durante el análisis de distribución de tamaños, se reveló una posible fragmentación del ADN en la preparación de librerías, ya que se obtuvieron muchas lecturas con tamaños inferiores a lo esperado. Se considera que este resultado se asocia a la preparación de librerías ya que se observaron los tamaños de banda esperados en la amplificación y el MinION genera lecturas de cadena completa (Laver et. al., 2015). Así mismo, la filtración de lecturas de acuerdo con tamaño y calidad reveló que la mayoría de las lecturas separadas por códigos de barras no tenían un *quality score* mayor o igual a Q=12, y un tamaño mayor o igual a 800 pb (Figura 3). De forma específica, se perdió el 61% de lecturas debido a la filtración. Sin embargo, si se utilizaban las lecturas sin filtrar, se corría el riesgo de determinar alelos incorrectos en los individuos, ya que mientras menor sea la calidad de las lecturas, se tiene mayor probabilidad de trabajar con secuencias incorrectas, debido a un *basecalling* no preciso (Laver et. al., 2015). De acuerdo con lo planteado, se considera que la filtración realizada es un paso necesario para garantizar análisis posteriores válidos, con lecturas de buena calidad y que poseen en su mayoría el gen íntegro.

En la investigación también se evaluó la viabilidad de la modificación de *primers* como método para añadir códigos de barras a los amplicones, de forma que estos pudieran ser secuenciados en conjunto. Para el etiquetado de los amplicones, el *primer* SPR (*forward*) fue modificado en su extremo 5´ con ocho nucleótidos, que constituyeron un código de barras para cada una de las 12 muestras amplificadas (Tabla 1). El tamaño y número de bandas observadas en la amplificación coincide con lo esperado para cada muestra (Figura 1), sugiriendo que los *primers* diseñados son adecuados para la amplificación del gen de la S-RNasa. De acuerdo con

lo obtenido, se considera que la presencia del código de barras no interfirió en la amplificación. Estos resultados coinciden con lo reportado por Berry et. al. (2011), quienes lograron un producto de amplificación adecuado con *primers* modificados. Sin embargo, a pesar de haber diseñado *primers* con baja probabilidad de formación de estructuras secundarias, en la amplificación sí se observó la presencia de dímeros, que se observan como bandas de bajo peso molecular (< 100 pb) al final de los carriles. No obstante, la formación de estos dímeros no afectó al producto amplificado. Estos resultados evidencian que, a nivel de amplificación, la metodología de etiquetado fue satisfactoria.

La efectividad de la metodología de etiquetado se evidenció también durante el proceso de separación de muestras utilizando la metodología robusta, ya que se encontraron de forma efectiva lecturas asociadas a cada uno de los códigos de barras. La resolución por códigos de barras inespecífica y no robusta, observada al utilizar un Barcode_Threshold de 75% y toda la región del *primer* con el código de barras, puede deberse a que este consistía solamente en 8 nucleótidos. Este tamaño es pequeño con respecto al tamaño del *primer* de 27 nucleótidos, compartidos por todas las lecturas. Por tanto, al utilizar un umbral de 75%, es posible que este valor se alcanzara considerando únicamente la región del *primer* para separar a una lectura en un código de barras determinado, lo cual lleva a una separación errónea. Por otro lado, la no identificación de los códigos de barras 7 y 8 pudiera explicarse por el propio algoritmo del programa, que, debido al tamaño y a la similitud existente entre estos, encontraba difícil la separación de algunas lecturas. De hecho, el programa describe que cuando una lectura posee identidad similar con dos códigos de barras, dicha lectura no se asigna a ningún grupo (PoreCamp Australia, 2017). De acuerdo con estos resultados, se recomienda no colocar la mayoría del *primer* en la base de datos del programa Porechop. Esto disminuiría la interferencia causada por los nucleótidos del *primer* durante la asignación de lecturas a un código de barras específico, posibilitando una separación más fiable.

En la resolución por códigos de barras robusta se utilizó un Barcode_Threshold de 100%, esto implica que una lectura debe poseer 100% de identidad con un código de barras para separarse en su grupo correspondiente. Este valor de separación estricto determinó una pérdida considerable de información, lo que disminuye la eficiencia en el manejo de lecturas y en el proceso global de análisis. Por tanto, se considera que el porcentaje de identidad que se utilice en el parámetro Barcode_Threshold debe tomar en cuenta la tasa de error del MinION, para evitar la pérdida de información asociada a errores de lectura en la región del código de barras.

Así mismo, tras la resolución por códigos de barras robusta, se observó una distribución heterogénea del número de lecturas asociadas a cada uno (Figura 3). En la investigación de Binladen et. al. (2007), los autores también detectaron un sesgo en la distribución de lecturas por códigos de barras, lo que se asocia a su composición de bases y a la metodología de separación del programa que se utilice. De forma específica, los autores encontraron que los *primers* modificados con más citosinas estuvieron fuertemente sobre-representados, mientras que los modificados con más timinas estuvieron sub-representados. Un resultado similar se observó en la presente investigación, donde existen algunos códigos de barras sobre-representados (como el 6 ,7 y 11), y otros sub-representados (como el 4 y el 10). Estos resultados sugieren que el diseño de códigos de barras no solo debe garantizar una secuencia específica para cada muestra, sino también que estos deben tener una distribución homogénea de nucleótidos en su secuencia cuando se comparen entre sí. A pesar de este sesgo de distribución, los experimentos y análisis realizados mostraron que modificar el extremo 5´ del *primer forward* es útil para etiquetar productos de PCR homólogos (Binladen et. al., 2007). La metodología utilizada permitió optimizar la secuenciación de amplicones en el MinION, al permitir la inclusión de varias muestras en una misma corrida de secuenciación.

En cuanto a los resultados del paso de agrupamiento, en este se esperaban un máximo de cuatro grupos por muestra, correspondientes a los 4 posibles alelos. Sin embargo, en todos los casos se obtuvo un mayor número de grupos. Este resultado pudiera relacionarse a la obtención de muchas lecturas que difieren en sitios puntuales, debido a la tasa de error del MinION y al procedimiento de *basecalling*, que no es 100% preciso (Kono & Arakawa, 2019). Según Laver et. al. (2015) la tasa de error del MinION después del *basecalling* es de alrededor de 38,2%. Otra característica del MinION asociable a lo observado es la baja profundidad de secuenciación, lo que se debe a la propia naturaleza de la tecnología nanoporo, que no lee una misma cadena dos veces (Malmberg et. al., 2019). No obstante, la profundidad de secuenciación depende del material de entrada. En este caso, el gen de la la S-RNasa fue amplificado antes de insertarse al secuenciador, existiendo múltiples copias de este. Por ello, la baja profundidad del MinION no es una explicación plausible para lo observado en el agrupamiento. La obtención de más grupos de los esperados también pudiera relacionarse con errores de procesividad de la polimerasa durante la amplificación, esto es poco probable tomando en cuenta que se utilizó la Taq Polimerasa Platinum, que reporta alta fidelidad de amplificación (Filges et. al., 2019).

La probabilidad de determinar los alelos correctos en cada muestra se maximizó al analizar solo los grupos representativos, y mediante la obtención de las secuencias consenso. Esto último permitió determinar la base más probable en cada posición del gen para cada uno de los grupos obtenidos (Okonechnikov et. al., 2012). Se considera que la obtención de secuencias consenso permite flexibilizar los valores de calidad con los que se trabaje, ya que en cada grupo se obtendrá la secuencia consenso más probable. Al permitir valores de calidad más bajos en las secuencias que se analicen, la pérdida de información será menor, lo que aumenta la eficiencia global del proceso de análisis.

En el paso final de determinación de alelos, se agruparon entre sí secuencias consenso provenientes de diferentes grupos, esto sugiere que se pudiera disminuir el porcentaje de identidad utilizado en el agrupamiento (90%). Utilizar este porcentaje de identidad probablemente también determinó la formación de más grupos de los esperados. La tendencia es que al aumentar el porcentaje de identidad se aumenta el número de grupos obtenidos (James et. al., 2018). No obstante, si bien esto garantiza alta similitud entre las lecturas de un mismo grupo, se corre el riesgo de perder información al no considerar lecturas no asociadas a un grupo representativo. Por otro lado, si se baja el porcentaje de identidad, se tiene el riesgo de colocar alelos distintos dentro de un mismo grupo. De acuerdo con lo planteado, se recomienda un rango de 85% a 90% en el parámetro -id de MeshClust (James et. al., 2018).

