UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
FACULTAD DE CIENCIAS MARINAS
INSTITUTO DE INVESTIGACIONES OCEANOLÓGICAS

ENSAMBLE Y ANOTACIÓN DEL GENOMA DE *DUNALIELLA SALINA*

T E S I S

QUE PARA CUBRIR PARCIALMENTE LOS REQUISITOS NECESARIOS PARA OBTENER EL GRADO DE

DOCTOR EN CIENCIAS EN ECOLOGÍA MOLECULAR y BIOTECNOLOGÍA

PRESENTA

DANTE ALBERTO MAGDALENO MONCAYO

ENSENADA, BAJA CALIFORNIA, MEXICO. OCTUBRE 2017

Resumen

La microalga verde *Dunaliella salina* tiene un gran potencial biotecnológico, produce hasta 10% de Beta-carotenos de peso seco y puede alcanzar una producción de lípidos de 35% de peso seco en condiciones de haloestrés, es el eucarionte más halotolerante que se conoce a la fecha. Debido a su potencial biotecnológico es importante la caracterización de los genomas nucleares y mitocondriales de las cepas de *Dunaliella salina*, para generar información base que sea útil en la manipulación genética de esta microalga. En este sentido en el presente trabajo se secuenció el genoma nuclear de *Dunaliella salina* con las plataformas de secuenciación PacBio e Illumina HiSeq 2500, se llevaron a acabo ensambles *de novo* usando algoritmos de traslape de grafos. Se secuenció el DNA mitocondrial con la plataforma MiSeq de Illumina de dos cepas de *Dunaliella salina* de Baja California, una aislada de una laguna hipersalina de San Quintín (SQ) y la otra de una laguna hipersalina de Guerrero negro (GN). Se ensamblaron los genomas mitocondriales *de novo* por medio del A5 *pipeline*. Se obtuvieron dos contigs que corresponden a los genomas completos mitocondriales, uno de 41,904 pares de bases (pb) para la cepa SQ y otro de 27,950 pb para la cepa GN. Se llevaron a cabo análisis de sintenia de con y se encontró que los genomas mitocondriales de las cepas SQ y GN presentan una arquitectura genómica distinta, contienen los mismos genes, pero se encuentran con un orden génico diferente entre cepas.

FACULTAD DE CIENCIAS MARINAS
INSTITUTO DE INVESTIGACIONES OCEANOLÓGICAS
POSGRADO EN ECOLOGÍA MOLECULAR y BIOTECNOLOGÍA

ENSAMBLE Y ANOTACION DEL GENOMA DE DUNALIELLA SALINA

T E S I S

QUE PARA CUBRIR PARCIALMENTE LOS REQUISITOS NECESARIOS
PARA OBTENER EL GRADO DE

DOCTOR EN CIENCIAS

PRESENTA

DANTE ALBERTO MAGDALENO MONCAYO

Aprobada por:

Dr. José Luis Stephano Hornedo
Director de tesis

Dra. Amelia Portillo López
Sinodal

Dra. Sawako Hori-Oshima
Sinodal

Dr. Gerardo Enrique Medina Basulto
Sinodal

Dr. José Manuel López Rodríguez
Sinodal

Agradecimientos

Agradezco al Consejo de Ciencia y Tecnología (CONACYT) por su patrocinio.

Al Laboratorio Meredith Gould por su patrocinio para realizar mi trabajo de tesis.

A la Facultad de Ciencias Marinas y a la UABC por su apoyo.

Al Doctor José Luis Stephano Hornedo por su excelente guía en esta tesis.

A los sinodales

Dra. Sawako Hori-Oshima
Dra. Amelia Portillo López
Dr. Gerardo Enrique Medina Basulto
Dr. José Manuel López Rodríguez

Por su apoyo y comprensión.
.

Contenido

Lista de tablas

Lista de figuras

I. Introducción

Las algas son eucariontes fotosintéticos (excluyendo a las plantas terrestres) las cuales presentan una gama amplia de morfologías celulares y ciclos de vida, se encuentran distribuidas en una gran diversidad de hábitats. Las algas se clasifican en Chlorophyta, Rhodophyta, Glaucocystophyta, Euglenophyta, Chlorarachniophyta, Heterokonta, Haptophyta, Cryptophyta (Bhattacharrya y Medlin 1998). En las algas se encuentran macroalgas y microalgas, las microalgas son microorganismos eucariontes protistas fotosintéticos, los cuales pueden presentar un crecimiento rápido y sobrevivir en condiciones adversas debido a su estructura unicelular y multicelular sencilla (Mata *et al*., 2010). Estos microorganismos son constituyentes importantes de una gran variedad de ecosistemas, distribuyéndose desde ambientes acuáticos marinos y de agua dulce hasta desiertos, fuentes termales, hielo y nieve. La clasificación de las microalgas tradicionalmente se ha realizado usando pigmentación, ciclo celular y estructuras celulares, sin embargo, también existen métodos moleculares como electroforesis en gel desnaturalizante y secuenciación de siguiente generación, los cuales han logrado detectar y clasificar nuevas especies de microalgas (Odjadjare et al., 2015).

Las microalgas verdes pertenecen al linaje Viridiplantae, en el cual se incluyen las algas verdes y las plantas terrestres (Leliaert *et al*., 2012), las plantas evolucionaron de un ancestro alga verde (Kenrick y Crane 1997). El mecanismo fotosintético de las microalgas es similar al de las plantas terrestres, pero debido

a su sencilla estructura celular, y el hecho que la mayoría de estos microorganismos habitan en ambientes acuosos en donde se presenta una mayor eficiencia a el acceso de agua, CO_2, y otros nutrientes, por lo general son más eficientes en la conversión de la energía solar a biomasa (Gouveia., 2011). Los organismos verdes se dividen en dos phyla, Chlorophyta y Streptophyta (Leliaert *et al*., 2012). La mayoría de las algas verdes se encuentran en el phylum Chlorophyta, el resto se encuentra en el phylum Streptophyta junto con todas las plantas. El phylum Clorophyta se encuentra dividido en cuatro clases, Sphaeropleales, Chaetophorales, Chaetopeltidales, Oedogoniales y Chlamydomonales (Brouard et al., 2010). Chlamydomonales contiene tres géneros de algas verdes ampliamente estudiados, *Chlamydomonas*, *Volvox* y *Dunaliella*.

Existe un gran interés en el cultivo de microalgas verdes para la obtención de productos de alto valor agregado, productos sintetizados de forma natural como los almidones, carotenoides y lípidos; también metabolitos con actividad antiviral, antifúngica e inmuno-moduladora (Deth, 1999). Las microalgas verdes son una gran fuente de generación de bioenergía ya que bajo condiciones adecuadas de cultivo producen una alta cantidad de lípidos para la elaboración de biodiesel con un rendimiento 100 veces mayor que otros organismos (Rittmann, 2008).

Muchas especies de algas presentan tasas de producción de biomasa que pueden sobrepasar aquellas de las plantas terrestres (Dismukes *et al*., 2008), y

muchas microalgas verdes tienen la habilidad de almacenar cantidades significantes de compuestos ricos en energía como almidón y triglicéridos (TG), los cuales pueden ser utilizados para la producción de diferentes biocombustibles como por ejemplo etanol y biodiesel.

Una de las microalgas verdes con potencial biotecnológico es *Dunaliella salina*. *D. salina* pertenece al phylum Chlorophyta del orden Volvocales y de la familia *Polyblepharidaceae*, es una microalga verde unicelular fotosintética, presenta movilidad por medio de dos flagelos. Su reproducción puede ser sexual y asexual, la reproducción sexual comienza con el contacto de flagelos de dos células provocando la fusión de las células (Oren 2005). La frecuencia de reproducción sexual en *D. salina* aumenta cuando disminuye la salinidad en el medio (Martinez *et al*., 1995). Morfológicamente se distingue por la falta de una pared celular rígida (Ben-Amotz y Avron, 1987), habita en ambientes marinos y en lagos salinos en donde puede sobrevivir en un rango de 0.5 % hasta 35% de salinidad, por lo cual es el eucarionte más halotolerante conocido hasta ahora (Hosseini-Tafreshi y Shariati, 2009). Para sobrevivir a las fluctuaciones de salinidad en el medio, *D. salina* comienza a producir y acumular glicerol (Seckbanch y Oren 2007), la ausencia de pared celular causa que los nivele osmóticos tengan gran influencia en la forma celular (Oren 2005).

D. salina es conocida por ser una fábrica de producción de Beta-caroteno, en donde hasta el 10% de peso seco celular es Beta-caroteno (Ben-Amotz *et al*., 1983). El Beta-caroteno es un pigmento que en *D. salina* se encuentra principalmente en una estructura llamada ojo (Kreimer 2009), esta estructura

contiene fotorreceptores que le permiten a la célula acercarse o alejarse de la luz. En la estructura ojo el Beta-caroteno junto con otros carotenoides, protegen a la célula de fotoblanqueo (Kreimer 2009). El color rojizo que muestra *D. salina* es debido a la alta producción y acumulación de Beta-caroteno (Oren 2005). Se ha reportado que *D. salina* puede producir hasta 35% de lípidos de su peso seco (Griffiths y Harrison, 2009), también que en respuesta a haloestres incrementa la producción de lípidos (Alhasan, Ghannoum *et al*., 1987).

A la fecha el DOE JGI (Department of Energy Joint Institute) se encuentra trabajando en el ensamble y anotación del genoma nuclear de *D. sailna* cepa CCAP 19/18 que se aisló en Australia, hasta el momento solo se han completado los genomas del cloroplasto y mitocondria (Smith *et al*., 2010). Esta tesis tiene como propósito secuenciar, ensamblar y anotar el genoma nuclear y mitocondrial de una cepa de *D. salina* aislada de una laguna costera de San Quintín Baja California México. La información que se genere con este trabajo permitirá tener un mejor entendimiento de los genes y rutas metabólicas que intervienen en la producción de Beta-caroteno y lípidos de una cepa de *D. salina* de la región, así también conocer la variabilidad que existe entre distintas cepas de *D. salina* que se encuentran separadas geográficamente.

La métodos de secuenciación de siguiente generación (NGS del inglés Next Generation Sequencing) o secuenciación de alto rendimiento son de gran importancia en la ciencias biológicas, ya que no se requiere clonar en bacterias fragmentos de DNA, en lugar de esto se preparan bibliotecas en un sistema libre de células, también se realiza la secuenciación de miles hasta millones de

fragmentos de DNA en paralelo y el resultado de la secuenciación se detecta sin la necesidad de hacer electroforesis, la información de las secuencias (También llamadas lecturas) se obtiene de manera cíclica y en paralelo (*Dijk et al*., 2014).

La gran cantidad de lecturas que se generan por NGS ha facilitado la secuenciación de genomas completos en cuestión de días y con un costo significativamente menor comparado con la secuenciación de primera generación como Sanger. El primer sistema de secuenciación NGS que salió al mercado fue el de pirosecuenciación (454) en 2005 por Roche, este sistema genera 200,000 lecturas de 110 pares de bases (pb) (*Margulies et al.,* 2005). En 2006 la plataforma de secuenciación solexa/Illumina sale al mercado, este sistema produce una mayor cantidad de lecturas que 454, hasta 30 millones de lecturas, pero con solo una longitud de 35 pb. Life Technologies en 2010 saco al mercado en sistema Personal Genome Machine (PGM), con este sistema se pueden obtener hasta 270 Mb de secuencias con una longitud de 100 pb (Metzker 2010). El error en las lecturas de las tecnologías de secuenciación antes mencionadas es cercano al 0.001%, lo cual las hace muy confiables para identificar variantes genéticas entre cepas de la misma especie, pero debido a la corta longitud de las lecturas que producen, aumenta la complejidad para ensamblar genomas completos. PacBio en 2010 lanza al mercado el sistema de secuenciación llamado PacBio RS, este sistema genera lecturas con un rango de longitud de 1,000 a 20,000 pb, lo cual hace de esta plataforma ideal para el ensamble de genomas completos (Eid *et al.,* 2009). La desventaja del sistema PacBio RS es el 18% de error que presenta en sus lecturas.

El desarrollo de las tecnologías de secuenciación de DNA ha incrementado el número de lecturas a millones por corrida de secuenciación, esta enorme cantidad de información dificulta en gran medida el ensamble de genomas, esto ha ocasionado que el ensamble de genomas de ser un problema biológico se convierta a un problema computacional. Para afrontar dicho problema se han desarrollado algoritmos que buscan ensamblar genomas utilizando estrategias, como ensamble *de novo* y ensamble por referencia.

Las lecturas obtenidas de la secuenciación son convertidas a una estructura de datos específica, para posteriormente ser el archivo de entrada del programa ensamblador, actualmente dos categorías de estructura de datos se utilizan principalmente en el ensamble de genomas, el modelo basado en cadenas (string-based model) y el modelo basado en gráficos (graph-based model) (Warren *et al*., 2007). Los ensambladores string-based model implementados con el algoritmo extensión codiciosa (greedy-extensión algorithm) son utilizados principalmente para el ensamble de genomas pequeños, mientras que los ensambladores graph-based model son utilizados cuando son genomas grandes o complejos en su arquitectura (Li *et al*., 2010) (Simpson *et al*., 2010). Dado que las lecturas de las plataformas NGS son de longitud corta, al momento de ensamblar genomas complejos con gran cantidad de secuencias repetitivas se vuelve un problema para los ensambladores ya mencionados, algunos ensambladores utilizan lecturas paired-end las cuales pueden unirse y obtener lecturas de mayor longitud que ayudan a resolver el problema de las secuencias repetitivas.

A pesar del avance tecnológico en los sistemas de secuenciación NGS, y el potencial biotecnológico que muestran las microalgas, a la fecha se encuentran pocos genomas secuenciados y anotados de estos microorganismos, por lo cual los trabajos enfocados en aumentar la producción de metabolitos en microalgas se han centrado en ingeniería bioquímica, esta estrategia se basa en controlar los nutrientes y las condiciones de cultivo.

También existen factores de estrés que pueden ocasionar el incremento de productos de interés para la elaboración de biocombustibles como los lípidos en microalgas, un ejemplo de esto se presenta en *Dunaliella salina* en condiciones de alta salinidad, en donde se ha observado que aumenta el contenido de TAGs (Takagi *et al*., 2006). Esto tiene sus desventajas, ya que la limitación de nutrientes y el estrés fisiológico que se requiere para acumular una alta cantidad de lípidos está asociado con la reducción de la división celular (Ratledge, 2002).

Una de las estrategias que se está considerando con gran interés es la utilización de ingeniería genética para incrementar la producción de lípidos en las microalgas, esto gracias a el rápido desarrollo que se ha estado dando en la biotecnología de microalgas. Se han terminado de secuenciar genomas completos de varias algas, como el alga roja *Cyanidioschyzon merolae* (Nozaki *et al*., 2007) y diatomeas como *Thalassiosira pseudonana* (Armbrust *et al*., 2004). También se han terminado de ensamblar y anotar los genomas de algunas microalgas como el genoma de la microalga verde *Chlorella variabilis* NC64A, el cual tiene un tamaño de 46.2 Mega bases (Mb) y el ensamble arrojó 413

andamios (Scaffolds) mayores a 1 kilo base (kb) con un conteo de 9,791 genes (Blanc *et al*., 2010). *Nannochloropis gaditana* una microalga marina es otro organismo de importancia biotecnológica del cual recientemente se obtuvo el ensamble y anotación completo de su genoma, obteniendo un tamaño estimado de 29 Mb con un contenido de G+C del 54.2%, un contiguo (contig) N50 de 404 y un scaffold N50 de 257, con un número predictivo de genes nucleares de 8,892 (Radakovits y Jinkerson *et al.,* 2012). También el genoma de la microalga verde *Chlamydomonas reinhardtii* se encuentra ensamblado y anotado (Merchant *et al*., 2007).

A la fecha el genoma de *Dunaliella salina* no se ha secuenciado, ensamblado y anotado, solo se han realizado trabajos de identificación de genes de manera individual (Ramos *et al*., 2008., Yan *et al*., 2004). Zhao *et al,* (2011) realizaron un estudio de EST (Expressed Sequence Tags) en *Dunaliella salina* y encontraron 4,118 secuencias únicas de las cuales el 56.1% presentó similitud con otras secuencias. En otro trabajo de transcriptoma se cultivó *Dunaliella salina* con 2.5 M de NaCl y se dio un cambio de salinidad a 3.4 M de NaCl por 5 horas para hacer el análisis de EST en condiciones de shock por aumento de salinidad (Alkayal *et al.,* 2010). En *Dunaliella tertiolecta* se tienen resultados del transcriptoma en condiciones de cultivo favorables para la producción de productos de almacenamiento de alta energía (Rismani *et al*., 2011), lo cual nos proporciona información de un organismo cercano a *Dunaliella salina* para poder inferir su metabolismo y tener una base de datos de secuencias potenciales

codificadoras a genes relacionados con las diferentes rutas metabólicas relacionadas con la producción de lípidos.

