UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias Biológicas y Ambientales

Ensamblaje del genoma del chocho (*Lupinus mutabilis*) variedad INIAP 450 Andino

.

Ana Gabriela Pupulin Concari

Ingeniería en Biotecnología

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

Quito, 19 de mayo del 2023

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias biológicas y Ambientales

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Ensamblaje del genoma del chocho (*Lupinus mutabilis*) variedad INIAP 450 Andino

Ana Gabriela Pupulin Concari

Nombre del profesor, Título académico	**María de Lourdes Torres, PhD**
Nombre del profesor, Título académico	**Milton Gordillo, MSc**

Quito, 19 de mayo del 2023

© DERECHOS DE AUTOR

Nombres y apellidos Ana Gabriela Pupulin Concari

Código: 00208155

Cédula de identidad: 0927705590

Lugar y fecha: Quito, 19 de mayo de 2023

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UNPUBLISHED DOCUMENT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

El chocho *(Lupinus mutabilis)* es una leguminosa andina de importancia económica, agronómica y ecológica en el Ecuador. A pesar de esto, aún no se ha publicado ninguna investigación genómica sobre la especie. La variedad de *L. mutabilis* INIAP 450 Andino es de gran interés debido a que presenta un rendimiento 183% superior al rendimiento promedio de los ecotipos cercanos. El objetivo de este estudio fue obtener el primer genoma de referencia de *L. mutabilis* a través de procesos de ensamblado y scaffolding a partir de las lecturas crudas obtenidas de la secuenciación con Oxford Nanopore Technologies. Se utilizaron diversas aplicaciones bioinformáticas para obtener el genoma de referencia del chocho, como Flye, SMARTdenovo, NtLink y RagTag y se evaluó el rendimiento de cada enfoque con QUAST y BUSCO. Se descubrió que el uso del algoritmo de SMARTdenovo, junto con una fase de pulido en Medaka y su posterior scaffolding con el algoritmo de NtLink dio como resultado un genoma de referencia de un tamaño de 588 Mb con un total de 1617 fragmentos y 93% de genes BUSCOs de Fabales completos. El genoma de referencia obtenido resultó ser cercano a los genomas de referencia de otras especies de *Lupinus* disponibles actualmente (*L. angustifolius* con 609 Mb y *L. albus* con 450 Mb). Esto demuestra la posibilidad de obtener genomas continuos y de alta calidad utilizando el flujo de trabajo demostrado en este reporte. El genoma de referencia de *L. mutabilis* obtenido a partir de la variedad INIAP 450 Andino representa un primer paso hacia el desarrollo de herramientas moleculares modernas específicas para el chocho que nos permitan realizar programas de mejoramiento genético en esta especie.

Palabras clave: Chocho, *Lupinus mutablis*, genoma de referencia, ONT, genómica, ensamblado, scaffolding.

ABSTRACT

Chocho (*Lupinus mutabilis*) is an Andean legume of economic, agronomic, and ecological importance in Ecuador. Despite this, no genomic research on the species has yet been published. The variety of *L. mutabilis* INIAP 450 Andino is of great interest because it presents a yield 183% higher than the average yield of nearby ecotypes. The objective of this study was to obtain the first reference genome of *L. mutabilis* through assembly and scaffolding processes from the raw reads obtained from sequencing with Oxford Nanopore Technologies. Various bioinformatics applications were used to obtain the chocho reference genome, such as Flye, SMARTdenovo, NtLink and RagTag and the performance of each approach was evaluated with QUAST and BUSCO. Using the SMARTdenovo algorithms, together with a polishing phase in Medaka and subsequent scaffolding with the NtLink algorithm, it was found that the reference genome was 588 Mb in size with a total of 1617 fragments and 93% of complete BUSCO genes of Fabales. The reference genome obtained turned out to be close to the reference genomes of other *Lupinus* species currently available (*L. angustifolius* with 609 Mb and *L. albus* with 450 Mb). This demonstrates the possibility of obtaining high-quality and continuous genomes using the workflow demonstrated in this report. The *L. mutabilis* reference genome obtained from the INIAP 450 Andino variety represents a first step towards the development of modern lupine-specific molecular tools that allow us to carry out breeding programs in this species.

Keywords: Chocho, *Lupinus mutablis*, reference genome, ONT, genomics, assembly, scaffolding.

TABLA DE CONTENIDOS

1 INTRODUCCION 12

1.1. Características de *Lupinus mutabilis* 12

1.1.1. Generalidades, origen y propiedades 12

1.1.2. Importancia 12

1.1.3. Variedad INIAP 450 Andino 13

1.1.4. Estudios moleculares preliminares de *L. mutabilis* en el Ecuador 13

1.2. Secuenciación y ensamblaje de genomas 14

1.3. Importancia de conocer el genoma completo de *Lupinus mutabilis* 15

1.4. Objetivos 15

2 METODOLOGÍA 16

2.1. Datos previos obtenidos de la secuenciación en MinION del genoma de L. mutabilis 16

2.2. Análisis bioinformático 16

2.2.1. Pre-procesamiento de las lecturas de la secuenciación de *L. mutabilis* 16

2.2.2. Ensamblaje del genoma 17

2.2.3. Pulido de genoma 17

2.2.4. Scaffolding del genoma 17

2.2.5. Evaluación de los ensamblajes y scaffoldings de los genomas 18

2.2.6. Evaluación de la integridad del ensamblaje del genoma 18

3 RESULTADOS 19

3.1. Datos previos de la secuenciación en MinION del genoma 19

3.2. Evaluación y Comparación de los genomas con QUAST 19

3.2.1. Ensamblaje del genoma con SMARTdenovo y Flye 19

3.2.2. Scaffolding del genoma con RagTag y NtLink 19

3.3. Evaluación de la integridad de los ensamblados y scaffoldings con BUSCO 20

3.3.1. Ensamblaje del genoma con SMARTdenovo y Flye 20

3.3.2. Scaffolding del genoma con RagTag y NtLink 21

3.4. Comparación de los resultados obtenidos a través del scaffolding con especies cercanamente emparentadas 21

4 DISCUSIÓN 23

4.1. Ensamblaje del genoma de *L. mutabilis* 23

4.2. Scaffolding del genoma de *L. mutabilis* después del pulido 23

4.3. Evaluación de la integridad del genoma con BUSCO 24

4.4. Comparación del genoma de referencia de *L. mutabilis* con respecto a especies cercanamente emparentadas 26

5 CONCLUSIONES 27

6 TABLAS 28

7 FIGURAS 30

8 REFERENCIAS 34

9 ANEXOS 41

INDICE DE TABLAS

Tabla 1. Estadísticas previas del desempeño de la secuenciación de *Lupinus mutabilis* 28

Tabla 2. Parámetros considerados para la evaluación del genoma de *L. mutabilis* después del ensamblado y scaffolding 28

Tabla 3. Porcentaje (%) de BUSCOs obtenidos para los ensamblados y scaffoldings 29

Tabla 4. Comparación del scaffolding entre referencias cercanamente emparentadas 29

INDICE DE FIGURAS

Figura 1. Flujo de trabajo del análisis bioinformático para el ensamblaje delgenoma de *Lupinus mutabilis*.....30

Figura 2. Resultado obtenido de BUSCO para los ensamblados.....31

Figura 3. Resultado obtenido de BUSCO para los scaffoldings.....32

Figura 4. Compilación de los resultados obtenidos de BUSCO para evaluación de integridad del genoma completo.....33

ÍNDICE DE ANEXOS

ANEXO 1. NanoPlot de los datos obtenidos a partir de la secuenciación de lecturas largas de ONT. 41

1 INTRODUCCION

1.1. Características de *Lupinus mutabilis*

1.1.1. Generalidades, origen y propiedades

Lupinus mutabilis, más conocido como chocho es una leguminosa endémica de la región andina que pertenece a la familia de las Fabáceas. Es originaria de los Andes centrales del norte de Perú y el sur de Ecuador y los datos sugieren que fue domesticada en esta región, en el departamento peruano de Cajamarca hace unos 2500 años (Eastwood et al., 2018). En el Ecuador, el chocho crece en los páramos, ecosistemas de climas templados y fríos (7-14ºC), en suelos arenosos y secos. El chocho se desarrolla entre los 2400-3600 m.s.n.m. con precipitaciones de 600 mm anuales (Peralta et al, 2012). '*Mutabilis*' se llama así porque muestra cambios de color de sus flores posteriores a la floración. Se cree que estos cambios orientan a los polinizadores hacia las flores no polinizadas, al tiempo que conservan la atracción más amplia de polinizadores (Eastwood & Hughes, 2018). A diferencia de la mayoría de los cultivos domesticados de legumbres, las semillas del chocho retienen un alto contenido de alcaloides (2,13 – 4,5%) que pueden llegar a ser tóxicos y deben ser rigurosamente eliminados antes del consumo de estas semillas (Eastwood & Hughes, 2018).