En la determinación de alelos a partir de las secuencias consenso, se comprobó la utilidad de realizar dendrogramas entre estas y las secuencias de referencia. Esto simplificó el proceso ya que se visualizó directamente con qué referencias se agruparon las secuencias consenso, sin la necesidad de programas más especializados para ello. Los tres dendrogramas realizados fueron consecuentes entre sí. Esto sugiere que para determinar los alelos del gen de la S-RNasa del capulí se pueden utilizar las secuencias provenientes del intrón I, de la región C2-C3 o la secuencia de todo el gen, siendo esto último lo óptimo. Los resultados de agrupamiento obtenidos en los árboles se comprobaron efectivamente mediante alineamientos (Figura 3) y CAPS *in silico* (Anexo E). Los alineamientos realizados evidenciaron que las secuencias de alelos obtenidas son equiparables a las referencias, generadas mediante secuenciación Sanger (Correa, 2018), que reporta una tasa de error menor que el MinION (Pfeiffer, et. al., 2018). Este resultado valida el manejo y el análisis de datos realizado.

El análisis bioinformático desarrollado, si bien evidenció ser útil en el proceso de determinación de alelos, posee ciertos puntos de mejora. Por ejemplo, se pudiera incluir un análisis más profundo de las lecturas obtenidas directamente de la secuenciación, que

determine menos pérdida de información. Sí existen programas que permiten el análisis de lecturas no procesadas, como por ejemplo NanoOK (Leggett et. al., 2016). Así mismo, para el análisis de calidad y distribución de tamaños, se recomienda un programa como el MinIONQC, especialmente diseñado para lecturas provenientes del secuenciador portable utilizado (Lanfear et. al., 2019). También se propone integrar en un solo programa las diferentes fases de análisis realizadas, ya que esto simplificaría la complejidad del estudio. El análisis desarrollado requiere de entrenamiento en varias interfaces y dependencias distintas, así como la transformación constante del formato de las lecturas.

Los resultados obtenidos con el modelo tetraploide utilizado, demostraron que el MinION es efectivo para determinar alelos, incluso cuando estos se visualizan como una misma banda en un gel de agarosa. Los resultados evidencian el potencial del MinION para discriminar entre productos aparentemente iguales. Esto puede extrapolarse a estudios de transcriptómica, para la identificación de diferentes variantes de *splicing* que estén presentes al mismo tiempo en una célula (Liu & Zhang, 2020). Este es un ejemplo de cómo la herramienta utilizada, en conjunto con un análisis adecuado, pueden ser empleados en diferentes investigaciones aplicadas. La optimización de la secuenciación mediante el etiquetado por modificación de *primers*, es también un sistema adaptable a estudios donde se quieran realizar secuenciaciones con múltiples productos de PCR homólogos. Sin embargo, para que la plataforma diseñada sea totalmente escalable y adaptable a otros estudios, se debe optimizar el uso de la cantidad de información proveniente del MinION.

CONCLUSIONES

El MinION, como dispositivo portable de secuenciación nanoporo, es efectivo para secuenciar amplicones y determinar alelos. Esto fue comprobado utilizando como modelo el sistema de alelos S del gen de la S-RNasa del capulí. En la investigación, se detectaron los alelos esperados para las 12 muestras utilizadas, y se lograron determinar dos nuevos alelos, no detectados en investigaciones previas en el capulí. Esto sugiere una mayor sensibilidad de la tecnología empleada a comparación de metodologías previas de haplotipado. Por tanto, se considera que la plataforma MinION es una herramienta que permite disminuir costos, tiempo y la necesidad de experiencia en la determinación de alelos en una especie tetraploide de interés comercial. En el estudio también se evidenció que el etiquetado por modificación de *primers* es un método factible para añadir códigos de barras nucleotídicos a los amplicones, esto permitió optimizar el uso de las celdas de secuenciación, disminuyendo el costo por muestra. El análisis bioinformático diseñado fue efectivo por diversas razones, que incluyen la determinación efectiva de alelos en una especie tetraploide y la obtención de secuencias equiparables a las referencias, obtenidas mediante secuenciación Sanger. Tanto la metodología de etiquetado, como el análisis bioinformático diseñado, son herramientas adaptables a investigaciones similares utilizando el MinION como secuenciador. No obstante, para lograr obtener los alelos adecuados en cada una de las muestras, fue necesario realizar análisis bioinformáticos estrictos, que determinaron una pérdida considerable de información. Esto implica que la herramienta diseñada es efectiva, pero no es eficiente en el uso de lecturas.

REFERENCIAS BIBLIOGRÁFICAS

Andrews, S., & Krueger, F. (2019). *FastQC: a quality control tool for high throughput sequence data.* Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Benítez, A., Portune, K., & Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the MinIONTM portable nanopore sequencer. *GigaScience*.

Berry, D., Mahfoudh, K., Wagner, M., & Loy, A. (2011). Barcoded Primers Used in Multiplex Amplicon Pyrosequencing Bias Amplification. *Applied and Environmental Microbiology*, 7846–7849.

Binladen, J., Thomas, M., Bollback, J., Panitz, F., Bendixen, C., Nielsen, R., & Willerslev, E. (2007). The Use of Coded PCR Primers Enables High-Throughput Sequencing of Multiple Homolog Amplification Products by 454 Parallel Sequencing. *Plos One*, doi: 10.1371/journal.pone.0000197.

Broothaerts, W., Janssens, G., Proost, P., & Broekaert, W. (1995). cDNA cloning and molecular analysis of two self-incompatibility alleles from apple. *Plant Molecular Biology*, 499–511.

Bushnell, B. (2015). *BBMap*. Retrieved from sourceforge.net/projects/bbmap/

Correa, L. (2018). *Caracterización molecular y diseño de marcadores moleculares CAPS para el gen de la S-RNasa en Prunus serotina subsp. capuli.* Quito: Trabajo de titulación presentado como requisito para la obtención de título Licenciada en Biología.

Currin, A., Swainston, N., & Dunstan, M. (2019). Highly multiplexed, fast and accurate nanopore sequencing for verification of synthetic DNA constructs and sequence libraries. *Synthetic Biology*, doi.org/10.1093/synbio/ysz025.

De Coster, W., D'Hert, S., Schultz, D., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data . *Bioinformatics, Volume 34, Issue 15*, 2666–2669.

Filges, S., Yamada, E., Ståhlberg, A., & Godfrey, T. (2019). Impact of Polymerase Fidelity on Background Error Rates in Next-Generation Sequencing with Unique Molecular Identifiers/Barcodes. *Scientific Reports*, doi: 10.1038/s41598-019-39762-6.

Genome Compiler Corporation. (2015). *Genome Compiler*. Retrieved from http://www.genomecompiler.com/

Gordillo, M., Tobar, J., Arahana, V., & Torres, M. (2013). Identificación de alelos S asociados con autoincompatibilidad en individuos de capulí (Prunus serotina subsp. capulí) mediante la amplificación del Intrón I del gen de la S-RNasa. *Avances en Ciencias e Ingenierías, 7*(1), B17-B23.

Guadalupe, J., Gutiérrez, B., Intriago, D., Arahana, V., Tobar, J., Torres, A., & Torres, M. (2015). Genetic diversity and distribution patterns of Ecuadorian capuli (Prunus serotina). *Biochemical Systematics and Ecology 60*, 67-73.

Heather, J., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 1-8.

James, B., Luczak, B., & Girgis, H. (2018). MeShClust: an intelligent tool for clustering DNA sequences . *Nucleic Acids Research*, doi.org/10.1093/nar/gky315.

Karamitros, T., & Magiorkinis, G. (2017). Multiplexed Targeted Sequencing for Oxford Nanopore MinION: A Detailed Library Preparation Procedure. *Next Generation Sequencing*, 43-51.

Konieczny, A., & Ausubel, F. (1993). A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers. *The Plant Journal, 4*(2), 403-10.

Kono, N., & Arakawa, K. (2019). Nanopore sequencing: Review of potential applications in functional genomics. *Development, Growth and Differentiation*, DOI: 10.1111/dgd.12608 .

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution*.

Lanfear, R., Schalamun, M., Kainer, D., Wang, W., & Schwessinger, B. (2019). MinIONQC: fast and simple quality control for MinION sequencing data. *Bioinformatics*, 523-525.

Laszlo, A., Derrington, I., & Ross, B. (2014). Decoding long nanopore sequencing reads of natural DNA. *Nature Biotechnoly*, 829–833.

Laver, T., Harrison, J., O'Neill, P., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification 3*, 1-8.

Leggett, R., Heavens, D., Caccamo, M., Clark, M., & Davey, R. (2016). NanoOK: multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. *Bioinformatics*, 142–144.

Liu, Q., Georgieva, D., Egli, D., & Wang, K. (2019). NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data. *BMC Genomics*, Article number: 78 .

Liu, W., & Zhang, X. (2020). Single-cell alternative splicing analysis reveals dominance of single transcript variant. *Genomics*, https://doi.org/10.1016/j.ygeno.2020.01.014.

Malmberg, M., Spangenberg, G., Daetwyler, H., & Cogan, N. (2019). Assessment of low-coverage nanopore long read sequencing for SNP genotyping in doubled haploid canola (Brassica napus L.). *Scientific Reports* , doi.org/10.1038/s41598-019-45131-0.