En el presente trabajo se secuenció el genoma nuclear y mitocondrial de la microalga *Dunaliella salina* SQ por medio de los sistemas de secuenciación PacBio, Illumina HiSeq 2500 e Illumina MiSeq. También se secuenció el genoma mitocondrial de una cepa aislada de las salinas de Guerrero Negro B.C que se le denominó *Dunaliella salina* GN, la secuenciación se llevó a cabo con el sistema Illumina MiSeq. El objetivo es realizar el ensamble *de novo* del genoma nuclear y ensamble por referencia de los genomas mitocondriales, para posteriormente hacer la anotación de estos e identificar los genes y la arquitectura de los genomas. Así también comparar el genoma mitocondrial con otras cepas de *Dunaliella salina* separadas geográficamente e identificar las variantes genéticas. La realización de este trabajo aportaría información base para que posteriormente se puedan llevar a cabo modificaciones genéticas de una manera más estratégica utilizando ingeniería genética en *Dunaliella salina* SQ, como aumentar la producción de lípidos o algún compuesto de interés biotecnológico que genere este microorganismo, así como también expresar proteínas recombinantes de interés médico o industrial en esta microalga utilizando secuencias reguladoras que promuevan una alta tasa de transcripción y traducción.

II. Justificación

Las microalgas son organismos de gran interés en la investigación ya que son un buen modelo de estudio para entender los procesos metabólicos, también son de interés biotecnológico debido a los productos que se obtienen de estas. Uno de los productos que se obtienen de las microalgas con gran potencial son los lípidos, ya que estos se pueden utilizar para la producción de biocombustibles como el biodiesel, el cual es una alternativa para remplazar a los combustibles a base de petróleo. *Dunaliella salina* es una microalga altamente productora de lípidos, también presenta un crecimiento en condiciones de alta salinidad lo cual nos da la ventaja de hacer cultivos a gran escala sin la problemática de la competencia por otros organismos. La microalgas que se utilizarán en este estudio fueron aisladas de las lagunas salinas de San Quintín Baja California México y de Guerrero Negro B.C. lo que facilita el escalamiento de cultivos a pozas abiertas en la zona para la producción de grandes cantidades de biomasa. Actualmente no se cuenta con la secuencia del genoma de *Dunaliella salina*, solo existen trabajos de transcriptomas y la identificación de algunos genes. En este sentido el presente trabajo tiene la importancia de contribuir con el desarrollo de conocimiento relacionado con la estructura del genoma, identificación de genes y secuencias reguladoras, todo esto con potencial uso biotecnológico.

III. Hipótesis

Con el ensamble y anotación del genoma de *Dunaliella salina* SQ y GN se resolverá la arquitectura del genoma nuclear y mitocondrial.

IV. Objetivo

Ensamblar y anotar el genoma nuclear de *Dunaliella salina* SQ y los genomas mitocondriales de *Dunaliella salina* SQ y GN.

IV.1 Objetivos particulares

1. Secuenciación del genoma nuclear de *Dunaliella salina* SQ con las plataformas PacBio e Illumina HiSeq 2500.
2. Ensamble *de novo* del genoma nuclear de *Dunaliella salina* SQ.
3. Secuenciación del genoma mitocondrial de *Dunaliella salina* SQ y *Dunaliella salina* GN con la plataforma Illumina MiSeq.
4. Ensamble por referencia del genoma mitocondrial de *Dunaliella salina* SQ y *Dunaliella salina* GN.
5. Ensamble *de novo* de los genomas mitocondriales de *Dunaliella salina* SQ y GN.
6. Anotación de los genomas mitocondriales de *Dunaliella salina* SQ y GN y análisis de variación genética de los organelos *D*. *salina* SQ y *D. Salina* GN.

V. Metodología

V.1 Secuenciación del genoma Nuclear de *Dunaliella salina*.

V.1.1. Extracción DNA total

Para la realización del objetivo 1 se cultivó Dunaliella *salina* SQ hasta la fase media exponencial en 500 mL de medio líquido de Johnson modificado a una concentración de 250 mM de NaCl (Feng *et al*., 2009). Posteriormente se llevó a cabo la extracción de DNA total utilizando el kit miniprep axyprep multisource genomic DNA (Axygene) siguiendo el protocolo del proveedor, se analizó integridad del DNA por medio de un gel de agarosa al 1% y se cuantificó usando un nanodrop. Se enviaron 21 microgramos de DNA total para hacer la secuenciación en UC San Diego Core Center con las plataformas PacBio e Illumina HiSeq 2500.

V.1.2 Ensamble de novo del genoma *Dunaliella salina* SQ.

El procesamiento y los análisis de los datos se llevaron a cabo en una Workstation con 8 núcleos 4.2 GHz, 32 GB de memoria RAM, 3 Tb de almacenamiento y enfriamiento líquido. El sistema operativo que se utilizó fue Bio-Linux basado en Ubuntu 14.04.

V.I.3 Preprocesamiento

Los datos obtenidos de la plataforma de secuenciación PacBio se descargaron del FTP que proporcionó el Core Center de UCSD, el formato de los

datos es fatsq compresos .gz, los datos se descomprimieron usando *gzip –d datos.fastq.gz* en la línea de comandos del sistema operativo Ubuntu 14.04. Una vez descompresos los datos, se analizaron los valores de calidad (QS) de los nucleótidos en las lecturas y la longitud de estas con el programa FastQC (Andrews 2010) posteriormente las lecturas con QS por debajo de 30 en una ventana de 20 nucleótidos se eliminaron utilizando el programa Strin Graph Assembler (SGA) (Simpson y Durbin 2010), el cual cuenta con un submódulo de preprocesamiento para QS. Los datos de la plataforma Illumina HiSeq 2500 fueron preprocesados de igual forma que los de PacBio.

V.1.4 Ensamble de novo con lecturas PacBio

Se utilizó el programa M.I.R.A versión 4.0 (Mimicking Intelligent Read Assembly) (Chevreux, 2013) para llevar a cabo el ensamble *de novo* con las lecturas de PacBio, se indicó al programa el tipo de algoritmo para ensamblar, en este caso es graph-based model, la longitud de las lecturas 20 Kb, formato de las lecturas, tipo de ensamble (En este caso es *de novo*) y un tamaño estimado del genoma a ensamblar, para D*unaliella salina* se estima 300 Mpb (Smith *et al*., 2010). El programa analiza la calidad de las lecturas y si constan con un QS superior a 30 por ventanas de 20 nucleótidos continua con el ensamble.

V.1.5 Ensamble de novo con lecturas Illumina HiSeq 2500

Posterior al preprocesamiento de las lecturas de Illumina Hiseq 2500, se les hizo cambio de formato de fastq a fasta con un script escrito en Python, ya las lecturas en formato fasta, se procedió a estimar el valor de K (K-mer) que presente una mayor diversidad en los datos y utilizar este valor como base para el ensamble *de novo*. Los valores de K evaluados fueron 21, 31, 51, 61, 71, 81 y 91. Esto se llevó a cabo con el programa Kmergenie (Chikhi y Medvedev, 2013), ya con el valor de K estimado se realizaron 4 ensambles *de novo* con distintos números de lecturas: 75 millones, 112 millones y 375 millones de lecturas que corresponden a una cobertura del genoma de 25x, 37.5x, 125x respectivamente. Los ensambles se llevaron a cabo con el programa Minia (Chikhi y Rizk, 2012), este programa es un ensamblador de lecturas de Ilumina basado en grafo de Bruijn, el programa toma como entrada las lecturas en formato fasta, un valor de K (K=51 para este caso) y da como salida un set de contigs en formato fasta. Los ensambles se evaluaron con QUAST (Quality Assessment Tool for Genome Assemblies) (Gurevich *et al*., 2013).

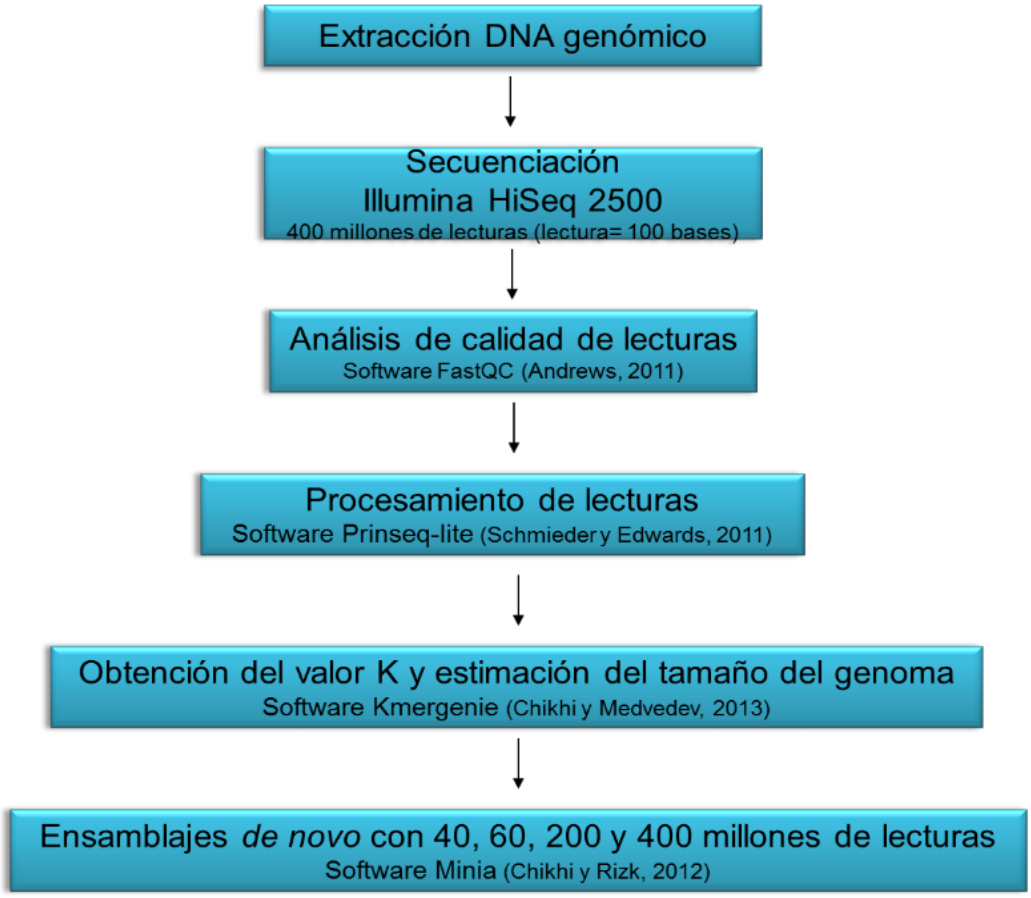

Figura 1. Diagrama de los pasos que se realizaron para ensamblar el genoma nuclear de *Dunaliella salina* SQ con lecturas de Illumina HiSeq 2500.

V.2 Secuenciación de los genomas mitocondriales de *Dunaliella salina* cepas SQ y GN.

V.2.1 Aislamiento de mitocondrias.

Para llevar a cabo el objetivo 3, se cultivaron las cepas SQ y GN de *Dunaliella salina* hasta alcanzar fase logarítmica en 200 ml de *medio Johnson modificado* a una concentración de 250 mM de NaCl, posteriormente los cultivos se incubaron a una temperatura de 4°C en obscuridad por un periodo de 48 horas, esto para llevar a cabo la disminución de los niveles de almidón. Los cultivos se centrifugaron a 3000 g durante 30 minutos, se descartó el sobrenadante y se homogeneizó por 5 minutos el pellet 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)), utilizando una micropipeta de 1000 µl. El homogeneizado se filtró por medio de dos capas de la membrana Miracloth (Merck) y el filtrado se colectó en tubos Falcon de 50 ml, posteriormente el filtrado se centrifugó 10 minutos a 3000 g y se resuspendió el pellet de mitocondrias en 10 ml de buffer de aislamiento frío. Enseguida se centrifugó por 10 minutos a 3000 g y el pellet se resuspendió en 10 ml de H2O destilada (Shi *et al*., 2012).

V.2.2 Extracción de DNA mitocondrial

La extracción del DNA de mitocondrias se llevó a cabo utilizando el kit AxyPrep Multisource Genomic DNA Miniprep siguiendo el protocolo del proveedor. Se analizó integridad del DNA por medio de un gel de agarosa al 1% y se cuantificó

usando un nanodrop. Se enviaron 3 microgramos de DNA mitocondrial de cada una de las cepas para hacer la secuenciación en IGM genomics center de UCSD, se utilizó el sistema MiSeq de Illumina con bibliotecas con tamaño de inserto de un rango de 400 a 900 pares de bases y lecturas de 300 bases paired-end.

V.2.3 Ensamble por referencia del genoma mitocondrial de *Dunaliella salina* SQ y *Dunaliella salina* GN.

La calidad de las lecturas se analizó con el software FastQC (Andrews 2010), las lecturas con un valor de calidad (QS) inferior a 30 se eliminaron. Se utilizó el programa Bowtie2 versión 2.2.6 de Langmead, Bowtie2 indexa los genomas de referencia utilizando un index FM basado en la Transformación Burrows-Wheeler, lo cual mantiene un uso bajo de memoria RAM, soporta tres tipos de alineamientos, espaciados, locales y con lecturas paired-end, tiene la capacidad de usar multi-procesadores para acelerar los ensamblajes, el formato de los archivos de alineamiento de salida son SAM (Sequence Alignment/Map). Como secuencia de referencia para el ensamblaje se utilizó el genoma de la mitocondria de *Dunaliella salina* (referencia NC_012930.1) y las lecturas paired-end de la cepa SQ, también se llevó a cabo el ensamblaje con las lecturas paired-end de la cepa GN usando como referencia la secuencia de NC_012930.1.

V.2.4 Ensamble *de novo* del genoma mitocondrial de *Dunaliella salina* SQ y GN.

El ensamblaje *de novo* se llevó a cabo por medio del pipeline A5-miseq, el ensamble consistió en 5 etapas:

1- Limpieza de lecturas:

En este paso secuencias adaptadoras y lecturas con valores de baja calidad fueron eliminadas utilizando el software Trimmomatic de (Lohse 2012), los errores en las lecturas fueron corregidos utilizando el algoritmo de corrección de errores basado en k-mers de SGA (Simpson y Durbin, 2012).

2- Ensamblaje de contigs:

Lecturas pareadas y no pareadas son utilizadas para el ensamblaje con el algoritmo IDBA-UD (Peng *et al*., 2013).

3- Andamios crudos (Crude scaffolding):

Los contigs son andamios con el tamaño disponible de inserto de la biblioteca secuenciada, en este caso los insertos son de 400 a 900 pares de bases.

4- Corrección de ensambles incorrectos:

Los ensambles incorrectos se detectan utilizando la información del tamaño de inserto de las lecturas forward y reverse pareadas (paired-end reads), en

ensambles donde lecturas pareadas que se encuentran a una distancia mayor de su inserto son desensamblados.

5- Andamios finales (Final scaffolding)

Se lleva a cabo una ronda de scaffolding con parámetros de reparación, también se generan los estadísticos de ensamblaje.

V.2.5 Anotación de los genomas mitocondriales de *Dunaliella salina* SQ y GN.

A los Andamios (scaffolds) obtenidos de los ensamblajes *de novo* se les realizó un BLAST (Basic Local Alignment Search Tool) esto para encontrar regiones de similitud local entre secuencias. El BLAST se llevó a cabo tomando los 10 scaffolds más grandes de cada ensamble, a cada uno de esos scaffolds se le comparó con secuencias de genes de mitocondria, los scaffolds que alinearon con genes se analizaron para estructurar el genoma completo de la mitocondria.

Una vez aislados los scaffolds correspondientes a mitocondria de la cepa SQ y GN, se identificaron los intrones, tRNAs y rRNA utilizando RNAweasel (http://megasun.bch.umon-treal.ca/RNAweasel/), RNAweasel predice estructuras complejas de RNA mitocondrial usando el algoritmo de búsqueda ERPIN, el cual es basado en perfiles de estructura secundaria de RNA. La

identificación de los genes codificantes para proteínas y los open reading frames se realizaron por medio de Mfannot y el resultado fue curado manualmente revisando los seis marcos de lecturas posibles y alineando otras secuencias de DNA mitocondrial de *Dunaliella salina* (referencia NC_012930.1). La curación manual de la anotación se realizó en el programa UGENE versión 1.18.0

VI. Resultados y Discusión

VI.1. Secuenciación del genoma Nuclear de *Dunaliella salina*.

VI.1.1 Extracción DNA total

Del cultivo de 500 ml de *Dunaliella salina* se obtuvo una concentración de DNA total de 223 ng/µl con un radio 260/280 de 1.8, 5 µl se corrieron en un gel de agarosa al 1% el resultado se muestra en la figura 2.

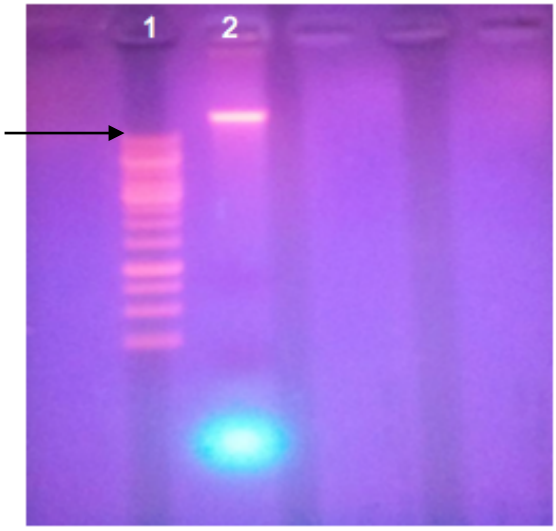

Figura 2. Electroforesis en gel de agarosa de la extracción de DNA total de *D. salina* **1)** Estándar molecular. **2)** Extracción de DNA total de *D. salina* SQ.