1.1.2. Importancia

El chocho presenta un potencial importante para la alimentación de la población humana y animal (Clements, et al., 2012). Entre sus características nutricionales más relevantes se destaca su elevado contenido de proteínas y aceites esenciales (hasta el 50% y 25%, respectivamente) (Molina-Poveda et al., 2013). En los últimos años, el cultivo del chocho ha ganado un interés creciente en todo el mundo como una alternativa sostenible y nutritiva a los cultivos de cereales tradicionales, como el arroz, la cebada y la soya (Carvajal-Larenas et al., 2016). Rico en proteínas y aceites esenciales, el chocho es considerado un superalimento con

el potencial de reemplazar la dependencia de la proteína animal. De hecho, varios estudios (Gabur y Petru, 2023; Gulisano et al., 2019) respaldan su potencial para ayudar en la transición a sistemas agrícolas más sostenibles.

1.1.3. Variedad INIAP 450 Andino

INIAP 450 Andino es una variedad mejorada de chocho obtenida mediante selección artificial a partir de cruces utilizando germoplasma procedente de Perú (INIAP, 2015). Durante el programa de mejoramiento iniciado en 1992, se realizaron pruebas de crecimiento y rendimiento bajo diversas condiciones ambientales para seleccionar los materiales más promisorios. Finalmente, en 1999, se logró obtener material homogéneo con buenos rendimientos en campo, y la variedad fue liberada para su comercialización (INIAP, 2015). Esta variedad se caracteriza por tener un hábito de crecimiento herbáceo y ser de ciclo precoz, es sensible a plagas y enfermedades foliares y radiculares (INIAP, 2015). A pesar de estos desafíos, el rendimiento de INIAP 450 Andino es notablemente alto, ya que, según el Instituto Nacional Autónomo de Investigaciones Agropecuarias, su producción es un 183% superior al rendimiento promedio de los ecotipos cercanos, presentando un rendimiento promedio entre 1.350 kg/ha y 1.500 kg/ha (INIAP, 2015).

1.1.4. Estudios moleculares preliminares de *L. mutabilis* en el Ecuador.

Al momento, los recursos moleculares para el fitomejoramiento de *L. mutabilis* son escasos y la mayor parte de las investigaciones moleculares hasta la fecha se han concentrado en descifrar su filogenia (Atchison et al., 2016). Recientemente, se han empleado marcadores basados en ADN para evaluar la diversidad genética entre las especies de *Lupinus* (Zoga et al., 2008) y para describir genéticamente *L. mutabilis* se han empleado 113 cebadores SSR y 118 InDel polimórficos de *L. luteus* (Osorio et al., 2018). En comparación con otras leguminosas, *L. mutabilis* no cuenta con gran cantidad de información genómica (Keller et al., 2017).

1.2. Secuenciación y ensamblaje de genomas

Contar con los genomas de referencia de las plantas es importante porque nos brindan información sobre la biología de éstas, así como sobre los fundamentos moleculares y genéticos de sus características de interés. El estudio del genoma de las plantas ha ayudado al descubrimiento rápido de genes esenciales que influyen en características importantes de los cultivos y ha servido como una plataforma importante para la selección genética y el desarrollo de cultivos agrícolas que contribuyan a la seguridad alimentaria (Wambugu et al., 2022).

Hoy en día existen varias metodologías que nos permiten decodificar los genomas de los organismos vivos, entre ellas el secuenciamiento por nanoporos. La secuenciación de ADN a través de nanoporos es un proceso en el que se lee el código de hebras simples de ADN a medida que se unen a través de orificios extremadamente pequeños (nanoporos) implantados en una membrana. Cuando el ADN pasa por el poro, se generan señales que pueden ser procesadas para leer cada nucleótido de ADN en forma secuencial (NIH, 2023). En 2014, Oxford Nanopore Technologies (ONT) creó la primera y más pequeña tecnología de secuenciación de nanoporos del mundo, utilizando el dispositivo MinION (Loose et al., 2015). Este equipo portátil y económico puede generar datos de secuenciación rápidamente, y se conecta a computadoras con requisitos mínimos de hardware para detectar nucleótidos sin necesidad de equipos de imagen (Loose et al., 2015).

El ensamblaje del genoma es el proceso de volver a ensamblar una gran cantidad de pequeñas secuencias de ADN para construir una representación de los cromosomas originales de donde proviene el ADN (Lu et al., 2016). El genoma del chocho se ensambló utilizando algoritmos de ensamblaje conocidos como de ensamblaje *de novo*. Estos algoritmos generan contigs (fragmentos creados a partir de lecturas individuales) basados en lecturas que se superponen (Sohn & Nam, 2016).

El scaffolding, que se realiza posterior al ensamblaje, se basa en generar scaffolds (contigs

que han sido ordenados y unidos de manera adecuada con respecto al genoma de tu organismo de interés) y se divide en tres pasos: 1) construcción de contigs mediante lecturas superpuestas, 2) generación de scafflods mediante unión ordenada de contigs y 3) relleno de espacios (Utturkar et al., 2014). Este procedimiento se puede realizar utilizando un genoma de referencia, que es un genoma previamente ensamblado y curado de la misma especie u otra especie emparentada cercanamente (Jung et al., 2020). Otra alternativa, es construir desde cero sin el uso de una referencia, empleando algoritmos de ensamblaje que te permitan utilizar tus lecturas crudas (Dominguez Del Angel et al., 2018).

1.3. Importancia de conocer el genoma completo de *Lupinus mutabilis*

Las razones para obtener un genoma de referencia *L. mutabilis* varían, pero incluyen estudios científicos, mejoras agrícolas y el empleo de lupinos como biorreactor para producir proteínas con uso terapéutico (Martins et al., 2016). Actualmente no hay evidencia de producción comercial de especies de lupino modificadas genéticamente (Eapen, 2008). Sin embargo, la investigación sobre la ingeniería genética de los lupinos se ha llevado a cabo en países como Australia donde el mejoramiento genético para mejorar las características agronómicas, en particular el rendimiento y la resistencia a las enfermedades, ha continuado desde la aparición de la primera variedad *de L. angustifolius*, Uniwhite (French & White, 2008). El enfoque principal del presente trabajo fue el ensamblaje y scaffolding del genoma de *L. mutabilis.*

1.4. Objetivos

El objetivo de este proyecto fue secuenciar, ensamblar y estructurar lecturas largas producidas por ONT para obtener el primer genoma de referencia de *Lupinus mutabilis.*

2 METODOLOGÍA

2.1. Datos previos obtenidos de la secuenciación en MinION del genoma de *L. mutabilis*

Para el presente estudio, se utilizó información obtenida de ensayos de secuenciamiento de genoma del chocho realizados previamente en el Laboratorio de Biotecnología Vegetal USFQ. Para los ensayos de secuenciación se utilizó como material de partida ADN de alto peso molecular (HMW-DNA) extraído a partir de hojas jóvenes de plantas de chocho de la variedad INIAP 450 Andino cultivadas en el cuarto de cultivo a 24 °C, con un fotoperiodo de 16 horas de luz y 8 horas de oscuridad. El protocolo utilizado fue el "High molecular weight gDNA extraction from plant leaves" reportado en la ONT Community (Belser, 2021) y para la reacción de secuenciación se utilizó el Ligation Kit LSK-109. Entre las principales métricas de los datos de secuenciación utilizados para los ensayos de ensamblado del genoma del chocho están: cantidad de información utilizada (41Gb), calidad media de las lecturas (Q-score 11.4), tamaño medio de las lecturas (7 Kb) y N50 (14.6 Kb).

2.2. Análisis bioinformático

El flujo de trabajo se encuentra ilustrado en la Figura 1.

2.2.1. Pre-procesamiento de las lecturas de la secuenciación de *L. mutabilis*

Para eliminar los adaptadores de los extremos de las lecturas utilizadas en la preparación de la biblioteca, se utilizó Porechop v. 0.2.4 (Wick et al., 2017) que reconoce la secuencia de los adaptadores y las remueve. Las lecturas con una calidad inferior a 7 se filtraron y eliminaron con Filtlong v. 0.2.1 (Wick & Menzel, 2021) que filtra las secuencias según la calidad y/o la longitud. Asimismo, se utilizó NanoPlot v. 1.30.1 (De Coster et al., 2018) para examinar los datos de secuenciación, ya que esta herramienta proporciona un archivo de resumen que contiene estadísticas y gráficos que simplifican elementos con respecto a lecturas de secuenciación.