Marquis, D. (2019). Prunus serotina Ehrh. *Black Cherry*, http://dendro.cnre.vt.edu/DENDROLOGY/USDAFSSilvics/66.pdf.

Metzker, M. (2010). Sequencing technologies-the next generation. *Nature Reviews Genetics*.

NCBI. (2019). *Basic Local Alignment Search Tool*. Retrieved from https://blast.ncbi.nlm.nih.gov/Blast.cgi

Okonechnikov, K., Golosova, O., & Fursov, M. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* , 1166-1167. doi:10.1093/bioinformatics/bts091.

ONT. (2020). *Nanopore Community*. Retrieved from Oxford Nanopore Technology: https://community.nanoporetech.com/

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J., & Mayer, G. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports volume 8*, https://doi.org/10.1038/s41598-018-29325-6.

PoreCamp Australia. (2017). *PoreCamp Australia*. Retrieved from https://porecamp-au.github.io/.

Primer3web. (2019). *Primer3web version 4.1.0 - Pick primers from a DNA sequence.* Retrieved from http://primer3.ut.ee/

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PlosOne*, doi:10.1371/journal.pone.0163962.

Stubbs, S., Blacklaws, B., & Yohan, B. (2020). Assessment of a multiplex PCR and Nanopore-based method for dengue virus sequencing in Indonesia. *Virology Journal*, https://doi.org/10.1186/s12985-020-1294-6.

van Dijk, E., Jaszczyszyn, J., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends in Genetics*, 666-681.

Varshney, R., Pandey, M., & Chitikineni, A. (2018). *Plant Genetics and Molecular Biology.* Springer.

Wu, J., Khan, M., Wu, J., Gao, Y., Wang, C., Korban, S., & Zhang, S. (2013). Molecular Determinants and Mechanisms of Gametophytic Self-Incompatibility in Fruit Trees of Rosaceae. *Journal Critical Reviews in Plant Sciences*, 53-68.

Zanchetta, C., & Manno, M. (2017). *MinION, GridION, how does Nanopore technology meet the needs of our users?* . Retrieved from https://get.genotoul.fr/wp-content/uploads/2017/12/02_171128_Zanchetta_Manno.pdf

Zhang, H., Chen, Q., Wu, Y., Wang, Y., Bei, X., & Xiao, L. (2018). The temporal resolution and single-molecule manipulation of a solid-state nanopore by pressure and voltage. *Nanotechnology 29*, 1-9 .

TABLAS

Tabla 1. *Primers forward* modificados, empleados para la amplificación por PCR de las 12 muestras de capulí

Muestra/Código de barras	Secuencia (5´-----3´)
1-pic023	**TCGCCTTA**GCTTTCCTTGTTCTTGGTTTTGCTTTC
2-imb011	**TTCTGCCT**GCTTTCCTTGTTCTTGGTTTTGCTTTC
3-pic002	**TGCCTCTT**GCTTTCCTTGTTCTTGGTTTTGCTTTC
4-car007	**TCCTCTAC**GCTTTCCTTGTTCTTGGTTTTGCTTTC
5-car011	**TAGATCGC**GCTTTCCTTGTTCTTGGTTTTGCTTTC
6-h014	**TATCCTCT**GCTTTCCTTGTTCTTGGTTTTGCTTTC
7-pic019	**ACAGGCGC**GCTTTCCTTGTTCTTGGTTTTGCTTTC
8-car005	**CATAGAGT**GCTTTCCTTGTTCTTGGTTTTGCTTTC
9-car003	**CTCTCGTC**GCTTTCCTTGTTCTTGGTTTTGCTTTC
10-car012	**TGCGAGAC**GCTTTCCTTGTTCTTGGTTTTGCTTTC
11-can009	**TAGCGAGT**GCTTTCCTTGTTCTTGGTTTTGCTTTC
12-c0017	**CCAAGTCT**GCTTTCCTTGTTCTTGGTTTTGCTTTC

Las secuencias corresponden al *primer forward* dirigido a la región SPR del gen de la S-RNasa. A este *primer* se le incluyen 8 nucleótidos adicionales como códigos de barras específicos para cada una de las 12 muestras analizadas. Estos códigos de barras específicos permitirán la identificación de los amplicones de cada una de las muestras después de la secuenciación.

Tabla 2. Alelos identificados con una resolución por códigos de barras inespecífica

Código de barras-muestra	Alelos identificados	Alelos reportados por Correa (2018)
1 – pic023	S13, S7	S8, S13
2 – imb011	S9, S13, S3, S7	S6, S8
3 – pic002	S9, S13, S3, S7	S3, S4
4 – car007	S9, S7, S8, S13	S5, S13
5 – car011	S9, S13, S8, SX	S7, S9
6 – h014	S9, S8	S8, S9
7 – pic019	NO	S4
8 – car005	NO	S12
9 – car003	S9, SY, S8	S9, S14
10 – car012	S9	S14
11 – can024	SY, SX	S8
12 – c0017	S9, S4, S8, S11, SX	S9

Se muestran los alelos obtenidos cuando se realizó una resolución por códigos de barras no robusta. En este caso, no se detectaron los alelos esperados y no se encontraron lecturas asociadas a los códigos de barras 7 y 8, evidencias indicativas de una resolución por códigos de barras inespecífica.

Tabla 3. Alelos identificados mediante un análisis adecuado

Código de barras-muestra	**Alelos identificados**	**Alelos reportados por Correa (2018)**
1 – pic023	S8, S13	S8, S13
2 – imb011	S6, S8	S6, S8
3 – pic002	S3, S4	S3, S4
4 – car007	S5, S13	S5, S13
5 – car011	S7, S9	S7, S9
6 – h014	S8, S9	S8, S9
7 – pic019	S4, S5, S9, SX	S4
8 – car005	S12	S12
9 – car003	S9	S9, S14
10 – car012	S14, SY	S14
11 – can024	S8	S8
12 – c0017	S9	S9

Los alelos se determinaron mediante dendrogramas de: todo el gen, la región C2-C3 y el intrón I. Todos los árboles fueron consecuentes en los alelos presentados, que coinciden además con lo reportado previamente. Los alelos denominados SX y SY no se habían reportado en estudios anteriores. En el Anexo F se encuentran las secuencias completas de ambos y la secuencia completa del alelo S6.

FIGURAS

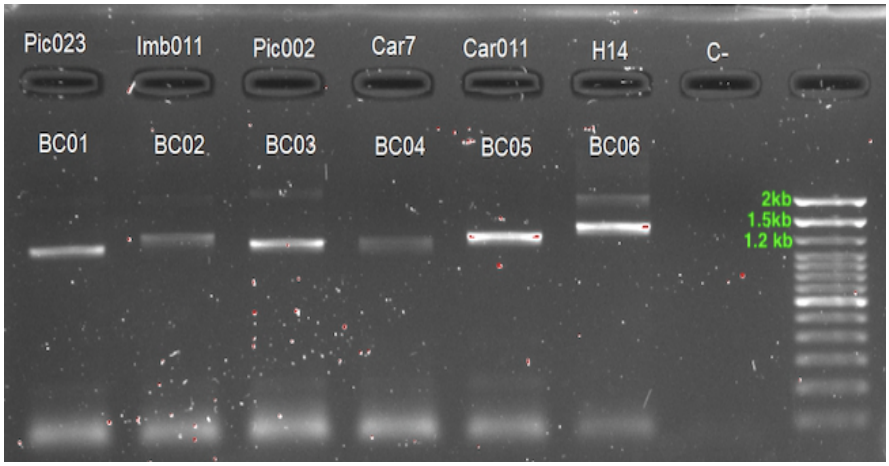


Figura 1. Amplificación del gen de la S-RNasa con los *primers* modificados

Se muestra el resultado de la PCR de 6 muestras: pic023, imb011, pic002, car007, car011, h014. La presencia del código de barras (BC) en el extremo 5´ del *primer forward* no afectó la calidad de la amplificación. Ladder: 100 pb (Promega).