En el carril 2 de la foto del gel de agarosa, podemos ver que se obtuvo DNA integro, ya que al compararlo con el estándar molecular en donde la banda superior es de 10 kb como lo indica la flecha, se observa una sola banda arriba de 10 kb con poca degradación. Ya comprobado la pureza e integridad del DNA

se enviaron 21 µg a UC San Diego Core Center para la secuenciación con PacBio e Illumina HiSeq 2500.

La secuenciación con el sistema PacBio arrojó 1,129,958 lecturas de longitud de 11,000 pb long consensus, esto corresponde a una cobertura de 41x del genoma de *Dunaliella salina* SQ ya que el tamaño estimado del genoma es de 300 Mpb (Smith *et al*., 2010), esta cobertura es suficiente para proseguir con el ensamble *de novo*. También se obtuvieron 74,481 lecturas con una longitud de 2,800 pb circular consensus, que corresponde a una cobertura de 0.6x lo cual es insuficiente para realizar un ensamble *de novo* o por referencia.

VI.1.2 Ensamble *de novo* del genoma *Dunaliella salina* SQ.

VI.1.2.1 Ensamble de novo con lecturas PacBio

Las lecturas Long consensus obtenidas del Sistema PacBio se analizaron con el programa FastQC para evaluar la calidad de las lecturas (QS), en la figura 3 se muestra el resultado del análisis de calidad de las lecturas provenientes del sistema de secuenciación PacBio de la cepa SQ, en la gráfica podemos observar el rango de los valores de calidad de todas las bases en cada una de las posiciones en el archivo Fastq. La línea roja central es el valor de la mediana, la caja de color amarillo representa el rango intercuartil (25-75%), los bigotes superior e inferior representan los puntos 10% y 90% y la línea azul representa la calidad media.

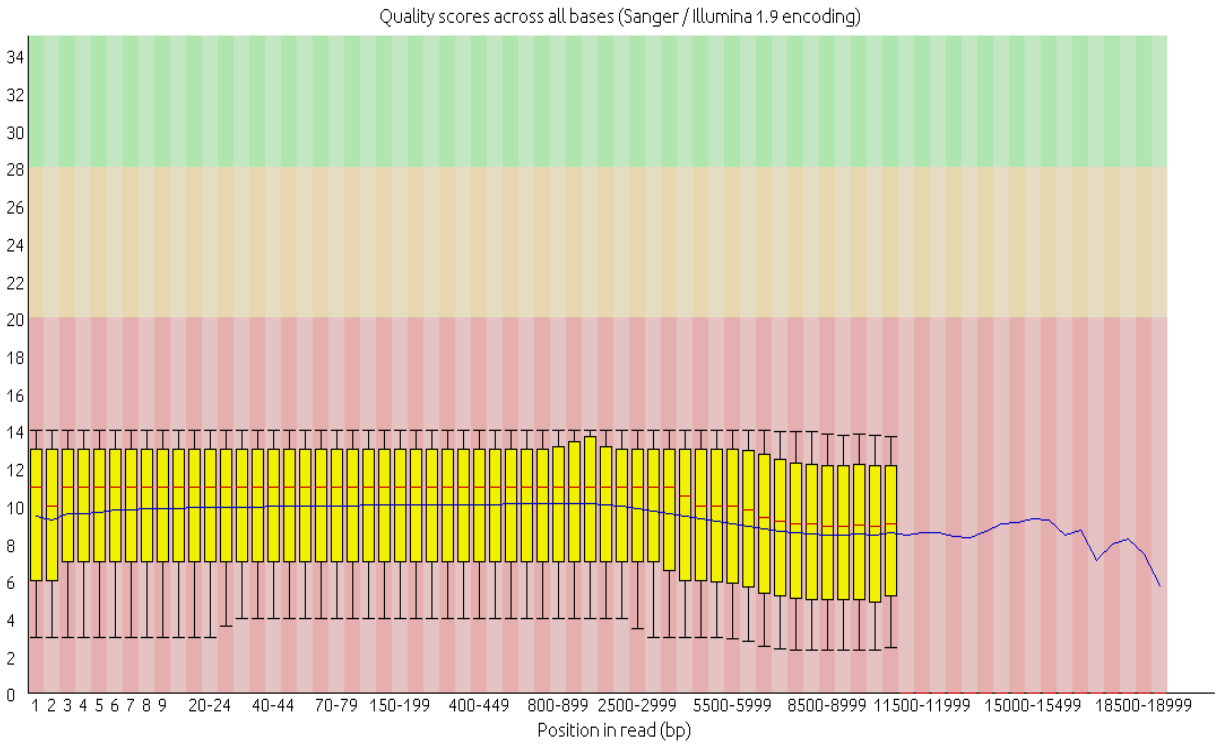

Figura 3. Valores de calidad de las lecturas long consensus generadas por el sistema PacBio de la cepa SQ.

En la figura 3 se observa que la media de los valores QS en las lecturas es de 10, esto significa que la probabilidad de error en el llamado de los nucleótidos es de 1 en 10 nucleótidos o el 90% de precisión en el llamado de los nucleótidos, también que el valor QS máximo es de 14 representado por el bigote superior. Con estos valores de QS obtenidos no es posible utilizar las lecturas para hacer un ensamble, ya que lo recomendable tener valores de QS superior a 30 en donde se tendría 1 error en cada 1000 nucleótidos o el 99.9% de precisión en el llamado de los nucleótidos. Debido a la baja calidad de las lecturas generadas con el sistema PacBio no se realizó el ensamble *de novo*.

Las lecturas circular consensus obtenidas del Sistema PacBio se analizaron con el programa FastQC para evaluar la calidad de las lecturas (QS), en la figura 4 se muestra la calidad de las lecturas del sistema de secuenciación PacBio de la cepa SQ, de igual forma que en la figura 3 podemos observar el rango de los valores de calidad de todas las bases en cada una de las posiciones en el archivo Fastq. La línea roja central es el valor de la mediana, la caja de color amarillo representa el rango intercuartil (25-75%), los bigotes superior e inferior representan los puntos 10% y 90% y la línea azul representa la calidad media.

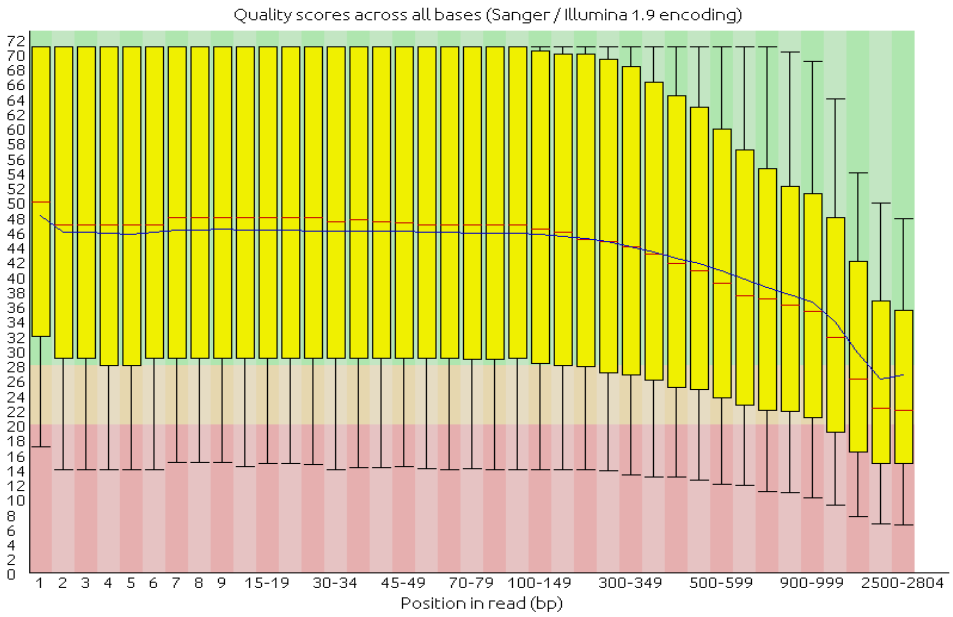

Figura 4. Valores de calidad de las lecturas circular consensus generadas por el sistema PacBio de la cepa SQ.

En la figura 4 se observa que la media de los valores QS en las lecturas es superior a 40, con esto se tiene una probabilidad de error en el llamado de los

nucleótidos menor de 1 en 10,000 nucleótidos o el 99.99% de precisión. Con estos valores de QS obtenidos es posible utilizar las lecturas para hacer un ensamble, ya que el QS es superior a 30, pero el número de lecturas circular consensus es de 74,481, con lo que se tiene una cobertura de 1.43x del genoma nuclear de *Dunaliella salina* y la cobertura mínima recomendable para llevar acabo un ensamble *de novo* es de 30x. Aun siendo de buena calidad las lecturas circular consensus, los resultados de un ensamble *de novo* no son válidos debido a que el número de lecturas es insuficiente.

Aun así, se llevaron a cabo ensambles *de novo* usando las lecturas de PacBio circular consensus, aunque no cubren los requisitos, esto para evaluar los resultados del ensamble, los resultados se muestran en la tabla 1.

Tabla 1. Estadísticos del ensamble *de novo* con secuencias circular consensus en el programa M.I.R.A usando las opciones Draft y accurate.

Ensamblaje	Draft	Accurate
# contigs	633	1084
Total consenso	1631994	2708110
Mayor contig	20956	42781
N50	3006	2884
N90	1496	1480
N95	1171	1167

En la tabla 1 vemos que el número de contigs es menor utilizando la opción draft que la opción accurate, mientras menos contigs es de mejor calidad el ensamble, pero al analizar el total consenso en draft se tiene 1,631,994 pb con el contig de mayor longitud de 20,956 y modo accurate 2,708,110 pb, con el contig de mayor longitud de 42,781. Los valores de longitud consenso total deben ser cercanos al tamaño estimado del genoma a ensamblar, que en este caso es de 300 Mb como lo reportó Smith et al 2010, por lo tanto, los resultados obtenidos en el ensamble están muy alejados de lo esperado.

La plataforma PacBio nos dio una baja profundidad de secuenciación, por tal motivo se decidió probar una plataforma de secuenciación que arroje un mayor número de lecturas por corrida, en este caso se seleccionó el sistema Illumina Hiseq 2500.

VI.1.2.2 Ensamble de novo con lecturas Illumina HiSeq 2500

Las lecturas generadas con el sistema Illumina HiSeq 2500 se analizaron con el programa FastQC para evaluar la calidad de las lecturas (QS), en la figura 5 se muestra el resultado del análisis de la cepa SQ, en la gráfica podemos observar el rango de los valores de calidad de todas las bases en cada una de las posiciones en el archivo Fastq. La línea roja central es el valor de la mediana, la caja de color amarillo representa el rango intercuartil (25-75%), los bigotes superior e inferior representan los puntos 10% y 90% y la línea azul representa la calidad media. La figura 6 muestra la distribución del valor QS en las

secuencias, y la figura 7 muestra la distribución de la longitud de las secuencias obtenidas por Illumina HiSeq 2500.

La secuenciación con el sistema Illumina Hiseq 2500 arrojó 400 millones de lecturas paired-end de 100 pb que corresponde a una cobertura de 133x, lo cual es suficiente para llevar a cabo el ensamble *de novo*. Enseguida se analizó la calidad de las lecturas.

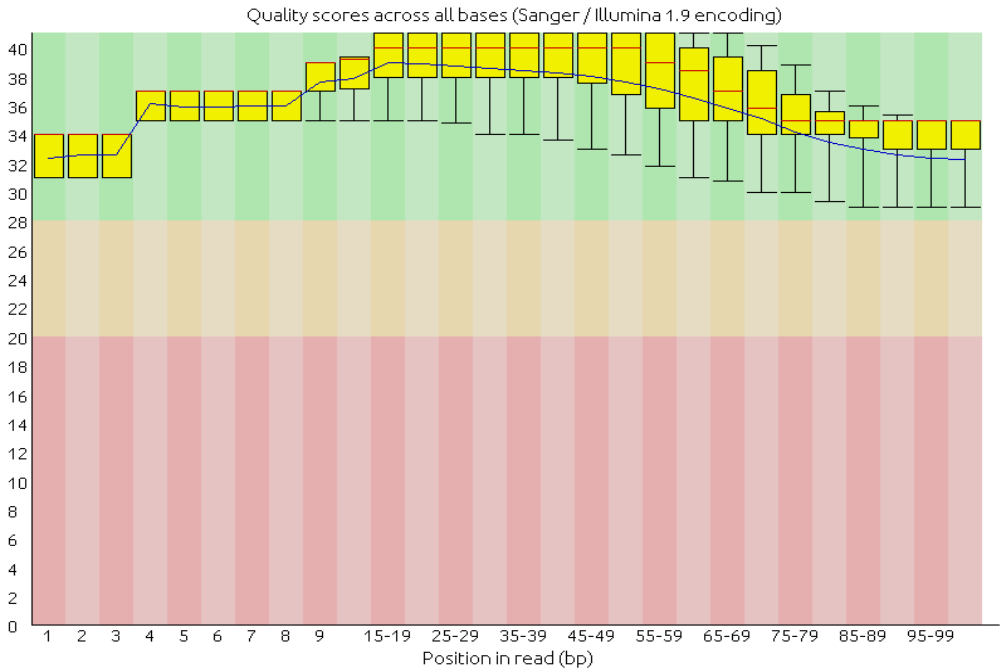

Figura 5. Valores de calidad de lecturas de 100pb generadas por el sistema Illumina HiSeq 2500 de la cepa SQ.

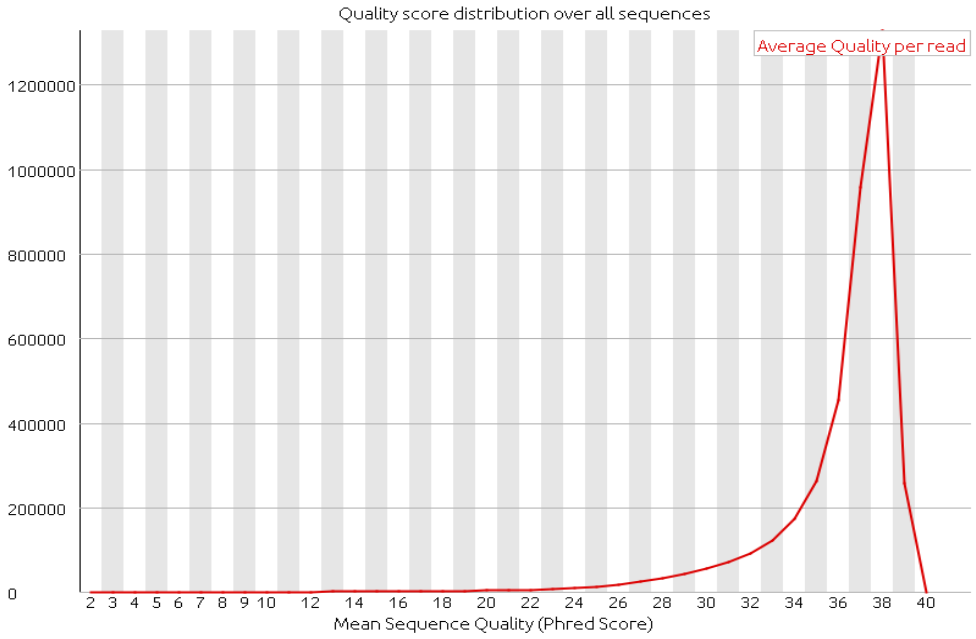

Figura 6. Distribución de los valores QS en las secuencias Illumina HiSeq 2500 de la cepa SQ.

