UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

FACULTAD DE INGENIERÍA, ARQUITECTURA Y DISEÑO

ENSAMBLE Y ANOTACIÓN DEL GENOMA DEL CLOROPLASTO DE *Dunaliella salina* DE GUERRERO NEGRO E IDENTIFICACIÓN DE REGIONES GENÓMICAS PARA LA EXPRESIÓN DE PROTEÍNAS RECOMBINANTES

TESIS

PARA CUBRIR LOS REQUISITOS NECESARIOS PARA OBTENER EL TÍTULO DE

BIOINGENIERO

PRESENTA:

MYRNA VANESSA CÁRDENAS BELTRÁN

ENSENADA, BAJA CALIFORNIA, 2021

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

FACULTAD DE INGENIERÍA, ARQUITECTURA Y DISEÑO

ENSAMBLE Y ANOTACIÓN DEL GENOMA DEL CLOROPLASTO DE *Dunaliella salina* DE GUERRERO NEGRO E IDENTIFICACIÓN DE REGIONES GENÓMICAS PARA LA EXPRESIÓN DE PROTEÍNAS RECOMBINANTES

TESIS

PARA CUBRIR LOS REQUISITOS NECESARIOS PARA OBTENER EL TITULO DE
BIOINGENIERO

PRESENTA:
MYRNA VANESSA CÁRDENAS BELTRÁN

Aprobada por:

Dr. Dante Alberto Magdaleno Moncayo
Director

Dra. Haydeé López Rodríguez
Codirector

Dra. Claudia Mariana Gómez Gutiérrez
Sinodal

Dr. Priscy Alfredo Luque Morales
Sinodal

Dr. Gerardo Salvador Romo CárdenasSinodal

Resumen

La microalga verde *Dunaliella salina* es un organismo con importantes aplicaciones biotecnológicas, tales como la producción de carotenoides, vitaminas, enzimas, ácidos grasos, entre otros productos. Para llevar a cabo la expresión de proteínas recombinantes en esta microalga, es necesario conocer su genómica. En el presente trabajo, se realizaron las dos estrategias existentes para obtener el ensamble del plastoma de la cepa aislada en una laguna hipersalina de Guerrero Negro, Baja California Sur, México. Se realizó el ensamble por referencia con las cepas aisladas en Australia (CCAP 19/18) y San Quintín, Baja California, México (SQ), respectivamente, seleccionando el alineamiento con la cepa CCAP 19/18, debido al porcentaje de concordancia entre las secuencias, constituyendo una longitud final de 263,577 pb. Por otra parte, se realizó la anotación del genoma ensamblado, en la que se determinaron 122 genes. Finalmente, se diseñó un casete para la expresión de proteínas recombinantes y se identificaron regiones potenciales para la inserción del casete, proponiéndolos con el fin de aplicarlos en trabajos futuros que tengan el objetivo de manipular genéticamente el cloroplasto de *D. salina* GN con fines biotecnológicos.

Palabras clave: *Dunaliella salina*, cloroplasto, ensamble, anotación, genoma, casete de expresión, proteínas recombinantes

Índice general

Índice de figuras

Índice de tablas

I. Introducción

I.1 Sistemas de expresión de proteínas recombinantes

Las proteínas recombinantes tienen diversas aplicaciones en las industrias farmacéutica, cosmética, alimenticia y química, así como en la agricultura y tratamiento de desechos.

Los sistemas de expresión de proteínas recombinantes son construcciones genéticas diseñadas para producir una proteína dentro o fuera de un entorno. Estos sistemas consisten de tres componentes esenciales para la producción de proteínas recombinantes. El primer componente a considerar es el entorno biológico para el sistema, en el que se encuentran la maquinaria y energía requeridas para la síntesis de proteínas. El segundo elemento es el vector, mismo que permite la introducción de la información genética en el huésped, conteniendo factores regulatorios para la replicación del material genético. Finalmente, el tercer componente se trata del casete de expresión, elemento que se incorpora al vector que contiene el marco de lectura abierto con los elementos necesarios para la transcripción y traducción de la proteína a producir (Fisher *et al.*, 2016).

Los principales sistemas de expresión de proteínas recombinantes utilizados comercialmente son los de tipo bacteriano y mamífero, sin embargo, se han empleado esfuerzos para que se utilicen en mayor frecuencia los sistemas de expresión en plantas (Shanmugaraj *et al.*, 2020).

Los sistemas de expresión en animales son de gran interés y están en continuo desarrollo, en comparación con los de tipo bacteriano, son de mayor costo (Clark y Pazdernik, 2016, p. 356). Como ejemplo de estos sistemas de expresión, se destaca el

constructo del gen que codifica para la β-caseína en la cabra, en cual el gen de esta proteína es reemplazado por el material genético de interés, manteniendo el promotor y elementos reguladores endógenos, permitiendo así la expresión de proteínas recombinantes por medio de la glándula mamaria de dicho animal (Shepelev *et al.*, 2018).

Las levaduras son sistemas de expresión que proveen la ventaja de expresar proteínas en un amplio rango de escalamiento, además de identificarse como un organismo seguro de utilizarse para tales fines. Se destacan dos tipos de sistemas de expresión en este grupo de organismos, los vectores plasmídicos episómicos que funcionan tanto en levaduras como en bacterias, conteniendo el material genético de interés, promotor y terminadores específicos y los vectores que además de contener las características anteriores, contienen secuencias que les confiere la capacidad de integrarse a los cromosomas de la levadura, ofreciendo la ventaja de no degradarse en comparación con los plásmidos episómicos, que pueden perderse durante las divisiones celulares (Clark y Pazdernik, 2016, p. 353).

Los sistemas de expresión de proteínas recombinantes en algas, ofrecen las ventajas de expresar una amplia variedad de proteínas recombinantes complejas, así como de requerir medios sencillos y económicos para cultivarse (Carrera *et al.,* 2018).

I.2 Sistemas de expresión de proteínas recombinantes en algas

Las algas son organismos eucariotas que pueden ser unicelulares o multicelulares. Son capaces de vivir en diferentes hábitats, desde lagunas de agua dulce hasta océanos profundos. Estos organismos contienen pigmentos tales como la clorofila que le permiten realizar la fotosíntesis (Manivasagan y Kim, 2015).

Existen diferentes especies de algas, las cuales se clasifican en los siguientes grupos: Rhodophyta, Glaucocystophyta, Chlorophyta, Chlorarachniophyta, Euglenophyta, Heterokonta, Haptophyta y Cryptophyta (Raven y Giordano, 2014).

Estos organismos se encuentran en formas macroscópicas (macroalgas) y microscópicas (microalgas) considerándose estas como las de mayor presencia dentro de la naturaleza. Las microalgas cultivables se pueden clasificar en algas rojas, crisófitas y algas verdes (Yan *et al.,* 2016).

En el caso específico de las algas del grupo Chlorophyta, existen diferentes procedimientos y sistemas de expresión dependiendo del organismo considerado para la modificación genética. Es posible transformar a este grupo de algas mediante bombardeo de micropartículas, electroporación y plásmido Ti (Gong *et al.,* 2011).

Considerados los métodos de transformación de algas, es posible adentrarse en los sistemas de expresión en algas. Primero, el vector que contiene el material genético a expresar, lleva consigo un promotor específico, un gen de interés y un terminador de transcripción fuerte, este vector se introduce por uno de los métodos de transformación anteriormente destacados, dependiendo del tipo de alga a transformar (Oey *et al.,* 2014).

Los transformantes primarios, se seleccionan por medio de marcadores que confieren alguna resistencia, para el caso de las algas es conveniente utilizar herbicidas. Los transformantes seleccionados se colocan en condiciones controladas de crecimiento. Finalmente se cultivan dichos organismos para obtener así grandes volúmenes de producción de proteínas recombinantes (Walker *et al.*, 2005).

I.3 *Dunaliella salina*

Entre las microalgas verdes, se encuentra un organismo con gran potencial biotecnológico, *Dunaliella salina* (*D. salina*), organismo reportado por primera vez hace aproximadamente 100 años. Se encuentra mayormente en ambientes hipersalinos, tales como lagos salados y salinas (Oren, 2005).

El ciclo de vida de *D. salina* es considerado complejo, debido a que los flagelos se fusionan formando un puente citoplasmático, creando un cigoto que germina al romperse la envoltura celular, liberando 32 células hijas haploides. Además, existe una etapa vegetativa que ciertas microalgas desarrollan debido a que se encuentran en ambientes de alta salinidad no óptimo para su crecimiento (Borowitzka y Siva, 2007).

Las principales aplicaciones biotecnológicas de *D. salina*, son la producción de carotenoides, vitaminas, antioxidantes, ácidos grasos poliinsaturados, reactores biológicos, minerales y enzimas. Además del rendimiento mencionado de estas moléculas, se identifica la capacidad de esta microalga para la remoción de metales pesados al implementarse en el tratamiento de aguas residuales (Hosseini *et al.*, 2009).

La expresión de proteínas recombinantes en esta microalga puede llevarse a cabo tanto en la mitocondria, cloroplasto y núcleo (Siddiqui *et al.,* 2020). Sin embargo, para fines objetivos, es necesario destacar la importancia y ventajas que confiere la expresión de proteínas recombinantes en los cloroplastos.

I.4 Expresión de proteínas recombinantes en cloroplastos

Primero, resulta importante subrayar que el genoma del cloroplasto es el que se ha estudiado en mayor medida en comparación con el nuclear y mitocondrial, ofreciendo así la capacidad de comprender el mecanismo del plastoma y de encontrar regiones específicas para llevar a cabo las modificaciones genéticas pertinentes (Bock, 2007).

Las células vegetales se componen de numerosos cloroplastos, mismos que contienen varios nucleoides, que a su vez incluyen abundantes copias de plastomas, esto permitiendo llevar a cabo la producción de proteínas recombinantes en gran medida (Kuroiwa, 1991). Los genes contenidos en estos organelos, codifican para obtener productos proteicos que están implicados en la fotosíntesis, así como en las vías metabólicas de la planta (Wicke *et al.,* 2011).

Los cloroplastos ofrecen la capacidad de expresar mayor cantidad de proteínas recombinantes en comparación con el núcleo, que al poseer sistemas de silenciamiento transgénico disminuye la capacidad de llevar a cabo modificaciones genéticas. Además, este organelo proporciona la capacidad de integración específica de los genes exógenos debido al mecanismo de doble recombinación homóloga que lo caracteriza (Scotti *et al.,* 2013).

Otro factor importante a considerar dentro de las modificaciones genéticas en los cloroplastos, es la capacidad que tiene para hospedar genes exógenos, debido a que este organelo sí puede retenerlos en contraste con el núcleo, que suele perder la información introducida (Meyers *et al.* ,2010).

Los genes exógenos son introducidos mediante un vector previamente diseñado, por algún método de transformación seleccionado e integrándolo en el genoma del cloroplasto mediante doble recombinación homóloga entre las secuencias homólogas en el vector y las secuencias homólogas en el cloroplasto (Lutz *et al.,* 2007).

La producción de proteínas recombinantes en los cloroplastos está habitualmente regulada por las condiciones de luz en la que se encuentre el cultivo, esto debido a que los procesos de transcripción para obtener estas proteínas se ven afectados por los elementos reguladores endógenos (Carrera *et al.*, 2018). Además, se deben de contemplar los efectos que tiene la optimización de codones para producir las proteínas de interés.

I.5 Genomas de la microalga *Dunaliella salina*

Actualmente, el genoma nuclear de *D. salina* se encuentra en proyecto de secuenciación (Smith *et al.*, 2010). Por otra parte, a la fecha solo se encuentran ensamblados dos genomas de cloroplastos de cepas de esta microalga, *D. Salina* CCAP (269,044 pb) (Smith *et al.*, 2010), *D. Salina* SQ (243,635 pb) (Lopez *et al.,* 2017). Dados los escasos cloroplastos ensamblados de *D. salina*, resulta un campo interesante por explorar.

I.6 Sistemas de secuenciación

Los genomas del cloroplasto de cepas de *D. salina* aisladas de lagunas hipersalinas, se han logrado ensamblar y anotar a partir de datos obtenidos por distintas plataformas de secuenciación de ADN. El proceso de secuenciación se refiere

a la determinación de la secuencia de los pares de bases (A, T, G, C) de una molécula de ADN.

A lo largo de los años, han existido diferentes métodos de secuenciación, como el método reportado por Maxam-Gilbert en 1977 (Maxam y Gilbert, 1977), en el cual, el rendimiento es bajo debido a que solo se obtiene información de unos pocos nucleótidos. En el mismo año, Sanger desarrolla otro método de secuenciación (Sanger *et al.,* 1977), procedimiento que genera mayor cantidad de datos y fue utilizado desde 1977 hasta aproximadamente el año 2005, sin embargo, debido a los altos costos y tiempos de operación se buscaron alternativas, surgiendo así diferentes tecnologías que permitieron el subsecuente desarrollo de las plataformas de secuenciación de la siguiente generación (Next Generation Sequencing, NGS por sus siglas en inglés) (Krishna *et al.*, 2019).

Las plataformas de NGS se consideran de mayor rendimiento, rápidas y económicas en comparación con el método de Sanger. Se dividen en segunda y tercera generación dependiendo del tiempo de aparición de dichas plataformas.

Existen diferentes metodologías entre las plataformas de secuenciación más utilizadas actualmente, las cuales se mencionan enseguida. El método de la plataforma Ion torrent, se basa en la detección de los iones de hidrógeno (Rothberg *et al.*, 2011), así como los cambios en los niveles de pH, considerándose indicadores de la incorporación de nucleótidos (Goodwin *et al.,* 2016).

La plataforma de secuenciación de tercera generación, Nanopore, trabaja mediante una membrana eléctrica nanométrica, en la que al aplicar voltaje, produce

residuos eléctricos que de manera proporcional indican la presencia de los nucleótidos en la secuencia analizada (Clarke *et al.*, 2009).

PacBio RS II, se basa en la tecnología SMRT (Single-Molecule Real-Time por sus siglas en inglés), misma que detecta los nucleótidos presentes en las secuencias analizadas mediante la localización de fluorescencia en dichas moléculas durante el análisis (Schadt *et al.,* 2010).

Illumina solexa sequencing, se rige mediante la tecnología de secuenciación por síntesis (Sequencing By Synthesis, SBS por sus siglas en inglés), en la que se utilizan cuatro nucleótidos modificados, así como de una DNA polimerasa específica que, al llevar a cabo el proceso de polimerización, permite la identificación de los nucleótidos presentes en la secuencia de ADN analizada mediante la fluorescencia emitida (Heather y Chain, 2016). De manera general, Illumina sequencing es la plataforma de secuenciación mayormente utilizada en comparación con las mencionadas, debido a las características de simplicidad que a esta plataforma le confiere (H. y G., 2018) y la gran cantidad de datos que se obtienen por corrida de secuenciación.

I.7 Ensamble de genomas

El ensamble del genoma, es el proceso en el que se ordenan las lecturas de ADN obtenidas durante la secuenciación hasta obtener la secuencia correcta del genoma de interés. Para llevar a cabo el ensamble, es necesario utilizar algún programa ensamblador en el que los datos de entrada sean las lecturas de nucleótidos registradas (Kalyanaraman *et al.*, 2011).

Actualmente existen dos estrategias para ensamblar un genoma, el ensamble por referencia, en donde las lecturas obtenidas mediante la secuenciación se alinean

con la secuencia de un genoma de referencia existente en bases de datos. Por otra parte, si no existe una secuencia de referencia, se utiliza la estrategia de ensamble *de novo*. Este proceso es consecutivo, debido a que las lecturas al ensamblarse se convierten en contigs, el ensamble de contigs en supercontigs y finalmente se fusionan en cromosomas. El proceso anteriormente descrito, no siempre logra ensamblar completamente todos los genomas, debido a la pérdida de información durante el proceso de secuenciación (Choudhuri, 2014).

Para lograr el ensamble *de no*vo, existen diferentes programas que funcionan bajo ciertos algoritmos. El algoritmo de tipo Greedy compara todos los fragmentos por pares de ADN y une aquellos que se superpongan entre sí. El proceso iterativo termina una vez que se logre la mayor superposición entre las lecturas (Miller *et al.,* 2010).

El algoritmo de gráficos de De Bruijn funciona mediante los pares de bases consecutivos identificados como k-mers, mismos que son representados como nodos, creando un arco entre ellos en caso de que se identifique alguna superposición (Compeau *et al.*, 2011).

Por su parte, el algoritmo de consenso de diseño de superposición (OLC del inglés Overlap-Layout-Consensus) genera un diseño a partir de las superposiciones encontradas entre las lecturas analizadas, creando un gráfico que permite obtener una secuencia de consenso como archivo de salida (Z. Li *et al.*, 2012). El gráfico de cadenas (string graph por su significado en inglés) es una variante del algoritmo anteriormente descrito debido a que tiene el mismo funcionamiento y además elimina las secuencias que no son necesarias para el ensamble (Khan *et al.*, 2018).

Determinar los genomas de los organismos, permite encontrar regiones genéticas para diversos usos, por ejemplo, integrar sistemas de expresión de proteínas recombinantes. En el presente trabajo, se pretende ensamblar y anotar el genoma del cloroplasto de *D. salina* GN aislada de la salina de Guerrero Negro B.C.S., México, utilizando datos de secuenciación obtenidos mediante la plataforma de secuenciación Illumina MiSeq y localizar regiones genómicas para la introducción de casetes de expresión de proteínas recombinantes.

II. Justificación

La microalga verde *Dunaliella salina* tiene un gran potencial biotecnológico. Las principales aplicaciones son la producción de carotenoides, vitaminas, antioxidantes, ácidos grasos, enzimas, entre otros productos. Este organismo puede expresar proteínas recombinantes en el cloroplasto, mitocondria y núcleo. El cloroplasto ofrece la ventaja de expresar proteínas en abundante cantidad, debido a los numerosos plastomas contenidos en las células vegetales.

Para llevar a cabo modificaciones genéticas que permitan la expresión de proteínas recombinantes, es necesario conocer la genómica del organismo. Actualmente, el genoma nuclear de *D. salina* se encuentra en proyecto de secuenciación, los genomas de las mitocondrias y cloroplastos de las cepas de San Quintín, México (SQ), Guerrero Negro, México (GN) y Australia, se encuentran en la base de datos NCBI, a excepción del plastoma de la cepa GN.

Debido a los escasos genomas secuenciados y ensamblados, en el presente trabajo se destaca el ensamble y anotación del genoma del cloroplasto de la cepa GN, con el fin de generar conocimiento que permita el análisis de la arquitectura genómica, reconocimiento de genes y secuencias reguladoras en dicho organelo. Aunado a lo anterior, dados los pocos plastomas reportados de esta microalga, se han realizado escasas modificaciones genéticas al cloroplasto de *D. salina*, por lo que, en este proyecto, se presenta el diseño de un casete para la expresión de proteínas recombinantes, así como la identificación de regiones potenciales para la inserción del casete diseñado, con el fin de implementarlo en trabajos futuros que tengan el objetivo de modificar el plastoma de *D. salina*.

III. Hipótesis

Utilizando métodos de ensamble y anotación de genomas, será posible conocer la secuencia y arquitectura genómica del cloroplasto de la cepa de *Dunaliella salina* aislada en Guerrero Negro, B.C.S., México, e identificar regiones genéticas para la inserción de casetes de expresión de proteínas recombinantes.