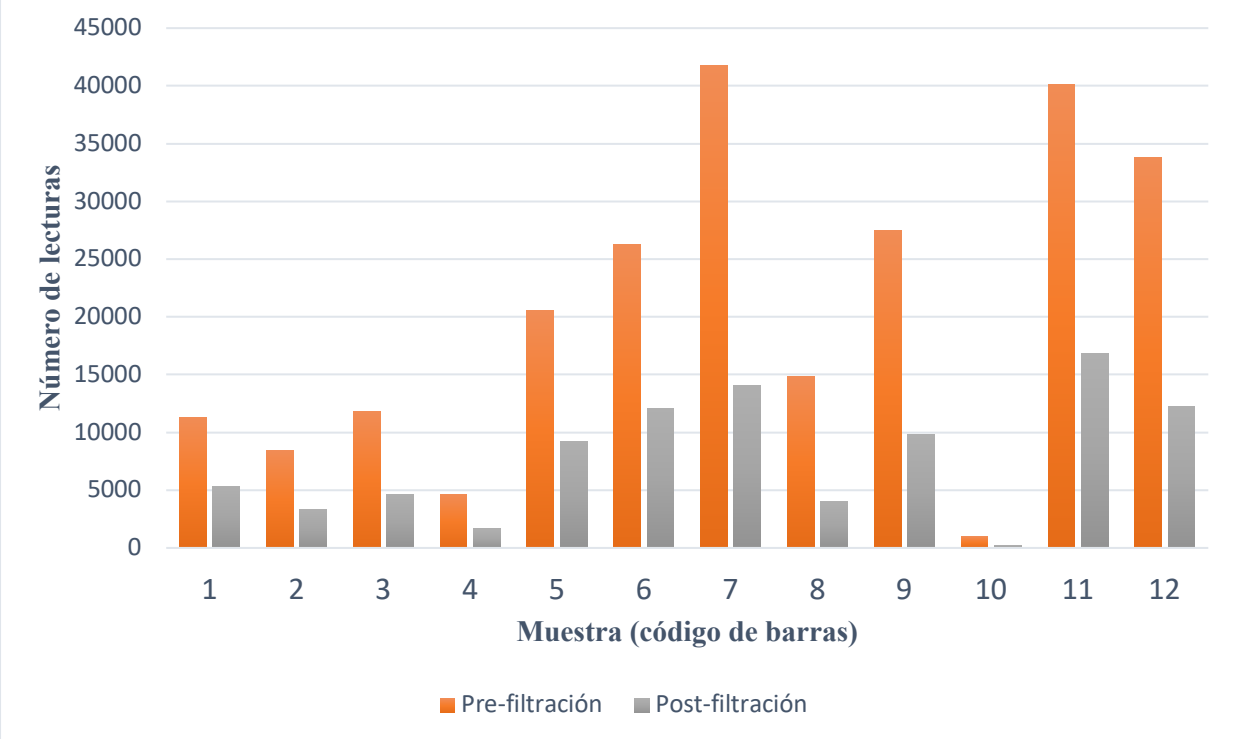


Figura 2. Número de lecturas luego de la resolución por códigos de barras (pre-filtración) y después de la filtración por calidad y tamaño

Se observa que no existe una distribución homogénea del número de lecturas por código de barras. Los umbrales utilizados en la filtración fueron: calidad: Q=12; tamaño: 800 pb. El paso de filtración eliminó el 61% de las lecturas crudas separadas por códigos de barras.

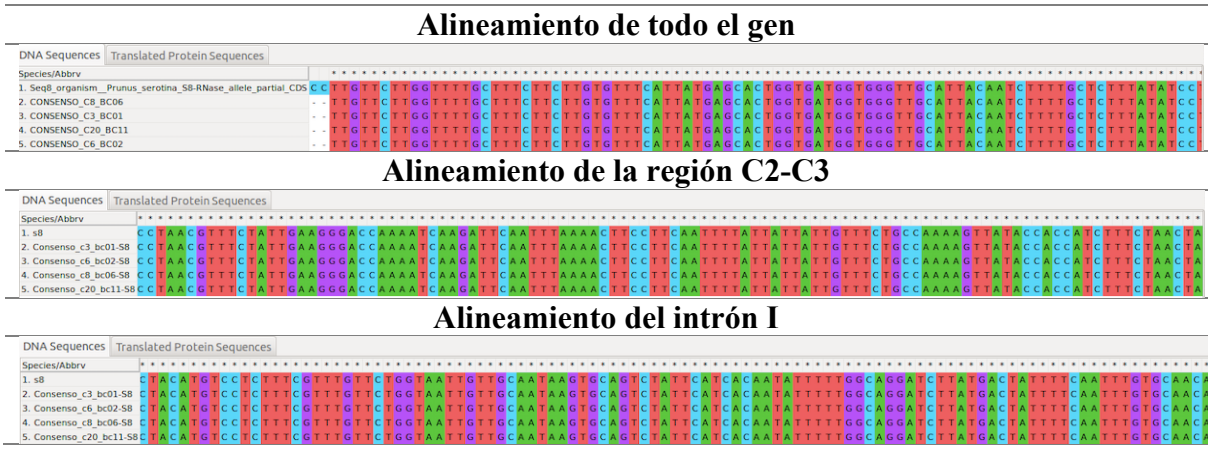
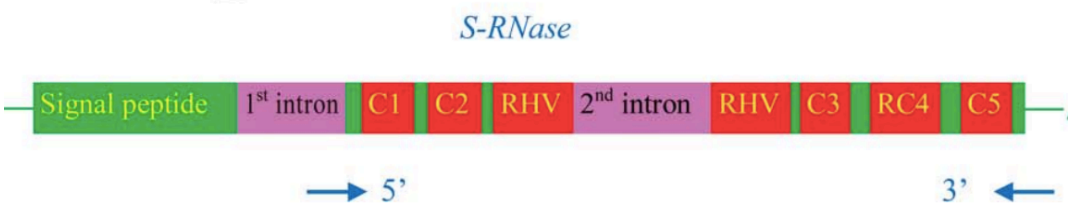


Figura 3. Ejemplo de los alineamientos realizados para comparar las secuencias obtenidas con las secuencias de referencia

En cada caso, los alineamientos se realizaron entre las secuencias de referencia reportadas previamente por Correa (2018) y las secuencias consenso que se agruparon a estas en los dendrogramas. En el ejemplo se muestran las secuencias consenso asociadas al alelo S8. Los resultados obtenidos con los demás alelos determinados son similares al presentado.

ANEXOS

ANEXO A: ESTRUCTURA DEL GEN DE LA S-RNASA



(Wu, et. al., 2013)

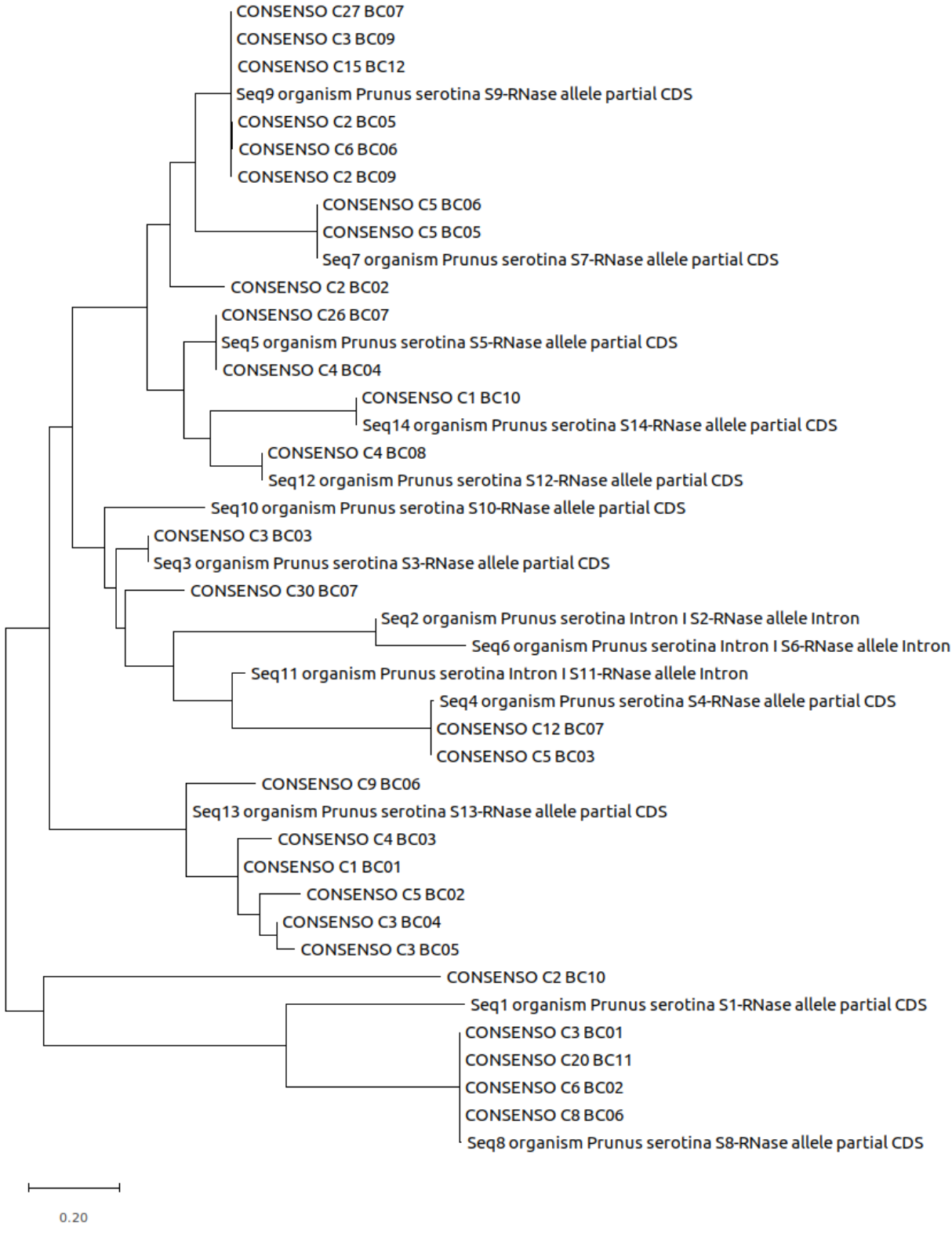
ANEXO B: MUESTRAS DE CAPULÍ UTILIZADAS PARA LA SECUENCIACIÓN DEL GEN DE LA S-RNASA