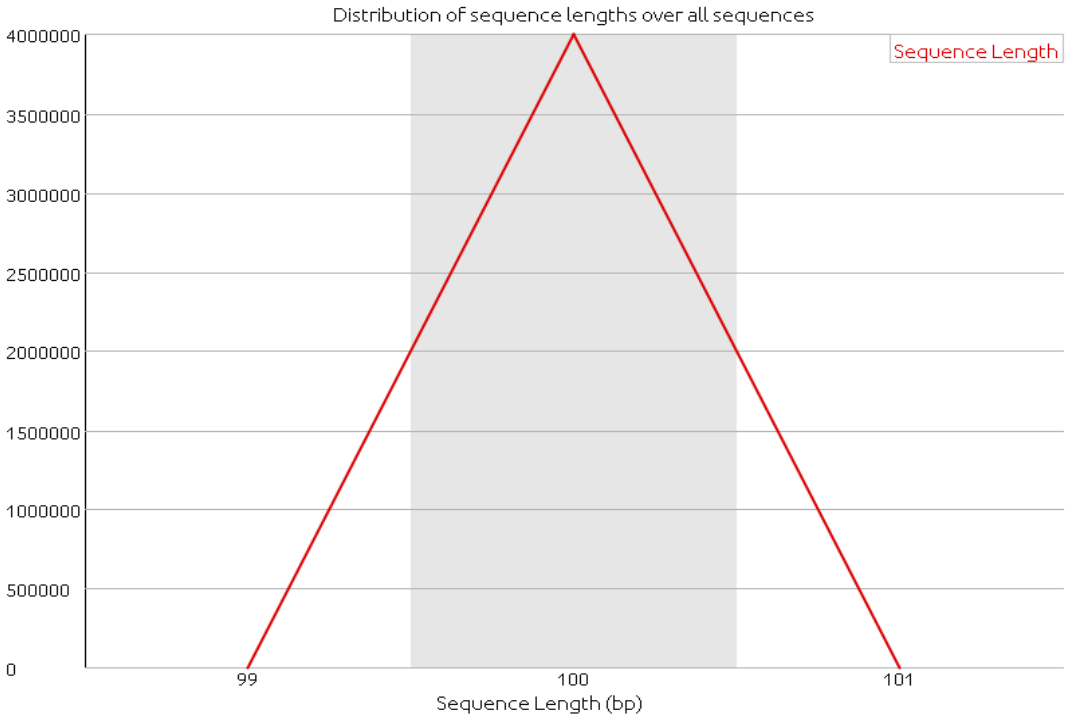

Figura 7. Distribución de la longitud en las secuencias Illumina HiSeq 2500 de la cepa SQ.

Las lecturas tienen una longitud de 100 pb y un QS superior a 30, se tienen 400 millones de lecturas paired-end de 100 pb con buena calidad para llevar a cabo el ensamble *de novo*.

En la figura 8 se muestra los resultados de las evaluaciones de los valores de K que se llevó a cabo con 400 millones de lecturas de Illumina Hiseq 2500. Se evaluaron los valores 21, 31, 41, 51,61, 71, 81 y 91.

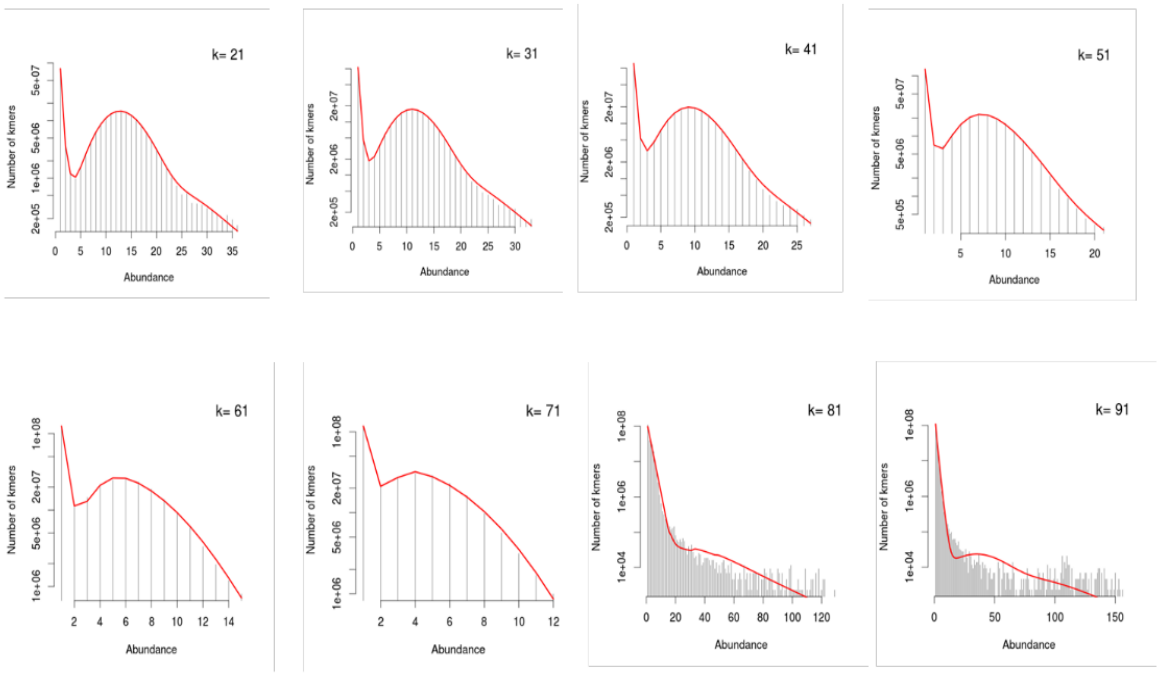

Figura 8. Histograma de abundancia de valores de k = 21, 31, 41, 51, 61, 71, 81 y 91.

El valor de K que presenta una mayor diversidad de k-mers es k= 51 con 23,556,690 k-mers, 51 fue el valor que se utilizó para hacer el ensamble *de novo* con las lecturas de Illumina HiSeq 2500.

Tabla 2. Estadísticos del ensamble *de novo* con secuencias Illumina, 3 ensambles con coberturas de 25x, 37x y 125x.

Ensamblaje	25x	37.5x	125x
# de contigs	80,140	64,483	61,744
Longitud total	90,804,110	118,879,431	128,525,703
Contig mas largo	12,825	18,119	22,632
% de GC	49.93	49.96	50
N50	1271	2507	2931

Los resultados del ensamble *de novo* muestran que el número de contigs disminuye cuando aumenta la cobertura, en donde al utilizar la cobertura de 125x se tienen 61,744 contigs mientras que con 25x de cobertura se obtienen 80,140 contigs. Al analizar la longitud total de todo el conjunto de contigs juntos, el ensamble con cobertura 125x es que tiene la mayor longitud con 128,525,703 pb, también contiene el contig más largo con 22,632 pb y la N50 más grande con 2931 pb.

De acuerdo a los resultados obtenidos el mejor ensamble fue en donde se utilizó la cobertura de 125x, y al comparar la longitud total obtenida con la esperada (Obtenida = 128,525,703 pb. Esperada= 300,000,000 pb) solo se alcanza un poco menos de la mitad de la longitud esperada. Por esta razón se realizaron otros ensambles con 40 millones de lecturas, 60 millones, 200 millones y 400 millones de lecturas, que corresponden a una cobertura de 13x, 20x, 66x y 133x respectivamente. Los ensambles se hicieron también con el programa minia y con 51 como valor de k. Los resultados se muestran en las tablas 3, 4, 5 y 6.

Tabla 3. Estadísticos del ensamble *de novo* con 40 millones de secuencias Illumina.

Ensamble	40 millones
Longitud total (>= 0 bp)	126792859
Longitud total (>= 1000 bp)	59116533
# contigs	80140
Longitud total	90804110
Contig con mayor longitud	12825
GC (%)	49.93
N50	1271
# N's por 100 kbp	0

Tabla 4. Estadísticos del ensamble *de novo* con 60 millones de secuencias Illumina.

Ensamble	60 millones
Longitud total (>= 0 bp)	139550357
Longitud total (>= 1000 bp)	102364317
# contigs	64483
Longitud total	118879431
Contig con mayor longitud	18119
GC (%)	49.96
N50	2507
# N's por 100 kbp	0

Tabla 5. Estadísticos del ensamble *de novo* con 200 millones de secuencias Illumina.

Ensamble	200 millones
Longitud total (>= 0 bp)	147597077
Longitud total (>= 1000 bp)	114180493
# contigs	61774
Longitud total	128525763
Contig con mayor longitud	22632
GC (%)	50.01
N50	2935
# N's por 100 kbp	0

Tabla 6. Estadísticos del ensamble *de novo* con 400 millones de secuencias Illumina.

Ensamble	400 millones
Longitud total (>= 0 bp)	140036209
Longitud total (>= 1000 bp)	105114440
# contigs	61191
Longitud total	120137651
Contig con mayor longitud	19387
GC (%)	49.98
N50	2732
# N's por 100 kbp	0

Los estadísticos de los 4 ensambles de las tablas 3, 4, 5 y 6 muestran que el mejor ensamble fue donde se usó 200 millones de lecturas, que corresponde a una cobertura de 66x. De igual manera se sigue obteniendo el valor de longitud total (147,597,077 pb) muy por debajo al valor esperado (300,000,000) del tamaño del genoma de *Dunaliella salina*.

Existen diferentes razones del porque no se logran obtener ensambles con valores cercanos al tamaño del genoma reportado de 300 millones de pb por Smith et al 2010. Como por ejemplo no tener una profundidad de secuenciación suficiente o cobertura, que se utilice el programa ensamblador incorrecto, que el genoma contenga regiones invertidas, un alto número de regiones repetitivas, una estimación incorrecta del tamaño del genoma o que tanto el tamaño como la

arquitectura del genoma dentro de las distintas cepas de *Dunaliella salina* sean diferentes.

Con referencia a la profundidad de secuenciación, se trabajó con 400 millones de secuencias que corresponde a tener secuenciado el genoma 166 veces (cobertura = 166x), lo mínimo recomendable es 15x (Bentley *et al*., 2008). Se utilizaron distintos números de lecturas, el mejor ensamble que fue el de 66x el cual esta solo a la mitad de la longitud de lo esperado que es 300 Mpb. Se utilizaron distintos ensambladores, sin embargo, los resultados reportados en este trabajo fueron los del ensamblador Minia que fue el programa que arrojó los mejores ensambles.

Smith et al 2010 junto con el DOE JGI (Department of Energy Joint Genome Institute) están trabajando en la secuenciación, ensamble y anotación del genoma nuclear de una cepa de *Dunaliella salina* CCAP 19/18, el trabajo comenzó desde el año 2006 y a la fecha no se ha logrado resolver la arquitectura del genoma, estos investigadores reportan una gran cantidad de secuencias repetitivas e invertidas las cuales complican los ensambles. En este sentido es posible que la cepa *Dunaliella salina* SQ contenga secuencias invertidas y repetitivas como la cepa de *Dunaliella salina* CCAP 19/18. Como no se logró completar el objetivo número 4, se prosiguió a ensamblar el genoma mitocondrial de la cepa *Dunaliella salina* SQ, también para este momento ya se contaba con otra cepa de *Dunaliella salina* que se aisló de las lagunas salinas de Guerrero Negro Baja California, a esta cepa se le denominó *Dunaliella salina* GN, también se ensambló el genoma mitocondrial de la cepa GN.

VI.2 Secuenciación de los genomas mitocondriales de *Dunaliella salina* cepas SQ y GN.

VI.2.1 Extracción DNA mitocondrial

Se obtuvo una concentración de 76 ng/µl con un radio 260/280 de 1.85 para el cultivo de *Dunaliella salina* SQ en un volumen final de 80 µl, en el caso de la cepa *Dunaliella salina* GN se obtuvo una concentración de 264 ng/µl con un radio 260/280 de 1.85 en un volumen final de 80 µl. Las mediciones se hicieron en un nanodrop y la integridad del DNA se evaluó haciendo una electroforesis en un gel de agarosa al 1% cargando 5µl de la muestra SQ y la muestra GN. En la figura 9 se muestra el resultado de la electroforesis en gel de agarosa.

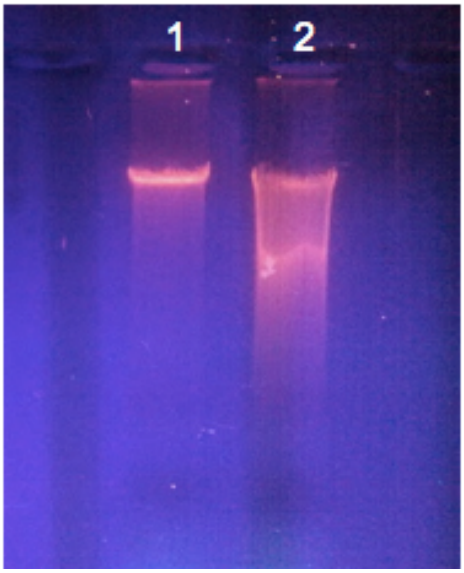

Figura 9. Electroforesis en gel de agarosa de extracción de DNA mitocondrial. **1)** DNA mitocondrial de *Dunaliella salina* SQ. **2)** DNA mitocondrial de *Dunaliella salina* GN.

En la figura 9 se muestra la integridad del DNA mitocondrial de la cepa SQ en el carril 1 y en el carril 2 de la cepa GN. En el caso del DNA de la cepa GN se observa degradación de la muestra, cuando se analizó en UCSD comentaron que la muestra se encontraba en condiciones adecuadas para realizar la secuenciación. 5 µg de DNA de cada cepa se enviaron a secuenciar a UC San Diego Core Center para la secuenciación con la plataforma Illumina MiSeq. Esta secuenciación arrojó 11,049,212 lecturas para la SQ y 15,562,756 lecturas para la cepa GN, lo cual corresponde a 117,001x y 164,795x respectivamente, esto si se asume que el tamaño del genoma mitocondrial de las cepas SQ y GN es similar al genoma mitocondrial de *Dunaliella salina* reportado por Smith en 2010 (Clave de acceso en NCBI NC_012930.1).

VI.2.2 Ensamble por referencia del genoma mitocondrial de *Dunaliella salina* SQ y *Dunaliella salina* GN.

La figura 10 y 11 muestran el resultado del análisis de calidad de las lecturas generadas de la secuenciación con MiSeq de Illumina de la cepa SQ y GN respectivamente, en la gráfica podemos observar el rango de los valores de calidad de todas las bases en cada una de las posiciones en el archivo Fastq. La línea roja central representa el valor de la mediana, la caja de color amarillo representa el rango intercuartil (25-75%), los bigotes superior e inferior representan los puntos 10% y 90% y la línea azul representa la calidad media.

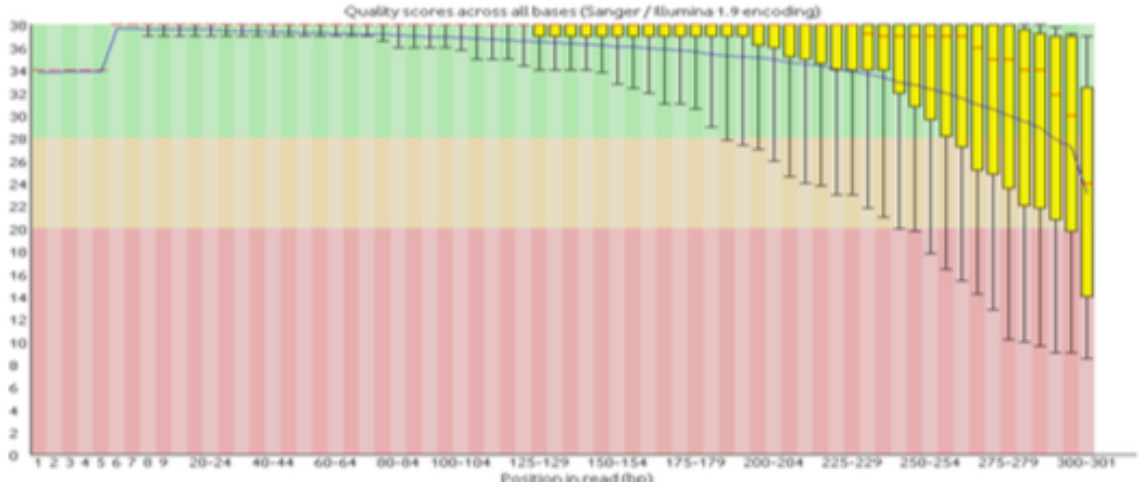

Figura 10. Valores de calidad de lecturas de 300 pb generadas por el sistema Illumina MiSeq del DNA mitocondrial de la cepa SQ.

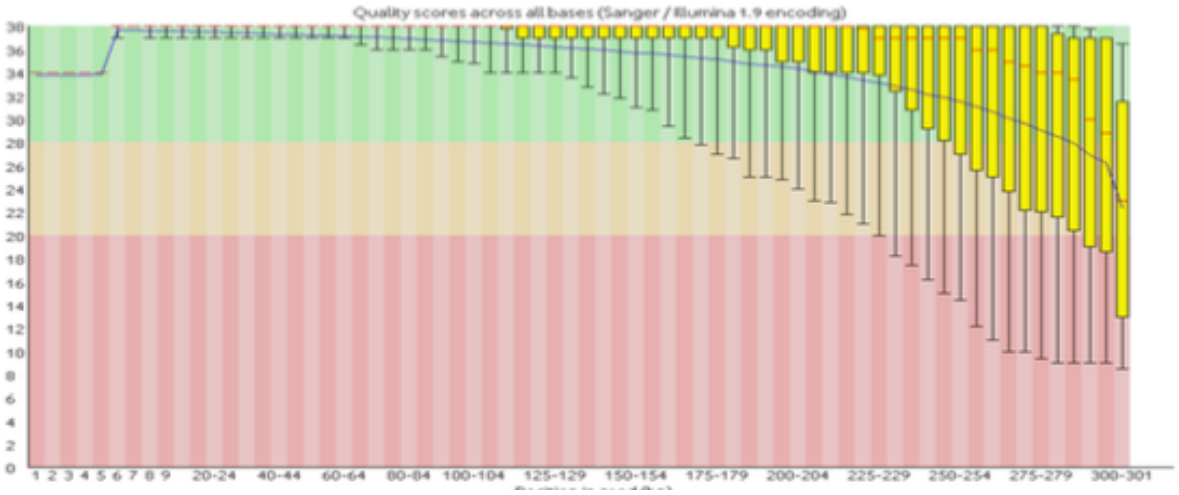

Figura 11. Valores de calidad de lecturas de 300 pb generadas por el sistema Illumina MiSeq del DNA mitocondrial de la cepa GN.