2.2.2. Ensamblaje del genoma

El ensamblado se realizó con dos programas de ensamblaje *de novo* de lecturas largas: SMARTdenovo v. 1.0.0 (Liu et al., 2021) y Flye v. 2.9.1 (Kolmogorov et al., 2019). Ambos son ensambladores rápidos *de novo* diseñados exclusivamente para lecturas largas de PacBio y Oxford Nanopore.

La diferencia entre SMARTdenovo y Flye, es que SMARTdenovo no contiene una etapa de corrección de errores. Asimismo, Flye emplea un método Bruijn Graph generalizado que genera contigs basados en superposiciones iniciales inexactas, luego construye gráficos que combinan los contigs en muchos ensamblajes alternativos posibles y finalmente genera contigs precisos (Kolmogorov et al., 2019).Por otro lado, el ensamblador SMARTdenovo utiliza un algoritmo gráfico de ensamblaje conocido como OLC (Overlap Layout Consensus) que consta de tres pasos: superposición de lecturas, creación de gráficos basados en la superposición de lecturas e inferencia de una secuencia de consenso (Wattam et al., 2017).

2.2.3. Pulido de genoma

El ensamblaje de SMARTdenovo se pulió con Medaka v. 1.7.2 (ONT Ltd., 2018) que genera secuencias de consenso y llama variaciones a partir de los datos de secuenciación de Oxford Nanopore. El resultado se crea utilizando redes neuronales, que comparan un ensamblaje inicial con respecto a las lecturas de secuenciación individuales (ONT Ltd., 2018). En esta etapa se corrigen errores que pudieron haberse generado durante el ensamblado debido a que los genomas suelen estar fragmentados e incompletos por la presencia de tramos repetitivos que dificultan la determinación precisa de la superposición (Hu et al., 2023).

2.2.4. Scaffolding del genoma

El scaffolding se llevó a cabo con dos programas: RagTag v. 2.1.0 (Alonge et al., 2022) y NtLink v. 1.3.8 (Coombe et al., 2021). RagTag filtra y combina alineaciones de todo el genoma para recuperar información a través de una forma integrada de filtrado de anclaje único

para eliminar las alineaciones repetitivas y utiliza un genoma de referencia (Alonge et al., 2022). RagTag genera alineaciones de genoma completo filtradas y fusionadas entre un conjunto de "ensamblado" y "referencia" (Alonge et al., 2022). En el caso de RagTag utilizamos como referencia los ensamblados de SMARTdenovo y Flye uno contra el otro. También se utilizó un método no requiere de un genoma de referencia para el proceso de scaffolding conocido como NtLink, el cual asigna pedazos de las lecturas largas (minimizadores) a los contigs obtenidos del ensamblaje para encontrar pares de contigs donde haya uniones sugeridas por la ordenación determinada por los minimizadores, lo cual sirve como evidencia para establecer el orden en que deben unirse los contigs (Coombe et al., 2021). Cabe recalcar que posterior a la etapa de scaffolding se debe realizar nuevamente un pulido con Medaka.

2.2.5. Evaluación de los ensamblajes y scaffoldings de los genomas

Se utilizó QUAST v. 5.2.0 (Gurevich et al., 2013) para evaluar ensamblajes específicos, así como para compararlos y contrastarlos. Las evaluaciones de QUAST presentadas en este reporte corresponden únicamente a métricas calculadas sin el uso de una referencia tanto para los ensamblados como para los scaffoldings.

2.2.6. Evaluación de la integridad del ensamblaje del genoma

Finalmente, se utilizó BUSCO v. 5.3.2 (Manni et al., 2021) para verificar la integridad del ensamblaje mediante la detección de genes altamente conservados contra una base de datos de referencia. El conjunto de datos del linaje seleccionado fue fabales_odb10 (Número de genomas: 10, número de genes BUSCO: 5366), y el predictor genético empleado fue metaeuk (Manni et al., 2021).

3 RESULTADOS

3.1. Datos previos de la secuenciación en MinION del genoma

Las estadísticas relevantes de la serie de secuenciación se muestran en la Tabla 1. Como resultado de los experimentos de secuenciación se produjo un total 41 GB de datos con una calidad promedio de 11.4 y un tamaño medio de lecturas de 7 Kb (Anexo 1). Adicionalmente, el N50 fue de 14.6 Kb, lo que sugiere que las lecturas de ese tamaño o más grandes contenían al menos el 50 % de los nucleótidos en el conjunto general de secuencias obtenidas (Mäkinen et al., 2012).

3.2. Evaluación y Comparación de los genomas con QUAST

3.2.1. Ensamblaje del genoma con SMARTdenovo y Flye

Las métricas obtenidas para ambos ensamblajes del genoma del chocho se muestran en la Tabla 2. Entre ellos, el genoma Flye fue el más fragmentado (21 376 contigs), mientras que el genoma de SMARTdenovo tuvo el genoma menos fragmentado (2 056 contigs). Finalmente, Flye mostró un tamaño de genoma inferior (552 Mb) comparado con el de SMARTdenovo (590 Mb) y un N50 igualmente inferior comparado con el de SMARTdenovo (554 Kb y 732 Kb, respectivamente).

3.2.2. Scaffolding del genoma con RagTag y NtLink

Las métricas obtenidas para ambos scaffolders del genoma del chocho se muestran en la Tabla 2. Cabe recalcar que, aunque se realizaron múltiples pruebas con RagTag utilizando como referencias los genomas de especies cercanamente emparentadas (*Lupinus albus* y *Lupinus angustifolius*), estos resultados no se presentan debido a la tendencia de RagTag de fusionar el ensamblado con la referencia utilizada. Para aprovechar esta característica de RagTag lo que se realizó fue un scaffolding del ensamblado de SMARTdenovo utilizando como referencia el ensamblado obtenido con Flye. De esta manera todos los datos utilizados para el scaffolding con RagTag se realizan con secuencias de *Lupinus mutabilis*. En este reporte

se presentan únicamente los mejores resultados, tanto para RagTag como para NtLink, basándonos en las métricas obtenidas por QUAST y BUSCO para una mejor comprensión de los resultados.

El genoma obtenido de RagTag de SMARTdenovo con referencia de Flye (RT_Sdn_rFlye) fue el menos fragmentado (1 445 contigs), mientras que el genoma obtenido de NtLink de SMARTdenovo (NtLink_SdN) tuvo el genoma más fragmentado (1 617 contigs) con 172 scaffolds más que el obtenido con RT_Sdn_rFlye. Finalmente, RT_Sdn_rFlye mostró un tamaño total superior (589 Mb) comparado con el de NtLink_SdN (588 Mb) y un N50 superior (1.1 Mb) comparado con el de NtLink_SdN (1 Mb) (Tabla 2).

3.3. Evaluación de la integridad de los ensamblados y scaffoldings con BUSCO

3.3.1. Ensamblaje del genoma con SMARTdenovo y Flye

En la Figura 2 se encuentra un gráfico comparativo de los BUSCOs obtenidos para ensamblados y scaffodings en conjunto. En la Figura 3 y Figura 4 se encuentran los resultados obtenidos de los BUSCOs para los ensamblados y scaffoldings por separado, respectivamente. El ensamblado obtenido con Flye tiene el mayor porcentaje de BUSCOs completos siendo estos del 90.8%, a diferencia de SMARTdenovo con 82.9% de BUSCOs completos (Tabla 3). Esto indica que las coincidencias de BUSCO se ubican dentro del rango esperado y que los genes de copia única correspondientes al orden de los Fabales se encuentran en una alta proporción (Seppey et al., 2019).

Por otro lado, el ensamblado obtenido con Flye presenta una menor cantidad de BUSCOs fragmentados, siendo estos del 1.6% y de BUSCOs perdidos del 7.6%, mientras que el ensamblado obtenido con SMARTdenovo presenta 2.7% de BUSCOs fragmentados y 14.4% de BUSCOs perdidos (Tabla 3).

Es importante mencionar que luego de haber analizado los ensamblados utilizando las herramientas de Quast y BUSCO, se determinó que el ensamblado que presentaba las mejores

características para pasar al proceso de scaffolding fue el ensamblado obtenido con SMARTdenovo dado que se encuentra alrededor de 10 veces menos fragmentado en comparación al ensamblado obtenido con Flye. Es por ello, que todos los resultados obtenidos luego de los procesos de scaffolding son comparados con respecto al ensamblado de SMARTdenovo. Para una mejor comprensión, se presentan solo los resultados considerados los mejores en base los parámetros previamente mencionados para cada uno de los scaffolders.