IV. Objetivos

IV.1 Objetivo general

- Ensamblar y anotar el genoma del cloroplasto de la cepa de la microalga *Dunaliella salina* aislada de Guerrero Negro, B.C.S., México e identificar regiones potenciales para la inserción de casetes de expresión de proteínas recombinantes.

IV.2 Objetivos particulares

- Ensamblar *de novo* el genoma del cloroplasto de *Dunaliella salina* GN.
- Ensamblar por referencia el genoma del cloroplasto de *Dunaliella salina* GN.
- Anotación del genoma del cloroplasto de *Dunaliella salina* GN.
- Llamado de SNPs (Single Nucleotide Polymorphism).
- Diseño *in silico* de un casete para la expresión de proteínas recombinantes en el cloroplasto de *Dunaliella salina* GN.
- Identificación de regiones potenciales para inserción de casetes de expresión de proteínas recombinantes.

V. Metodología

V.1 Extracción de ADN

Se utilizaron secuencias genómicas obtenidas de la siguiente manera:

La muestra del cloroplasto fue obtenida de la cepa de *D. salina,* aislada en Guerrero Negro, Baja California Sur, México (Magdaleno y Stephano, 2017).

En dicho estudio, se cultivó la muestra de la microalga hasta la fase media exponencial en 200 mL de medio líquido de Johnson modificado (Feng *et al.,* 2009) (10 mM KNO_3, 50 mM $NaHCO_3$, 5 mM $MgSO_4 \cdot 7H_2O$, 0.4 mM KH_2PO_4, 2µM $FeCl_3 \cdot 6H_2O$, 5 µM EDTA, 7 µM $MnCl_2 \cdot 4H_2O$, 1 µM $CuCl_2 \cdot 2H_2O$, 1 µM $ZnCl_2$, 1 µM $CoCl_2 \cdot 6H_2O$, 1 µM $(NH_4)\ Mo_7O_{24} \cdot 4H_2O$, 185 µM H_3Bo_3, 0.2 mM $CaCl_2$), a una concentración de 250 mM de NaCl.

El cultivo se incubó a una temperatura de 4 °C en oscuridad por 48 horas, para disminuir los niveles de almidón. Después de la incubación, se procedió a centrifugar el cultivo a 3000 g por 30 minutos, descartando el sobrenadante y homogeneizándolo por 5 minutos en 30 mL del buffer de aislamiento frío (1.25 M NaCl, 50 mM Tris-HCl (pH 8.0), 5 mM EDTA, 0.1% BSA (w/v), 0.1% b-mercaptoethanol (v/v).

El homogenado se filtró a través de dos capas de la membrana Miracloth (Merck) y el resultado se colocó en tubos falcon de 50 mL. El filtrado se centrifugó a 3000 g por 10 minutos y se resuspendió el pellet en 10 mL de buffer de aislamiento frío. Se centrifugó por 10 minutos a 3000 g y finalmente se resuspendió el pellet en 10 mL de H_2O destilada (Shi *et al.*, 2012).

El procedimiento anteriormente descrito, se realizó con el objetivo de aislar mitocondrias, sin embargo, durante el proceso se identificó que también se aislaron

cloroplastos. La extracción del ADN mitocondrial y cloroplástico se realizó mediante el protocolo establecido por el kit miniprep Axyprep multisource genomic DNA (Axygene). La integridad del ADN extraído se verificó por medio de un gel de agarosa al 1% y se cuantificó usando un nanodrop.

Finalmente, se separaron 3 microgramos del ADN total para secuenciarse en la plataforma Illumina MiSeq en IGM genomics center de UCSD, con bibliotecas de tamaño de inserto con un intervalo de 400 a 900 nucleótidos y lecturas de 300 bases paired-end.

V.2 Preprocesamiento

Los análisis de los datos fueron realizados en una computadora portátil de 4 núcleos, 8 GB de memoria RAM y 500 GB de disco duro. El sistema operativo utilizado fue Windows Subsystem for Linux, basado en Ubuntu 20.04.

La plataforma Illumina (Shen *et al.*, 2005), utiliza la tecnología de secuenciación por química de síntesis (Sequencing by Synthesis, SBS por sus siglas en inglés). El software que integra SBS (Ambardar *et al.,* 2016), almacena cada ciclo secuenciado en archivos de llamada de base binaria (BCL, Binary Base Call, por su significado en inglés). Para utilizar los BCL generados en herramientas de tratamientos de datos actualizados, deben de convertirse a formato FASTQ.

El formato FASTQ es un archivo que contiene los datos de la secuenciación. En una corrida de lectura única, se genera un archivo de lectura 1 FASTQ (R1). Por otra parte, si se realiza una lectura de extremo emparejado, se generan dos archivos, FASTQ (R1) y FASTQ (R2) respectivamente, comprimiéndolos con la extensión *.fastq.gz.

Los archivos generados se conforman de cuatro líneas, en el que la primera línea muestra la información referente a lectura de la secuenciación, presentando el informe dependiendo del software que se haya utilizado para convertir el archivo de formato BCL a FASTQ.

La segunda línea brinda información sobre la secuencia, conformándose de A, T, G, C o N en caso de que no se identifique al nucleótido secuenciado. El separador se encuentra en la tercera fila, representado por un signo de más (+). Por último, la cuarta línea indica los puntajes de calidad de las llamadas de base.

Las secuencias crudas se sometieron al programa FastQC (Andrews, 2010). Este software proporciona un informe de la calidad de las lecturas que permite identificar problemas que hayan surgido durante la secuenciación o en el material de biblioteca inicial.

FastQC realiza el análisis de los datos mediante diferentes secciones. La forma interactiva de este programa, muestra las evaluaciones de los análisis en el lado izquierdo de la pantalla, considerando diferentes símbolos y colores dependiendo de las calificaciones obtenidas, en las que, el módulo se considera aprobado si muestra una marca verde, indica la revisión de los datos si reporta un triángulo naranja y finalmente, la sección se considera reprobada si muestra una cruz roja. Las evaluaciones finales deben de considerarse solo como un indicativo para el tratamiento de los datos.

Una vez contemplados los resultados de las evaluaciones de calidad, se introdujeron las lecturas crudas en el programa Trimmomatic (Bolger *et al.*, 2014).

Esta herramienta de línea de comandos, permite realizar recortes y eliminar adaptadores en los archivos de secuenciación de formato FASTQ de Illumina.

En este caso específico, se consideró un tratamiento estándar reportado en Trimmomatic (Bolger *et al.*, 2014), en el que se eligieron algunas opciones para lograr el filtrado de las lecturas, mismas que se mencionan a continuación.

Se seleccionó ILLUMINACLIP, comando que retira los adaptadores de secuenciación, así como algunas lecturas específicas de Illumina. La opción AVGQUAL, elimina las secuencias en las que el nivel de calidad se encuentre por debajo del definido por el usuario, en este caso, se consideró el puntaje 30.

Se definió el nivel 25 en SLIDINGWINDOW, comando que escanea las lecturas y las recorta en caso de que la calidad esté por debajo de la especificada.

Por último, la opción MAXINFO, recorta secuencias considerando la longitud de la lectura y la tasa de error de secuenciación, definiendo este parámetro en al menos 100 pares de bases y tasa de error en 0.3 para lograr una mayor exactitud en la secuencia.

Una vez obtenidas las lecturas limpias, se introdujeron en diferentes programas para realizar el ensamble *de novo* y por referencia respectivamente.

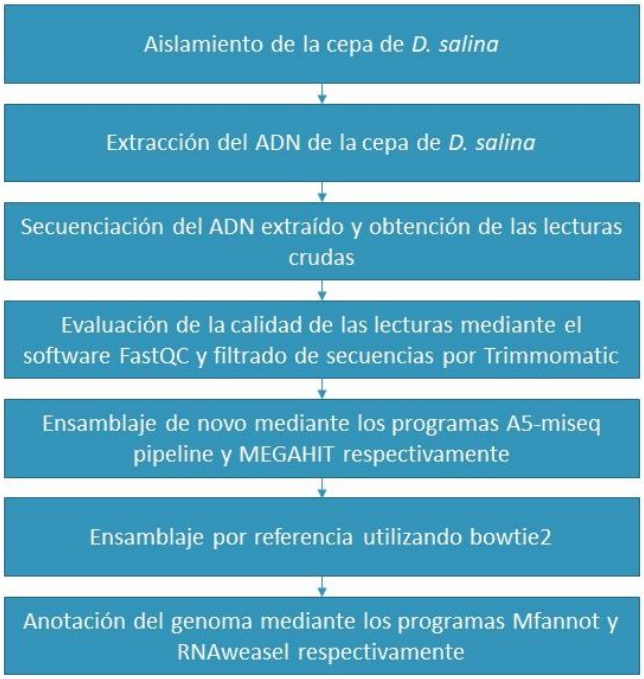


Figura 1. Diagrama en el que se describen los pasos para realizar el ensamble y anotación del genoma del cloroplasto de *D. salina* GN.

V.3 Ensamble *de novo* del genoma del cloroplasto de *Dunaliella salina* GN

Para realizar el ensamble *de novo*, se introdujeron las lecturas limpias al pipeline A5-miseq (Coil *et al.,* 2015). El proceso consiste de cinco pasos que a continuación se describen:

1. Para realizar la limpieza en las lecturas, se eliminan las de baja calidad y los adaptadores de secuencias mediante el programa Trimmomatic. Los errores se corrigen con el algoritmo de SGA (Simpson y Durbin, 2012).
2. Los ensambles de contigs se realizan mediante el algoritmo IDBA-UD (Peng *et al.,* 2012).
3. El andamio de los contigs se realiza con la biblioteca secuenciada.

4. Los errores en los ensambles se corrigen mediante la detección de la asignación errónea de los pares de bases en las longitudes esperadas.
5. El andamio final se realiza mediante una última ronda de andamiaje. En este último paso, también se generan estadísticas referentes al ensamble y estimaciones de calidad.

V.4 Ensamble por referencia del genoma del cloroplasto de *Dunaliella salina* GN

Las lecturas de secuenciación previamente tratadas se sometieron a diferentes programas con el objetivo de obtener el ensamble por referencia del cloroplasto de *D. salina* GN.

Primero, los plastomas de las cepas de *D. salina* CCAP 19/18 y SQ (con números de acceso en GenBank de NC_016732.1 y KX530454.1, respectivamente), se indexaron mediante el software Bowtie2 versión 2.4.1 (Langmead y Salzberg, 2012), programa que indexa los genomas de referencia con un índice basado en la transformada de Burrows-Wheeler. Se sometieron las secuencias indexadas y las lecturas limpias al mismo software, con el objetivo de obtener las correspondientes alineaciones, seleccionando el comando -S para generar ambos archivos en formato SAM (Sequence Alignment/Map, por su significado en inglés).

Los archivos SAM se convirtieron a formato BAM (Binary Aligment/Map) mediante el programa Samtools versión 1.10 (Li *et al.*, 2009), indexando también los ficheros creados por medio del mismo software. Finalmente, los ensambles generados se visualizaron en Tablet viewer (Milne *et al.*, 2013).

Por otra parte, se realizó una segunda ronda de ensambles por referencia mediante el programa BWA (Burrows-Wheeler Aligner, por su significado en inglés) versión 0.7.17-r1188 (H. Li, 2013). Este paquete de software, permite alinear secuencias con genomas de referencia, mediante tres diferentes algoritmos, BWA-backtrack (H. Li y Durbin, 2009), BWA-SW (H. Li y Durbin, 2010) y BWA-MEM (H. Li, 2013).

Se seleccionó el comando index en el programa BWA, para realizar la indexación de los genomas de referencia de las cepas de D. salina CCAP 19/18 y SQ.

Para realizar las correspondientes alineaciones de los plastomas de referencia con las lecturas de secuenciación, se seleccionó a BWA-MEM. Este algoritmo permite efectuar alineamientos de secuencias largas con genomas grandes, esto en comparación con los otros dos algoritmos que permiten ensambles con lecturas más cortas. Los ensambles generados se visualizaron en el programa Tablet viewer.

Por otra parte, con el objetivo de identificar la proximidad evolutiva de las cepas GN, SQ y CCAP 19/18, se realizaron dos análisis filogenéticos moleculares mediante el programa MEGA-X (Kumar *et al.,* 2018), software que provee herramientas para realizar análisis comparativos de secuencias moleculares.

Para realizar ambos alineamientos, se consideró el análisis evolutivo por máxima verosimilitud (Evolutionary analysis by maximum likelihood, por su significado en inglés).

En el primer análisis, se introdujo la secuencia de nucleótidos del gen *petA* de las cepas GN, SQ, CCAP 19/18 y de otros 9 organismos del Chlamydomonadales.

Finalmente, para realizar el segundo análisis, se consideraron las secuencias genómicas del cloroplasto de las cepas GN, SQ, CCAP 19/18 y de otros 9 organismos del orden Chlamydomonadales.

V.5 Anotación del genoma del cloroplasto de *Dunaliella salina* GN

Para realizar la anotación del genoma del cloroplasto de *D. salina* GN, se consideró el programa MFannot (Beck y Lang, 2010), software que permite anotar los genomas de los cloroplastos y mitocondrias. Se introdujo la secuencia ensamblada por referencia con la cepa CCAP 19/18, con el objetivo de generar un archivo de formato sequin, en el que se identificaron las coordenadas de los genes encontrados.

Por otra parte, se introdujo la secuencia del genoma del cloroplasto en el programa RNAweasel (Lang *et al.,* 2007), para identificar a los intrones del grupo I y II, ARNt (ARN de transferencia) y ARNr (ARN ribosomal).

Finalmente, se realizó la curación manual de los genes anotados mediante el software UGENE versión 35.0 (Okonechnikov *et al.*, 2012).

V.6 Llamado de SNPs (Single Nucleotide Polymorphism)

El llamado de variantes se refiere a la identificación de diferencias entre nucleótidos al comparar dos genomas. Específicamente, el reconocimiento de polimorfismos de un solo nucleótido (SNPs, Single Nucleotide Polymorphism por su significado en inglés) indica las bases nitrogenadas distintas entre dos secuencias, una muestra y una de referencia (EMBL-EBI, 2019).

Para acumular las lecturas alineadas en todas las posiciones dentro de los genomas ensamblados, se utilizó el comando mpileup en el programa Samtools, considerando las alineaciones previamente generadas en formato BAM y a las secuencias de referencia (con número de acceso en GenBank de NC_016732.1 y KX530454.1, respectivamente) en formato FASTA.

Por otra parte, para llamar variantes de las lecturas acumuladas, se utilizó el comando call en el programa BCFtools versión 1.10.2 (Li, 2011). Los archivos generados en formato VCF (Variant Calling Format, por su significado en inglés) se convirtieron a FASTQ mediante la opción vcf2fq en la herramienta vcfutils (Li, 2011).

Por último, se utilizó el comando -filter en el programa Seqret versión 6.6.0.0 (Madeira *et al.*, 2019), para convertir los archivos de formato FASTQ a FASTA.

V.7 Diseño de un casete para la expresión de proteínas recombinantes

Se revisaron artículos en los que se modificó el genoma del cloroplasto de *D. salina* o en microalgas con proximidad evolutiva, con el objetivo de elegir a los promotores, elementos reguladores endógenos (5' UTR y 3' UTR, untranslated region por sus siglas en inglés), marcadores de selección de cepas transformadas y genes informadores, elementos necesarios para el casete de expresión.

Con los componentes seleccionados, se diseñaron dos vectores, el primero para la selección de las cepas transformadas y un segundo constructo con el fin de expresar la proteína recombinante.

Por último, mediante la investigación de las correspondientes secuencias dependiendo de los elementos elegidos para el diseño, se realizaron ambos vectores en el software SnapGene (de Insightful Science; disponible en snapgene.com).

V.8 Identificación de regiones potenciales para inserción de casetes de expresión de proteínas recombinantes

Se revisaron diferentes artículos en los que se lograron modificaciones genéticas en el cloroplasto de *D. salina*, para identificar regiones intergénicas potenciales con el objetivo de insertar el casete de expresión previamente diseñado.

VI. Resultados y discusiones

VI.1 Extracción de ADN

La extracción del ADN del cloroplasto de *D. salina* GN, se realizó en otro estudio (Magdaleno y Stephano, 2017). Mediante el procedimiento realizado, se obtuvo una concentración del ADN de 264 ng/µL con un radio de 260/280 de 1.85 en un volumen final de 80 µL.

Las mediciones se realizaron en un nanodrop y la integridad del ADN se identificó mediante una electroforesis en gel de agarosa al 1%, en el que se cargaron 5 µL de las cepas GN y SQ. El resultado se muestra en la figura 2.

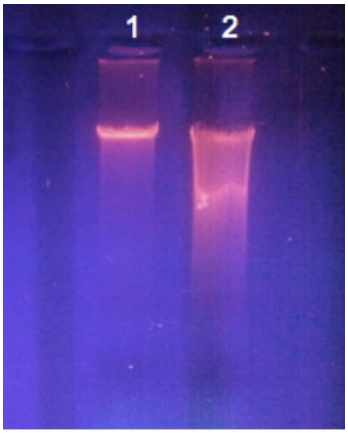


Figura 2. Electroforesis en gel de agarosa de la extracción de ADN. 1) ADN de *Dunaliella salina* SQ. 2) ADN de *Dunaliella salina* GN.

Mediante la figura 2, es posible observar la integridad y degradación del ADN analizado en el carril 2 correspondiente a la cepa de *D. salina* GN, características que se reportaron como adecuadas para llevar a cabo la secuenciación en el estudio en el que se realizó este procedimiento.

En la metodología descrita en este documento, se declara que, en el trabajo de Magdaleno et al 2017, inicialmente se llevó a cabo el procedimiento con la finalidad de aislar mitocondrias de las cepas SQ y GN, sin embargo, en el estudio se menciona que al realizar el ensamble *de novo* de los genomas mitocondriales de ambas cepas, se generaron contigs que se alinearon con el genoma del cloroplasto de *D. salina* CCAP 19/18, razón por la que se concluyó que también se aislaron cloroplastos.

Después de la extracción del ADN, se llevó a cabo la secuenciación del material genético extraído, obteniendo 15,562,756 secuencias para la cepa GN, lo que corresponde a 17,353x el tamaño del genoma secuenciado, considerando la longitud del plastoma de la cepa CCAP 19/18 (269,044 pb), reportada por Smith en 2010.

La cobertura mínima recomendada para realizar el ensamble *de novo*, es de al menos 50x (Illumina, 2010), por lo que se infiere que es posible ensamblar el genoma mediante esta estrategia, considerando la cobertura estimada de 17,353x del plastoma de *D. salina* GN.

VI.2 Preprocesamiento

La secuenciación del cloroplasto de *D. salina* arrojó dos achivos paired-end. Estas bibliotecas se analizaron en el software FastQC para obtener las evaluaciones de calidad en cada una de ellas. Los estadísticos básicos generados por este programa se muestran en la Tabla 1.

Tabla 1. Estadísticos básicos de calidad generados por el programa FastQC, a partir de los archivos paired-end del cloroplasto de *Dunaliella salina* GN.