Provincia	Accesiones	Estudio de origen de las muestras
Pichincha	Pic023, Pic002, Pic019	(Guadalupe, et. al., 2015)
Imbabura	Imb011	(Guadalupe, et. al., 2015)
Carchi	Car007, Car011, Car005, Car003, Car012	(Guadalupe, et. al., 2015)
Chimborazo	H014	(Guadalupe, et. al., 2015)
Cañar	Can009	(Guadalupe, et. al., 2015)
Cayambe	C0017	(Baquero, 2018)

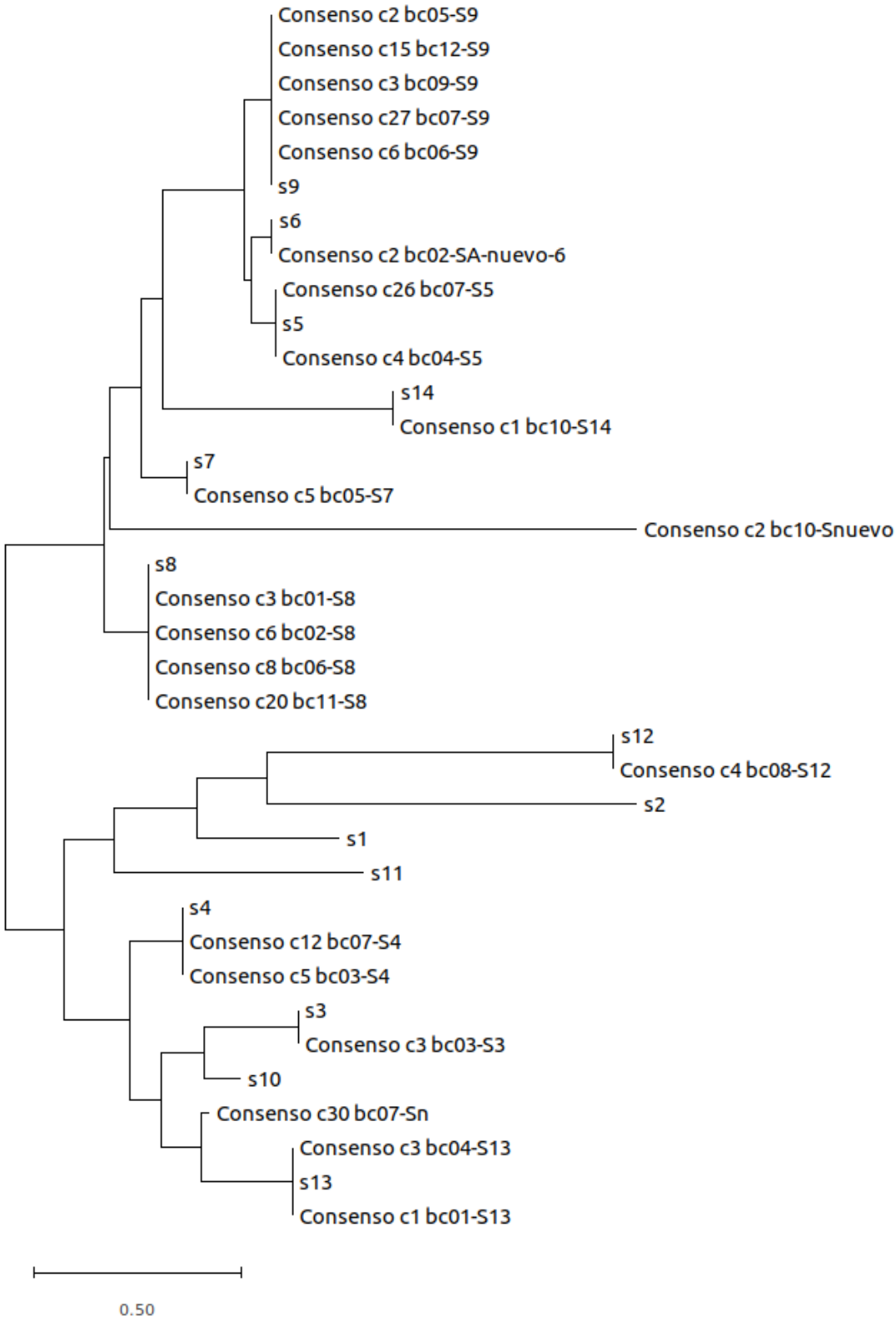
ANEXO C: *PRIMERS* ESPECÍFICOS PARA LAS REGIONES SPR (*FORWARD*) Y C5 (*REVERSE*)

Primer	**Secuencia (5´------- 3´)**
Forward	CCTTGTTCTTTGTTTTGCTTTCTTCTT
Reverse	GTTACATGAAGTGGTATTTTG

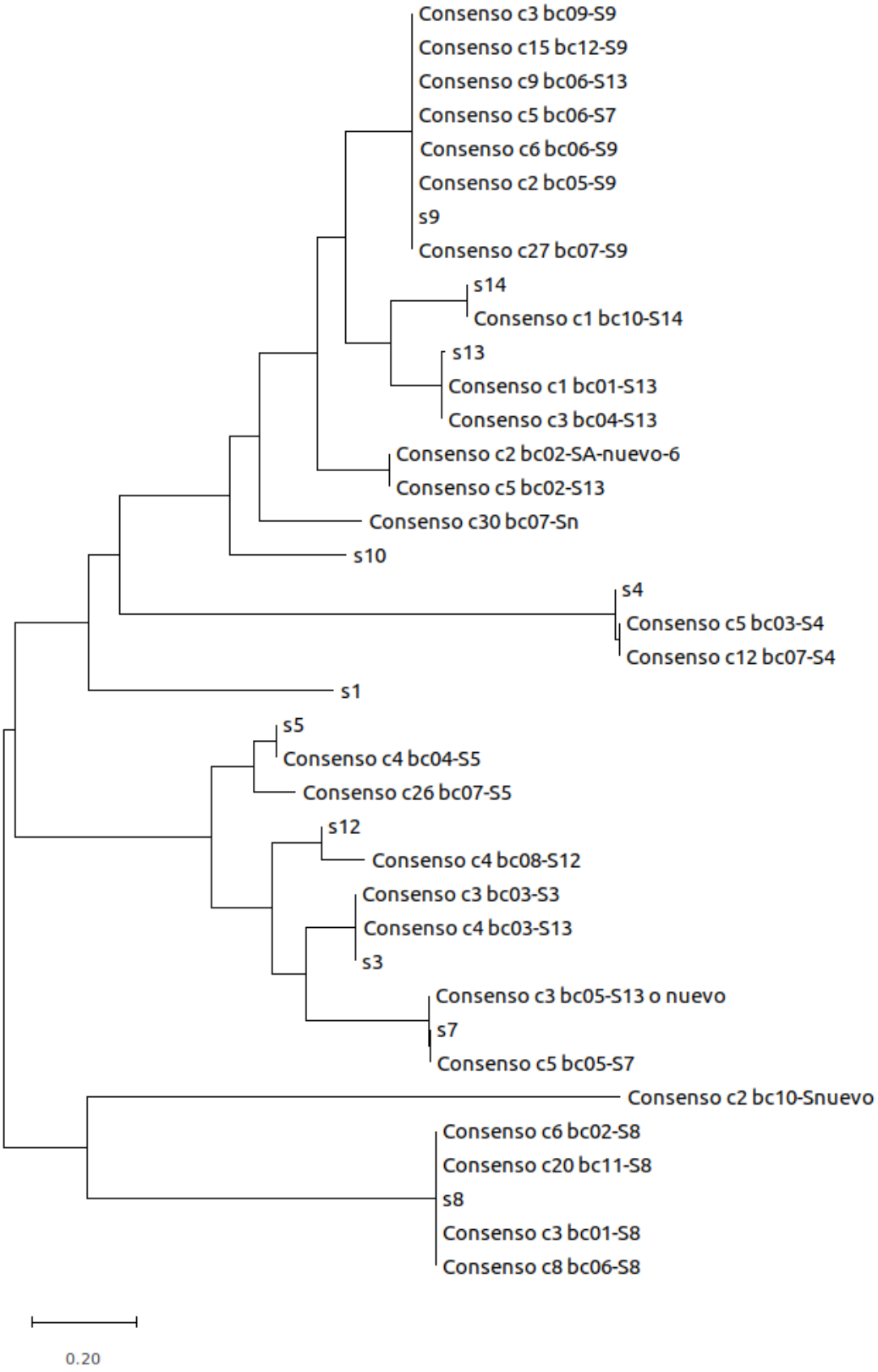
ANEXO D: DENDROGRAMAS REALIZADOS PARA OBTENCIÓN DE ALELOS



Dendrograma realizado con las secuencias de referencia y las secuencias consenso, utilizando todo el gen