3.3.2. Scaffolding del genoma con RagTag y NtLink

En cuanto al scaffolding hecho con RT_Sdn_rFlye se conservó el número de BUSCOs completos con respecto al ensamblado del 82.9% con SMARTdenovo (Tabla 3) así como el mismo porcentaje de BUSCOs fragmentados y perdidos (2.7% y 14.4%, respectivamente). Mientras que, para el scaffolding realizado con NtLink_SdN se obtuvo un porcentaje de BUSCOs completos del 93.2% (Tabla 3), siendo este un incremento del 10.3% con respecto al ensamblado de SMARTdenovo. Esto indica que se logró obtener un genoma más completo en términos de contenido de genes conservados gracias al proceso de scaffolding (Waterhouse et al., 2018). Por consiguiente, los resultados obtenidos con NtLink presentan una menor cantidad de BUSCOs fragmentados, siendo estos del 0,7% y de BUSCOs perdidos del 6,1% (Tabla 3), mientras que el ensamblado obtenido con RagTag presenta BUSCOs fragmentados del 2.7% y de BUSCOs perdidos del 14.4% (Tabla 3).

3.4. Comparación de los resultados obtenidos a través del scaffolding con especies cercanamente emparentadas

Una vez obtenido el primer genoma de referencia de *Lupinus mutabilis* a partir del NtLink_SdN, realizamos comparaciones con respecto a los genomas de referencia de *Lupinus albus* y *Lupinus angustifolius* publicados en el NCBI. Respecto al tamaño del genoma, los resultados obtenidos para nuestro genoma de referencia de *L. mutabilis* (588 Mb) se encuentran en un intermedio entre *L. angustifolius* (609 Mb) y *L. albus* (450 Mb) (Tabla 4). Asimismo,

respecto al número de fragmentos el genoma de *L. mutabilis* presenta un total de 1 617, en comparación con los 14 387 y 89 fragmentos reportados para *L. angustifolius* y *L. albus*, respectivamente (Tabla 4). Respecto a la métrica N50, el genoma de *L. mutabilis* se encuentra en un valor intermedio (1Mb) en comparación con los valores de N50 reportados para *L. angustifolius* (702 Kb) y *L. albus* (21Mb) (Tabla 4). Finalmente, el porcentaje de GC obtenido es bastante similar entre las 3 especies, alrededor del 35%.

4 DISCUSIÓN

4.1. Ensamblaje del genoma de L. mutabilis

El genoma de *Lupinus mutabilis* se ensambló utilizando métodos de ensamblaje *de novo* que crean contigs basados en lecturas superpuestas (Amarasinghe, et al. 2020) obtenidas a partir de lecturas largas. La tasa promedio de errores de secuenciamiento para lecturas largas utilizando ONT es del 15% y por tanto es indispensable incluir una ronda de pulido del genoma (Tan et al., 2018). Por esta razón, se empleó Medaka v. 1.7.2. que utiliza datos de secuenciación para generar secuencias de consenso y llamadas de variantes (Salzberg, et al. 2011).

Es importante recordar que la efectividad de una técnica de ensamblaje depende de una serie de variables, como la cobertura y la complejidad del genoma (Chen, Erickson & Meng, 2020). Para el ensamblado de genomas completos a partir de lecturas largas, Jung et al. (2020), informan que Flye es una de las mejores alternativas de ensambladores. Sin embargo, estudios que ensamblaron el genoma de *Macadamia* con Flye dieron como resultado un genoma altamente fragmentado (Sharma et al., 2021), lo cual se asemeja a los resultados obtenidos en este estudio de 21 376 contigs con respecto al ensamblado realizado con Flye de *L. mutabilis*.

Por otra parte, Schmidt et al. (2017) muestran que la reconstrucción del genoma de *Solanum pennellii* obtenido usando SMARTdenovo es estructuralmente muy similar a la de la referencia *S. pennellii*. Aunque previo al estudio realizado para este reporte no existía un genoma de referencia para *L. mutabilis*, los resultados obtenidos indican que el algoritmo de SMARTdenovo produce un genoma más extenso, con menos fragmentos y un valor de N50 superior en comparación con el ensamblaje realizado con Flye. Estos hallazgos sugieren que, para los datos analizados en este estudio, el algoritmo SMARTdenovo fue el más adecuado.

4.2. Scaffolding del genoma de *L. mutabilis* después del pulido

Estudios han demostrado que el scaffolding de lecturas largas de mijo perla (*Cenchrus*

americanus) obtenidas a través de secuencias ONT mostró una mejora en comparación con el genoma de referencia anterior de mijo perla (Varshney et al. 2017) en términos de continuidad del genoma al usar RagTag como scaffolder (Salson et al. 2023). Por otro lado, estudios confirman que Ntlink funciona de manera eficiente y genera ensamblajes finales de alta calidad con lecturas largas de plantas (arroz) y animales (humano). Esto debido a que los minimizadores que utiliza NtLink para realizar una asignación entre el ensamblado de destino de entrada y las lecturas proporcionadas exhiben gran utilidad en el mapeo rápido y preciso de ensamblajes de lecturas largas (Coombe et al. 2021). En este caso, la evaluación de los genomas del chocho, indica que el algoritmo de NtLink es el adecuado pues produce el genoma con mayor cantidad de BUSCOs completos (93%), aunque se encuentre ligeramente más fragmentado que el genoma obtenido a partir de RagTag (Tabla 2).

4.3. Evaluación de la integridad del genoma con BUSCO

Usando BUSCO, fue posible evaluar la calidad del ensamblaje en términos de contigüidad de genes y recuento total de genes. La resolución de este análisis depende del número de genes empleados. El linaje, en este caso fabales_odb10, que consta de 10 genomas y 5366 BUSCO, se caracteriza por ser de alta resolución, lo que arroja un alto grado de confianza para estas evaluaciones (Waterhouse et al. 2018). Según Waterhouse et al. (2018), los genes completos son aquellos cuyas longitudes de alineamiento están dentro de dos desviaciones estándar de la longitud media del grupo BUSCO. Para que el ensamblaje sea considerado de buena calidad y continuidad al menos el 90 % de todos los genes BUSCO deben estar presentes en el ensamblado (Jung et al. 2020). En el genoma de referencia de *L. mutabilis* producido en este estudio (scaffolding usando NtLink de SMARTdenovo) se identificó el 93.2% de los genes BUSCO completos (Tabla 3), lo cual indica que estos resultados presentan buena calidad y continuidad.

Por otro lado, el 0.7% de los genes BUSCO que se recuperaron se encontraban incompletos, es decir, que sus secuencias estaban fragmentadas. Estos resultados son indicativos de que las coincidencias de BUSCO se han puntuado dentro del rango de puntuaciones, pero no dentro del rango de alineaciones de longitud con el perfil de BUSCO (Jentoft, 2017), lo cual podría indicar que el gen está parcialmente presente por errores de ensamblaje. También podría deberse a que los pasos de búsqueda de secuencias y predicción de genes de BUSCO no lograron producir un modelo de gen de longitud completa, aunque el gen completo podría estar presente en el ensamblaje (Jentoft, 2017).

Dado que los genes BUSCO son genes que evolucionan bajo el control de una sola copia, los genes duplicados deberían ser poco comunes. Los genes duplicados obtenidos para el genoma de referencia producido en este estudio corresponden a un valor de 19,3% (Tabla 3). La duplicación de genes BUSCO en un genoma podría denotar un ensamblaje potencial de diferentes haplotipos, una duplicación reciente del genoma completo o artefactos técnicos que pudieron haber surgido durante el proceso de ensamblaje (Teh, 2017). Según Hufnagel et al. (2017), el genoma del lupino blanco (*L. albus*), cercanamente emparentado con *L. mutabilis*, está cargado de duplicaciones de genes y elementos repetitivos, es decir, presenta extensos bloques de duplicación dentro de su propio genoma.