Archivo	**Secuencias totales**	**Longitud de la secuencia**	**%GC**	**Media de QS**
Forward (R1)	7,781,378	35-301	46	28
Reverse (R2)	7,781,378	35-301	47	24

La puntuación de calidad Phred (Q score, por su significado en inglés), indica la probabilidad de que el secuenciador llame de forma incorrecta a una base determinada (Illumina, 2011).

El módulo de calidad de secuencia por base en el software FastQC, muestra mediante gráficos, los intervalos de puntajes de calidad en las bases de los archivos paired-end introducidos en dicho programa.

Estos gráficos se componen de la línea roja central que representa la mediana, los diagramas de cajas y bigotes en cada una de las posiciones de la lectura, en los que, la caja amarilla corresponde al intervalo intercuartil (25-75%), los bigotes superior e inferior representan el 10% y 90% respectivamente y, por último, el trazo azul muestra la calidad media de las secuencias. El fondo de la gráfica se divide en verde, naranja y rojo, correspondiendo a buena, aceptable y baja calidad, respecto a las llamadas de base.

Las evaluaciones obtenidas de los archivos paired-end analizados mediante este módulo, se muestran en las figuras 3 y 4 respectivamente.

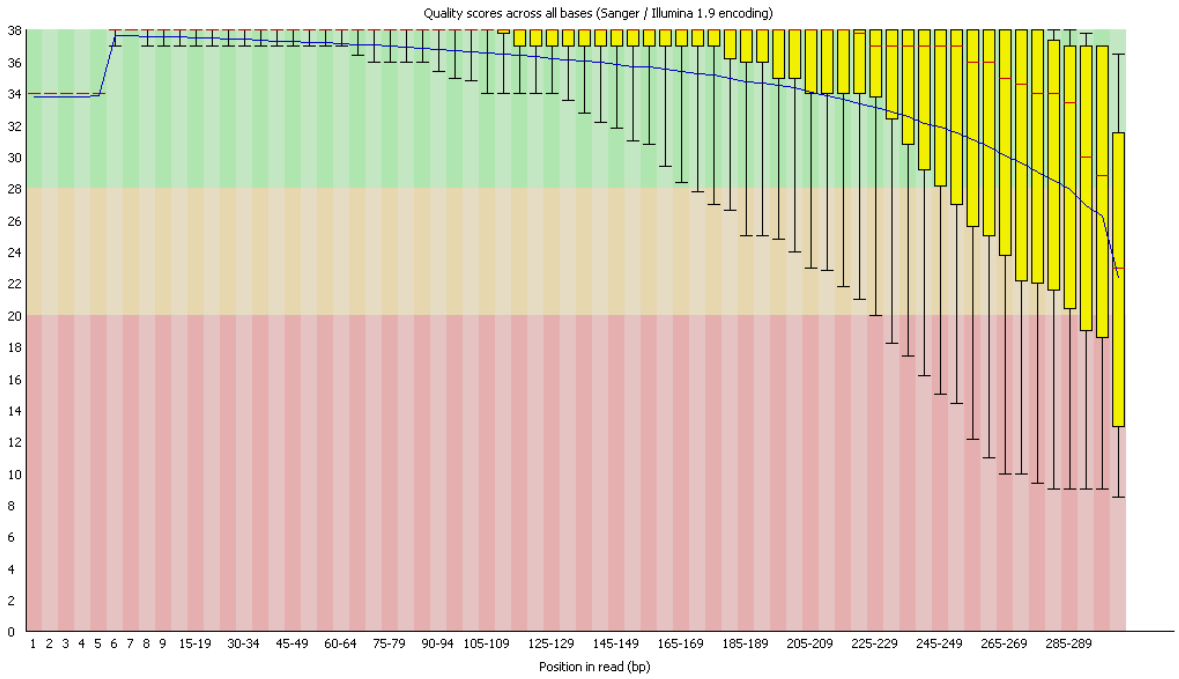


Figura 3. Calidad de secuencia por base en el archivo R1 del cloroplasto de *Dunaliella salina* GN.

En la figura 3, se observan los valores de calidad de cada nucleótido (base) de la lectura forward (R1), la línea azul que representa la media e inicia en los 34 puntos y desciende hasta los 22.

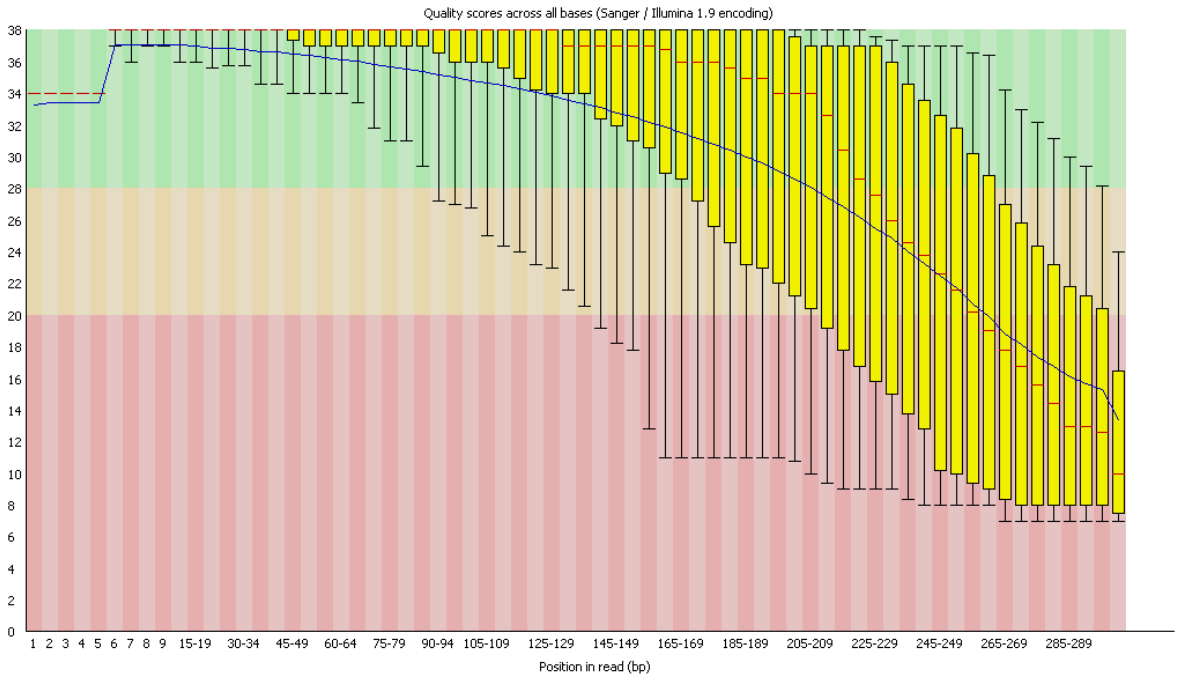


Figura 4. Calidad de secuencia por base en el archivo R2 del cloroplasto de *Dunaliella salina* GN.

En la figura 4, se observan los valores de calidad de cada nucleótido (base) de la lectura reverse (R2), la línea azul que representa la media e inicia en los 37 puntos y desciende hasta los 12.

El valor mínimo considerado para trabajar con lecturas de ADN es 30, debido a que corresponde a una llamada incorrecta por cada 1000 realizadas. El programa FastQC muestra lecturas con valores inferiores a 30 en las lecturas R1 y R2, por lo cual es necesario descartar las lecturas que presentan valores de calidad por debajo de 30, antes de comenzar con el ensamble *de novo* y por referencia.

Por otra parte, las puntuaciones de calidad por secuencia permiten evaluar si algún subconjunto de las secuencias tiene valores de calidad bajos, esto sucediendo debido a factores tales como deficiencias en las lecturas. Los resultados en este módulo, se pueden observar en las figuras 5 y 6, imágenes que representan a la calidad en los archivos R1 y R2 respectivamente.

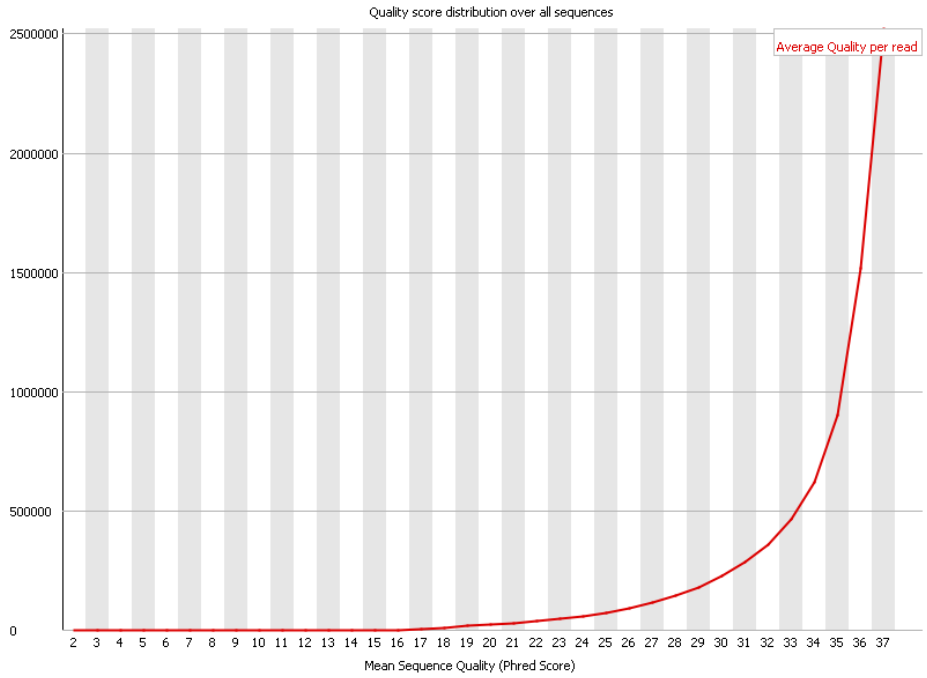


Figura 5. Puntuaciones de calidad por secuencia en el archivo R1 del cloroplasto de *Dunaliella salina* GN.

En la figura 5 se muestra el resultado del módulo de evaluación de puntuaciones de calidad por secuencia en el archivo R1. En el gráfico, se observa que más de 2,500,000 secuencias presentan puntajes mayores a 30. Por su parte, el resto de las secuencias, que representan la menor parte en la lectura, reportan valores que van desde 17 hasta 29 puntos, razón por la que se aprobó la fase.

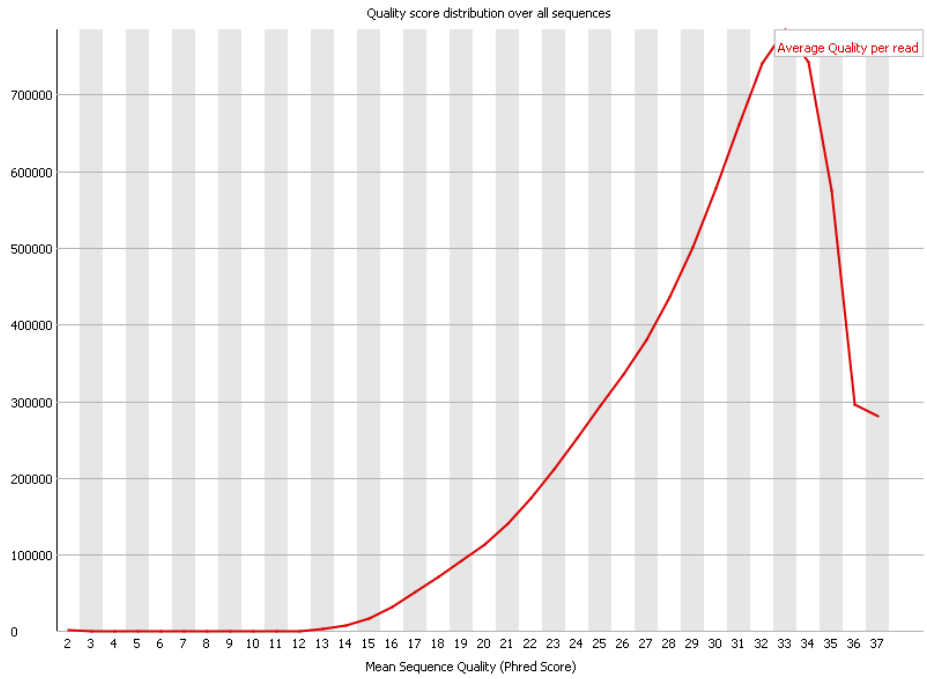


Figura 6. Puntuaciones de calidad por secuencia en el archivo R2 del cloroplasto de *Dunaliella salina* GN.

FastQC evaluó como aprobado el módulo de puntuaciones de calidad por secuencias en el archivo R2. En la figura 6 se muestra que más de 700,000 secuencias (la mayoría de las secuencias en la lectura) presentaron valores que van desde 31 hasta 34 puntos de calidad media, razón por la que se aprobó esta sección.

Por su parte, el módulo de evaluación de calidad correspondiente a la distribución de la longitud de la secuencia, presenta mediante gráficos el tamaño de los fragmentos en las secuencias analizadas. Los resultados de los archivos paired-end analizados, se muestran en las figuras 7 y 8.

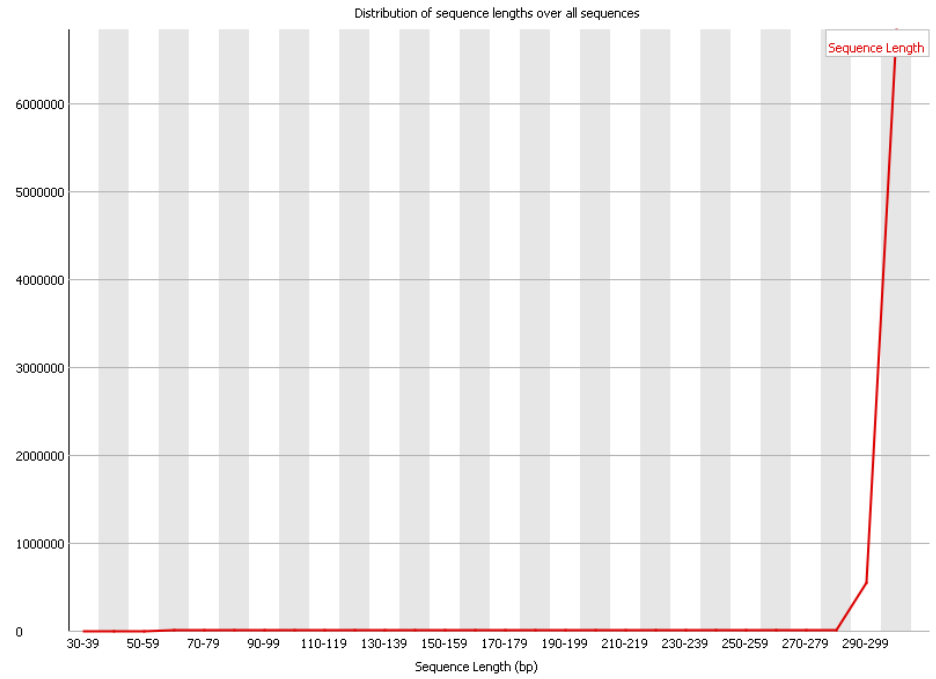


Figura 7. Distribución de los tamaños de los fragmentos de las secuencias en el archivo R1 del cloroplasto de *Dunaliella salina* GN.

La figura 7 corresponde a la distribución de los tamaños de los fragmentos de las secuencias del archivo R1, indicando que la longitud aproximada en ellos va desde 290 hasta 299 nucleótidos. Debido a que existen variaciones entre las longitudes de las secuencias, FastQC indicó la revisión este módulo, sin embargo, los desarrolladores indican que este aviso se puede ignorar debido a que es algo que sucede con frecuencia (“Sequence Length Distribution,” Babraham Bioinformatics).

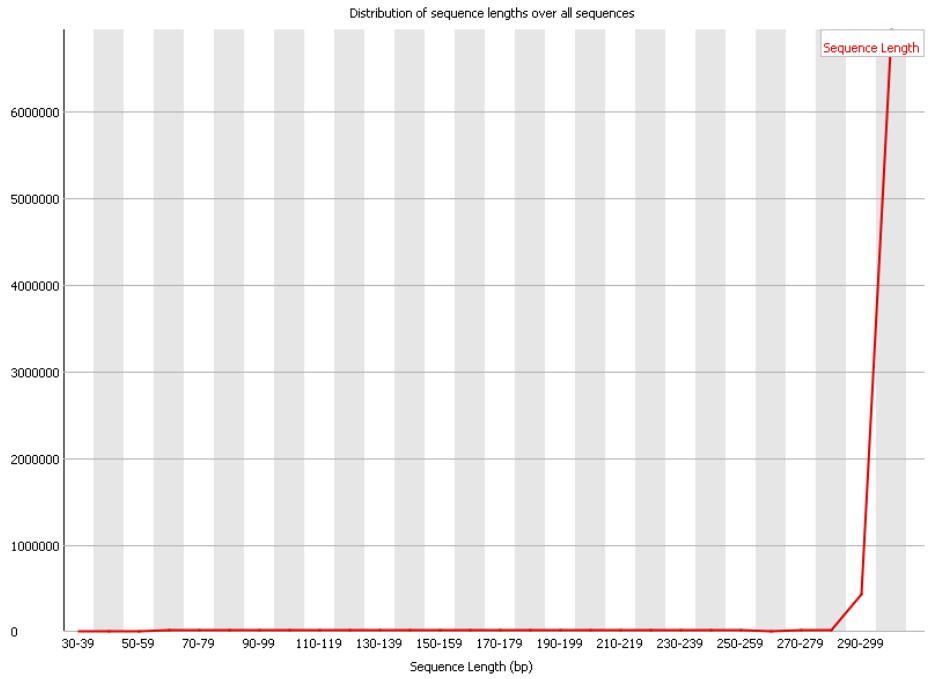


Figura 8. Distribución de los tamaños de los fragmentos de las secuencias en el archivo R2 del cloroplasto de *Dunaliella salina* GN.

El programa FastQC indicó la revisión de este módulo, debido a que se identificó variación en la longitud de las secuencias en la lectura R2. En la figura 8, se observa que el tamaño de los fragmentos de las secuencias es de aproximadamente 290 hasta 299 nucleótidos.

En otra instancia, la deduplicación se refiere a la compresión de datos con el objetivo de eliminar copias duplicadas. El módulo de evaluación del grado de duplicación en los archivos, muestra mediante un gráfico el número de secuencias y las copias que presentan cada una de ellas. En este esquema, la línea azul representa a todas las lecturas en relación con los niveles de duplicación. Por su parte, el trazo rojo indica a las secuencias deduplicadas respecto a la proporción que le corresponde a la

secuencia. Las evaluaciones obtenidas se muestran en las figuras 9 y 10 respectivamente.