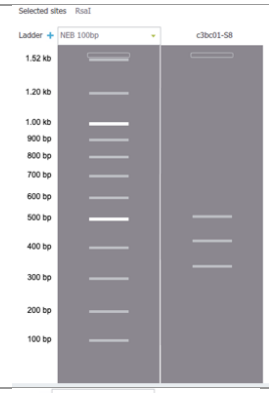
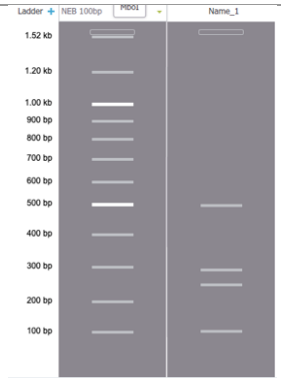
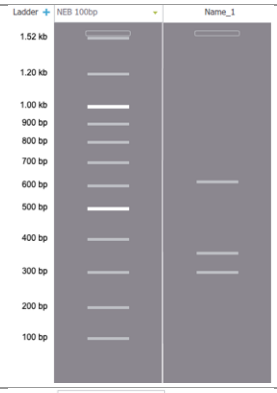
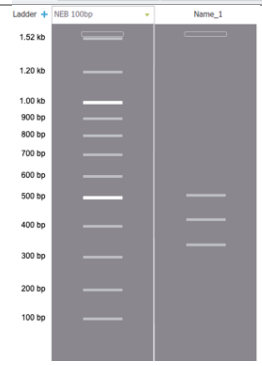
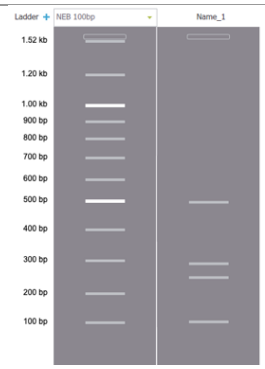
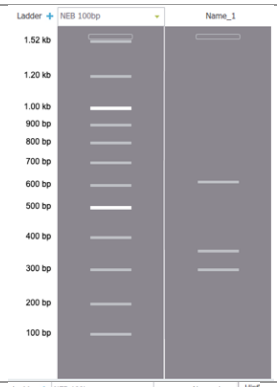
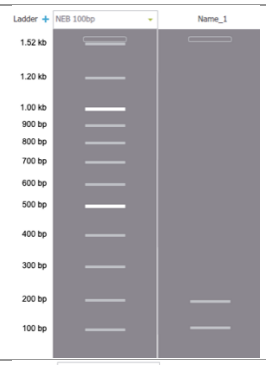
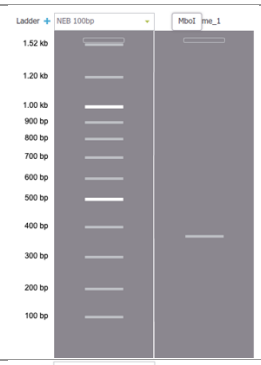
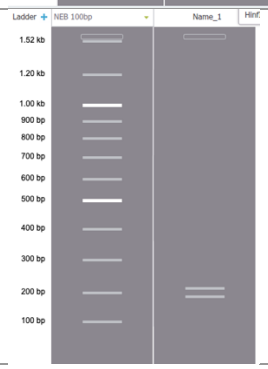
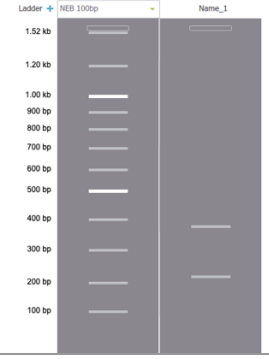
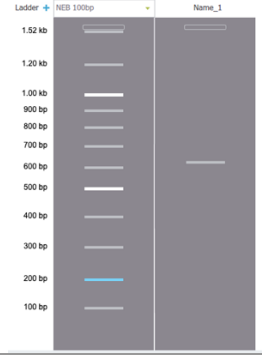
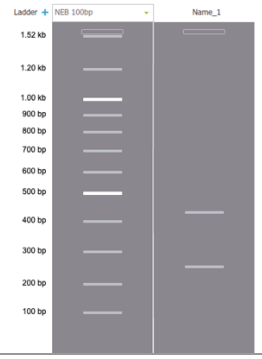


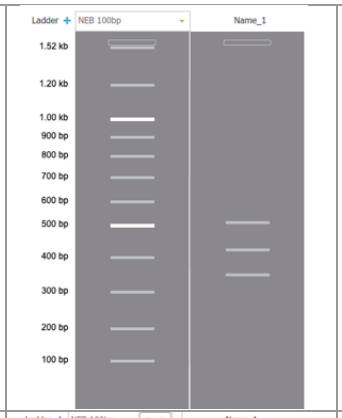
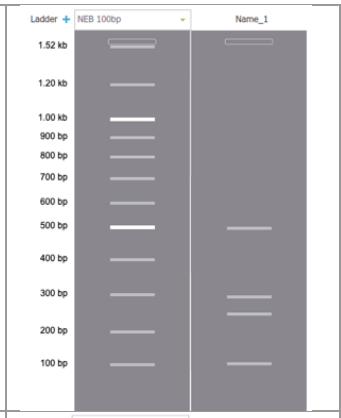
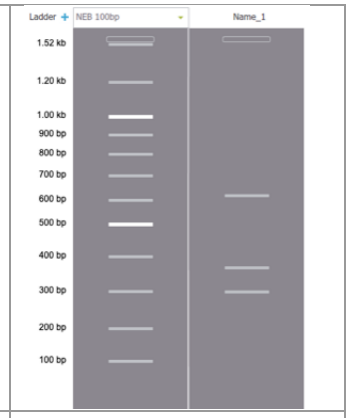
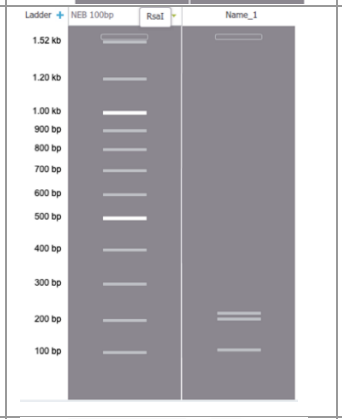
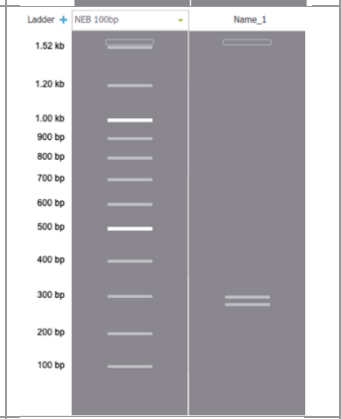
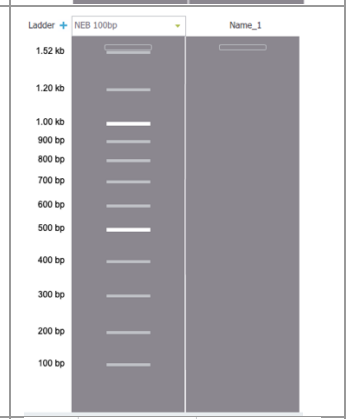
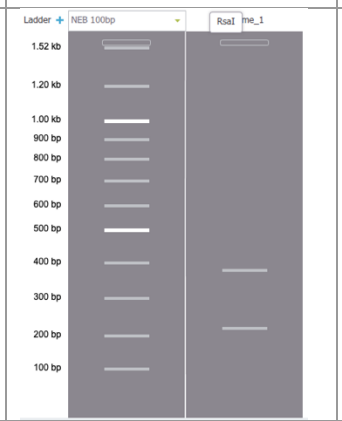
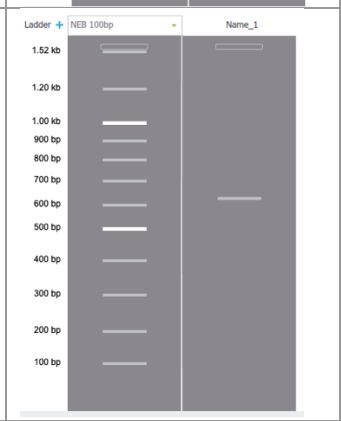
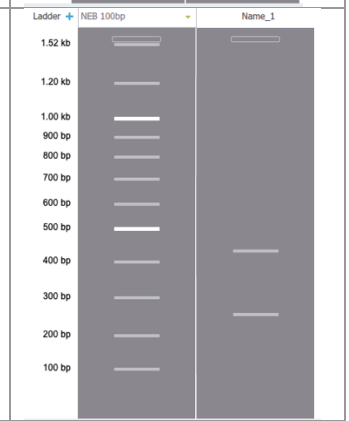
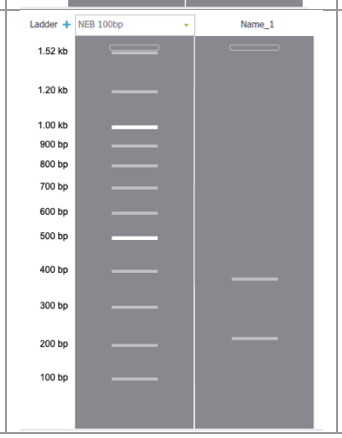
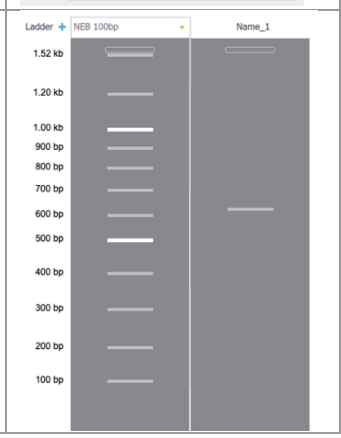
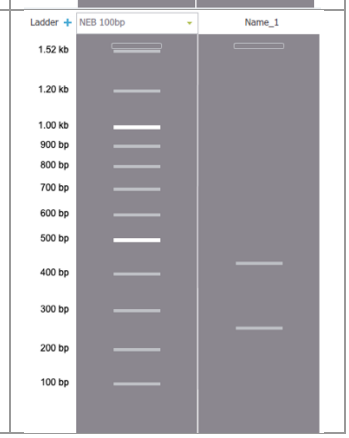
Dendrograma realizado con las secuencias de referencia y las secuencias consenso, utilizando las secuencias del intrón I.