Respecto a los genes BUSCO perdidos, que para el genoma de referencia obtenido corresponden a un valor de 6.1%, se considera que no se hubo coincidencias significativas en absoluto o las coincidencias de BUSCO se calificaron por debajo del rango de puntajes para el perfil de BUSCO (Simão et al., 2015). Esto podría indicar que faltan estos ortólogos o que el paso de búsqueda de secuencias de BUSCO no logró identificar ninguna coincidencia significativa. No obstante, el bajo porcentaje de genes BUSCO perdidos sugiere que se logró un ensamblaje completo y de buena calidad del genoma de *L. mutabilis* (Tabla 3). Es importante considerar que el umbral de BUSCO se adapta al hecho de que incluso los genes

bien conservados pueden perderse en algunos linajes, lo cual es otra posible explicación a los resultados en cuanto al porcentaje de BUSCOs perdidos obtenidos (Jauhal & Newcomb, 2021).

4.4. Comparación del genoma de referencia de *L. mutabilis* con respecto a especies cercanamente emparentadas

El genoma de referencia obtenido para *L. mutabilis* se encuentra en un rango intermedio para las métricas tamaño del genoma, N50 y número de fragmentos obtenidos en comparación con los genomas reportados en las bases de datos para las especies cercanamente emparentadas *L. albus* y *L. angustifolius* (Tabla 4), lo cual es indicativo de un ensamblaje apropiado. Igualmente, el contenido de GC del genoma de referencia obtenido presenta un valor del 35% que se considera apropiado y dentro del rango esperado para plantas (GC del 33.6% al 47.5%) (Smarda et al., 2012).

Por otro lado, aunque para la métrica N50 el valor obtenido (1 Mb) haya estado en un rango intermedio entre *L. albus* y *L. angustifolius* podría considerarse bastante bajo e inusual dado un alto contenido de BUSCOs completos para el ensamblaje de *L. mutabilis*. Sin embargo, Jauhal & Newcomb (2021) afirman que se podría producir una puntuación BUSCO alta a partir de ensamblajes con un valor N50 bajo, como se observa a partir de los resultados obtenidos (Tabla 2 y Tabla 3). Cabe recalcar que, aunque el valor obtenido a través de la métrica N50 es bastante bajo, estudios afirman que el N50 no es la indicación ideal para cuantificar la precisión del genoma (Jauhal & Newcomb, 2021). Además, estudios realizados por Jung et al. (2020) afirman que un N50 de un valor superior o igual a 1 Mb es suficiente para pasar al nivel de ensamblaje cromosómico y anotación. Es importante mencionar que la calidad de un genoma es determinada por un conjunto de métricas y no por parámetros analizados individualmente.

5 CONCLUSIONES

Este estudio proporciona el primer genoma de referencia de *Lupinus mutabilis*, una especie andina semidomesticada de Ecuador con gran potencial nutricional, mediante el secuenciamiento con ONT, que permitió generar 41 Gb de información, y posterior ensamblaje y scaffolding del genoma.

Se evaluaron diferentes métodos de ensamblaje y se determinó que el ensamblaje con SMARTdenovo logró proporcionar un genoma menos fragmentado (2 056 contigs) y más continuo (N50: 732 kb) en compración con el ensamblaje realizado con Flye (21 376 contigs; N50: 554 kb). A pesar de tener menor cantidad de BUSCOs completos presentes con el ensamblaje con SMARTdenovo (82.9%) se optó por escogerlo como el mejor ensamblado por presentar una menor fragmentación del genoma.

Se evaluaron diferentes métodos de scaffolding y se obtuvo que al combinar el ensamblaje en SMARTdenovo seguido del proceso de scaffolding con NtLink se obtiene el genoma del chocho de mejor calidad en cuanto a número de contigs (1 617) y BUSCOs completos (93.2%). Siendo este último criterio, el que permitió estimar de manera más adecuada la continuidad del ensamblaje. Adicionalmente, se observó que el genoma de referencia de *L. mutabilis* obtenido se encuentra en un intermedio en cuanto a tamaño del genoma, N50 y número de contigs con respecto a los genomas de referencia de *L. albus* y *L. angustifolius* que se encuentran en el mismo nivel de ensamblaje (scaffold).

Finalmente, para perfeccionar el genoma de referencia obtenido se recomienda incorporar lecturas cortas que permitan obtener un menor número de fragmentos y eventualmente poder llegar a nivel cromosómico. Asimismo, se considera la posibilidad de realizar investigaciones futuras como anotación funcional para identificar genes de interés y potencialmente hacer mejoramiento genético en esta especie.

6 TABLAS

Tabla 1. Estadísticas previas del desempeño de la secuenciación de *Lupinus mutabilis*

Estadística	Valor
Información Obtenida (GB)	41
Número de lecturas (M)	5.9
Calidad promedio de las lecturas (X)	11.4
Promedio de la longitud de las lecturas (Kb)	7
N50 (Kb)	14.6

Datos proporcionados por: Milton Gordillo

Tabla 2. Parámetros considerados para la evaluación del genoma de *L. mutabilis* después del ensamblado y scaffolding

Ensamblador	Tamaño (Mb)	Número de contigs	N50 (Kb)	Contig más largo (pb)
Flye	552	21 376	554	4 472 302
SMARTdenovo	590	2 056	732	4 427 805
RT_SdN_rFlye	589	1 445	1 107	5 080 222
NtLink_SdN	588	1 617	1 065	6 512 614

RT_SdN_rFlye: RagTag de SMARTdenovo con referencia de Flye

NtLink_SdN: NtLink de SMARTdenovo

Tabla 3. Porcentaje (%) de BUSCOs obtenidos para los ensamblados y scaffoldings

Parámetro	**Flye**	**SMARTdenovo**	**RT_SdN_rFlye**	**NtLink_SdN**
BUSCOs completos	90.8	82.9	82.9	93.2
-BUSCOs de una copia	73.3	70.2	70.5	73.9
-BUSCOs duplicados	17.5	12.7	12.4	19.3
BUSCOs fragmentados	1.6	2.7	2.7	0.7
BUSCOs perdidos	7.6	14.4	14.4	6.1
BUSCOs totales	100.00	100.00	100.00	100.00

Tabla 4. Comparación del scaffolding entre referencias cercanamente emparentadas

Estadísticos	***L. mutabilis***	***L. angustifolius***	***L.albus***
Tamaño del Genoma	588 MB	620 MB	458 MB
Número de fragmentos	1 617	13 573	89
Métrica N50	1 Mb	21 Mb	17 Mb
Porcentaje GC	35	33.5	33.7