Las lecturas tienen una longitud de 300 pb y una media QS superior a 22, se tienen 11,049,212 de lecturas paired-end de la cepa SQ y 15,562,756 lecturas paired-end de la cepa GN. Acercándose al final del extremo 3’ de las lecturas se

observa que los valores de QS disminuyen, esto se corrigió al fusionar las lecturas forward y reverse, ya que a las lecturas reverse se les aplica reverse and compose y el extremo 5’ de las lecturas reverse es parte del extremo 3’de las lecturas forward.

VI.2.3 Resultado ensamblaje por referencia de la mitocondria de la cepa SQ y GN

En la figura 12 se muestra el resultado gráfico del ensamblaje por referencia de las lecturas de la cepa SQ, estas lecturas se alinearon con el genoma de la mitocondria referencia NC_012930.1. En la parte superior de la imagen, podemos observar regiones con alta cobertura (color azul), a lo largo de los 28,331 nucleótidos del genoma de referencia se presentaron gaps lo cual indica que las lecturas de la cepa SQ no alinearon contra el genoma de referencia, en este caso las zonas sin alineamiento corresponden aproximadamente a la mitad del genoma de referencia, esto no era de esperarse ya que el genoma de referencia es de la mitocondria de la misma especie (*Dunaliella salina*), pero de la cepa CCAP 19/18. Para el caso de la cepa GN, se obtuvo una cobertura mayor del genoma de referencia, pero se siguieron presentando gaps de gran longitud. Algo que podemos destacar con estos resultados es que el genoma mitocondrial de la cepa GN tiene mayor similitud al genoma mitocondrial de la cepa CCAP 19/18 que la cepa SQ. Esto debido a que el ensamble por referencia de la cepa SQ tiene un mayor número de gaps.

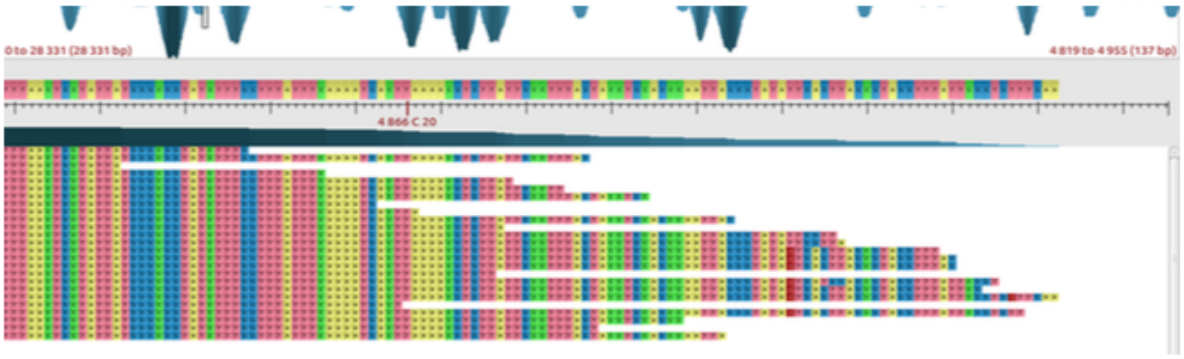

Figura 12. Ensamble por referencia usando el genoma de la mitocondria de *Dunaliella salina* NC_012930.1 como referencia y lecturas MiSeq de la cepa SQ.

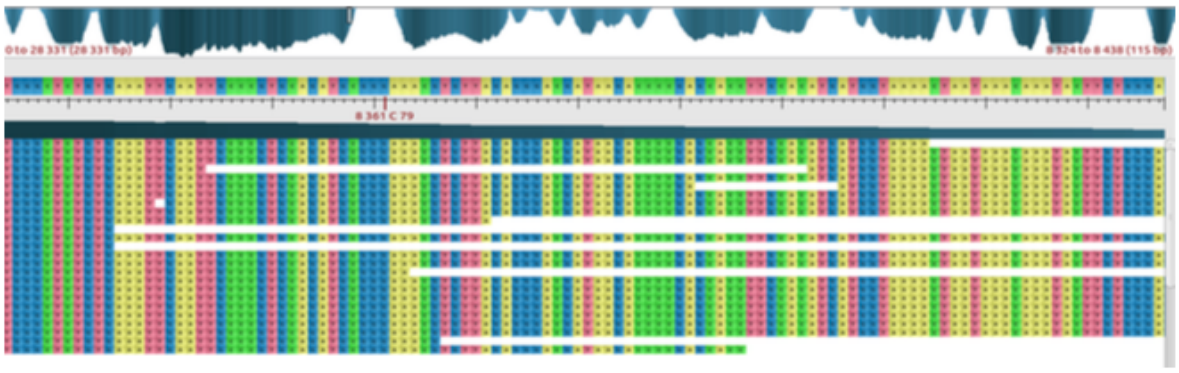

Figura 13. Ensamble por referencia usando el genoma de la mitocondria de *Dunaliella salina* NC_012930.1 como referencia y lecturas MiSeq de la cepa GN.

La razón por la cual se decidió llevar a cabo los ensamblajes por referencia fue con la premisa de que genomas de orgánulos como la mitocondria y el cloroplasto de organismos de la misma especie no presentan grandes variaciones en la arquitectura genómica, solo polimorfismo de nucleótidos

sencillo (SNP's). Por lo tanto, los resultados obtenidos pueden ser causados por tres motivos:

1. Que las bibliotecas generadas previo a la secuenciación tuvieran un sesgo y las lecturas obtenidas no provengan de una biblioteca aleatoria.
2. Que el método de ensamble o software utilizado no sea el adecuado.
3. Que la arquitectura de los genomas sea diferente sin importar que los orgánulos provengan de la misma especie.

Para el caso 1 se llevó a cabo un análisis de secuencias duplicadas con el programa FastQC, si una biblioteca es diversa se espera que la mayoría de las secuencias ocurran solo una vez en el conjunto final de datos, si se presenta un nivel bajo de duplicación es indicativo de que existe un alto nivel de cobertura de la secuencia objetivo (lo que se está secuenciando), pero si se encuentra un alto nivel de duplicación es indicativo de sesgo en la generación de las bibliotecas. En las figuras 14 y 15 se muestran el grado de duplicación de cada secuencia en la biblioteca de la cepa SQ y GN.

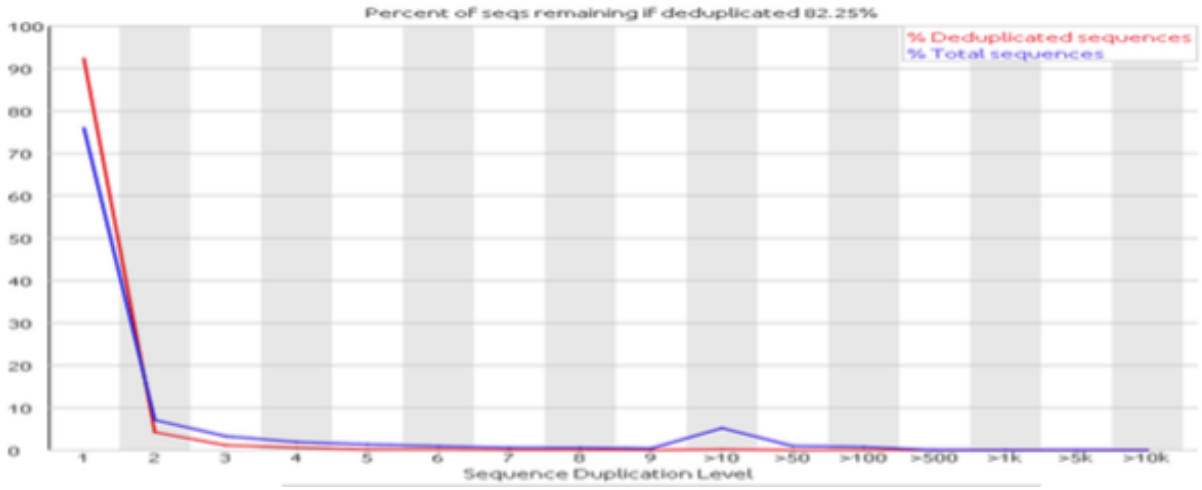

Figura 14. Grado de duplicación de las secuencias de la biblioteca de la cepa SQ.

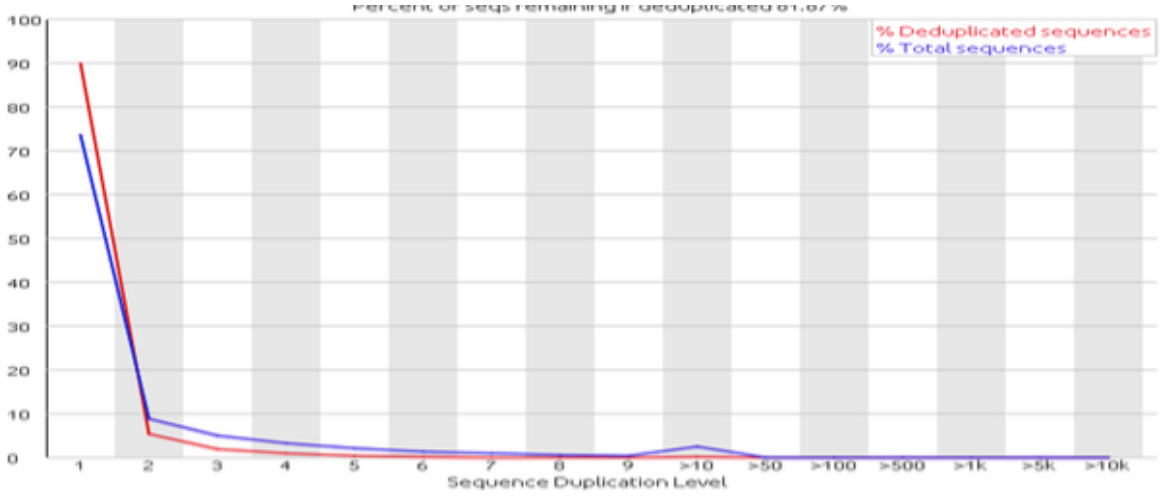

Figura 15. Grado de duplicación de las secuencias de la biblioteca de la cepa GN.

Como se observa en las gráficas 14 y 15, las bibliotecas de las cepas SQ y GN muestran un bajo nivel de duplicación, por lo tanto, se concluye que los ensambles incompletos no son ocasionados por sesgo en la generación de la biblioteca. Con esto se descarta que la extracción de DNA realizada en el laboratorio Meredith Gould o la preparación de las bibliotecas para la secuenciación que se realizó en el IGM genomics center de UCSD se procesara de forma incorrecta. Para el caso 2 se utilizaron los programas BWA y MIRA 4.1 para realizar los ensambles, los resultados fueron los mismos que en el caso de Bowtie2, concluyendo que no es problema del software utilizado la obtención de ensambles incompletos.

Se encontró un método para la reconstrucción de genomas mitocondriales de organismos no modelos que utiliza secuencias de Illumina MiSeq, el programa se llama MITObim (Bachmann, 2013), este programa al igual que Bowtie2 y BWA realiza el ensamble por referencia, pero los segmentos en donde no alienan secuencias en el genoma de referencia, los aísla y sigue buscando en las lecturas que restan de la secuenciación. A este método el autor lo llama *a baiting and iterative mapping approach*. En la figura 16 se muestra el resultado obtenido con este método.

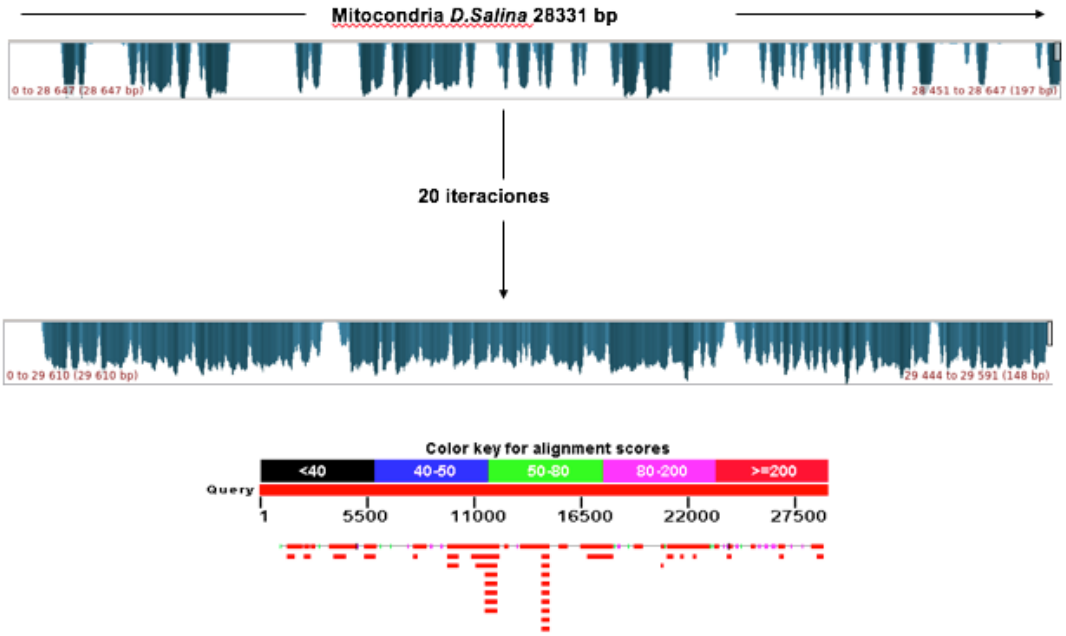

Figura 16. Ensamble por referencia usando el programa MITObim con genoma de la mitocondria de *Dunaliella salina* NC_012930.1 como referencia y lecturas MiSeq de la cepa SQ.

En la figura 16 se muestra el resultado gráfico del ensamblaje por referencia de las lecturas de la cepa SQ usando el programa MITObim, estas lecturas se alinearon con el genoma de la mitocondria referencia NC_012930.1. En la parte superior de la imagen, podemos observar regiones con alta cobertura (color azul), a lo largo de los 28331 nucleótidos del genoma de referencia se presentaron gaps lo cual indica que las lecturas de la cepa SQ no alinearon contra el genoma de referencia. En la imagen media se muestra el alineamiento después de 20 iteraciones, aquí se ve claramente cómo se rellenaron los gaps que se tenían antes de las 20 iteraciones obteniendo la cobertura casi total del genoma de referencia. Con este nuevo ensamble se realizó un blast para corroborar que

correspondía a un genoma mitocondrial, en la imagen inferior de la figura 16 se observa el resultado del alineamiento. Se observa que el alineamiento no es continuo, y se fragmenta el ensamble cuando se alinea con el genoma de referencia. Los gaps que se rellenaron con las iteraciones o son nuevas secuencias en el genoma mitocondrial o el programa realizó alineamientos incorrectos.

Smith et al 2010 no explica que plataforma de secuenciación utilizó en su trabajo, este trabajo utilizó Illumina, es conocido que Illumina tiene problemas para secuenciar zonas del DNA que contienen porcentajes altos de Guaninas y Citosinas (GC), por esta razón se decidió utilizar el lenguaje de programación R con las paqueterías Seqinr y Bioconductor para analizar por segmentos de 5000 pb el porcentaje de GC del genoma de referencia de la cepa CCAP 19/18 y realizar una gráfica de los porcentajes de GC. Esto con el objetivo de explicar porque no se pueden cerrar los gaps cuando se realizan los ensambles por referencia, si las zonas gap contienen un porcentaje de GC alto podría ser que el sistema MiSeq no logró secuenciar estas zonas. En la figura 17 se muestra el resultado del análisis.

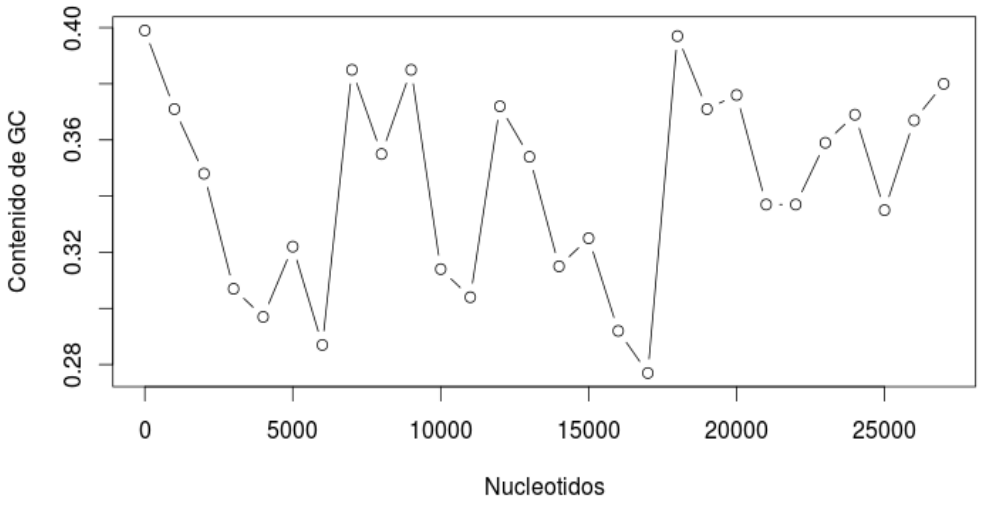

Figura 17. Porcentaje de GC en segmentos de 5000 pb del genoma completo de la mitocondria de *Dunaliella salina* cepa CCAP 19/18.