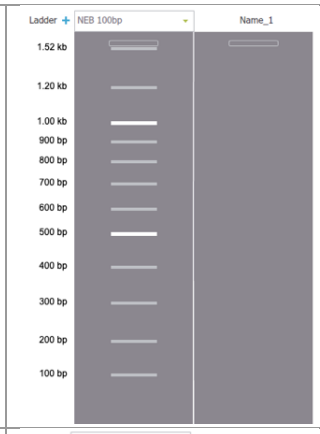
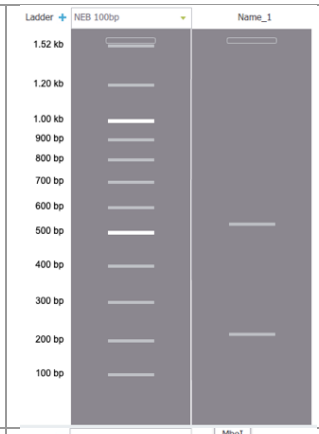
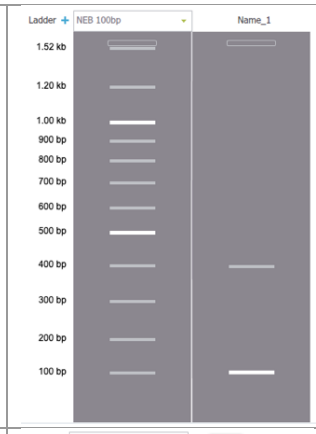
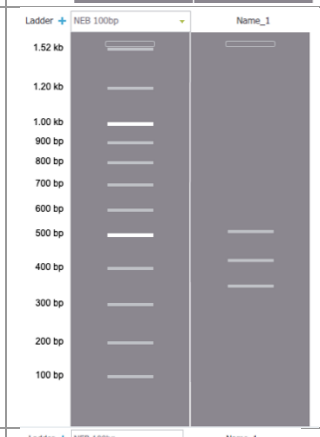
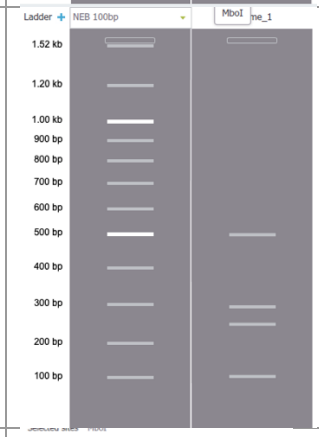
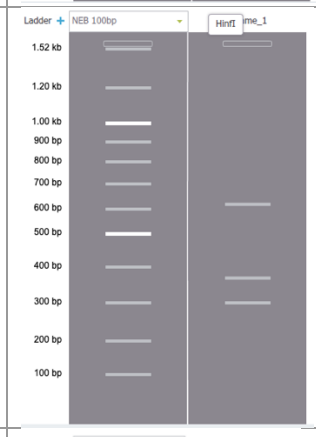
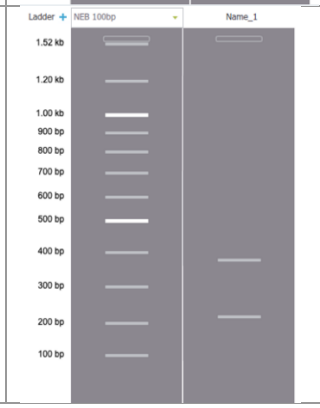
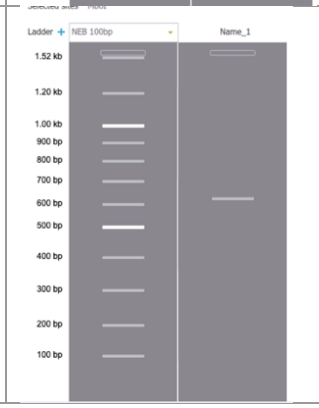
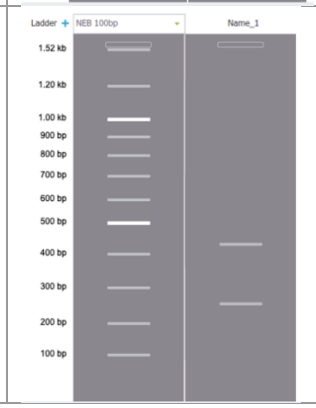


Dendrograma realizado con las secuencias de referencia y las secuencias consenso, utilizando las secuencias de la región C2-C3.

ANEXO E: PATRONES CAPS OBTENIDOS *IN SILICO* PARA COMPROBAR LOS ALELOS DETERMINADOS

	Enzima de Restricción		
Consenso y Alelo identificado	**RsaI**	**MboI**	**HinfI**
C3bc01-S8			
C6bc02-S8			
C4bc04-S5			
C2bc05-S9			

C8bc06-S8			
C30bc07-Nuevo Patrón			
C27bc07-S9			
C3bc09-S9			

C2bc10-Nuevo Patrón			
C20bc11-S8			
C15bc12-S9			

Se muestran algunos ejemplos de los patrones CAPS obtenidos *in silico* utilizando el programa Genome Compiler. La región sometida a la restricción *in silico* fue la C2-C3. Los patrones obtenidos coinciden con los reportados para los alelos detectados según la investigación de Correa (2018), a excepción de C30bc07 y C2bc10, que representan los alelos nuevos detectados en el presente estudio. La nomenclatura CAbc0B indica el consenso A del código de barras B.

ANEXO F: SECUENCIAS COMPLETAS DE LOS DOS ALELOS NUEVOS DETECTADOS Y DEL ALELO S6

SX (CONSENSO_C30_BC07)