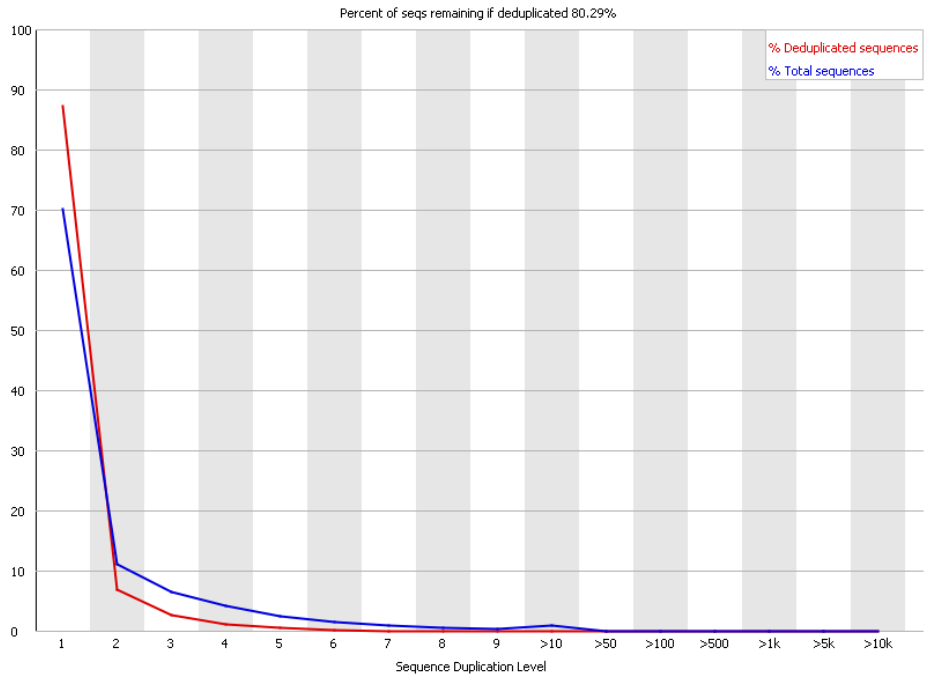


Figura 9. Grado de duplicación en las secuencias del archivo R1 del cloroplasto de *Dunaliella salina* GN.

En la figura 9, se observa el resultado del módulo de duplicación de las secuencias en el archivo R1. Este apartado se aprobó por FastQC, debido a que los niveles se mantuvieron bajos tanto en el trazo azul que representa a la lectura, así como en las secuencias deduplicadas representadas con la línea roja, indicando un nivel de cobertura alto para el genoma objetivo.

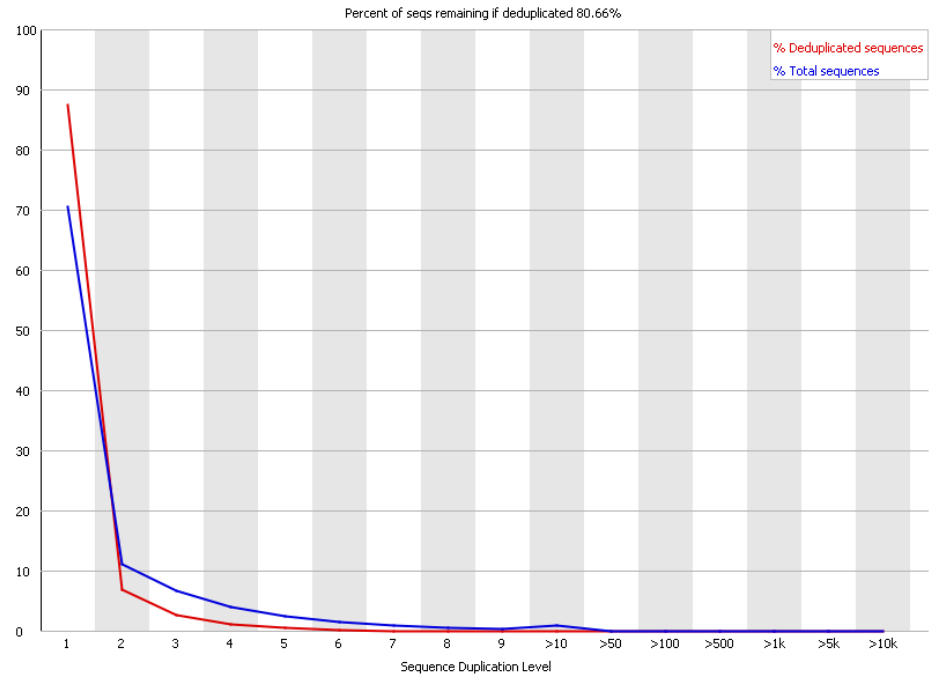


Figura 10. Grado de duplicación en las secuencias del archivo R2 del cloroplasto de *Dunaliella salina* GN.

FastQC indicó que la evaluación de la duplicación en el archivo R2 fue aprobada. En el gráfico 10, se muestra que las líneas caen en el nivel 2, grado que se considera bajo en las secuencias duplicadas y deduplicadas respectivamente, por lo que se evaluó a la biblioteca como diversa.

Una vez obtenidos y analizados los resultados de evaluaciones de calidad en las secuencias crudas, se sometieron las lecturas al programa Trimmomatic, con el objetivo de mejorar la calidad en ellas.

Los archivos procesados se introdujeron al programa FastQC para obtener las evaluaciones de calidad después del tratamiento de los archivos. Los estadísticos básicos de calidad de las secuencias tratadas se muestran en la tabla 2.

Tabla 2. Estadísticos básicos de calidad generados por el programa FastQC a partir de los archivos paired-end procesados del cloroplasto de *Dunaliella salina* GN.

Archivo	**Secuencias totales**	**Longitud de la secuencia**	**%GC**	**Media de QS**
R1	4,479,343	1-301	44	36
R2	4,479,343	1-301	43	36

Después del tratamiento de las lecturas, se obtuvieron 8,958,686 secuencias para la cepa GN, lo que corresponde a 9,989x el tamaño del genoma secuenciado, considerando la longitud del plastoma de la cepa CCAP 19/18 (269,044 pb), reportada por Smith en 2010.

El resultado del módulo de calidad de secuencias por base en los archivos paired-end procesados, se muestran en las figuras 11 y 12.

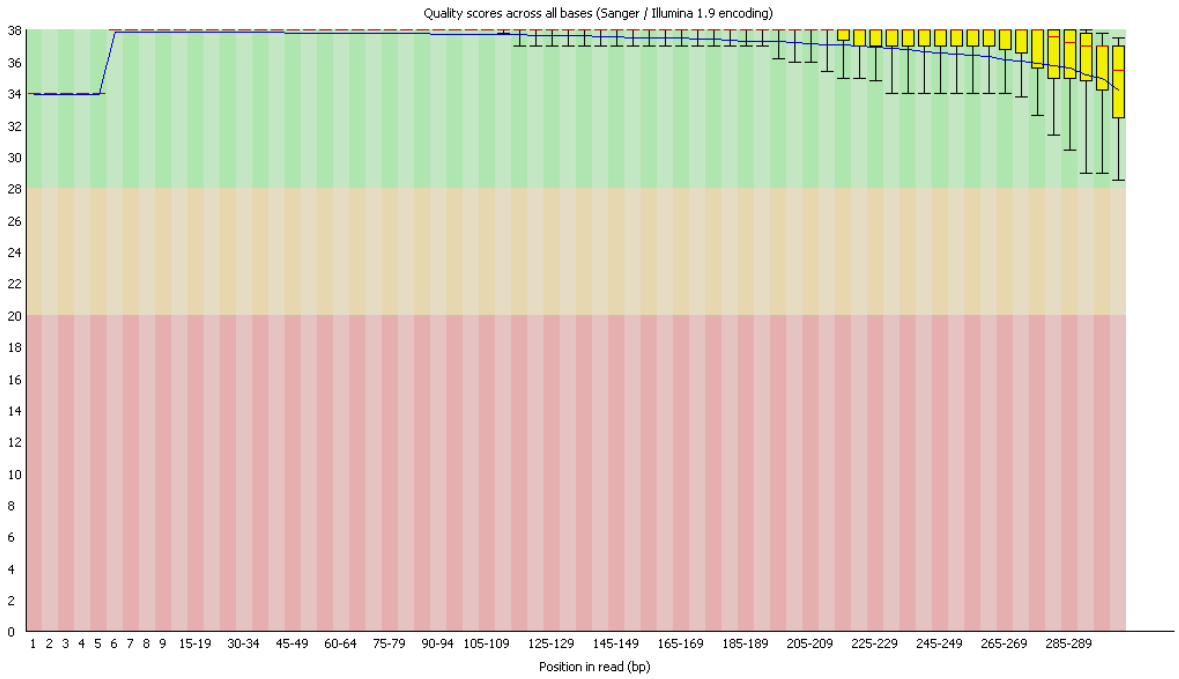


Figura 11. Calidad de secuencia por base en el archivo R1 procesado del cloroplasto de *Dunaliella salina* GN.

En el caso del archivo R1 procesado, FastQC reportó que el módulo de calidad de las secuencias por base fue aprobado, debido al tratamiento realizado mediante Trimmomatic. En la figura 11, es posible observar que la línea azul que representa la media, comienza en 38 y desciende hasta aproximadamente 34 puntos, superando el nivel mínimo de 30, requerido para realizar ensambles de calidad.

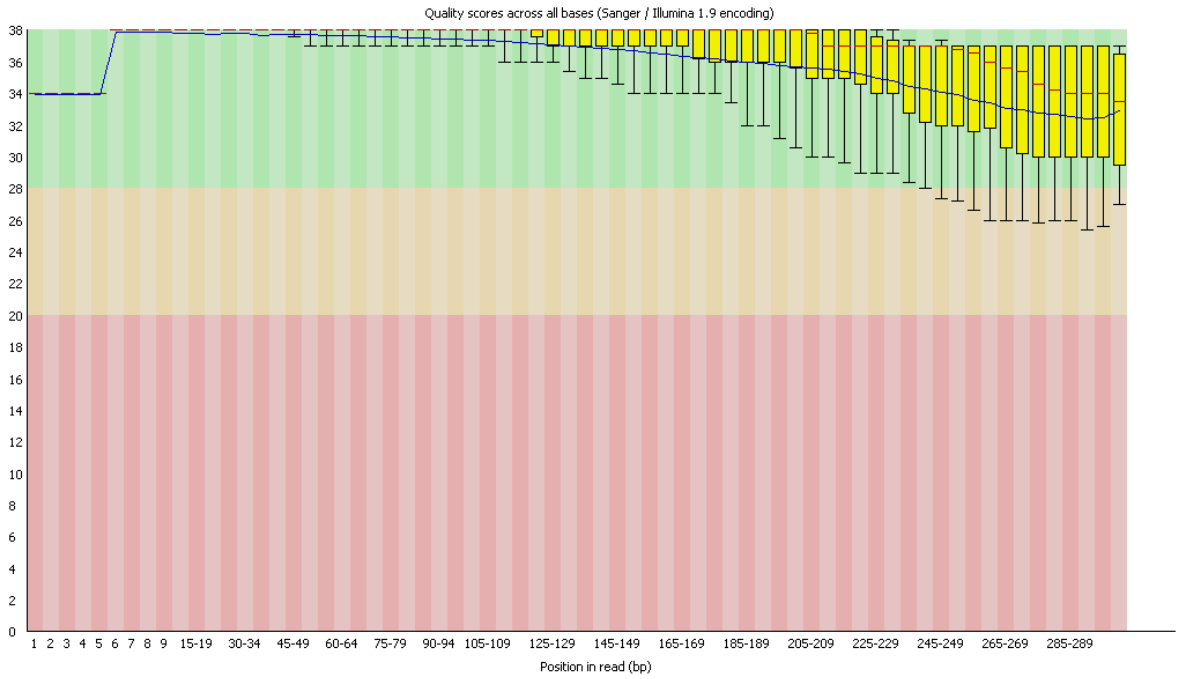


Figura 12. Calidad de secuencia por base en el archivo R2 procesado del cloroplasto de *Dunaliella salina* GN.

La calidad por base en las secuencias tratadas del archivo R2 mostrada en la figura 12, se evalúa como aprobada por el programa FastQC, debido a que la media en este módulo inicia en 38 puntos de calidad y desciende hasta aproximadamente 34, puntuaciones logradas debido al tratamiento realizado en las lecturas crudas.

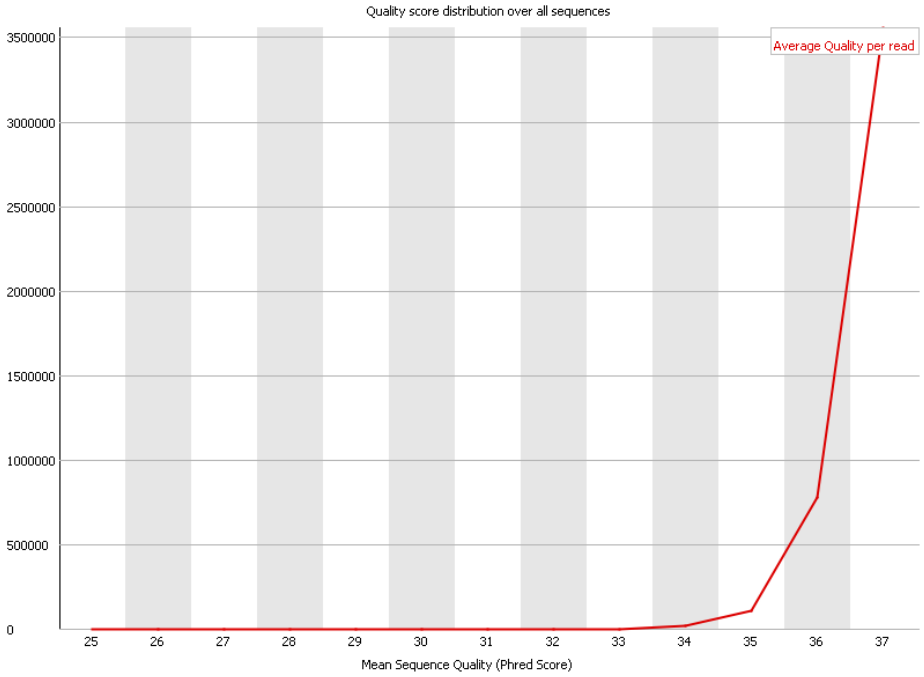


Figura 13. Puntuaciones de calidad por secuencia en el archivo R1 procesado del cloroplasto de *Dunaliella salina* GN.

El módulo de las puntuaciones de calidad por secuencia en el archivo R1 procesado que se muestra en la figura 13, es aprobado al igual que en la evaluación previa al tratamiento de las secuencias crudas, ya que mantiene la calidad por encima de los 30 puntos.

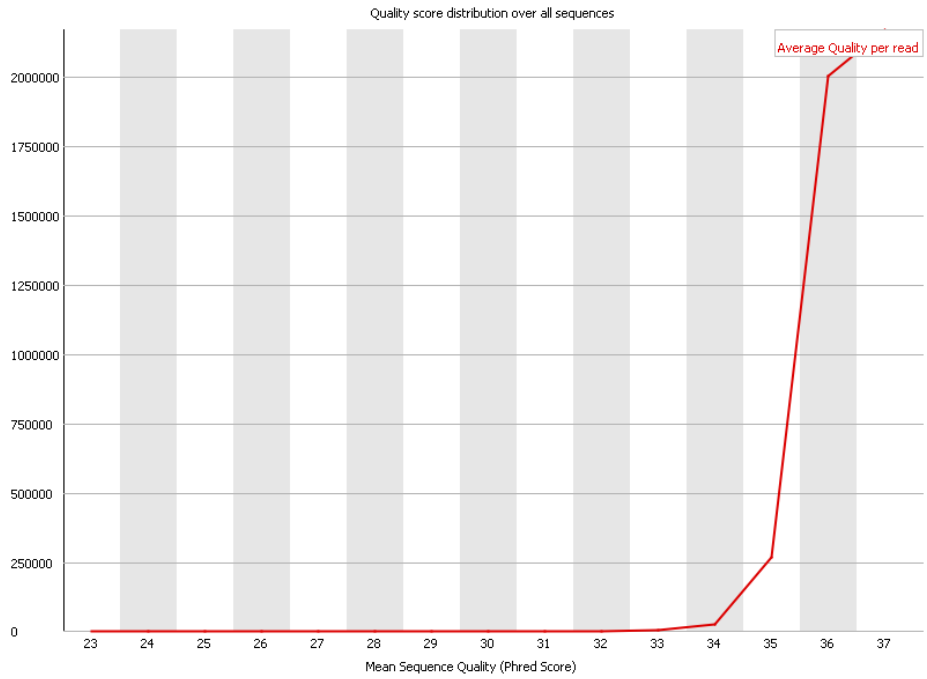


Figura 14. Puntuaciones de calidad por secuencia en el archivo R2 procesado del cloroplasto de *Dunaliella salina* GN.

En la figura 14, se muestra la distribución de los puntajes de calidad en las secuencias del archivo R2 procesado, en la que los niveles abarcan desde 34 hasta 37 puntos, por lo que el programa calificó este módulo como aprobado, al igual que en la evaluación anterior previo al tratamiento de las secuencias.

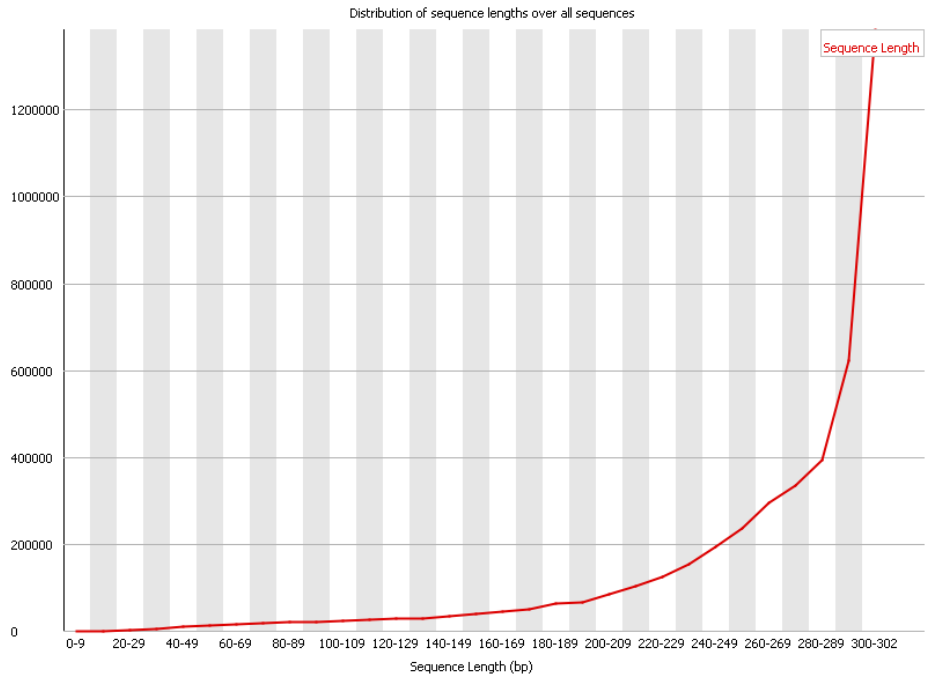


Figura 15. Distribución de los tamaños de los fragmentos de las secuencias en el archivo R1 procesado del cloroplasto de *Dunaliella salina* GN.

FastQC sugirió la revisión del módulo de distribución de los tamaños de los fragmentos en las secuencias del archivo R1 procesado, debido a la diferencia entre los tamaños de los fragmentos, misma evaluación obtenida previo al tratamiento de los datos. A pesar de que las evaluaciones permanezcan igual, existen diferencias entre los tamaños. En la figura 15, se identifican fragmentos con tamaños desde 40 hasta 300 pares de bases, en comparación con el análisis anterior, que reportó longitudes de aproximadamente 290 nucleótidos.