Al comparar las zonas con porcentajes altos de GC con los gaps del alineamiento por referencia, algunos gaps coinciden con el alto porcentaje de GC, pero no es algo constante. Por lo tanto, no se puede concluir que la razón de no tener completo el genoma mitocondrial de la cepa SQ sea porque la plataforma Illumina tuvo problemas para secuenciar el genoma.

De acuerdo a los resultados obtenidos se decidió hacer ensambles *de novo* con las lecturas Illumina Miseq y buscar contigs de mitocondria en el ensamble resultante, esto para intentar estructurar los genomas mitocondriales de las cepas SQ y GN y evaluar si existe diferente arquitectura genómica entre cepas de la misma especie de *Dunaliella salina*.

VI.2.4 Ensamble de novo de los genomas mitocondriales de *Dunaliella salina* SQ y GN.

Los ensambles se llevaron a cabo con el pipeline A5-miseq. Los resultados de los ensambles *de novo* se muestran en las tablas 7 y 8.

Tabla 7. Estadísticos del ensamble *de novo* con 11 millones de secuencias de Illumina MiSeq cepa SQ.

Ensamble *de novo* cepa SQ	11 millones de lecturas
Contigs (>= 1000 bp)	4,602
Longitud total (>= 1000 bp)	6,823,422
# contigs	15,624
Longitud total	15,393,537
Contig con mayor longitud	125,298
GC (%)	47.22
N50	949
# N's por 100 kbp	0

Tabla 8. Estadísticos del ensamble *de novo* con 11 millones de secuencias de Illumina MiSeq cepa GN.

Ensamble *de novo* cepa GN	15 millones de lecturas
Contigs (>= 1000 bp)	1,907
Longitud total (>= 1000 bp)	3,762,779
# contigs	6,109
Longitud total	6,991,323
Contig con mayor longitud	183,356
GC (%)	46.82
N50	1,055
# N's por 100 kbp	0

Los contigs obtenidos del ensamble *de novo* para la cepa SQ en formato fasta se ordenaron de mayor a menor longitud, a los 10 contigs de mayor longitud se realizó un BLAST y alineamientos con genes de mitocondriales de la *Dunaliella salina* CCAP 19/18. El mismo procedimiento se llevó a cabo para el ensamble de la cepa GN.

El resultado del BLAST de la cepa SQ, los primeros 3 contigs de longitud de 125,298 pb (contig 1), 101,617 pb (contig 2) y 41,904 pb (contig 3) dieron positivo para *Dunaliella salina*, pero lo interesante fue que los contigs 1 y 2 alinearon con la secuencia del genoma del cloroplasto de *Dunaliella salina* cepa CCAP 19/18, con que podemos concluir que, en el procedimiento del aislamiento de mitocondrias, también se aislaron cloroplastos. El contig número 3 de longitud 41,904 pb dio positivo para genoma de mitocondria. Para el caso de la cepa GN, el resultado del BLAST fue positivo para *Dunaliella salina* los primeros dos contigs de longitud 183,356 pb (contig 1) y 27,950 pb (contig 2), el contig 1 alineo con genoma de cloroplasto y el contig 2 con genoma de mitocondria. Con estos resultados se prosiguió a anotar y estructurar los genomas mitocondriales.

VI.3. Anotación de los genomas mitocondriales de *Dunaliella salina* SQ y GN.

A los contigs que dieron positivo para genoma de mitocondria de la cepa SQ y GN, se les realizó la búsqueda de genes que se encuentran presentes en los genomas mitocondriales de microalgas del orden chlamydomonales, los reportados son 12 (Smith *et al*., 2010). El resultado fue que el contig 3 de la cepa SQ y el contig 2 de la cepa GN contienen los genes esperados, 7 genes codificantes para proteínas, 3 genes de tRNA y 2 de genes de rRNA. Ya con estos resultados se continuó con la anotación para terminar de estructurar los genomas y los resultados se muestran en las figuras 18 y 19.

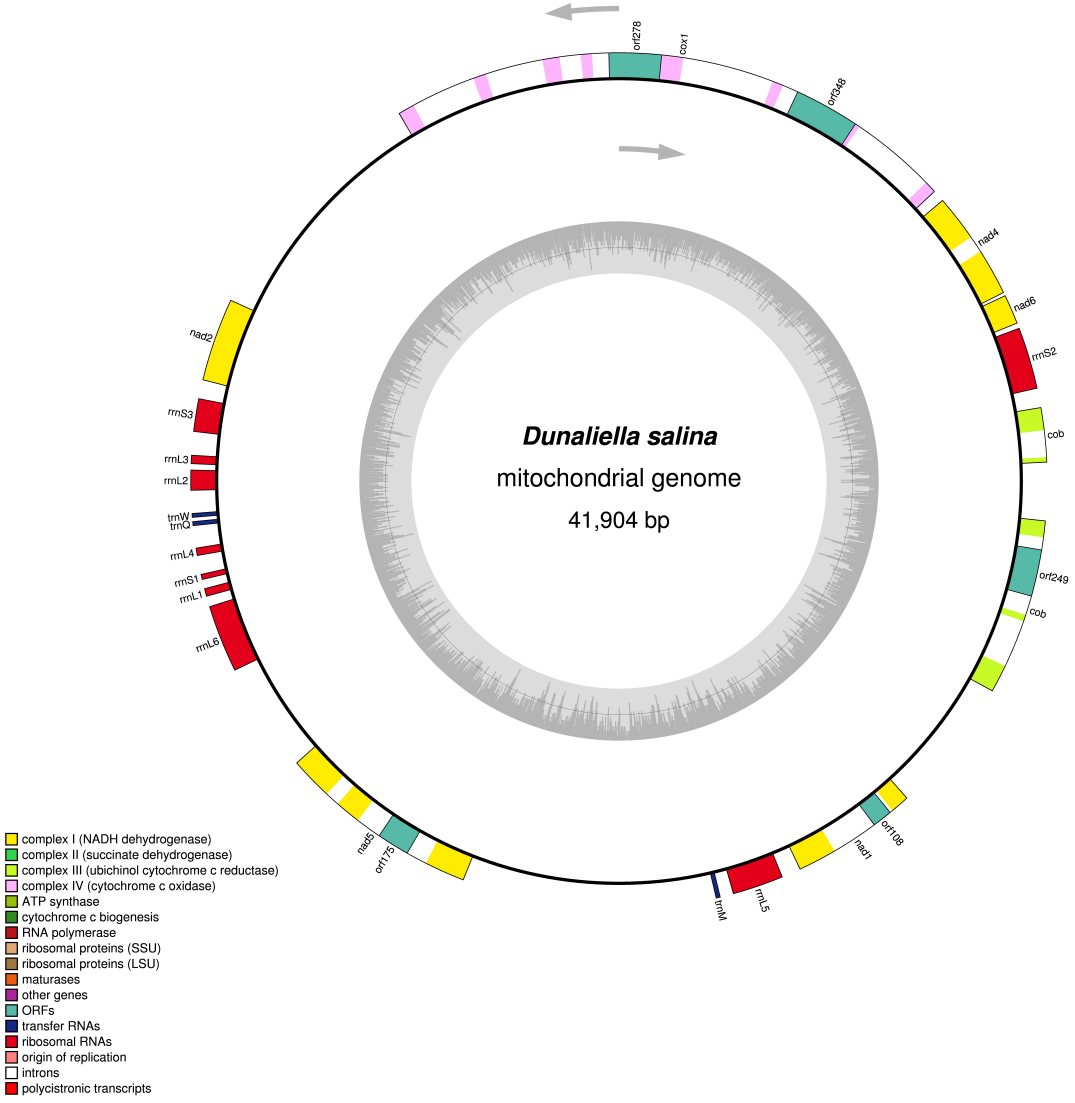

Figura 18. Arquitectura del genoma mitocondrial de *Dunaliella salina* SQ.

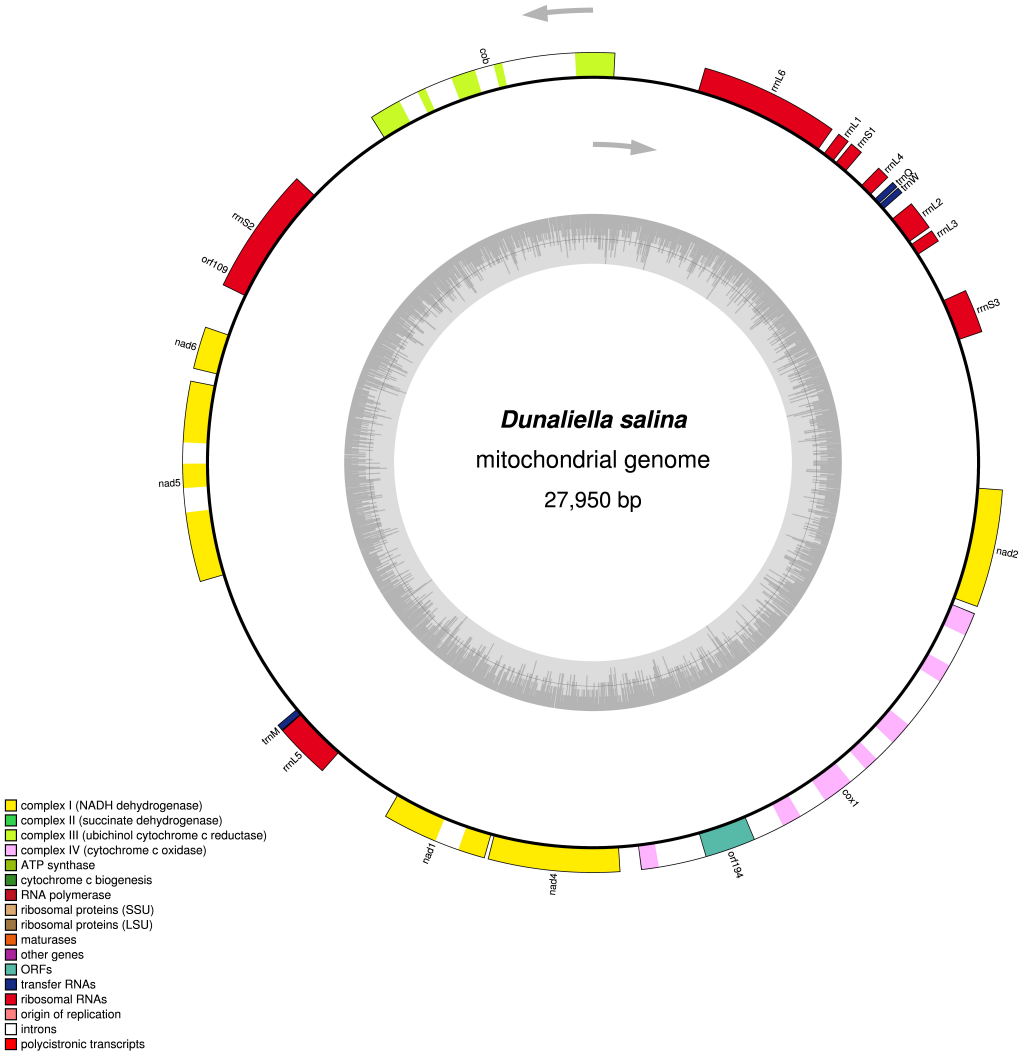

Figura 19. Arquitectura del genoma mitocondrial de *Dunaliella salina* GN.

El genoma mitocondrial de *Dunaliella salina* SQ muestra una estructura circular de 41,904 pb, contiene 7 genes codificantes para proteínas, 15 intrones, 2 genes de RNA ribosomal y 3 genes de RNA de transferencia. Los contenidos de A, G, T y C son 32.6%, 16.7%, 35.5% y 15.2% respectivamente, con un 31.85% de contenido GC. Todos los genes codificantes para proteínas comienzan con el codón ATG y terminan con el codón TAA, excepto el gen *nad6* que termina con el codón TAG. El intrón del gen *nad1* contiene un iorf (Intronic Open Reading Frame) GIY-YIG homing-endonucleasa, *cob* presenta un iorf LAGLIDADG homing-endonucleasa, *cox1* dos iorf LAGLIDADG homing-endonucleasa y *nad5* un iorf GIY-YIG homing-endonucleasa. En tabla 9 se muestran los nombres de los genes y el número de intrones que contienen.

Tabla 9. Genes mitocondriales de *Dunaliella salina* SQ.

Tipo de gen	Nombre del gen	# intrones
Codifica proteína	nad 6	0
Codifica proteína	nad 4	1
Codifica proteína	cox 1	7
Codifica proteína	cob	4
Codifica proteína	nad 2	0
Codifica proteína	nad 5	2
Codifica proteína	nad 1	1
RNA transferencia	trnM	0
RNA transferencia	trnW	0
RNA transferencia	trnQ	0
RNA ribosomal	rrnS	0
RNA ribosomal	rrnL	0

El genoma mitocondrial de *Dunaliella salina* GN muestra una estructura circular de 27,950 pb, contiene 7 genes codificantes para proteínas, 13 intrones, 2 genes de RNA ribosomal y 3 genes de RNA de transferencia. Los contenidos de A, G, T y C son 31.3%, 17.8%, 35.5% y 16.0% respectivamente, con un 33.76% de contenido GC. Todos los genes codificantes para proteínas comienzan con el codón ATG y terminan con el codón TAA, excepto el gen *nad6* que termina con el codón TAG. Un intrón del gen *cox1* contiene un iorf LAGLIDADG homing-endonucleasa y *nad5* un iorf GIY-YIG homing-endonucleasa. En tabla 10 se muestran los nombres de los genes y el número de intrones que contienen.

Tabla 10. Genes mitocondriales de *Dunaliella salina* GN.

Tipo de gen	Nombre del gen	# intrones
Codifica proteína	*nad* 6	0
Codifica proteína	*nad* 4	0
Codifica proteína	*cox* 1	6
Codifica proteína	*cob*	4
Codifica proteína	*nad* 2	0
Codifica proteína	*nad* 5	2
Codifica proteína	*nad* 1	1
RNA transferencia	*trnM*	0
RNA transferencia	*trnW*	0
RNA transferencia	*trnQ*	0
RNA ribosomal	*rrnS*	0
RNA ribosomal	*rrnL*	0

Los genomas mitocondriales ensamblados y anotados de las cepas SQ y GN se encuentran en la base de datos de NCBI con los números de acceso KX641169 y KX641170 respectivamente.

Comparando las tablas 9 y 10 se observa que los genomas mitocondriales de las cepas de *Dunaliella salina* SQ y GN contienen los mismos genes, pero varían en el número de intrones, también al analizar las figuras 19 y 18 vemos que los genes de RNA ribosomal están fragmentados y dispersos por todo el genoma. Los 12 genes se encuentran codificados en la misma hebra, esto no es único para *Dunaliella salina*, se ha visto en otras algas verdes del orden chlamydomonales (Smith *et al*., 2010).

Algo importante a destacar es el tamaño de los genomas, si se comparan el tamaño de los genomas mitocondriales de *Dunaliella salina* GN (tamaño genoma = 27,950 pb) con *Dunaliella salina* CCAP 19/18 (tamaño genoma = 28,300 pb) la cepa CCAP 19/18 es 1.2% más grande que GN, pero al comparar estos genomas con el genoma mitocondrial de *Dunaliella salina* SQ (tamaño genoma = 41,904 pb), el genoma de la mitocondria de la cepa SQ es 32.4% más grande que CCAP 19/18 y 33.2% más grande que la cepa GN.

Al momento de realizar el ensamble por referencia del genoma mitocondrial de la cepa SQ, el genoma de referencia que se utilizó fue el de la cepa CCAP 19/18, los resultados que se muestran en la figura 12 donde el ensamble presenta una gran cantidad de gaps, es debido a la diferencia de los tamaños de los genomas y las diferencias en secuencias no codificantes, ya que los dos genomas contienen los mismos genes. Para el caso del ensamble por

referencia de la cepa GN (Figura 13) se obtuvo un ensamble con un número menor de gaps que en el caso anterior, esto resultado fue debido a que el genoma de referencia como el genoma a ensamblar presentan un tamaño similar (350 pb más grande el de la cepa CCAP 19/18), y los gaps son el resultado la diferencias en secuencias no codificantes. Del Vasto *et a*l., 2015 ensambló y anotó el genoma mitocondrial de *Dunaliella salina* CCM-UDEC 001 que se aisló del desierto de Atacama chile, el genoma mitocondrial esta cepa contiene los mismos genes que las cepas ya aquí mencionadas, pero al igual que la cepa SQ, presenta un tamaño grande de 47,397 pb.