7 FIGURAS

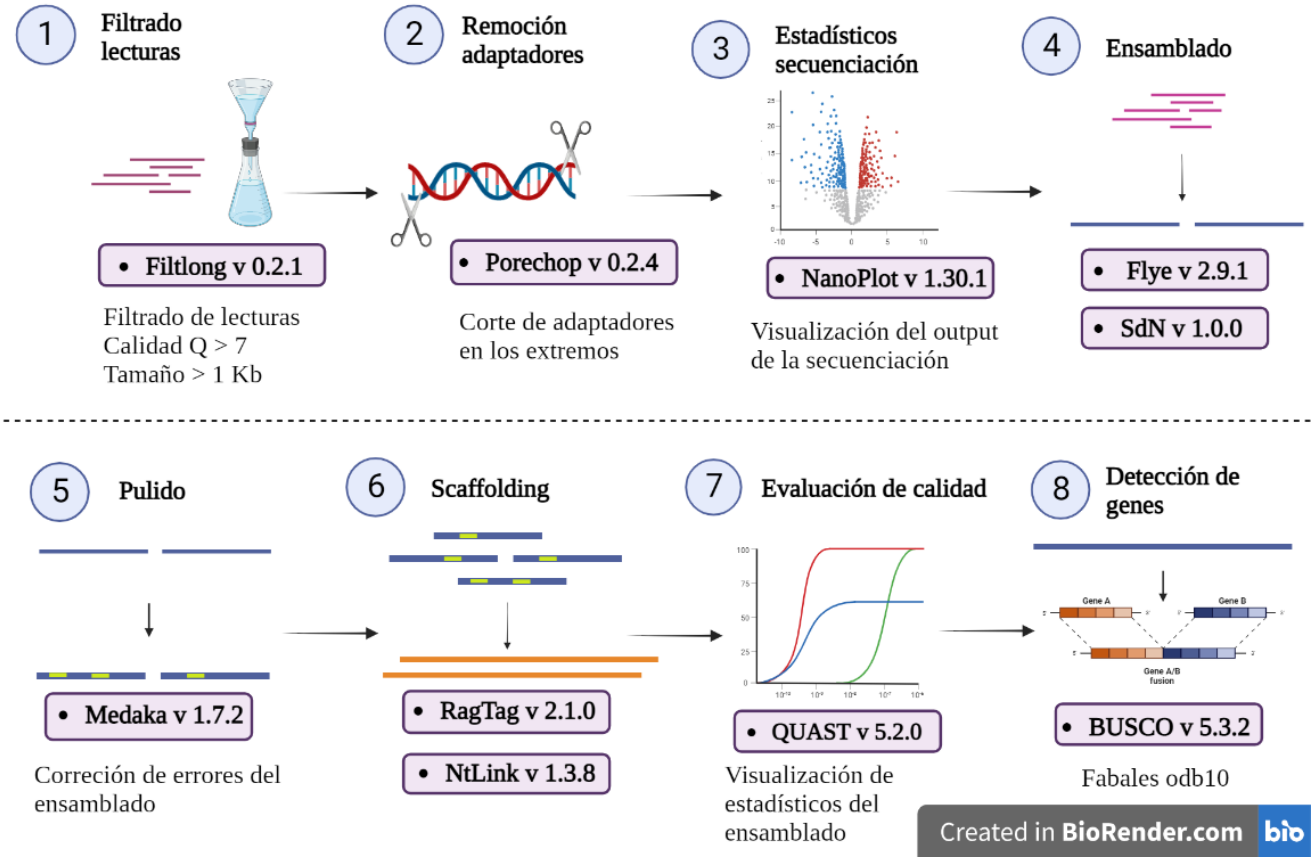

Figura 1. Flujo de trabajo del análisis bioinformático para el ensamblaje delgenoma de *Lupinus mutabilis.* *Esquema realizado en BioRender.*

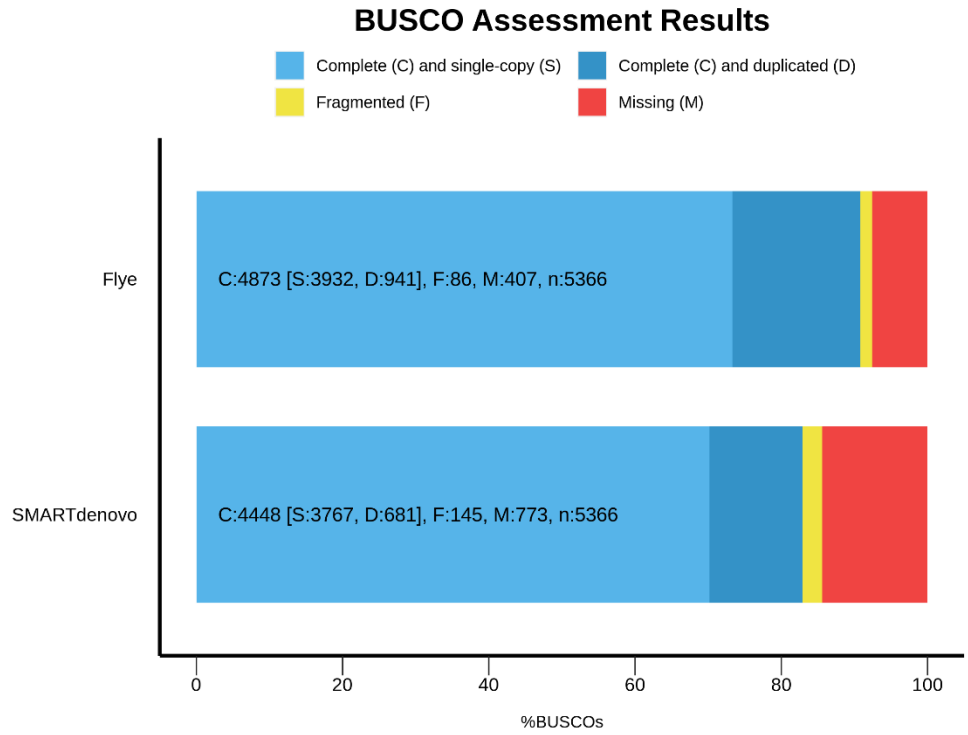

Figura 2. Resultado obtenido de BUSCO para los ensamblados.

Se presentan los porcentajes obtenidos de BUSCOs completos, fragmentados y perdidos para SMARTdenovo y Flye.

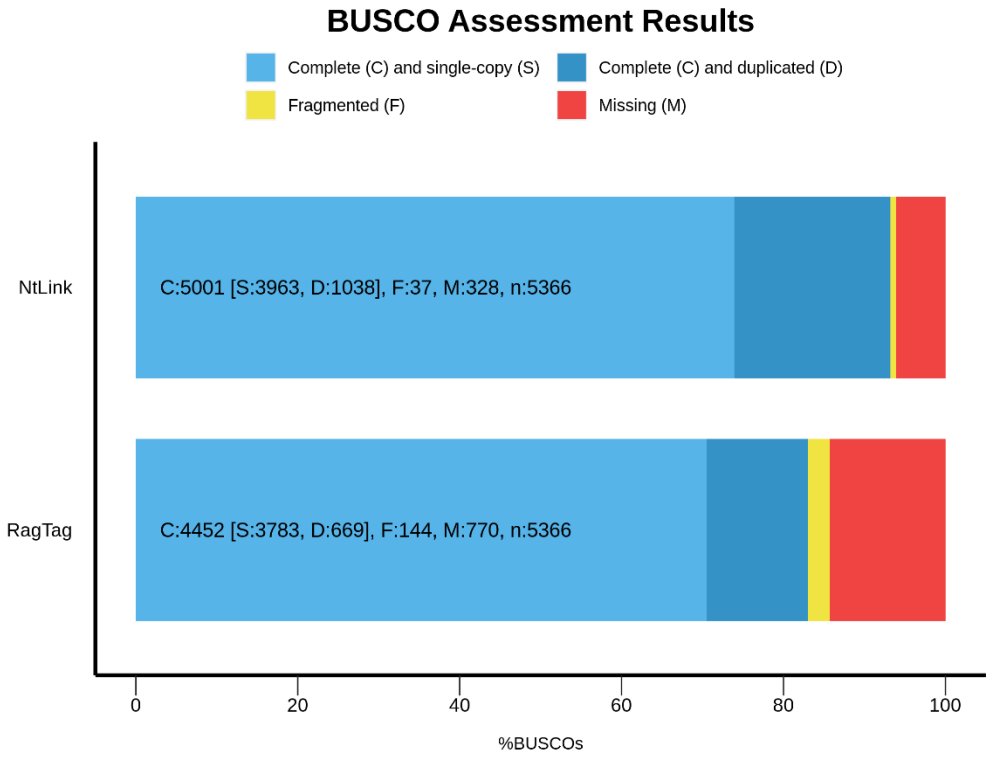

Figura 3. Resultado obtenido de BUSCO para los scaffoldings.

Se presentan los porcentajes obtenidos de BUSCOs completos, fragmentados y perdidos para NtLink y RagTag.

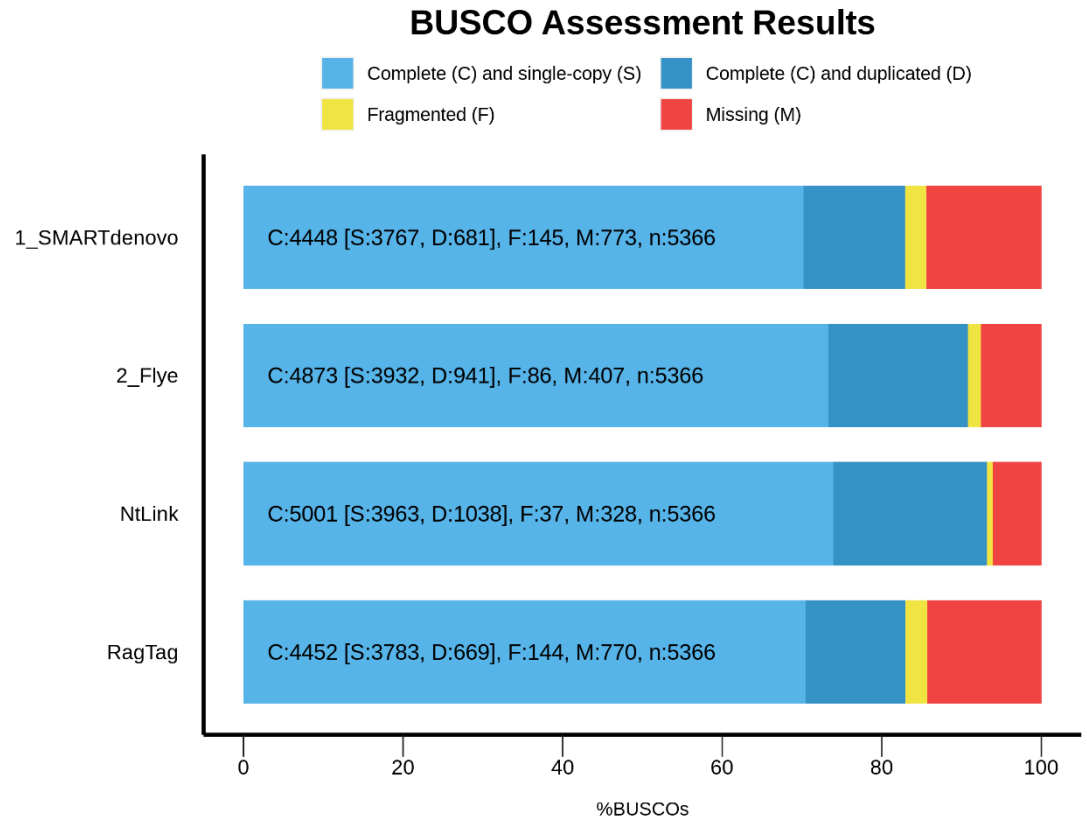

Figura 4. Compilación de los resultados obtenidos de BUSCO para evaluación de integridad del genoma completo.