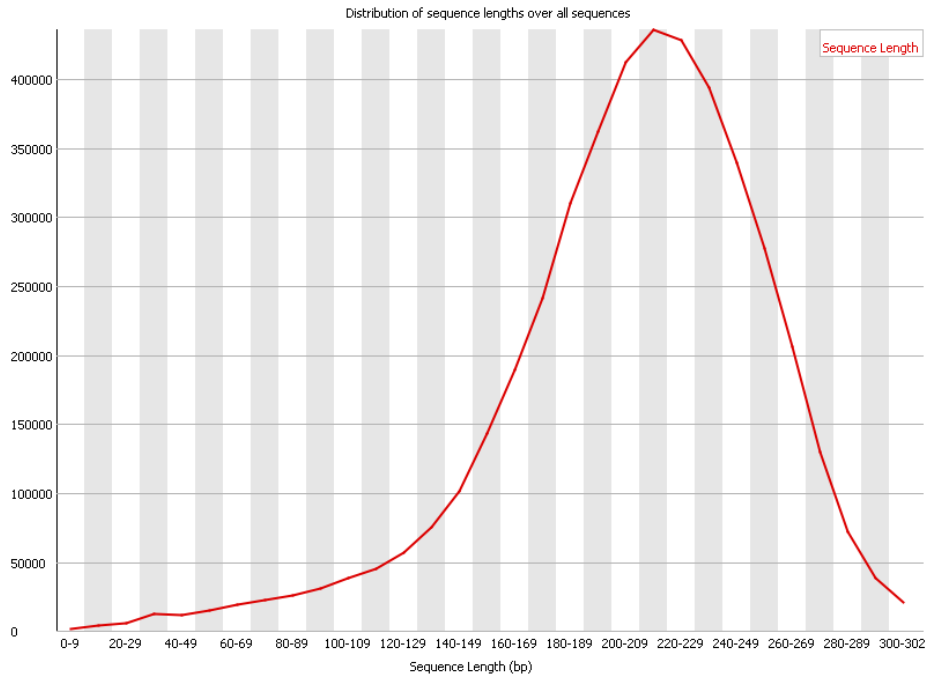


Figura 16. Distribución de los tamaños de los fragmentos en las secuencias del archivo R2 procesado del cloroplasto de *Dunaliella salina* GN.

El programa FastQC indicó la revisión del módulo de la distribución de los tamaños de los fragmentos en las secuencias del archivo R2 procesado, debido a que se reportaron diferentes longitudes de las secuencias. En la evaluación anterior al tratamiento del archivo R2 original, se reporta la misma calificación, sin embargo, estas difieren, debido a que el primer análisis presenta valores que van incrementando conforme a la longitud de los fragmentos, llegando hasta 300 nucleótidos, en comparación con este caso representado en la figura 16, en la que se identifica una campana que comienza en 100, llega a la cúspide con 220 y termina en 300 pares de bases.

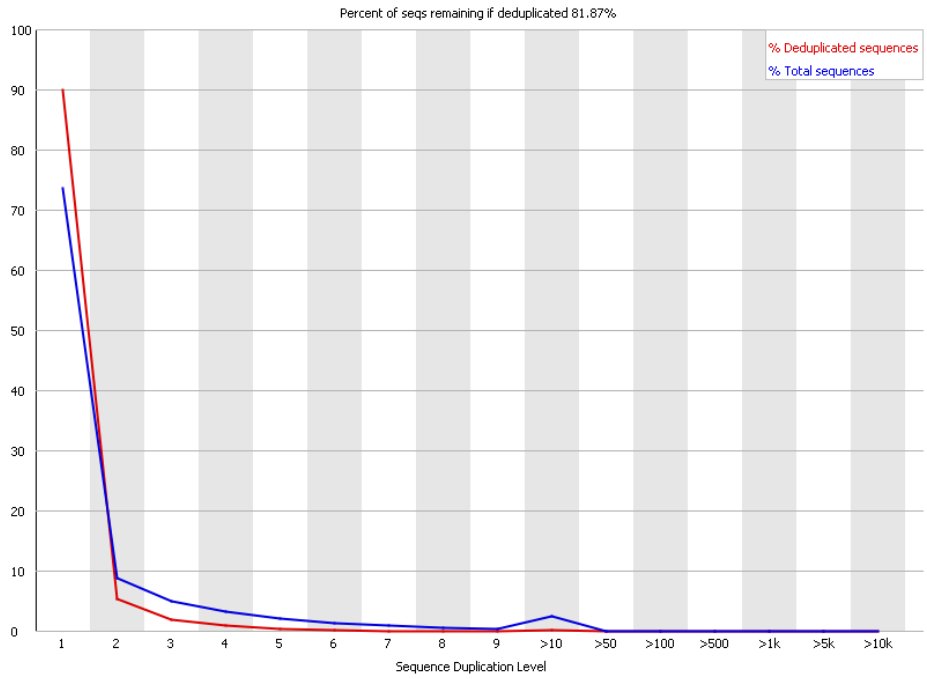


Figura 17. Grado de duplicación en las secuencias del archivo R1 procesado del cloroplasto de *Dunaliella salina* GN.

El módulo de duplicación de las secuencias en el archivo R1 procesado fue aprobada por el software FastQC. En la figura 17, la línea azul y roja indican que las secuencias tanto duplicadas como deduplicadas comienzan a descender en el nivel 2, grado considerado bajo, calificación que indica que existe una alta cobertura en la biblioteca secuenciada.

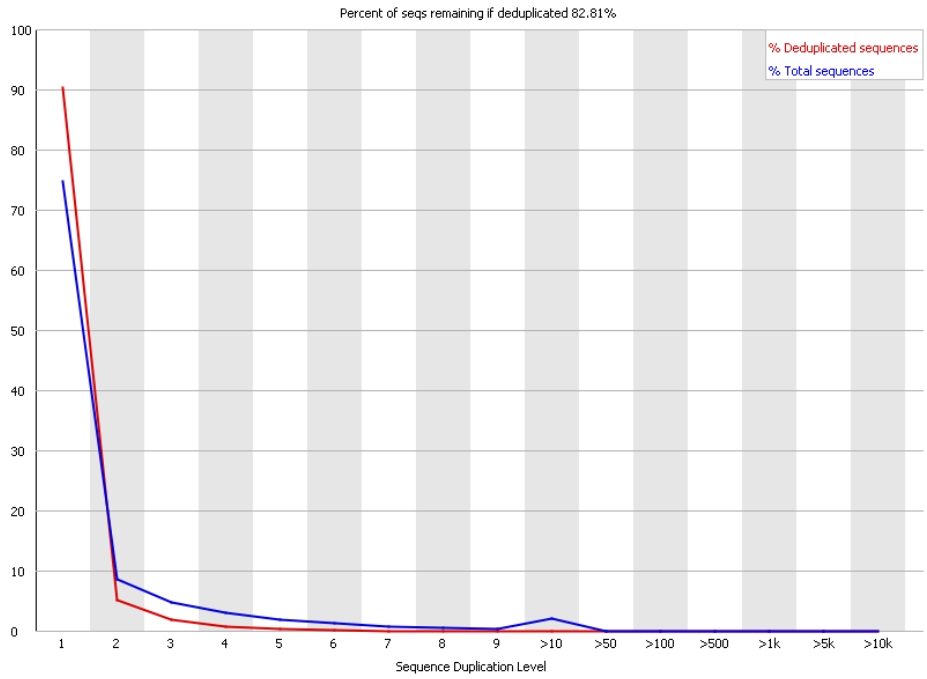


Figura 18. Grado de duplicación en las secuencias en el archivo R2 procesado del cloroplasto de *Dunaliella salina* GN.

En el caso de la duplicación de las secuencias en el archivo R2 procesado, el software indica una calificación aprobatoria para este módulo. La figura 18, muestra los resultados de esta sección, en la que las líneas comienzan a descender en el nivel 2, conservando el mismo grado de duplicación desde antes de que la biblioteca se procesara.

Una vez considerados y discutidos los resultados en cuanto al pretratamiento de las secuencias, se procedió a introducirlas en diferentes programas para realizar los ensambles *de novo* y por referencia respectivamente.

VI.3 Ensamble *de novo* del genoma del cloroplasto de *Dunaliella salina* GN

Para realizar el ensamble *de novo* del cloroplasto de *D. salina* GN, se introdujeron las secuencias procesadas al pipeline a5-miseq. Los pasos logrados en el algoritmo del programa utilizado fueron los siguientes:

1. Limpieza en las lecturas: se eliminaron las lecturas de baja calidad con el programa Trimmomatic

2. Los errores en las lecturas se corrigieron con el algoritmo de SGA.

3. Al llegar al apartado de ensamble de contigs mediante IDBA-UD, el proceso se detuvo, mencionando que tales secuencias no fueron generadas.

```
Using Long Clipping Sequence: 'TTTTTTTTTTCAAGCAGAAGACGGCATACGA'
Using Long Clipping Sequence: 'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG'
Using Long Clipping Sequence: 'TCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT'
ILLUMINACLIP: Using 3 prefix pairs, 16 forward/reverse sequences, 0 forward onl
y sequences, 0 reverse only sequences
Input Reads: 15562756 Surviving: 15367594 (98.75%) Dropped: 195162 (1.25%)
TrimmomaticSE: Completed successfully
[a5] '/home/myrna/a5_miseq_linux_20160825/bin'/sga preprocess -q 25 -f 20 -m 35
  --pe-mode=0  dunaliella.s1/GN_S2_L001_R1_001.fastq.trim.fastq > dunaliella.s1
/GN_S2_L001_R1_001.fastq.both.pp 2> /dev/null
[a5] sga index -d 1060037 -t 4 dunaliella.s1/dunaliella.pp.fastq > dunaliella.s
1/index.out 2> dunaliella.s1/index.err
[a5] '/home/myrna/a5_miseq_linux_20160825/bin'/sga correct -t 4 -p dunaliella.p
p -o dunaliella.s1/GN_S2_L001_R1_001.fastq.both.pp.ec.fastq dunaliella.s1/GN_S2
_L001_R1_001.fastq.both.pp > dunaliella.s1/raw1.correct.out
[timer - sga::correct] wall clock: 19868.22s CPU: 76935.72s
Running cat dunaliella.s1/GN_S2_L001_R1_001.fastq.both.repair.fastq >> dunaliel
la.s1/dunaliella.ec.fastq
Done merging libraries
[a5_s2] Building contigs from dunaliella.ec.fastq with IDBA
[a5_s2] Building contigs from dunaliella.ec.fastq with IDBA
[a5] dunaliella.s2/dunaliella.ec.fasta has 4320729967 bytes of FastA sequence d
ata
[a5] '/home/myrna/a5_miseq_linux_20160825/bin'/idba_ud500 --num_threads 4 -r du
naliella.s2/dunaliella.ec.fasta  -o dunaliella.s2/dunaliella --mink 35 --maxk 3
00 --min_pairs 2
Killed
[a5] Error building contigs with IDBA
(base) myrna@myrna-VirtualBox:~$ ls -a
```

Figura 19. Final del proceso de ensamble *de novo* del cloroplasto de *Dunaliella salina* GN, en el pipeline a5-miseq.

Los resultados indican que aunque fue posible realizar las respectivas limpiezas de lecturas mediante Trimmomatic y la corrección de las secuencias con el algoritmo de SGA, la construcción de contigs con IDBA (Iterative De Bruijn Graph) no fue lograda, debido a que como se establece en el programa, IDBA presentó un error al intentar estimar el tamaño de inserción causando un bloqueo antes del andamio, en el que los contigs de ser generados se habrían utilizado, sin embargo, al no generarse alguno, el programa se cerró de forma automática.

Aunado a la razón anterior, existen algunas otras situaciones que pudieron no haber permitido el logro de este objetivo. La cobertura mínima del genoma recomendada para realizar un ensamble *de novo* es de 50x, sin embargo, en el presente estudio se utilizan secuencias que representan 9,989x el tamaño del plastoma ensamblado. Illumina establece que una amplia cobertura puede implicar un mayor uso de memoria (Illumina, 2010), por lo que se infiere que la capacidad de la computadora utilizada (8 GB de RAM y 4 núcleos), no fue suficiente para obtener el genoma.

Por otra parte, se infiere la presencia de regiones repetitivas en la secuencia a ensamblar, tal y como se presentan en el genoma del cloroplasto de la cepa CCAP 19/18 reportada por Smith en 2010, mismas regiones que complican el proceso de ensamble *de novo*.

Entre las diferentes razones anteriormente consideradas, se destaca el hecho de que el ensamble se realizó en una computadora de 8 GB de RAM, en comparación con otros ensambles logrados en sistemas con 32 GB de RAM (Lopez *et al.*, 2017;

Magdaleno y Stephano, 2017) en una duración menor a 16 horas como la aquí reportada.

Al no lograr este objetivo, se establece que es necesario utilizar las mismas herramientas que las reportadas en estudios en los que se lograron ensambles, para poder considerar comparaciones entre los resultados obtenidos.

VI.4 Ensamble por referencia del genoma del cloroplasto de *Dunaliella salina* GN

Las lecturas pretratadas se sometieron a diferentes programas con el objetivo de obtener el ensamble por referencia del genoma del cloroplasto de *D. salina* GN.

Se indexaron los genomas de las cepas CCAP 19/18 y SQ, respectivamente, mediante el programa Bowtie2, con el que se generaron 6 archivos en formato BT2 por cada genoma de referencia, ficheros en los que se encuentran divididos los índices de las secuencias. Se introdujeron los archivos indexados y secuencias de referencia en el mismo programa, obteniendo los alineamientos en formato SAM.

Por otra parte, los ficheros generados se convirtieron a su forma binaria (BAM) mediante la herramienta Samtools.

Los alineamientos generados en formato SAM se visualizaron en el programa Tablet viewer. Los resultados se presentan en las figuras 20 y 21.

Figura 20. Ensamble por referencia del genoma del cloroplasto de *Dunaliella salina* GN con la cepa CCAP 19/18, generado por el programa Bowtie2.

Figura 21. Ensamble por referencia del genoma del cloroplasto de *Dunaliella salina* GN con la cepa SQ, generado por el programa Bowtie2.

Al realizar los dos ensambles por referencia mediante el programa Bowtie2, considerando los genomas de los cloroplastos de *D. salina* de las cepas CCAP 19/18 y SQ respectivamente, se encontraron diferencias considerables en cuanto a la cobertura de bases. Al comparar los alineamientos realizados, se identificó una concordancia mayor de 98.458% con la cepa de Australia, en comparación con la del

plastoma de San Quintín, constando de 70.047%. Por otra parte, el porcentaje de discordancia con la cepa de Australia es de 1.747%, cantidad menor que la de la cepa SQ, siendo de 5.432%. Los resultados se presentan en la tabla 3.

Existe una mayor proximidad geográfica de las zonas en las que fueron aisladas las cepas SQ y GN en comparación con la cepa CCAP 19/18, sin embargo, al observar los resultados de los alineamientos, se identifica la gran similitud existente entre las secuencias de las cepas de Australia y Guerrero Negro, México.

En la figura 20, se muestra el alineamiento casi total con la cepa CCAP 19/18. Las zonas sin cobertura (gaps) observadas, se analizaron mediante el programa UGENE, utilizando el genoma de referencia para identificar a dichas regiones. Las áreas identificadas correspondían a secuencias no codificantes en el plastoma de la microalga de Australia. Estos resultados se muestran en la tabla 4.

Por otra parte, en la figura 21, se presenta el ensamble por referencia a la cepa SQ, alineamiento con un número mayor de gaps en comparación con el plastoma de la microalga de Australia. Al analizar las zonas identificadas, se encontraron regiones genéticas codificantes y no codificantes.

Una vez considerados y analizados los resultados de los alineamientos generados por el programa Bowtie2, se realizó una segunda ronda de ensambles por referencia mediante el algoritmo BWA.

Al implementar el comando index en los genomas de referencia de las cepas CCAP 19/18 y SQ, se obtuvieron 5 archivos por genoma indexado. Los ficheros generados son de tipo binario, denominados BWT, PAC y SA, así como también los documentos de texto AMB.

Mediante la implementación del comando bwa mem, los genomas de referencia y las secuencias procesadas, se obtuvieron dos alineamientos distintos en formato SAM.

Los ensambles generados se analizaron mediante Tablet viewer, mismos alineamientos que se pueden observar en las figuras 22 y 23.

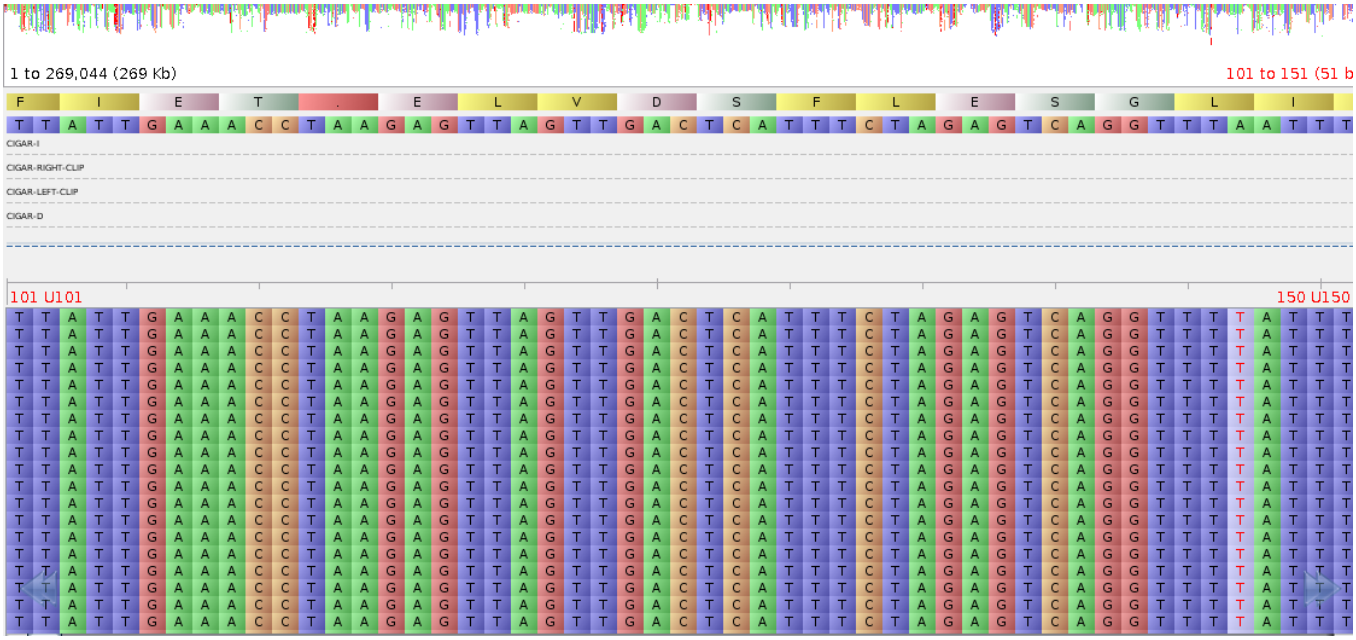


Figura 22. Ensamble por referencia del genoma del cloroplasto de *Dunaliella salina* GN con la cepa CCAP 19/18, generado por el programa BWA.