En este sentido es importante destacar que los genomas mitocondriales de las cuatro cepas de *Dunaliella salina* reportadas a la fecha, contienen los 12 genes presentes en las algas verdes del orden Chlamydomonales, la misma distribución de los genes dentro del genoma de las cepas, pero se puede encontrar variaciones en los tamaños de casi un 50% entre cepas, las variaciones se presentan en regiones no codificantes, las cuales no están expuestas a selección natural.

VI.4 Análisis de variación genética

Para tener una mayor certidumbre que los genomas de las cepas SQ y GN pertenecen a la especie *salina* del género *Dunaliella*, se realizaron análisis filogenéticos moleculares, utilizando como base la secuencia de nucleótidos del

gen *cox1* de las cepas SQ y GN, la secuencia del gen *cox1* de 10 organismos del orden Chlamydomonales incluyendo las cepas de *Dunaliella salina* CCAP 19/18 y CCM-UDEC 001.

El alineamiento se realizó en el programa MEGA 7 utilizando el método de análisis Maximun Likelihood, el árbol filogenético se muestra en la figura 20.

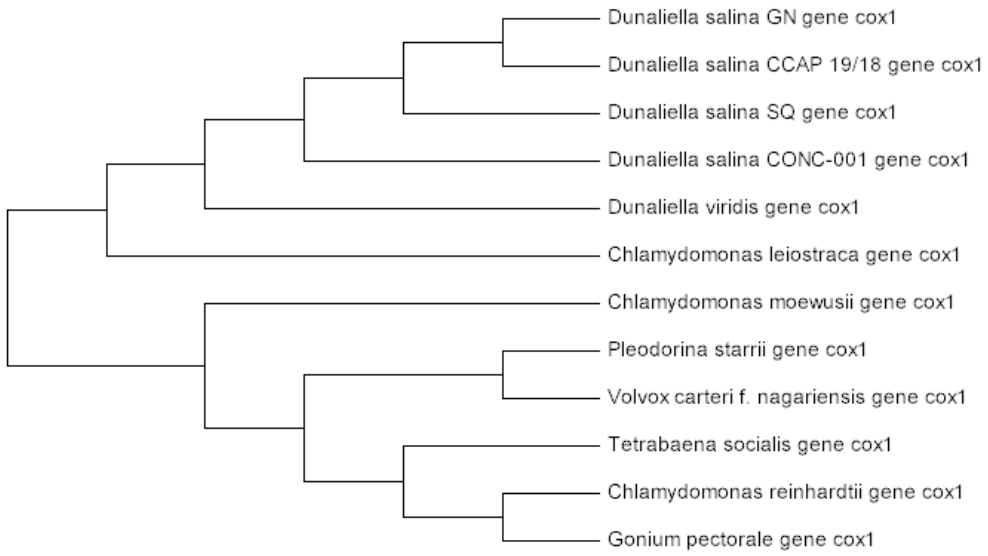

Figura 20. Análisis filogenético molecular por el método Maximum Likelihood en MEGA7. Genes cox 1 de 10 organismos del orden Chlamydomonales y genes cox1 de las cepas SQ y GN.

Con el resultado del árbol filogenético de los genes cox1, se observa que las cepas SQ y GN se agrupan junto con las otras cepas de *Dunaliella salina*,

también se observa que la cepa GN y CCAP 19/18 se encuentran más relacionadas evolutivamente que la cepa SQ. Esto nos ayuda a responder por qué no se logró obtener el ensamble por referencia cuando de utilizó el genoma mitocondrial de la cepa CCAP 19/18 con las lecturas de la cepa SQ. También responde la razón de un mejor resultado en el ensamble por referencia con la cepa GN comparado con el ensamble de la cepa SQ, ya que GN y CCAP 19/18 tienen una mayor cercanía evolutiva que SQ y CCAP 19/18 si nos basamos en el gen marcador *cox1*.

Se realizó también el análisis utilizando como base la secuencia de aminoácidos de la proteína cox1 de las cepas SQ y GN, con las secuencias de aminoácidos de proteínas cox1 de 10 organismos del orden Chlamydomonales incluyendo las cepas de *Dunaliella salina* CCAP 19/18 y CCM-UDEC 001. El alineamiento se realizó en el programa MEGA 7 utilizando el método de análisis Maximun Likelihood, el árbol filogenético se muestra en la figura 21.

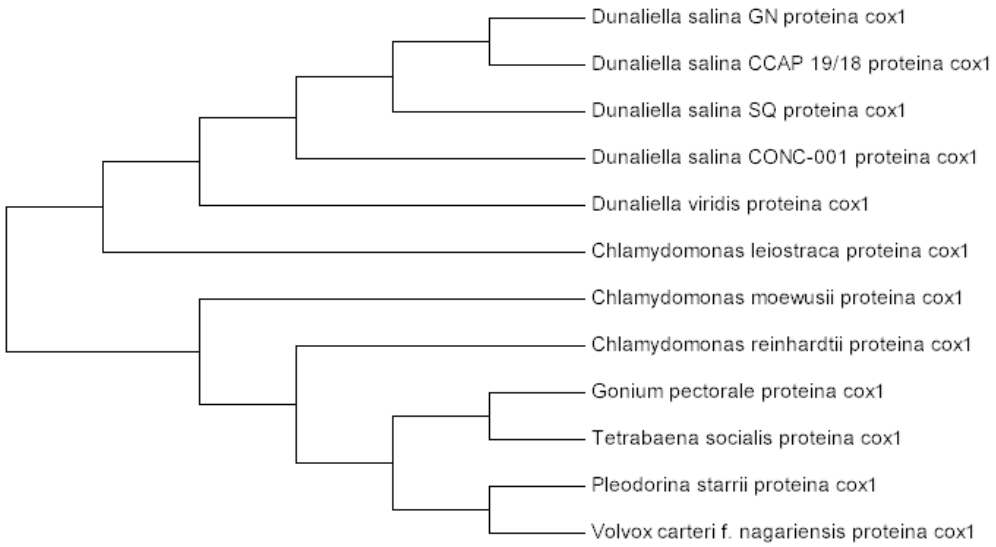

Figura 21. Análisis filogenético molecular por el método Maximum Likelihood en MEGA7. proteínas cox 1 de 10 organismos del orden Chlamydomonales y proteínas cox1 de las cepas SQ y GN.

De igual manera que el análisis de genes *cox1*, el resultado del análisis del árbol filogenético realizado con secuencias de proteínas cox1, nos muestra que las cepas SQ y GN se agrupan junto con las otras cepas de *Dunaliella salina*, también se observa que la cepa GN y CCAP 19/18 se encuentran más relacionadas evolutivamente que la cepa SQ.

Se llevó acabo el alineamiento de los genomas completos y los genomas de 10 organismos del orden Chlamydomonales incluyendo las cepas de *Dunaliella salina* CCAP 19/18 y CCM-UDEC 001.

El alineamiento se realizó en el programa MEGA 7 utilizando el método de análisis Maximun Likelihood, el árbol filogenético se muestra en la figura 22.

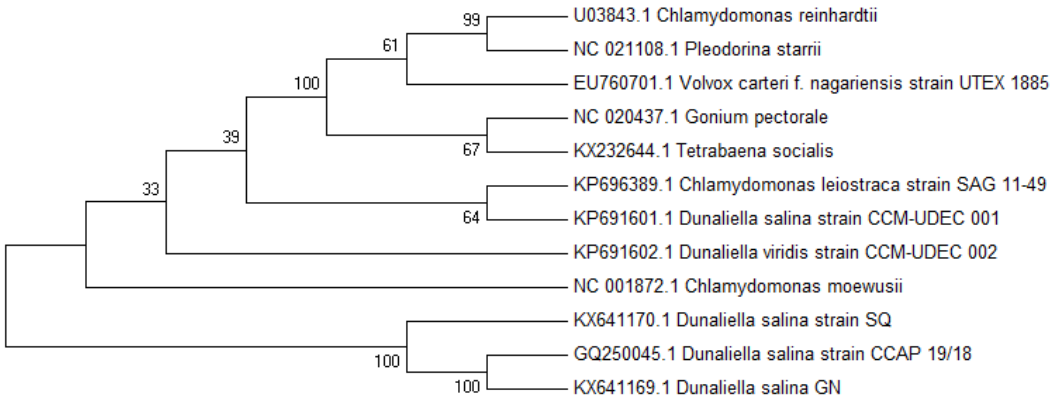

Figura 22. Análisis filogenético molecular por el método Maximum Likelihood en MEGA7. Genomas mitocondriales de 10 organismos del orden Chlamydomonales y los genomas mitocondriales de las cepas SQ y GN de *Dunaliella salina*.

En la figura 22 se observa cómo sigue agrupando las cepas SQ y GN con la cepa CCAP 19/18 de *Dunaliella salina* usando las secuencias de los genomas completos.

Enseguida se realizaron análisis de porcentaje de identidad de los genes y proteínas entra las cepas SQ, GN, CCAP 19/18 y CCM-UDEC 001 para tener mayor evidencia de cuales cepas se encuentran más cercanas evolutivamente.

Los análisis se llevaron a cabo en el programa UGENE versión 1.18.0 utilizando *Pairwise Identity* con el algoritmo Smith-Waterman. En la tabla 11 se muestran los resultados de la comparación de 7 genes y proteínas de la cepa CCAP 19/18 con las cepas SQ y GN.

	Pairwise Identity (%) CCAP 19/18 vs SQ		Pairwise Identity (%) CCAP 19/18 vs GN	
Gene	DNA	Proteína	DNA	Proteína
cob	91	96	99	99
cox1	90	96	98	99
nad1	91	98	98	100
nad2	89	89	98	99
nad4	91	93	100	100
nad5	94	97	99	99
nad6	83	81	97	99

Tabla 11. Análisis *Pairwise Identity* con el algoritmo Smith-Waterman. Porcentaje de identidad de secuencias codificantes y proteínas entre CCAP 19/18 con SQ y GN.

Los resultados que se presentan en la en la tabla 11 muestran que los porcentajes de identidad de los genes y proteínas entre la cepa GN y CCAP 19/18 son superiores a SQ y CCAP 19/18. Presentando desde el 98% hasta el 100% de identidad a nivel de secuencia de nucleótidos y aminoácidos entre CCAP

19/18 y GN. Para el caso de CCAP 19/18 y SQ presentan desde el 83% al 98 % de identidad. Con esto se demuestra nuevamente que la cepa GN es evolutivamente más cercana a CCAP 19/18 que a SQ, aun estando separadas geográficamente a una mayor distancia, ya que GN y SQ fueron aisladas en Baja California México y la cepa CCAP 19/18 se aisló de Australia.

En la tabla 12 se muestran los resultados de la comparación de 7 genes y proteínas de la cepa CCM-UDEC 001 con las cepas SQ y GN.

	Pairwise Identity (%) CCM-UDEC 001 vs SQ		Pairwise Identity (%) Pairwise Identity (%) CCM-UDEC 001 vs GN	
Gene	DNA	Proteína	DNA	Proteína
cob	89	93	88	94
cox1	90	96	88	94
nad1	88	96	88	96
nad2	95	98	90	91
nad4	81	83	79	83
nad5	89	92	88	92
nad6	77	78	78	79

Tabla 12. Análisis *Pairwise Identity* con el algoritmo Smith-Waterman. Porcentaje de identidad de secuencias codificantes y proteínas entre CCM-UDEC 001 con SQ y GN.

En la tabla 13 se muestran los resultados de la comparación de 7 genes y proteínas de la cepa SQ con la cepa GN.

	Pairwise Identity (%) SQ vs GN	
Gene	DNA	Proteína
cob	91	96
cox1	90	96
nad1	91	98
nad2	89	89
nad4	91	93
nad5	94	97
nad6	84	81

Tabla 13. Análisis *Pairwise Identity* con el algoritmo Smith-Waterman. Porcentaje de identidad de secuencias codificantes y proteínas entre SQ y GN.

Los resultados en las tablas 12 y 13 muestran que existe una mayor relación evolutiva entre las cepas SQ y GN que entre CCM-UDEC 001 con SQ y GN.

Con los resultados obtenidos en este trabajo se muestran que los genomas mitocondriales de las cepas SQ y GN contienen los mismos genes y la misma distribución que las cepas de *Dunaliella salina* CCAP 19/18 y CCM-UDEC 001, pero varían en el número de intrones, también al analizar las figuras 19 y 18 vemos que los genes de RNA ribosomal están fragmentados y dispersos por todo

el genoma como en las cepas reportadas (CCAP 19/18 y CCM-UDEC 001), los 12 genes se encuentran codificados en la misma hebra como se ha observado en *Dunaliella salina* (Smith *et al*., 2010). Aunado con los resultados obtenidos de los alineamientos de los genes cox1, los alineamientos de los genomas completos y el porcentaje de identidad de las 7 secuencias codificantes, se tiene evidencia suficiente para decir que los genomas de SQ y GN pertenecen a mitocondrias de cepas de *Dunaliella salina*.

VI.4.1 Análisis de sintenia

Se realizaron análisis de sintenia para evaluar la conservación del orden génico y la orientación a lo largo del genoma, ya que un elevado grado de sintenia es indicativo de proximidad filogenética. Los análisis se llevaron a cabo con el programa mauve (Darling *et al.,* 2004). Se comparó la sintenia de los genomas entre CCAP 19/18 con SQ, CCAP 19/18 con GN, CCM-UDEC 001 con SQ, CCM-UDEC 001 con GN y SQ con GN.

Los Resultados se muestran en las figuras 23, 24, 25, 26 y 27.

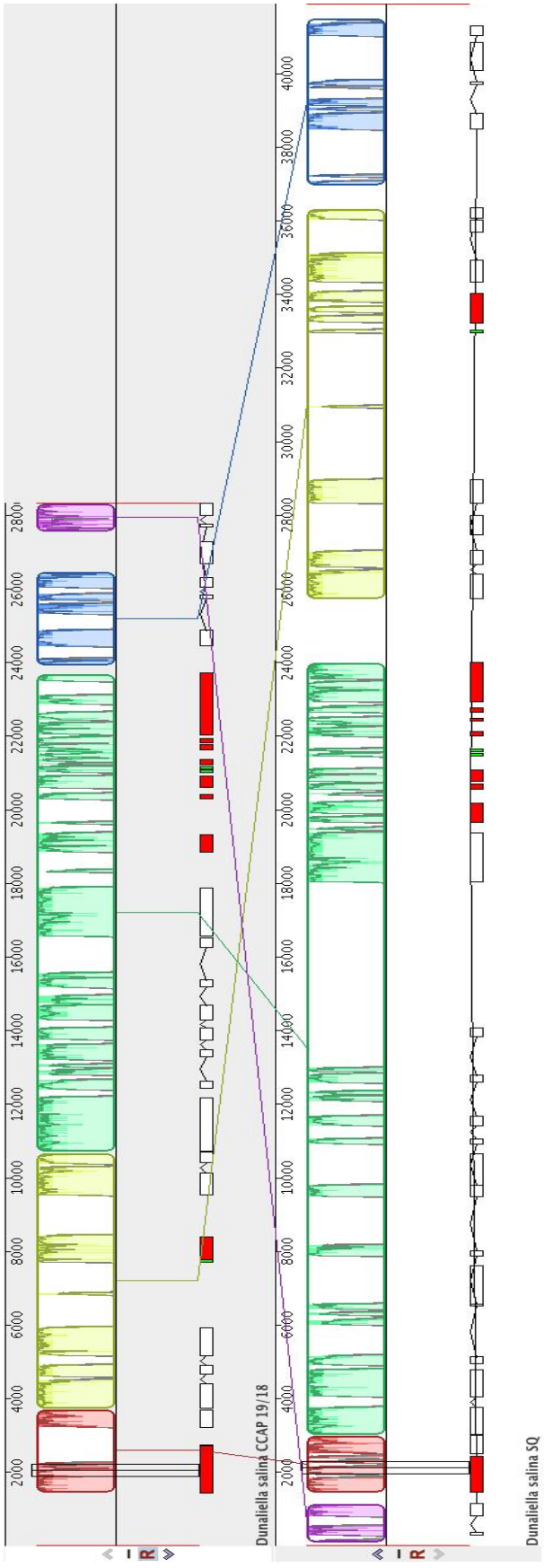

Figura 23. Análisis de sintenia entre genomas de *Dunaliella salina* CCAP 19/18 y *Dunaliella salina* SQ