Se presentan las puntaciones obtenidas de BUSCOs completos, fragmentados y perdidos para Flye, SMARTdenovo, NtLink y RagTag.

8 REFERENCIAS

Alonge, M. et al. (2022). "Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing." *Genome Biology*. https://doi.org/10.1186/s13059-022-02823-7

Amarasinghe, S.L., Su, S., Dong, X. et al. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 21, 30. https://doi.org/10.1186/s13059-020-1935-5

Atchison, G. W., Nevado, B., Eastwood, R. J., Contreras-Ortiz, N., Reynel, C., Madriñán, S., et al. (2016). Lost crops of the Incas: Origins of domestication of the Andean pulse crop tarwi, *Lupinus mutabilis*. *Am. J. Bot.* 103, 1592–1606. doi: 10.3732/ajb.1600171

Carvajal-Larenas, F. E., Linnemann, A. R., Nout, M. J. R., Koziol, M., & van Boekel, M. A. J. S. (2016). *Lupinus mutabilis*: Composition, Uses, Toxicology, and Debittering. *Critical Reviews in Food Science and Nutrition*, 56(9), 1454-1487. https://doi.org/10.1080/10408398.2013.772089

Chen, Z., Erickson, D. L., & Meng, J. (2020). Benchmarking Long-Read Assemblers for Genomic Analyses of Bacterial Pathogens Using Oxford Nanopore Sequencing. *International Journal of Molecular Sciences*, 21(23), 9161. doi:10.3390/ijms21239161

Clements, J.C., Wilson, J., Sweetingham, M.W., Quealy, J., Francis, G. (2012). Male Sterility in three crop *Lupinus* species. *Plant Breeding* 131: 155-163

Coombe L, Li JX, Lo T, Wong J, Nikolic V, Warren RL and Birol I. (2021). LongStitch: high-quality genome assembly correction and scaffolding using long reads. *BMC Bioinformatics* 22, 534. https://doi.org/10.1186/s12859-021-04451-7

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018).

NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics,* 34(15), 2666–2669. doi:10.1093/bioinformatics/bty14

Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., … Lantz, H. (2018). Ten steps to get started in Genome Assembly and Annotation. *F1000Research*, 7, 148. doi:10.12688/f1000research.13598.

Eapen, S. (2008). Advances in development of transgenic pulse crops. *Biotechnology Advances* 26: 162-168

Eastwood, R. J., & Hughes, C. E. (2018). 878. *LUPINUS MUTABILIS*. *Curtis's Botanical Magazine*, 35(2), 134–148. doi:10.1111/curt.12233

French, B. & White, P. (2008). Environmental influences on lupin growth. Chapter 3. In: P White, B French, A Mclarty, eds. Producing lupins, Edition 2. *Department of Agriculture and Food,* Perth, Western Australia pp 27-36.

Gabur, I. & Petru, D. (2023). Chapter 16 - Pearl lupin (*Lupinus mutabilis*): a neglected high protein and oil content crop. Neglected and Underutilized Crops. *Academic Press*, https://doi.org/10.1016/B978-0-323-90537-4.00015-6.

Gulisano, A., Alves, S., Martins, J. N., & Trindade, L. M. (2019). Genetics and Breeding of *Lupinus mutabilis*: An Emerging Protein Crop. *Frontiers in Plant Science*, 10. doi:10.3389/fpls.2019.01385

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). Quast: Quality Assessment Tool for Genome Assemblies. *Bioinformatics*, 29(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hu, J., Wang, Z., Liang, F., Liu, S., Ye, K., & Wang, D. (2023). NextPolish2: A repeat-aware polishing tool for genomes assembled using HiFi long reads. https://doi.org/10.1101/2023.04.26.538352

Hufnagel, B., Marques, A., Marande, W., Sallet, E., Sorriano, A. et al. (2018). Genome sequence of white lupin, a model to study root developmental adaptations. *12th Congress of the International Plant Molecular Biology*, Montpellier, France.

INIAP. (2015). INIAP 450 ANDINO VARIEDAD DE CHOCHO (*Lupinus mutabilis Sweet*). Programa Nacional de Leguminosas y Granos Andinos. Estación Experimental Santa Catalina. *Plegable Divulgativo No. 169*

Jauhal, A. & Newcomb, R. (2021). Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO. *Molecular Ecology Resources*, 21(5), 1416–1421. doi:10.1111/1755-0998.13364

Jentoft, S. (2017). Whole genome sequencing data and de novo draft assemblies for 66 teleost species. *Scientific Data*, 4, 160132. https://doi.org/10.1038/sdata.2016.132

Jung H, Ventura T, Chung JS, Kim W-J, Nam B-H, Kong HJ, et al. (2020) Twelve quick steps for genome assembly and annotation in the classroom. *PLoS Comput Biol* 16(11): e1008325. https://doi.org/10.1371/journal.pcbi.1008325

Keller, J., Rousseau-Gueutin, M., Martin, G. E., Morice, J., Boutte, J., Coissac, E., et al. (2017). The evolutionary fate of the chloroplast and nuclear rps16 genes as revealed through the sequencing and comparative analyses of four novel legume chloroplast genomes from *Lupinus. DNA Res.* 24, 343–358. doi: 10.1093/dnares/dsx006

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546. doi:10.1038/s41587-019-0072-8

Liu, H., Wu, S., Li, A., & Ruan, J. (2021). SMARTdenovo: A de novo assembler using long noisy reads. *Gigabyte* 1–9. https://doi.org/10.46471/gigabyte.15

Loose M, Tyson JR, de Cesare M, Brown BL, Jain M, et al. (2015). MinION analysis and

reference consortium: phase 1 data release and analysis. *F1000Res.*

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*, 14(5), 265–279. doi:10.1016/j.gpb.2016.05.004

Mäkinen, V., Salmela, L. & Ylinen, J. (2012). Normalized N50 assembly metric using gap-restricted co-linear chaining. *BMC Bioinformatics* 13, 255. https://doi.org/10.1186/1471-2105-13-255

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing genomic data quality and beyond. *Current Protocols*, 1, e323. doi: 10.1002/cpz1.323

Martins, J. M. N., Talhinhas, P., Sousa, R. B. D. (2016). Yield and seed chemical composition of *Lupinus mutabilis* in Portugal. *Revista de Ciências Agrárias* 39, 518–525. doi: 10.19084/RCA16079

Molina-Poveda, C., Lucas, M., & Jover, M. (2013). Evaluation of the potential of Andean lupin meal (*Lupinus mutabilis* Sweet) as an alternative to fish meal in juvenile Litopenaeus vannamei diets. *Aquaculture,* 410-411, 148-156. https://doi.org/10.1016/j.aquaculture.2013.06.007

NIH. (2023). *SECUENCIACIÓN DE ADN A TRAVÉS DE NANOPOROS*. Recuperado de: https://www.genome.gov/es/genetics-glossary/Nanopore-DNA-Sequencing.

Oxford Nanopore Technologies Ltd. (2018). *Medaka. PyPI*. Recuperado de: https://pypi.org/project/medaka/

Osorio, C. E., Udall, J. A., Salvo-Garrido, H., Maureira-Butler, I. J. (2018). Development and characterization of InDel markers for *Lupinus luteus L.* (Fabaceae) and cross-species amplification in other Lupin species. *Electron. J. Biotechnol.* 31, 44–47. doi: 10.1016/j.ejbt.2017.11.002

Peralta, E., N. Mazón, A. Murillo, M. Rivera, D. Rodríguez, L. Lomas, C. Monar. (2012). Manual Agrícola de Granos Andinos: Chocho, Quinua, Amaranto y Ataco. *Cultivos, variedades y costos de producción.* Tercera edición. Publicación Miscelánea No. 69. programa Nacional de Leguminosas y Granos andinos. Estación Experimental Santa Catalina. INIAP. Quito, Ecuador. 68 p.