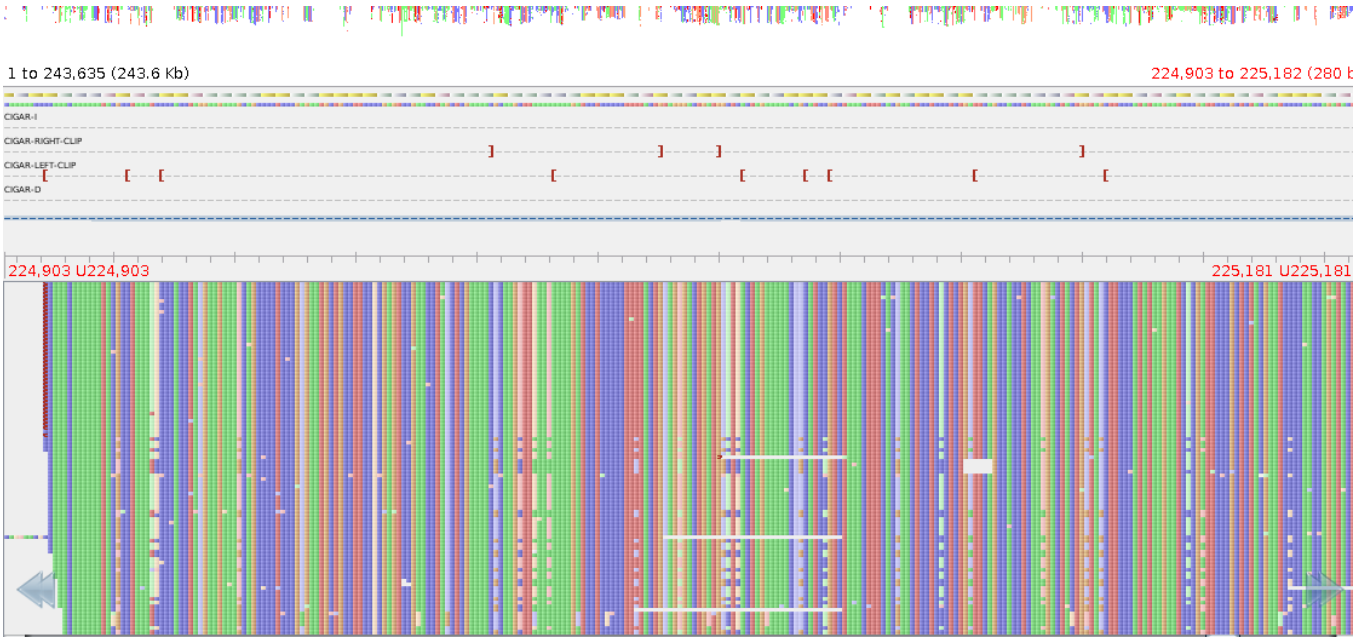


Figura 23. Ensamble por referencia del genoma del cloroplasto de *Dunaliella salina* GN con la cepa SQ, generado por el programa BWA.

La figura 23, representa el ensamble por referencia con el plastoma de la cepa CCAP 19/18 mediante el programa BWA. En ella, es posible observar que se obtuvo una cobertura casi total del genoma, constando de 98.71% con la secuencia de referencia. Por otra parte, el porcentaje de discordancia es de 3.069%.

El ensamble por referencia con la cepa SQ utilizando el programa BWA, se observa en la imagen 24. En esta figura, se identifican gaps entre las secuencias ensambladas, mismas zonas que corresponden a 81.557% en cuanto a la cobertura del genoma original y una discordancia del 6.307%.

Considerando los resultados con respecto a las cepas anteriormente mencionadas, la cepa CCAP 19/18 representa una mayor cobertura y un menor porcentaje de discordancia con el genoma del cloroplasto aquí ensamblado, razones por las que se eligen los alineamientos generados mediante el plastoma de la cepa de Australia, para cuestiones de análisis y anotación de la secuencia obtenida.

Tabla 3. Tabla comparativa de los resultados encontrados en cada uno de los ensambles generados por los programas Bowtie2 y BWA respectivamente.

Ensamble realizado en	Número del genoma de referencia en GenBank	Cantidad de secuencias	Porcentaje de bases con cobertura	Porcentaje de discordancia con el genoma de referencia
Bowtie2	NC_016732.1	895,873	98.458%	1.747%
BWA	NC_016732.1	1,507,697	98.71%	3.069%
Bowtie2	KX530454.1	501,551	70.047%	5.342%
BWA	KX530454.1	1,906,312	81.557%	6.307%

Tabla 4. Gaps identificados en el ensamble de *Dunaliella salina* GN, considerando las regiones en el genoma del cloroplasto de la cepa CCAP 19/18.

Región repetitiva	Región del gap identificado	%GC	Gen
Si	134,200 – 135,791	16% G, 15% C	*rrnL* (región intrónica)
No	188,409 – 189,321	22% G, 17% C	Región no codificante
No	210,461 – 210,610	16% G, 12% C	*psbA* (Región intrónica)
No	211,449 – 212,020	15% G, 19% C	*psbA* (Región intrónica)
Si	260,578 – 262166	15% G, 16% C	*rrnL* (Región intrónica)

Al comparar las diferencias que existen entre los alineamientos generados, resulta importante destacar que los genomas de los cloroplastos tienen arquitecturas genómicas generalmente conservadas (Daniell *et al.*, 2016).

Considerando lo anterior, pueden existir diferentes razones en cuanto a las diferencias entre los resultados obtenidos, tales como que las bibliotecas generadas previo a la secuenciación tengan sesgo, que el algoritmo empleado no sea el adecuado (Alkan *et al.*, 2011) y que la arquitectura de los genomas difiera a pesar de que provengan de la misma especie.

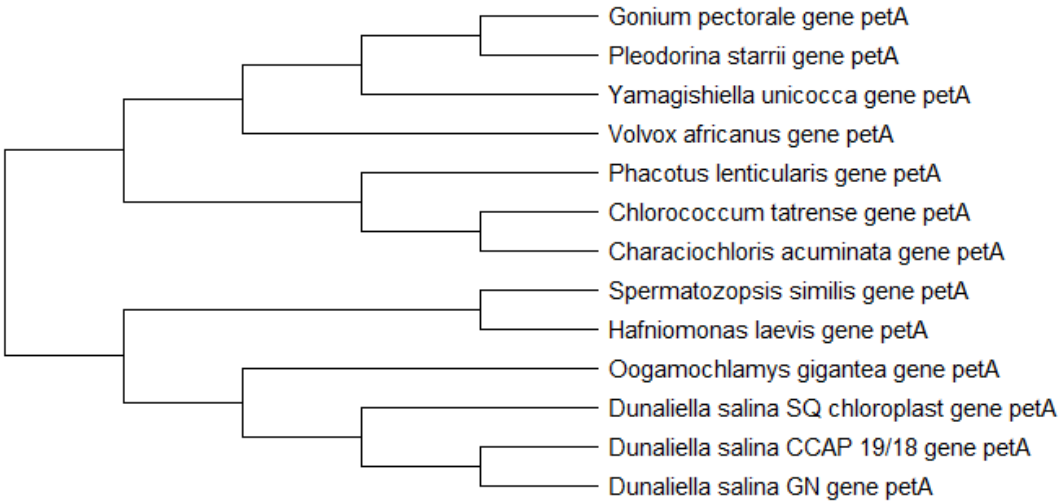
En cuanto a las razones anteriormente mencionadas con respecto a las diferencias entre los genomas alineados, es preciso considerar que las evaluaciones de duplicación de las bibliotecas forward y reverse resultaron aprobadas, puesto que como es posible observar en las figuras 9 y 10 antes del procesamiento y 18 y 19 después del procesamiento de las secuencias, los niveles de duplicación son bajos, por lo que se descarta que la identificación de gaps se deba a que el protocolo seguido para la extracción del ADN con el objetivo de preparar las lecturas haya sido incorrecto.

En el alineamiento obtenido mediante el programa BWA, se observa que los gaps se presentan en las mismas zonas que en el ensamble realizado por Bowtie2, por lo que se rechaza la idea de que la presencia de gaps se deba a que el software utilizado no sea el correcto para el procedimiento.

Por otra parte, se puede inferir una posible diferencia entre las arquitecturas de los genomas de los cloroplastos analizados, debido a que en el estudio en el que se extrajo el ADN del plastoma aquí analizado (Magdaleno y Stephano, 2017), el genoma de la mitocondria de la cepa GN presentó mayor similitud con la cepa de Australia en comparación con la de San Quintín, México.

La plataforma que se utilizó para la secuenciación del plastoma en el presente estudio es Illumina Miseq. Existen artículos que destacan que esta tecnología puede ocasionar que no exista cobertura en las zonas de regiones repetitivas (Alkan *et al.*, 2011; Peona *et al.*, 2020), razón por la que se infiere la presencia de algunos gaps dentro del genoma aquí analizado, tal y como se muestra en la tabla 4.

Además de la situación anteriormente descrita, se considera que Illumina Miseq, es una herramienta que presenta deficiencias en cuanto a la secuenciación de regiones con contenido %GC alto (Chen *et al.*, 2013). Considerando los diferentes niveles de GC, correspondiendo a bajo (20-40%), medio (41-59%) y alto (60-80), al analizar los contenidos de G y C de los gaps encontrados en el genoma ensamblado



mediante las secuencias de referencia, es posible percatarse de que los niveles son bajos, por lo que se descarta como una probable razón de fallo durante el ensamble.

Figura 24. Análisis filogenético del gen marcador *petA* de las cepas SQ, GN y 10 organismos del orden Chlamydomonadales por el método de Maximum likelihood, mediante el software MEGA-X.

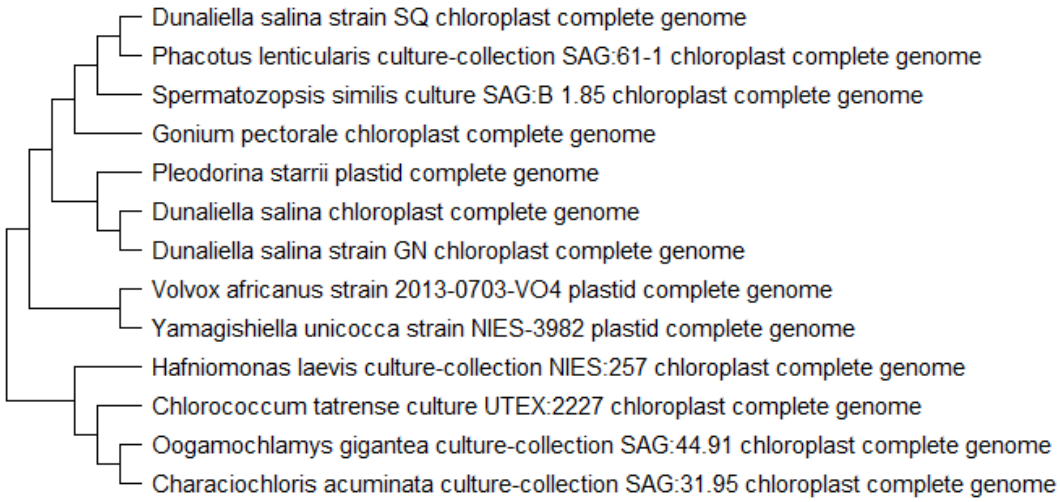


Figura 25. Análisis filogenético de los genomas de los cloroplastos de las cepas SQ, GN y 10 organismos del orden Chlamydomonadales, mediante el software MEGA-X.

La figura 24 representa el análisis filogenético del gen marcador petA de las cepas SQ, GN, CCAP 19/18 y otros 9 organismos del orden Chlamydomonadales mediante el software MEGA-X. En ella, se observa la proximidad evolutiva que existe entre la cepa GN y CCAP 19/18, a diferencia de la cepa SQ que se encuentra más alejada evolutivamente.

El análisis filogenético de los genomas del cloroplasto de las cepas SQ, GN, CCAP 19/18 y otros 9 organismos del orden Chlamydomonadales se observa en la figura 25. En esta figura, se identifica la mayor relación evolutiva existente entre las cepas GN y CCAP 19/18, reafirmando así el mejor resultado de los alineamientos, misma razón por la que se comprende el bajo porcentaje de concordancia con la cepa SQ.

VI.5 Anotación del genoma del cloroplasto de *Dunaliella salina* GN

Una vez ensamblado el genoma del cloroplasto de *D. salina* GN, se introdujo la secuencia en diferentes programas con el objetivo de anotarla.

Al introducir la secuencia FASTA en Mfannot, el programa reportó 134 genes con sus respectivas coordenadas genéticas. Por su parte, RNAweasel informó 57 intrones, entre ellos, 17 del grupo I y 10 del grupo II, así como 30 ARNt's y 4 ARNr's.

Al localizar las regiones genéticas en las que el nuevo genoma ensamblado difería con la secuencia de referencia, se realizó una curación manual mediante el programa UGENE, para detectar las diferencias entre los codones de inicio y terminación respectivamente en las posiciones análogas dentro de ambos genomas. Los resultados obtenidos se muestran en la tabla 5.

Tabla 5. Tabla comparativa de los genes anotados con la cantidad de intrones, considerando los codones de inicio y terminación, respectivamente. Los codones subrayados en amarillo corresponden a los que difieren con los del genoma de referencia.

Región de la secuencia	Tipo de gen codificante	Nombre	Cantidad de intrones	Codón de inicio	Codón de terminación
LSC (long sequence copy)	Proteína	*ycf1*	0	ATG	TAA
	Proteína	*rpl6*	0	ATG	TAA
	Proteína	*rpl14*	0	ATG	TAA
	Proteína	*rpl5*	0	ATG	TAA
	Proteína	*rps8*	0	ATG	TAA
	Proteína	*psaA*	3	ATG	CTC
	Proteína	*petD*	0	ATG	TAA
	Proteína	*petA*	0	GCT	TAA
	Proteína	*rpoC1*	0	ATG	TAA
	ARNt	*trnP*	0	CGG	CGA
	ARNt	*trnV*	0	AGG	CTA
	Proteína	*tufA*	0	ATG	TAA
	ARNt	*trnE*	0	GCC	GTA

	Proteína	*rpl36*	0	ATG	TAA
	Proteína	*petB*	0	ATG	TAA
	ARNt	*trnH*	0	GCG	GCC
	Proteína	*rps12*	0	ATG	TAA
	Proteína	*psaJ*	0	ATG	TAA
	Proteína	*atpI*	0	ATG	TAA
	Proteína	*psbJ*	0	ATG	TAA
	ARNt	*trnM*	0	GCA	GCA
	ARNt	*trnN*	0	TCT	GAG
	Proteína	*chlL*	1	ATG	TAG
	Proteína	*orf440*	0	ATG	TAG
	Proteína	*chlL*	0	TTT	GTT
	ARNt	*trnL*	0	GCC	GCA
	Proteína	*clpP*	0	ATG	TAA
	ARNt	*trnM*	0	AGC	CAA
	Proteína	*petL*	0	ATG	TAA
	Proteína	*psaC*	0	ATG	TAA
	ARNt	*trnF*	0	GCC	GCA
	Proteína	*psbD*	1	ATG	TAA
	Proteína	*orf315*	0	ATG	TAA
	ARNt	*trnT*	0	GCC	GCT
	Proteína	*rps4*	0	ATG	TAA
	ARNt	*trnG*	0	GCC	GCT
	ARNt	*trnR*	0	GAG	TCA
	ARNt	*trnS*	0	GGA	CCG
	Proteína	*rpl20*	0	ATG	TAA
	ARNt	*trnC*	0	GGC	CCT
	Proteína	*psbK*	0	ATG	TAA
	ARNt	*trnG*	0	GCG	GCT
	ARNt	*trnW*	0	ACG	GTG
	Proteína	*atpA*	1	ATG	TAA
	Proteína	*orf276*	0	ATG	TAA
	Proteína	*psbI*	0	ATG	TAA
	Proteína	*cemA*	0	ATG	TAA
	Proteína	*rpl23*	0	ATG	TAA
	Proteína	*rpl2*	0	ATG	TAG
	Proteína	*rps19*	0	ATG	TAA
	Proteína	*atpB*	1	ATG	TAA
	Proteína	*orf374*	0	ATG	TAA
	Proteína	*atpB*	0	AGT	TAA
	Proteína	*ftsH*	0	ATG	TAA

	Proteína	*psbC*	0	GTG	TAA
	Proteína	*orf262*	2	ATG	TAA
	Proteína	*orf130*	0	ATG	TGA
	ARNt	*trnR*	0	GAA	TCA
	Proteína	*chlB*	0	ATG	TAA
IR_A (Inverted repeat)	ARNr	*rrnS*	4	AAA	TGT
	ARNt	*trnI*	0	GGG	CCA
	ARNt	*trnA*	0	GGG	CCA
	ARNr	*rrnL*	7	ACG	GAT
	Proteína	*orf143*	0	ATG	TAG
	Proteína	*orf243*	0	ATG	TAA
	Proteína	*orf230*	0	ATG	TAA
	ARNr	*rrn5*	0	CCT	GGG
SSC (Small sequence copy)	Proteína	*atpH*	0	ATG	TAA
	Proteína	*atpF*	0	ATG	TAA
	ARNt	*trnQ*	0	TGG	CAG
	ARNt	*trnY*	0	GGG	CCA
	Proteína	*rps11*	0	ATG	TAA
	Proteína	*psaB*	1	ATG	TAA
	Proteína	*orf121*	0	ATG	TAA
	Proteína	*ccsA*	0	ATG	TAA
	Proteína	*psbZ*	0	ATG	TAA
	Proteína	*psbM*	0	ATG	TAA
	Proteína	*rps14*	0	ATG	TAA
	Proteína	*rps7*	0	ATG	TAA
	Proteína	*atpE*	0	ATG	TAA
	Proteína	*rbcL*	0	ATG	TAA
	ARNt	*trnL*	0	GGG	CCA
	Proteína	*ycf3*	0	ATG	TAA
	Proteína	*ycf4*	0	ATG	TAA
	Proteína	*rps9*	0	ATG	TAA
	Proteína	*psbE*	0	ATG	TAA
	ARNt	*trnE*	0	GCC	GTA
	ARNt	*trnI*	0	GGG	CCA
	Proteína	*psbH*	0	ATG	TAA
	Proteína	*psbN*	0	ATG	TAA
	Proteína	*psbT*	0	ATG	TAG
	Proteína	*psbB*	0	ATG	TAA
	ARNt	*trnD*	0	GGG	CCG
	Proteína	*rps2*	0	ATG	TAA
	Proteína	*rps18*	0	ATG	TAA

	Proteína	*psbA*	5	ATG	TAA
	Proteína	*orf191*	0	ATG	TAA
	Proteína	*orf167*	0	ATG	TAG
	Proteína	*orf293*	0	ATG	TAA
	Proteína	*orf222*	0	ATG	TAC
	ARNt	*trnK*	0	GGG	CCA
	Proteína	*ycf12*	0	ATG	TAA
	ARNt	*trnM*	0	GCC	GCA
	Proteína	*rpoA*	0	ATG	TAG
	ARNt	*trnS*	0	GGA	CCG
	Proteína	*chlN*	0	ATG	TAA
	Proteína	*rpoBb*	0	ATG	TAA
	Proteína	*rpoBa*	0	ATG	TAA
	Proteína	*psbF*	0	ATG	TAA
	Proteína	*psbL*	0	ATG	TAA
	Proteína	*petG*	0	ATG	TAA
	Proteína	*rps3*	0	ATG	TAA
	Proteína	*rps7*	0	ATG	TAA
	Proteína	*rpoC2*	0	ATG	TAA
IR_B (Inverted repeat)	ARNr	*rrn5*	0	CCT	GGG
	ARNr	*rrnL*	7	ACG	GAT
	Proteína	*orf230*	0	ATG	TAA
	Proteína	*orf234*	0	ATG	TAA
	Proteína	*orf143*	0	ATG	TAG
	ARNt	*trnA*	0	GGG	CCA
	ARNt	*trnI*	0	GGG	CCA
	ARNr	*rrnS*	4	AAA	TGT
	Total	122 genes	37 intrones		

Para anotar el plastoma de *D. salina* GN, se consideró el genoma ensamblado por referencia a la cepa CCAP 19/18, con una longitud de 269,044 pares de bases, eliminando los nucleótidos desconocidos reportados con las letras N, M, Y, R y K. Al eliminar estas bases, se obtuvo una secuencia final de 263,577 pb.