TTGTTCTTGGTTTTGCTTTCTTCTTGTGTTTCATTATGAGCACTGGTGGGTTGCATT
ACAATTTCTTTGCTCTTTATATTCTGTCATATAGTTAGTATTGCATTTTATACTTAATT
TTTATGTTTAGAGAAATATGTATGCATTTATCAGGTAAGGGAGGACAACGTTCTTTG
GATGAATAACTATTTGGGAATTACATTTCTGCATGGTTTCTTTGGTCTACTCTGATA
GTTGTTGCAATAAGTGCAGTGTTCATCATTTCAAGCTAAAACTAGGTACTTTATACA
TCAAATCCTTATTTAAGATACCATTAACCTTCTCACAATAATTTTCGCAGGATCTTAT
GTGTATTTTCAATTTGTGCAACAATGGCCACCGACGACCTGCAGACTTAGCAGCA
AACCTTCCAACCAACACCGGCCATTACAAAGATTCACCATCCATGGCCTATGGCCA
AGTAACTATTCAAACCCAAGGAAGCCTAGTAATTGCAATGGATCACAATTTGACGC
AAGGGAAGTGGTACGTATTGTTTCATTATTTTCTAACTTACTCTTTGGCATTTAGTTT
TTTAGGTTTTTTTTAAAAGGAATTAGTGTTTAGAAAATTAGATTGTCATGTGAAGAT
TTTAATAAATAAATAAACCTTTTTCAATAAGCCTTGGGTGTTATAGATTAAATTTTGA
TGTTGGTTCTTAGTTAGACACATTATTTTAAATATATAGTTAAGTAAAAAATGGTAA
GTACATATTAATATACCATCGAAAATATAATGGATCTGCTAATCTAATTATATGACCTA
CCATTTTGTACTAATGCATATATGCAAAACATTGTACATCAATTTTTTTTTTTTAAAG
CAAGGCTATAATATATTATTGGGGATTAAACTCAAAATTAATGCTCGGATTTAATGA
GACAAAAAACAATCTTTATATTTTTAGTACAAGCGACAATATAAATTACACAAGAG
TATCATTCAAGAATGAAAATCTAATTATCATTTATTCAATTTACTTTTTCTCAAAATAT
GTGTCTACATGGTTTGGATGTCTCAGTCCCCCCCCAATTGCGATCAAAACTGAAGA
TATCTTGGCCCGACGTGGAAAGTGGCAATGATACACAATTTTGGGAAGGCGAATG
GAACAAACATGGTACTTGTTCCGAAGAGACACTTGACCAAACGCAATACTTCGCG
CGATCCCACGCATTTTGGAACATGCGCAATATTACGGAGATCCTTAAAAACGCATC
AATCGTACCACATCCGACAAAAACATGGAAATACTCGGACATAGTGGCACCCATTA
AAGCAGCAACTAAAAGAATCCCCTCCTTCGTTGCAAACGTGATCCAGCACAGATT
AAGAGCGGGCCGAAGACTCAGCTGTTACATGAAGTGGTATTTT

SY(CONSENSO_C2_BC10)

TTGTTCTTGGTTTTGCTTTCTTCGACATTTGTTTCAAGTGTTTTTGCTTAACTCTGTT
CATTTTGGTTGTTTAATAAAAATTTTGCTTAACTCAAACTAATGAAGAGTTTATTATT
GAATCTTCTTGGGTTGCATGTACTTGTTAATTCGAGCTACAAATTAACAAACTTTAT
GATGGATAGCTCACATTTGTCTGCGAAAGGAATTTCTTTTGGTTGTTAGTTCTGGTT
TCTTGTTGGGTTAAGTTCATTTTTTGTTTTACTTAATTGAAATTACACTAAATTTAAA
ATAGCGGGCTTCATGAGGGTTCTTACTGTTGAATACCTTTTGATACTCGTTGATCTG
CAGGACATTAAAAGGGAGACGTGGATGCTTACTGATGAAGGAAAGGCATATACTG
CTACTGGATCACCTGAAGTGCAACTGTTCCTGGCCATACCACCAGAGGGTATTCCA
AAAGAAGAATTGCAGGTGATTATTATGGCTCTCGCCTTTAATTTAGAAAATCTGATA
CGATGCATGATTCATTGGTTCTATTCAATTATGTTCTGTGCTTGTAGTTTCTCTTTGT
TCACGAGTTATTAATCTTTGATGCTACATGTGTTTTGAAACTGATGTTTTATTCAAG
AAAAGCTGGATCCATCAGTTTTCAAAATAGGTTGCGCCCAGGCTGCAAAGAATAA
ATGGGTGGAGATGGGAAAACAACTTGTCACCAGAAAGGTAACACCTCACTAACC
AAATCTCTATAAGTAATAATGTCTGTTAGGAGTGTTTTAGATATGTTACGTAGGGGC
TATGAGCTTCCTGGAATTTATCCTAATACAGTATTTTGTTCCAGTTTATGTGAATAAG
TAGGCCCTTTTGTCAATTTATTTTGGCGTCATGGTTAATGAAGCCATGGATACTTAT
CTTGTGTCAGGAGTCCTTCATGTTCTCCATTGACAATCCAGCCAAAGACTTGTTAT
TGGGATAAAAATTTATCTCTGATACACTAAAATATAATTTTTGTAGAATCTGTGATAG

TATTTAGTGTGTTTTCTTTCCCCTTCATTGATAATTATTATATGTTTCTCGTTTGACTG
TTTTCAAGATTTATTCTTTTATTTGTGTGTCTAATGGACTCCGATATTATTGTTTGTTT
TATTATGCGGTTCGTTGGTGCTTTTAAACTTTCATTTGATAGTGTTCATAAATGTTCA
TCAGCTTTGTTCTGTTTGTGTGTCTTTGTCTTTTTGTTTTGAAGTGATTCTGGATGT
GTGTGGAATCCACAAATGTCTGATGATGCTAGTATATCCATGTTTTGCCTCAGTTGT
TACATGAAGTGGTATTTG

S6 (CONSENSO_C2_BC02)

TTGTTCTTGGTTTTGCTTTCTTCTTTTGTTACGTTATGAGCAGTAGTGGTGGGTTGC
ATTACAATCTTTTGCTATATCCTATATGCATATAATCAGCATTGCATTTTTCTACTTTT
ATTTGTTGTTCAGAGAACTATTGTGTGTATTCGATGATGTGTCAGGTGACATGCGGT
GTATTGAATTAACCCACATATTTTTCATTTAATCTAACGCACAACTTTCTTTGGATGA
GCAAGTATTTGGAATTGTTTTTCCCCATGTCCTCTTTTTTGTTTTCATCATCTTTTGT
TTATTCTGATAATTGGTTGCAATAAGTGCAGTCTATTCATCATGATAATTTTGGCAGG
ATCTTATGACTATTTTCAATTTGTGCAACAATGGCCACCGACGAACTGCAGAGTTC
GCGTCAAGCGACCTTGCTCCAATCCCCGGCCATTACAATATTTCACCATCCATGGC
CTATGGCCAAGTAATTATTCAAACCCAAGGATGCCAAGTAATTGCACTGGGTCGCA
ATTTAAGAAACAGAATTTGGTATGTATTTTTTTCACTTTGTTTTTAGAAAATTAGATT
GTCATCTGAAAATAATATAAACTTTTCAATAAATTTTGGGTGTAACATAAAATTTTAT
GCTGGCACACATCGTAAGTATAAAGTTGGAAGTACATTTATTTACGTATTTTATAAT
AATATATCGAATCATTTAATAAATGGATTTACCACTCCGTACTATAATCGAAATATTG
TACTTATATGAGATGATAAAAATAAAAATATAATTTAGGACAGATTTAATGAAAAAA
GAAGTTCTTGTCCAATAATGAAAATCTAACTCTCCCTTGCGTTTTTACCTTTTTCTC
CTCAGTACCCTTATATGCAATCCAAACTGAAGATATCTTGGCCGGACGTGGAAAGT
GGGAATGATACAAAATTTTGGGAAGGCGAATGGAACAAACATGGTACATGTTCCG
CACGGACACTTAACCTAATGCAATACTTCCAACGATCCCACGCAATGTGGAAATCA
CACAATATTACAGAGATCCTTAAAAATGCTTCAATCGTACCACATCCGACACAAAC
ATGGAAGTACTCGGACATAGAATCACCCATTAAAAGAGCAACTAAAAGAACACCC
GTCCTTCGTTGCAAACGTGATCCAGTACAGGCGAATACCCAGCTGTTACATGAAGT
GGTATTTTAAC

ANEXO G: COMANDOS UTILIZADOS (ANÁLISIS BIOINFORMÁTICO)

Paso	Programa	Comandos
1-Análisis general de parámetros de la secuenciación (distribución de tamaño de lecturas, contenido de GC, calidad de las lecturas)	Fastqc	$fastqc *.fastq -o input.fastqc
2-Resolución por códigos de barras	Porechop	$porechop -i input_reads.fastq.gz -b output_dir --barcode_threshold 100
3-Filtración	NanoFilt	$gunzip -c input.fastq.gz | NanoFilt -q 12 | gzip > input-hq-reads.fastq.gz $gunzip -c input-hq-reads.fastq.gz | NanoFilt -l 800 | gzip > input-hq-l800-reads.fastq.gz
4-Obtención de estadísticas y conversión a .fasta	SeqKit	$seqkit stat input-hq-l800-reads.fastq.gz $seqkit fq2fa input-hq-l800-reads.fastq > input-hq-l800-reads.fasta
5-Reordenamiento de las lecturas	BBMap	$sortbyname.sh in=file.fa out=sorted.fa length descending
6-Agrupamiento	MeshClust	$meshclust input-hq-l800-reads-sorted.fasta [--id 0.90] [--kmer 3] [--delta 5] [--output output.clstr] [--iterations 20] [--align] [--sample 3000] [--pivot 40] [--threads TMAX]
7-Selección y etiquetado	BBMap, SeqKit	$filterbyname.sh in=reads.fa out=selected.fa names=names.txt include=t $seqkit replace -p .+ -r "seq_{nr}"
Otros comandos de utilidad		
Función	**Programa/Paquete**	**Comando**
Unir varios archivos	---	$cat *.fasta > merged.fasta
Separar varios archivos	pyfasta	$pyfasta split -n X original.fasta X: número de lecturas dentro del archivo original