Salson, M. et al. (2023). Improved assembly of the pearl millet reference genome using Oxford Nanopore long reads and optical mapping. *G3 Genes|Genomes|Genetics*, Volume 13, Issue 5. https://academic.oup.com/g3journal/article/13/5/jkad051/7073532

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., … Yorke, J. A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, 22(3), 557–567. doi:10.1101/gr.131383.111

Schmidt, M. H.-W., Vogel, A., Denton, A. K., Istace, B., Wormit, A., Geest, H. van de, Bolger, M. E., Alseekh, S., Maβ, J., Pfaff, C., Schurr, U., Chetelat, R., Maumus, F., Aury, J.-M., Fernie, A. R., Zamir, D., Bolger, A. M., & Usadel, B. (2017). Reconstructing the Gigabase Plant Genome of Solanum pennellii using Nanopore Sequencing (p. 129148). *bioRxiv*. https://doi.org/10.1101/129148

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. *Gene Prediction*, 227–245. doi:10.1007/978-1-4939-9173-0_14

Sharma, P., Al-Dossary, O., Alsubaie, B., Al-Mssallem, I., Nath, O., Mitter, N., Margarido, G. R. A., Topp, B., Murigneux, V., Masouleh, A. K., Furtado, A., & Henry, R. J. (2021). Improvements in the Sequencing and Assembly of Plant Genomes (p. 2021.01.22.427724). bioRxiv. https://doi.org/10.1101/2021.01.22.427724

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212. doi:10.1093/bioinformatics/btv351

Smarda, P., Bureš, P., Smerda, J., & Horová, L. (2012). Measurements of genomic GC content in plant genomes with flow cytometry: A test for reliability. *The New Phytologist*, 193(2), 513-521. https://doi.org/10.1111/j.1469-8137.2011.03942.x

Sohn, J., & Nam, J.-W. (2016). The present and future ofde novowhole-genome assembly. *Briefings in Bioinformatics*, bbw096. doi:10.1093/bib/bbw096

Tan, M. H., Austin, C. M., Hammer, M. P., Lee, Y. P., Croft, L. J., & Gan, H. M. (2018). Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (*Amphiprion ocellaris*) genome assembly. *GigaScience,* 7(3). doi:10.1093/gigascience/gix137

Teh, B., Lim, K., Yong, C. et al. (2017). The draft genome of tropical fruit durian (*Durio zibethinus*). *Nat Genet* 49:1633–1641. https://doi.org/10.1038/ng.3972

Utturkar, S., Klingeman, D., Land, M., Schadt, C., Doktycz, M., Pelletier, D. et al. (2014). Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics*. 30:2709–16.

Varshney, R., Shi, C., Thudi, M., Mariac, C., Wallace, J., Qi, P., Zhang, H., Zhao, Y., Wang, X., Rathore, A., et al. (2017). Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nat Biotechnol*. 35(10):969–976. http://dx.doi.org/10.1038/nbt.3943.

Wambugu, P.W.; Henry, R.; Browne, L (2022). Supporting in situ conservation of the genetic diversity of crop wild relatives using genomic technologies. *Mol. Ecol.*

Wattam, A. R., Brettin, T., Davis, J. J., Gerdes, S., Kenyon, R., Machi, D., … Yoo, H.

(2017). Assembly, Annotation, and Comparative Genomics in PATRIC, the All Bacterial Bioinformatics Resource Center. *Methods in Molecular Biology*, 79–101. doi:10.1007/978-1-4939-7463-4_4

Waterhouse, R., Seppey, M., Simao, F. et al. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* 35:543–548. https:// doi.org/10.1093/molbev/msx319

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics*, 3(10). doi:10.1099/mgen.0.000132

Wick, R. & Menzel, P. (2021). *Filtlong.* Recuperado de: https://github.com/rrwick/Filtlong

Zoga, M., Pawelec, A., Galek, R., Sawicka-Sienkiewicz, E. (2008). "Morphological, cytological and molecular characteristics of parents and interspecific hybrid (*Lupinus mutabilis* LM-13 × *Lupinus albus* sensu lato)" In Lupins for health and wealth. *Proceedings of the 12th International Lupin Conference*, Fremantle, Western Australia.

9 ANEXOS

ANEXO 1. NanoPlot de los datos obtenidos a partir de la secuenciación de lecturas largas de ONT.

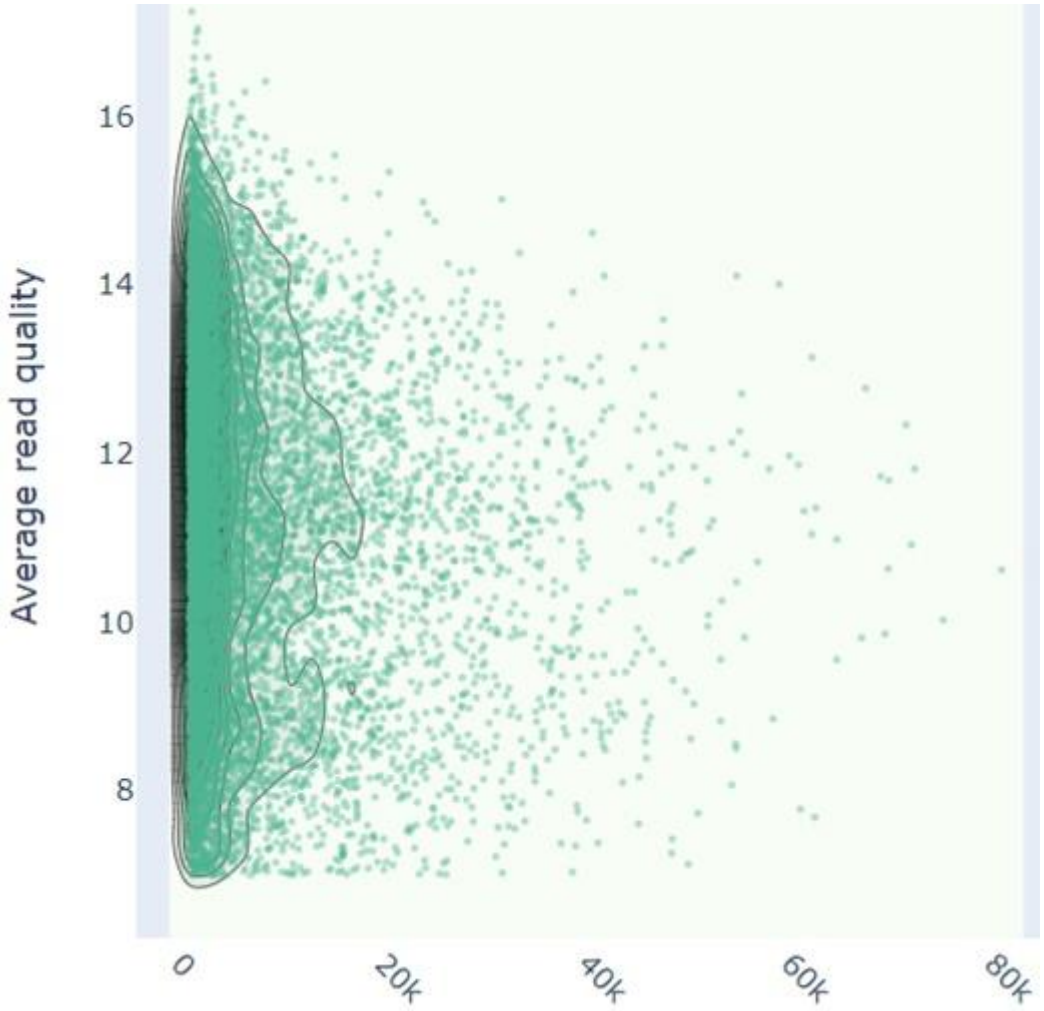

Diagrama de puntos: Longitud de lectura frente a la calidad de lectura promedio para secuencias adquiridas mediante tecnología ONT. Se muestra la distribución de todas las lecturas. La longitud de lectura en kb se muestra en el eje de abscisas y la calidad de lectura promedio, con límites de calidad de 0 a 16, se muestra en el eje de ordenadas.

Datos proporcionados por: Milton Gordillo.