Esta secuencia final, se sometió a los programas RNAweasel y MFannot respectivamente, para obtener las coordenadas correspondientes a las secuencias

genéticas, identificando 122 genes y 37 intrones, siendo el mismo número en el genoma de referencia (Smith *et al.*, 2010).

Se identifican 86 genes que codifican para proteína, 30 para RNA de transferencia y 6 para ARN ribosomal.

El genoma ensamblado presenta la estructura cuatripartita típica de los plastomas, compuesta de una región de copia única grande (LSC, long sequence copy por su significado en inglés) de 128,080 pb, dos copias de la región invertida (inverted repeat, IR por sus siglas en inglés) de 11,115 pb cada una y una región de copia única sencilla (SSC, single sequence copy por su significado en inglés) de 112,182 pb. Los genes integrados en las regiones se muestran en la tabla 5.

El mapa del genoma del cloroplasto de *D. salina* GN se dibujó en el programa OGDRAW (Greiner *et al.*, 2019). El resultado se muestra en la figura 26.

Los contenidos A, G, C y T son 33.75%, 16.13%, 15.91% y 34.19% respectivamente. Se reporta 32.1% de contenido GC.

Mediante la tabla 5, es posible observar que los codones de inicio y terminación son ATG y TAA en los genes que codifican para proteína. De igual manera, se identifican los codones específicos en cada uno de los genes que codifican para ARNt y ARNr.

Se encontraron algunas excepciones en cuanto a los codones de terminación, tal y como sucede en el gen orf143, que termina en TAG, sin embargo, esto se debe a que es un codón de terminación de menor frecuencia.

Por otra parte, se observa el codón de terminación TAC en el gen *orf222*, esto ocurriendo debido a los pares de bases faltantes por los gaps generados durante el ensamble.

Al identificar los gaps en los dos ensambles realizados con respecto a los genomas de referencia, de SQ y CCAP 19/18 respectivamente, fue posible constatar que el número de regiones sin cobertura fue menor con referencia a la secuencia de Australia.

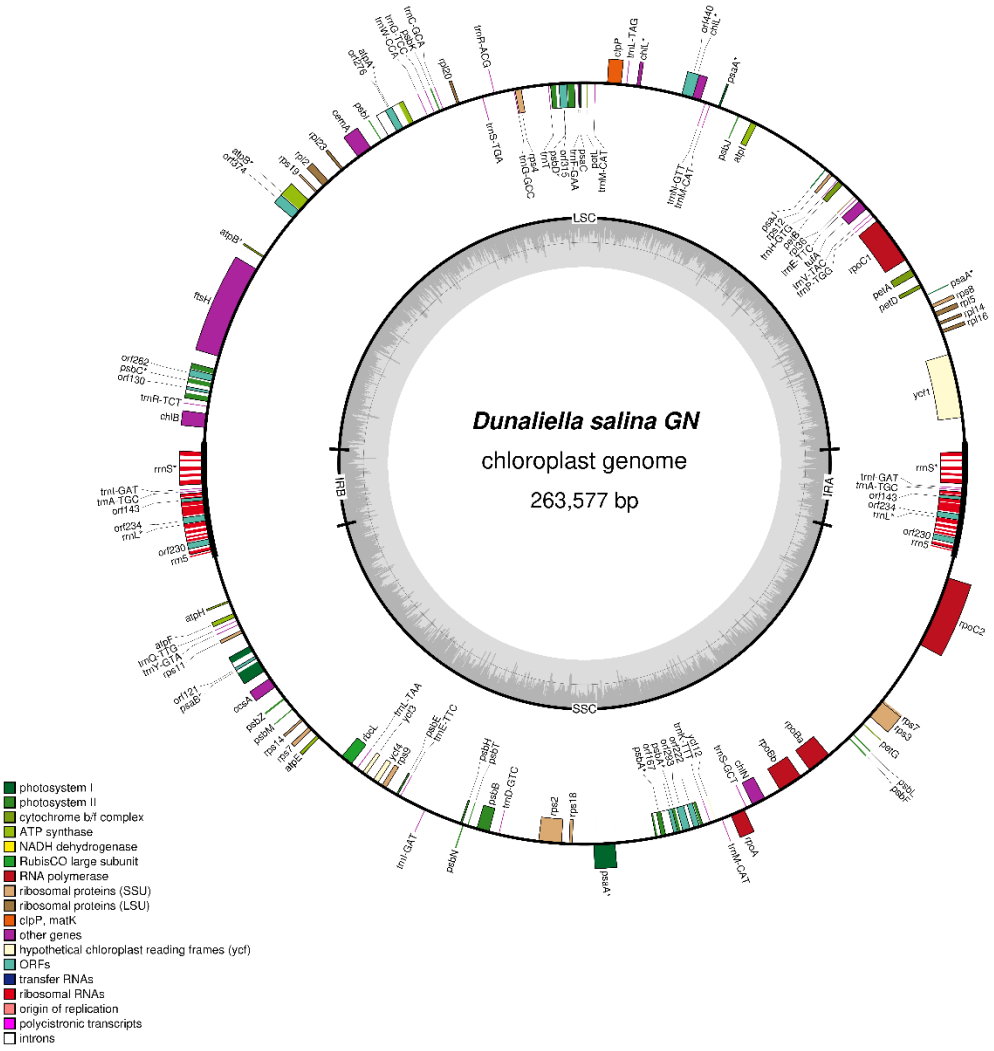
El número de gaps con respecto a la cepa SQ, se debe a la diferencia entre la longitud de los plastomas, en comparación con la cepa de Australia, en la que los gaps identificados en el genoma ensamblado corresponden a regiones no codificantes en la secuencia de referencia, mismas que no están expuestas a selección natural.

Las diferencias entre las longitudes de los genomas en los cloroplastos, a menudo se deben por expansiones o contracciones de las IR (Wang *et al.*, 2018). Debido a lo anterior, resulta importante destacar que los tamaños de las regiones repetitivas de las cepas CCAP 19/18, SQ y GN son de 14,408 pb, 17,144 pb y 11,115 pb, respectivamente. Se identifica que la región invertida de la cepa SQ es considerablemente mayor que las de las otras dos cepas.

Los maquinaria de reparación del ADN puede estar implicada en la evolución de los plastomas (Smith y Keeling, 2015). Las condiciones ambientales específicas a las que están expuestas las diferentes cepas son capaces de provocar un impacto en los mecanismos de recombinación homóloga de los cloroplastos, terminando por ocasionar desestabilizaciones de los genomas (Maréchal y Brisson, 2010), razones por las que se infiere la diferencia entre las longitudes de los genomas del cloroplasto de

las cepas de *D. salina* CCAP 19/18, SQ y GN, ya que se localizan geográficamente en distintas zonas.

A pesar de las diferencias entre las secuencias anotadas debido a los gaps, se



identifican a los mismos genes entre los plastomas de *D. Salina* anotados.

Figura 26. Arquitectura del genoma del cloroplasto de *Dunaliella salina* GN.

VI.6 Llamado de SNPs (Single Nucleotide Polymorphisms)

El ensamble por referencia con la cepa CCAP 19/18 reportó menor cantidad de gaps en comparación con la cepa SQ, por lo que, para fines de análisis, se seleccionó al alineamiento generado a partir de la cepa de Australia.

En la tabla 6, se muestran a los SNPs identificados en el genoma de *D. salina* GN, con respecto al plastoma de referencia CCAP 19/18. En dicha tabla, es posible observar que se hallaron SNPs en todos los tipos de genes.

Tabla 6. SNPs potenciales identificados en la secuencia ensamblada.

Gen	**Coordenadas**		**SNPs**		**Cantidad de SNPs por gen**
Nombre	Ensamble	Referencia	Ensamble	Referencia	
ycf1	3971	3972	C	A	6
	3972	3973	T	G	
	5749	5750	A	C	
	6178	6179	T	A	
	8305	8306	G	A	
	9321	9322	T	G	
rpl14	13240	13242	T	C	2
	13297	13299	A	T	
psaA	196394	199432	T	A	4
	196655	199694	A	T	
	196938	199977	G	A	
	197188	200227	T	C	
petA	19273	19453	A	T	5
	19276	19457	A	T	
	19294	19474	G	A	
	19420	19600	C	T	
	19896	20076	C	T	
rpoc1	23872	24051	G	A	1
petB	32340	32552	C	A	1

rps12	33850	34062	A	G	1
chlL	57351	57567	T	A	1
psbD	65424	65656	C	G	1
orf315	65424	65656	C	G	1
atpA	83082	83289	C	T	10
	83530	83787	C	T	
	83688	83945	A	T	
	83744	84001	T	C	
	83945	84202	C	A	
	84030	84287	G	T	
	84042	84299	A	G	
	84146	84403	A	T	
	84861	85118	T	G	
	85419	85676	G	A	
orf276	84861	85118	T	G	2
	84970	85227	G	A	
cemA	90045	90322	T	G	1
rpl2	94644	94921	G	A	1
rps19	96220	96497	C	G	4
	96229	96506	A	G	
	96265	96542	T	C	
	96274	96551	G	A	
atpB	98848	99139	A	C	6
	98919	99242	T	C	
	99154	99482	C	A	
	99431	99754	T	C	
	99444	99767	A	C	
	100047	100370	T	A	
orf374	99159	99482	C	A	4
	99431	99754	T	C	
	99444	99767	A	C	
	100047	100370	T	A	
atpB	105430	105754	T	G	1
ftsH	106936	107262	A	T	2
	114294	114619	A	G	
psbC	120030	120379	G	A	7
	120077	120426	A	C	
	120564	120913	C	T	
	121467	121816	A	T	
	121492	121841	G	T	
	121772	122121	G	T	

	122146	122495	T	C	
orf262	120030	120379	G	A	2
	120077	120426	A	C	
orf130	121492	121841	G	T	1
rrn5	132780	133130	T	G	1
psaB	151211	153187	C	T	2
	151980	153956	A	C	
ccsA	155503	157500	G	A	1
rps7	161581	163580	C	T	7
	161677	163676	C	T	
	161685	163683	G	A	
	161686	163684	T	A	
	161687	163685	T	A	
	161688	163687	C	A	
	161791	163790	A	T	
trnI	177176	179234	G	T	1
rps2	190759	193788	C	T	1
rps18	193969	197007	A	G	1
psbA	203222	206309	T	C	7
	203889	206975	T	G	
	203991	207078	T	C	
	203992	207079	G	C	
	205747	208834	C	T	
	207014	210101	A	C	
	207196	210283	T	C	
orf293	207014	210101	A	C	1
rpoA	212718	216609	T	A	3
	212719	216610	T	A	
	212720	216611	T	A	
chlN	214813	218704	A	T	8
	214816	218707	G	A	
	214822	218713	G	A	
	214840	218731	T	C	
	214851	218740	C	T	
	214854	218743	G	A	
	214935	218826	G	A	
	214975	218866	C	T	
rpoBb	220157	224048	C	T	1
rpoBa	223606	227496	C	G	3
	223607	227497	T	A	
	223608	227498	T	A	

rpoC2	240479	244373	T	A	10
	240480	244374	T	A	
	240481	244375	T	A	
	241462	245356	A	G	
	242096	245889	T	A	
	243116	247010	T	C	
	243204	247098	G	T	
	244618	248513	G	C	
	244663	248558	A	G	
	245922	249817	T	A	
trnI	258438	263905	G	A	2
	258454	263921	C	T	
Total de SNPs identificados					113

En la tabla 6, se identifican la mayor cantidad de SNPs en los genes *atpA* y *rpoC2*, reportando 10 nucleótidos en cada gen respectivamente.

La cobertura mínima para inferir SNPs es de 30x (Illumina, 2010), por lo que al considerar la cobertura de 9,989x utilizada para llamar a las variantes, los SNPs reportados se pueden tratar como potenciales para trabajos futuros en los que se considere el genoma del cloroplasto aquí ensamblado.

VI.7 Diseño de un casete para la expresión de proteínas recombinantes en el cloroplasto de *Dunaliella salina* GN

Para realizar el diseño del casete de expresión, se revisaron diferentes artículos en los que previamente se realizaron modificaciones genéticas al cloroplasto de *D. salina* o microalgas con proximidad evolutiva a este organismo. A continuación, se muestran los vectores realizados mediante el programa SnapGene presentados en las figuras 27 y 28 respectivamente, así como la descripción de dichos plásmidos en las tablas comparativas 7 y 8.

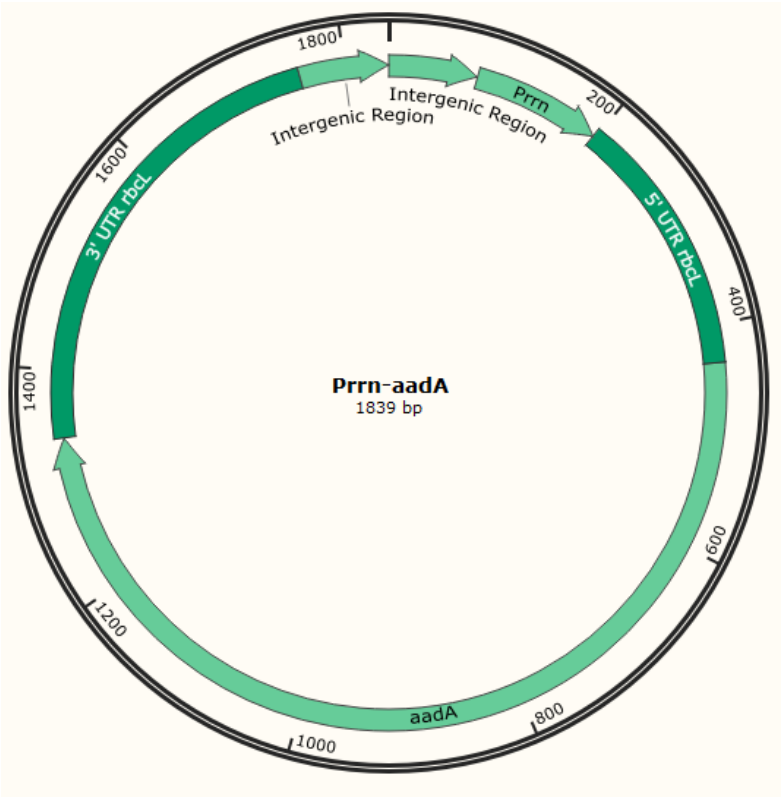


Figura 27. Vector diseñado para la selección de las cepas de *Dunaliella salina* GN transformadas.

Tabla 7. Elementos elegidos para el plásmido de la selección de las cepas transformadas.

	Región intergénica	**Promotor**	**5' UTR**	**Gen**	**3'UTR**
Elemento	*trnI-trnA*	*Prrn*	*rbcL*	*aadA*	*rbcL*
Utilizado en el cloroplasto de	*D. salina 1009*	*D. salina 1009*	*D. salina CCAP 19/18*	*C. reinhardtii*	*D. salina CCAP 19/18*
Longitud (pb)	160	117	236	906	420
Referencia	(D. Li *et al.*, 2011)	(D. Li *et al.*, 2011)	(Talebi *et al.*, 2014)	(Goldschmidt-clermont, 1991)	(Talebi *et al.*, 2014)

El promotor Prrn, se utilizó previamente para la transformación de la cepa de *D. salina* 1009 (Talebi *et al.*, 2014), con el objetivo de expresar la proteína verde fluorescente mejorada (Enhanced Green Fluorescent Protein, EGFP por sus siglas en inglés).

Los elementos reguladores 5' UTR y 3' UTR seleccionados que pertenecen al gen *rbcL*, se utilizaron en el cloroplasto de *D. salina* CCAP 19/18 (Talebi *et al.*, 2014), con el objetivo de expresar los genes *ME y AccD*, secuencias implicadas en la producción de lípidos.

El elemento elegido para la selección es *aadA*, gen implicado en la resistencia a la espectinomicina, previamente utilizado para la transformación del plastoma de *Chlamydomonas reinhardtii* (*C. reinhardtii*) (Goldschmidt-clermont, 1991).

Al considerar los elementos elegidos que se utilizaron anteriormente con el fin de transformar a cepas de *D. salina* y microalgas con proximidad evolutiva, se infiere que, de utilizar el vector aquí presentado en trabajos futuros, permitirá la selección exitosa de las cepas transformadas de *D. salina* GN.

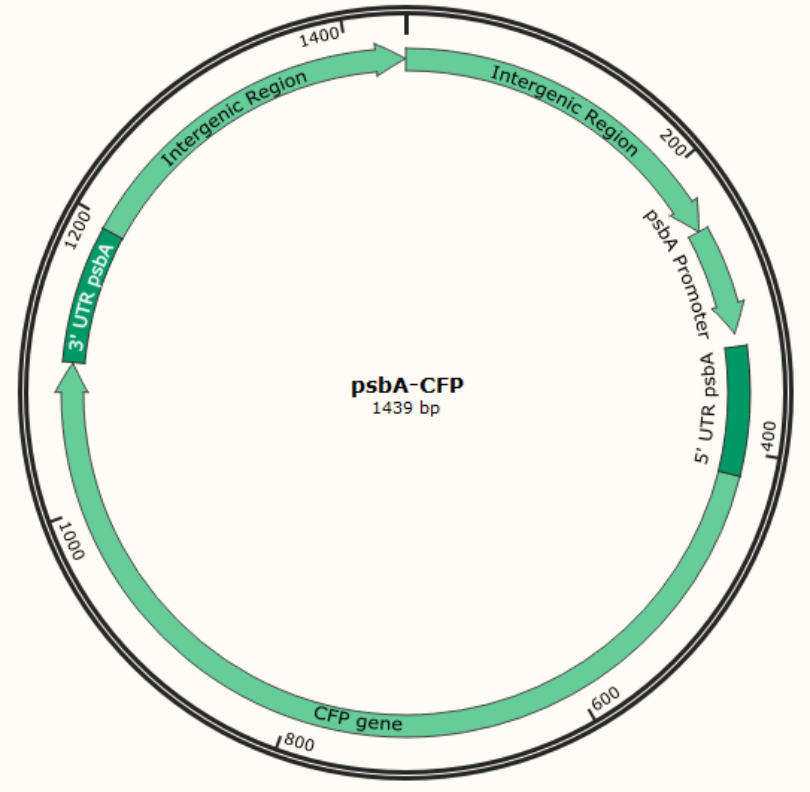


Figura 28. Vector diseñado para la expresión del gen informador en *Dunaliella salina* GN.

Tabla 8. Elementos elegidos para el plásmido de expresión del gen informador.

	Región intergénica	**Promotor**	**5' UTR**	**Gen**	**3'UTR**
Elemento	chlb y rrnS	psbA	psbA	CFP	psbA
Utilizado en el cloroplasto de	D. salina CCAP 19/18	*C. reinhardtii*	*D. salina 1009*	*C. reinhardtii*	*D. salina 1009*
Longitud (pb)	482	74	88	684	93
Referencia	(Talebi *et al.*, 2014)	(Ishikura *et al.*, 1999)	(D. Li *et al.*, 2011)	(Lim *et al.*, 2013)	(D. Li *et al.*, 2011)

En el segundo vector, el promotor elegido pertenece al gen *psbA*, elemento anteriormente utilizado para la modificación del cloroplasto de *C. reinhardtii* (Ishikura *et al.*, 1999), con la finalidad de expresar la proteína β-glucuronidasa, gen proveniente de la bacteria *Escherichia coli*.

Por otra parte, las secuencias 5' UTR y 3' UTR elegidas provienen del gen *psbA*, utilizadas para la modificación de la cepa 1009 de *D. salina* (D. Li *et al.*, 2011), con el objetivo de expresar la proteína verde fluorescente mejorada (Enhanced Green Fluorescent Protein, EGFP por sus siglas en inglés).

Finalmente, el gen informador elegido es la proteína cyan fluorescente (CFP por sus siglas en inglés), utilizado anteriormente para la expresión en el cloroplasto de *C. reinhardtii* (Lim *et al.*, 2013).

Los elementos elegidos para el diseño del gen informador, se utilizaron anteriormente en cepas distintas de *D. salina* y organismos con proximidad evolutiva, teniendo éxito en las transformaciones, por lo que se infiere que la implementación de este plásmido permitirá expresar proteínas de interés en trabajos futuros que tengan el objetivo de modificar el plastoma de *D. salina.*

VI.7 Identificación de regiones potenciales para inserción de casetes de expresión de proteínas recombinantes

Se seleccionaron dos regiones intergénicas para la inserción de los vectores diseñados, considerando que se utilizaron para anteriores modificaciones genéticas en el cloroplasto de otras cepas de *D. salina.*

En el plásmido diseñado para la expresión del gen informador, se seleccionó a la región *trnA-trnI*, secuencia intergénica que consta de 3005 pares de bases, identificándose en las coordenadas 174,126 y 177,130 del genoma aquí ensamblado. La región potencial elegida se utilizó con anterioridad para fines de expresar el gen de la proteína verde fluorescente mejorada en la cepa de *D. salina* 1009 (D. Li *et al.*, 2011).

En el caso del vector para la expresión de proteínas recombinantes, se seleccionó a la región intergénica *chlb* y *rrnS*, que se encuentra entre las coordenadas 125,459 y 128,080 del genoma de la cepa GN, constando de 2,622 pares de bases de longitud. Esta secuencia se consideró anteriormente en la cepa de *D. salina* 1009, con el objetivo de expresar los genes *ME* y *AccD*, genes implicados en la producción de lípidos (Talebi *et al.*, 2014).

Por lo anteriormente constatado, es posible inferir que la identificación de dichas regiones potenciales utilizadas en otras cepas de *D. salina*, permitirá la recombinación homóloga de estos vectores diseñados para la expresión de proteínas recombinantes en trabajos futuros.

VII. Conclusiones

El ensamble *de novo* del cloroplasto de *Dunaliella salina* de Guerrero Negro, no se logró debido a la baja capacidad del sistema de cómputo utilizado.

El ensamble por referencia con *D. salina* CCAP 19/18 presentó cantidad menor de gaps en comparación con la cepa SQ, esto debido a la proximidad evolutiva que existe entre las cepas de Australia y Guerrero Negro, México.

El genoma del cloroplasto de *D. salina* GN, contiene 122 genes, de los cuales 86 codifican para proteína, 30 para ARN ribosomal y 6 para ARN de transferencia, mismos genes encontrados en la cepa CCAP 19/18, originaria de Australia.

Las regiones intergénicas seleccionadas (125,459-128,080 y 174,126-177,130) son potenciales para la inserción del casete diseñado, debido a que son regiones que no afectan el funcionamiento del organismo si se reemplazan por un casete de expresión.

Referencias

Alkan, C., Sajjadian, S., y Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nature Methods*. https://doi.org/10.1038/nmeth.1527

Ambardar, S., Gupta, R., Trakroo, D., Lal, R., y Vakhlu, J. (2016). High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian Journal of Microbiology*. Springer India. https://doi.org/10.1007/s12088-016-0606-4

Bock, R. (2007). Structure, function, and inheritance of plastid genomes. *Topics in Current Genetics*, *19*, 29–63. https://doi.org/10.1007/4735_2007_0223

Bolger, A. M., Lohse, M., y Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Borowitzka, M. A., y Siva, C. J. (2007). The taxonomy of the genus Dunaliella (Chlorophyta, Dunaliellales) with emphasis on the marine and halophilic species. *Journal of Applied Phycology*, *19*(5), 567–590. https://doi.org/10.1007/s10811-007-9171-x

Carrera Pacheco, S. E., Hankamer, B., y Oey, M. (2018). Optimising light conditions increases recombinant protein production in Chlamydomonas reinhardtii chloroplasts. *Algal Research*, *32*, 329–340. https://doi.org/10.1016/j.algal.2018.04.011

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., y Hwang, C. C. (2013). Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PLoS ONE*, *8*(4),

62856. https://doi.org/10.1371/journal.pone.0062856

Choudhuri, S. (2014). Additional Bioinformatic Analyses Involving Nucleic-Acid Sequences. In *Bioinformatics for Beginners* (pp. 157–181). Elsevier. https://doi.org/10.1016/b978-0-12-410471-6.00007-4

Clark, D. P., y Pazdernik, N. J. (2016a). Chapter 10 - Recombinant Proteins. In D. P. Clark y N. J. B. T.-B. (Second E. Pazdernik (Eds.) (pp. 335–363). Boston: Academic Cell. https://doi.org/https://doi.org/10.1016/B978-0-12-385015-7.00010-7

Clark, D. P., y Pazdernik, N. J. (2016b). Chapter 3 - Recombinant DNA Technology. In D. P. Clark y N. J. B. T.-B. (Second E. Pazdernik (Eds.) (pp. 63–95). Boston: Academic Cell. https://doi.org/https://doi.org/10.1016/B978-0-12-385015-7.00003-X

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., y Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, *4*(4), 265–270. https://doi.org/10.1038/nnano.2009.12

Coil, D., Jospin, G., y Darling, A. E. (2015). A5-miseq: An updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics*, *31*(4), 587–589. https://doi.org/10.1093/bioinformatics/btu661

Compeau, P. E. C., Pevzner, P. A., y Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, *29*(11), 987–991. https://doi.org/10.1038/nbt.2023

Daniell, H., Lin, C. S., Yu, M., y Chang, W. J. (2016). Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. *Genome Biology*. BioMed

Central Ltd. https://doi.org/10.1186/s13059-016-1004-2

Feng, S., Xue, L., Liu, H., y Lu, P. (2009). Improvement of efficiency of genetic transformation for Dunaliella salina by glass beads method. *Molecular Biology Reports*, *36*(6), 1433–1439. https://doi.org/10.1007/s11033-008-9333-1

Fisher, D. I., Mayr, L. M., y Roth, R. G. (2016). Expression Systems. In *Encyclopedia of Cell Biology* (Vol. 1, pp. 54–65). Elsevier Inc. https://doi.org/10.1016/B978-0-12-394447-4.10009-4

Goldschmidt-clermont, M. (1991). Transgenic expression of aminoglycoside adenine transferase in the chloroplast: A selectable marker for site-directed transformation of chlamydomonas. *Nucleic Acids Research*, *19*(15), 4083–4089. https://doi.org/10.1093/nar/19.15.4083

Goodwin, S., McPherson, J. D., y McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*. Nature Publishing Group. https://doi.org/10.1038/nrg.2016.49

Greiner, S., Lehwark, P., y Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Research*, *47*(W1), W59–W64. https://doi.org/10.1093/nar/gkz238

H., G., y G., M. T. (2018). Next-generation sequencing platforms for latest livestock reference genome assemblies. *African Journal of Biotechnology*, *17*(39), 1232–1240. https://doi.org/10.5897/ajb2018.16605

Heather, J. M., y Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*. Academic Press Inc. https://doi.org/10.1016/j.ygeno.2015.11.003

Hosseini Tafreshi, A., y Shariati, M. (2009). Dunaliella biotechnology: Methods and applications. *Journal of Applied Microbiology*. John Wiley y Sons, Ltd. https://doi.org/10.1111/j.1365-2672.2009.04153.x

Illumina. (2010). *De Novo Assembly Using Illumina Reads*.

Illumina. (2010). *Calling Sequencing SNPs*.

Illumina. (2011). *Quality Scores for Next-Generation Sequencing*. Recuperado de http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/

Ishikura, K., Takaoka, Y., Kato, K., Sekine, M., Yoshida, K., y Shinmyo, A. (1999). Expression of a foreign gene in Chlamydomonas reinhardtii chloroplast. *Journal of Bioscience and Bioengineering*, *87*(3), 307–314. https://doi.org/10.1016/S1389-1723(99)80037-1

Kalyanaraman, A., Hammond, K., Nieplocha †, J., Krishnan, M., Palmer, B., Tipparaju, V., … Gustafson, J. L. (2011). Genome Assembly. In *Encyclopedia of Parallel Computing* (pp. 755–768). Springer US. https://doi.org/10.1007/978-0-387-09766-4_402

Khan, A. R., Pervez, M. T., Babar, M. E., Naveed, N., y Shoaib, M. (2018). A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective. *Evolutionary Bioinformatics*, *14*, 117693431875865.

https://doi.org/10.1177/1176934318758650

Kumar, S., Stecher, G., Li, M., Knyaz, C., y Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Kuroiwa, T. (1991). The Replication, Differentiation, and Inheritance of Plastids with Emphasis on the Concept of Organelle Nuclei. *International Review of Cytology*, *128*(C), 1–62. https://doi.org/10.1016/S0074-7696(08)60496-9

Lang, B. F., Laforest, M. J., y Burger, G. (2007). Mitochondrial introns: a critical view. *Trends in Genetics*, *23*(3), 119–125. https://doi.org/10.1016/j.tig.2007.01.006

Li, D., Han, X., Zuo, J., Xie, L., He, R., Gao, J., … Cao, M. (2011). Construction of rice site-specific chloroplast transformation vector and transient expression of EGFP gene in Dunaliella Salina. *Journal of Biomedical Nanotechnology*, *7*(6), 801–806. https://doi.org/10.1166/jbn.2011.1339

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Recuperado de http://arxiv.org/abs/1303.3997

Li, H., y Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.

https://doi.org/10.1093/bioinformatics/btp324

Li, H., y Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., … Fan, W. (2012). Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Briefings in Functional Genomics*, *11*(1), 25–37. https://doi.org/10.1093/bfgp/elr035

Lim, J. M., Ahn, J. W., Hwangbo, K., Choi, D. W., Park, E. J., Hwang, M. S., … Jeong, W. J. (2013). Development of cyan fluorescent protein (CFP) reporter system in green alga Chlamydomonas reinhardtii and macroalgae Pyropia sp. *Plant Biotechnology Reports*, *7*(3), 407–414. https://doi.org/10.1007/s11816-013-0274-3

Lopez, H., Magdaleno, D., y Stephano, J. (2017). The complete chloroplast genome of the green microalgae *Dunaliella salina* strain SQ. *Mitochondrial DNA Part B*, *2*(1), 225–226. https://doi.org/10.1080/23802359.2017.1310610

Lutz, K. A., Azhagiri, A. K., Tungsuchat-Huang, T., y Maliga, P. (2007). A guide to choosing vectors for transformation of the plastid genome of higher plants. *Plant Physiology*, *145*(4), 1201–1210. https://doi.org/10.1104/pp.107.106963

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., … Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, *47*(W1), W636–W641. https://doi.org/10.1093/nar/gkz268

Magdaleno Moncayo, D. A., y Stephano Hornedo, J. L. (2017). *Ensamble y anotación del genoma de Dunaliella salina*.

Manivasagan, P., y Kim, S.-K. (2015). Chapter 34 - An Overview of Harmful Algal Blooms on Marine Organisms. In S.-K. B. T.-H. of M. M. Kim (Ed.) (pp. 517–526). Boston: Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-800776-1.00034-0

Maréchal, A., y Brisson, N. (2010). Recombination and the maintenance of plant organelle genome stability. *New Phytologist*. John Wiley y Sons, Ltd. https://doi.org/10.1111/j.1469-8137.2010.03195.x

Maxam, A. M., y Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, *74*(2), 560 LP – 564. https://doi.org/10.1073/pnas.74.2.560

Meera Krishna, B., Khan, M. A., y Khan, S. T. (2019). Next-generation sequencing (NGS) platforms: An exciting era of genome sequence analysis. In *Microbial Genomics in Sustainable Agroecosystems: Volume 2* (pp. 89–109). Springer Singapore. https://doi.org/10.1007/978-981-32-9860-6_6

Meyers, B., Zaltsman, A., Lacroix, B., Kozlovsky, S. V., y Krichevsky, A. (2010). Nuclear and plastid genetic engineering of plants: Comparison of opportunities and challenges. *Biotechnology Advances*. Elsevier. https://doi.org/10.1016/j.biotechadv.2010.05.022

Miller, J. R., Koren, S., y Sutton, G. (2010). Assembly algorithms for next-generation

sequencing data. *Genomics*. NIH Public Access. https://doi.org/10.1016/j.ygeno.2010.03.001

Oey, M., Ross, I. L., y Hankamer, B. (2014). Gateway-Assisted Vector Construction to Facilitate Expression of Foreign Proteins in the Chloroplast of Single Celled Algae. *PLoS ONE*, *9*(2), e86841. https://doi.org/10.1371/journal.pone.0086841

Okonechnikov, K., Golosova, O., Fursov, M., Varlamov, A., Vaskin, Y., Efremov, I., … Tleukenov, T. (2012). Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics*. Oxford Academic. https://doi.org/10.1093/bioinformatics/bts091

Oren, A. (2005). A hundred years of Dunaliella research: 1905-2005. *Saline Systems*, *1*(1), 2. https://doi.org/10.1186/1746-1448-1-2

Peng, Y., Leung, H. C. M., Yiu, S. M., y Chin, F. Y. L. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, *28*(11), 1420–1428. https://doi.org/10.1093/bioinformatics/bts174

Peona, V., Blom, M. P. K., Xu, L., Burri, R., Sullivan, S., Bunikis, I., … Suh, A. (2020). Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13252

Raven, J. A., y Giordano, M. (2014). Algae. *Current Biology*, *24*(13), R590–R595. https://doi.org/10.1016/j.cub.2014.05.039

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., … Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, *475*(7356), 348–352. https://doi.org/10.1038/nature10242

Sanger, F., Nicklen, S., y Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463–5467. https://doi.org/10.1073/PNAS.74.12.5463

Schadt, E. E., Turner, S., y Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, *19*(R2), R227–R240. https://doi.org/10.1093/hmg/ddq416

Scotti, N., Bellucci, M., y Cardi, T. (2013). The chloroplasts as platform for recombinant proteins production. In *Translation in Mitochondria and Other Organelles* (Vol. 9783642392801, pp. 225–262). Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39426-3_10

Sequence Length Distribution. (n.d.). Recuperado de https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3 Analysis Modules/7 Sequence Length Distribution.html

Shanmugaraj, B., Bulaon, C. J. I., y Phoolcharoen, W. (2020). Plant molecular farming: A viable platform for recombinant biopharmaceutical production. *Plants*. MDPI AG. https://doi.org/10.3390/plants9070842

Shen, R., Fan, J. B., Campbell, D., Chang, W., Chen, J., Doucet, D., … Oliphant, A. (2005). High-throughput SNP genotyping on universal bead arrays. *Mutation Research -*

Fundamental and Molecular Mechanisms of Mutagenesis. Elsevier. https://doi.org/10.1016/j.mrfmmm.2004.07.022

Shepelev, M. V., Kalinichenko, S. V., Deykin, A. V., y Korobko, I. V. (2018). Production of recombinant proteins in the milk of transgenic animals: Current state and prospects. *Acta Naturae*. Acta Naturae. https://doi.org/10.32607/20758251-2018-10-3-40-47

Shi, C., Hu, N., Huang, H., Gao, J., Zhao, Y.-J., y Gao, L.-Z. (2012). An Improved Chloroplast DNA Extraction Procedure for Whole Plastid Genome Sequencing. *PLoS ONE*, *7*(2), e31468. https://doi.org/10.1371/journal.pone.0031468

Siddiqui, A., Wei, Z., Boehm, M., y Ahmad, N. (2020). Engineering microalgae through chloroplast transformation to produce high-value industrial products. *Biotechnology and Applied Biochemistry*, *67*(1), 30–40. https://doi.org/10.1002/bab.1823

Simpson, J. T., y Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, *22*(3), 549–556. https://doi.org/10.1101/gr.126953.111

Smith, David R., Lee, R. W., Cushman, J. C., Magnuson, J. K., Tran, D., y Polle, J. E. W. (2010). The Dunaliella salina organelle genomes: Large sequences, inflated with intronic and intergenic DNA. *BMC Plant Biology*, *10*. https://doi.org/10.1186/1471-2229-10-83

Smith, David Roy, y Keeling, P. J. (2015). Mitochondrial and plastid genome

architecture: Reoccurring themes, but significant differences at the extremes. *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences. https://doi.org/10.1073/pnas.1422049112

Talebi, A. F., Tohidfar, M., Bagheri, A., Lyon, S. R., Salehi-Ashtiani, K., y Tabatabaei, M. (2014). Manipulation of carbon flux into fatty acid biosynthesis pathway in Dunaliella salina using AccD and ME genes to enhance lipid content and to improve produced biodiesel quality. *Biofuel Research Journal*, *1*(3), 91–97. https://doi.org/10.18331/BRJ2015.1.3.6

Variant calling | EMBL-EBI Train online. (n.d.). Recuperado de https://www.ebi.ac.uk/training/online/glossary/variant-calling

Walker, T. L., Purton, S., Becker, D. K., y Collet, C. (2005). Microalgae as bioreactors. *Plant Cell Reports*. Springer. https://doi.org/10.1007/s00299-005-0004-6

Wang, X., Zhou, T., Bai, G., y Zhao, Y. (2018). Complete chloroplast genome sequence of Fagopyrum dibotrys: genome features, comparative analysis and phylogenetic relationships. *Scientific Reports*, *8*(1), 12379. https://doi.org/10.1038/s41598-018-30398-6