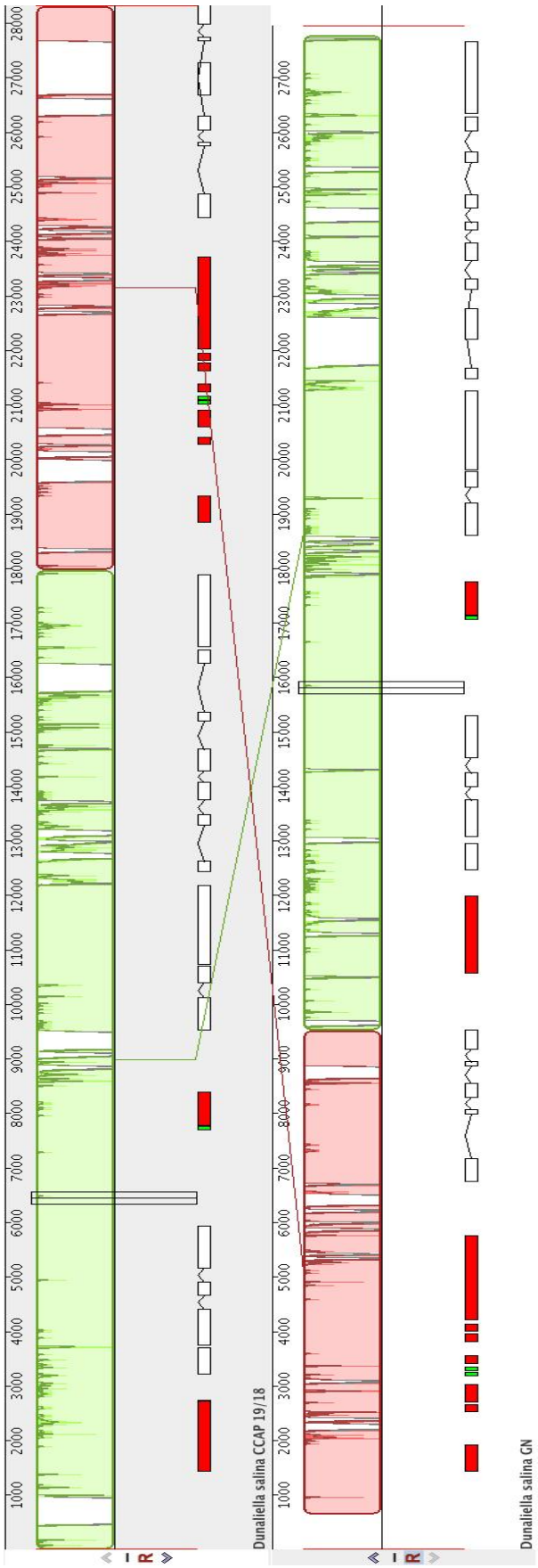

Figura 24. Análisis de sintenia entre genomas de *Dunaliella salina* CCAP 19/18 y *Dunaliella salina* GN

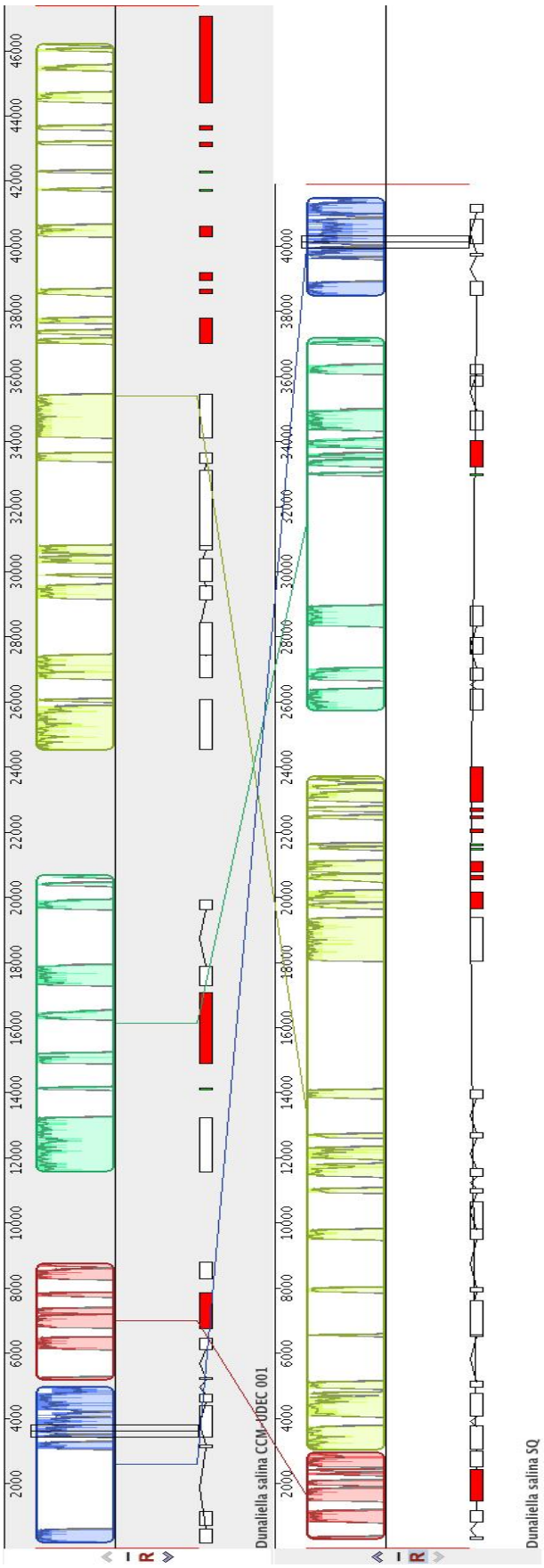

Figura 25. Análisis de sintenia entre genomas de *Dunaliella salina* CCM-UDEC 001 y *Dunaliella salina* SQ

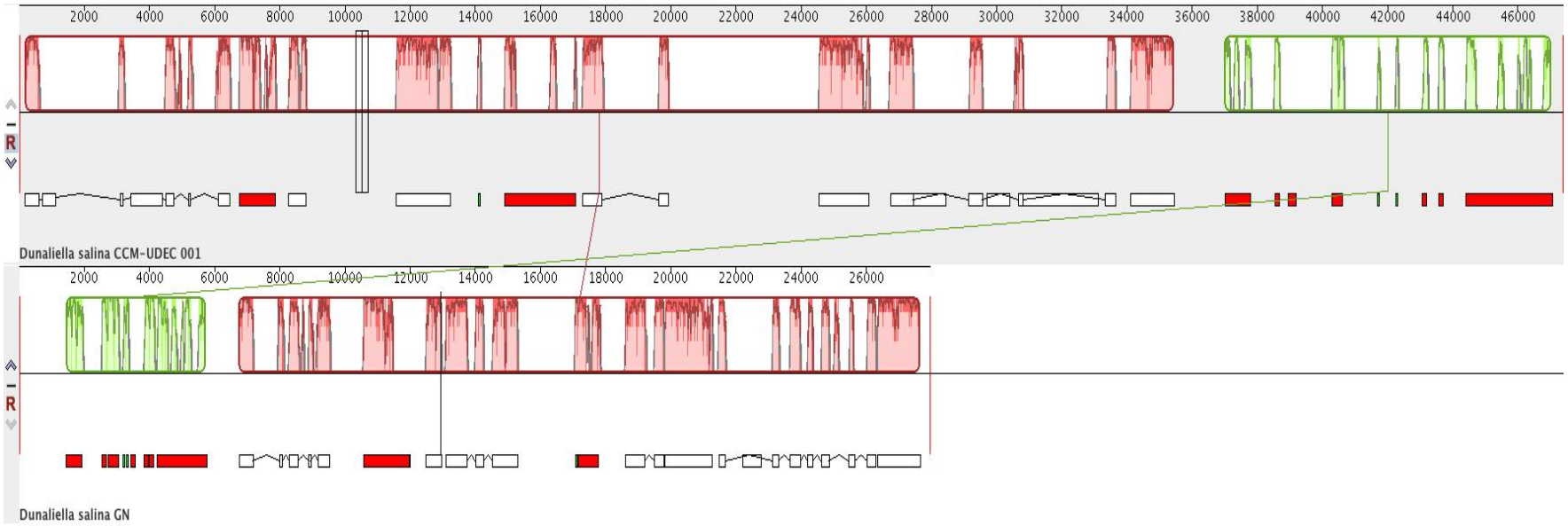

Figura 26. Análisis de sintenia entre genomas de *Dunaliella salina* CCM-UDEC 001 y *Dunaliella salina* GN

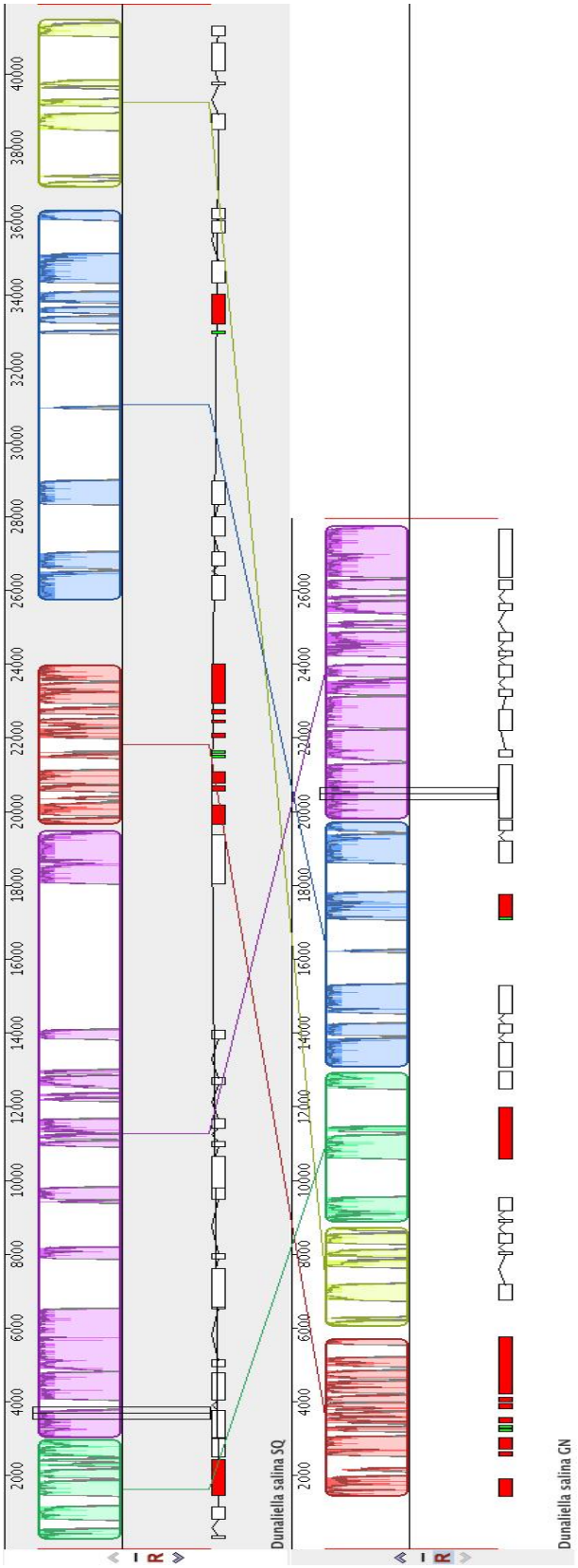

Figura 27. Análisis de sintenia entre genomas de *Dunaliella salina* SQ y *Dunaliella salina* GN

El resultado del análisis de sintenia entre genomas de *Dunaliella salina* CCAP 19/18 y *Dunaliella salina* SQ, que se muestra en la figura 23, presenta una arquitectura distinta de los genomas, se observa cómo se presentó un reacomodo de las regiones de color amarillo y verde. Las cuales en la cepa CCAP 19/18 se encuentra de izquierda a derecha región amarilla y región verde. Y en la cepa SQ de izquierda a derecha se encuentra primero la región verde y enseguida la amarilla. También se observan regiones que no comparten (zonas blancas) las cuales corresponden a secuencias intergénicas. Podemos observar que la cepa SQ contiene un mayor número de secuencias intergénicas, que son la razón de que el genoma mitocondrial sea de mayor tamaño que el de la cepa CCAP 19/18.

El resultado del análisis de sintenia entre genomas de *Dunaliella salina* CCAP 19/18 y *Dunaliella salina* GN, que se muestra en la figura 24, presenta una arquitectura similar en los genomas mitocondriales, lo que significa un alto grado de sintenia que es indicativo de proximidad filogenética. El resultado que se muestra en la figura 25, presenta una arquitectura distinta entre los genomas de la cepa CCM-UDEC 001 y la cepa SQ, se vuelve a observar el reacomodo de las regiones color amarillo y verde, también la diferencia entre las regiones intergénicas. La figura 26 muestra la sintenia de los genomas entre las cepas CCM-UDEC 001 y la cepa GN, aquí vemos que la arquitectura genómica es similar, pero las regiones intergénicas presentan una gran variación, e inclusive se observa como la cepa CCM-UDCE 001 contiene una mayor cantidad de

secuencias intergénicas, lo cual hace que su genoma sea aproximadamente el doble de tamaño que el de la cepa GN. El resultado de sintenia entre los genomas de las cepas SQ y GN que se muestra en la figura 27, es de bajo grado, ya que entre estas cepas se observa un reordenamiento de tres regiones. La cepa SQ de izquierda a derecha se distribuyen color azul, amarillo y verde, en la cepa GN de izquierda a derecha se distribuyen amarillo, verde y azul. También se observa una gran diferencia en regiones intergénicas, presentando una mayor cantidad la cepa SQ.

Integrando los resultados obtenidos en este trabajo, podemos decir que las cepas de *Dunaliella salina* presentan un patrón constante de 12 genes, todos los genes se encuentran en la misma hebra, 7 de ellos codificantes para proteínas, 3 para RNA de transferencia y dos para RNA ribosomal, también que los genes de RNA ribosomal están fragmentados y dispersos por todo el genoma. Pero las cepas de *Dunaliella salina* presentan una gran diversidad de arquitecturas genómicas y de tamaño de secuencias intergénicas, esto provoca que los tamaños de los genomas presenten variación de hasta el doble de tamaño entre cepas. Estos resultados sirven como referencia base para la identificación de *Dunaliella salina*, ya que no solo se deben de utilizar marcadores como el gen cox1, sino también los alineamientos completos de los genomas para poder distinguir entre cepas de *Dunaliella salina*, esto debido a que la mayor variación entre cepas es en los tamaños de las regiones no codificantes y la arquitectura de los genomas.

VI.4.2 Análisis de radios dn/ds

Se llevó a cabo un análisis de sustituciones sinónimas en sitios sinónimos (ds) y sustituciones no-sinónimas en sitios no-sinónimos (dn), esto para evaluar si es selección neutral, selección negativa o selección positiva el mecanismo causante de la variación genómica entre cepas de *Dunaliella salina*.

Para esto se alienaron los 7 genes codificantes entre las cepas de *Dunaliella salina* en el programa UGENE versión 1.18.0. Los resultados se muestran en las tablas 14, 15, 16, 17, 18, 19 y 20.

Tabla 14. Análisis dn/ds del gen cox1 de las cepas CCAP 19/18, CCM-UDEC 001, SQ y GN.

Comparación Gen cox1	ds	dn	dn/ds
SQ vs GN	0.4685	0.0206	0.04397
SQ vs CCAP_19/18	0.4331	0.0180	0.04156
SQ vs CONC-001	0.5081	0.0226	0.04448
GN vs CCAP_19/18	0.0596	0.0051	0.08557
GN vs CONC-001	0.5879	0.0348	0.05919
CCAP_19/18 vs CONC-001	0.6026	0.0301	0.04995
Promedio	0.4433	0.0219	0.05412

Tabla 15. Análisis dn/ds del gen cob de las cepas CCAP 19/18, CCM-UDEC 001, SQ y GN.

Comparación Gen cob	ds	dn	dn/ds
SQ vs GN	0.3300	0.0199	0.0603
SQ vs CCAP_19/18	0.3297	0.0187	0.0567
SQ vs CONC-001	0.4370	0.0361	0.0826
GN vs CCAP_19/18	0.0000	0.0011	nan
GN vs CONC-001	0.5360	0.0397	0.0741
CCAP_19/18 vs CONC-001	0.5355	0.0409	0.0764
Promedio	0.3614	0.0261	0.0583

Tabla 16. Análisis dn/ds del gen nad1 de las cepas CCAP 19/18, CCM-UDEC 001, SQ y GN.

Comparación Gen nad1	ds	dn	dn/ds
SQ vs GN	0.3606	0.0180	0.04992
SQ vs CCAP_19/18	0.3846	0.0180	0.04680
SQ vs CONC-001	0.5843	0.0343	0.05870
GN vs CCAP_19/18	0.0553	0.0000	0.00000
GN vs CONC-001	0.6076	0.0290	0.04773
CCAP_19/18 vs CONC-001	0.6186	0.0290	0.04688
Promedio	0.4352	0.0214	0.04167

Tabla 17. Análisis dn/ds del gen nad2 de las cepas CCAP 19/18, CCM-UDEC 001, SQ y GN.

Comparación Gen nad2	ds	dn	dn/ds
SQ vs GN	0.3437	0.0446	0.12976
SQ vs CCAP_19/18	0.3900	0.0498	0.12769
SQ vs CONC-001	0.2228	0.0048	0.02154
GN vs CCAP_19/18	0.0601	0.0078	0.12978
GN vs CONC-001	0.3629	0.0434	0.11959
CCAP_19/18 vs CONC-001	0.4104	0.0496	0.12086
Promedio	0.2983	0.0333	0.10821

Tabla 18. Análisis dn/ds del gen nad4 de las cepas CCAP 19/18, CCM-UDEC 001, SQ y GN.

Comparación Gen nad4	ds	dn	dn/ds
SQ vs GN	0.3000	0.0311	0.10367
SQ vs CCAP_19/18	0.3000	0.0311	0.10367
SQ vs CONC-001	0.5292	0.0763	0.14418
GN vs CCAP_19/18	0.0000	0.0000	nan
GN vs CONC-001	0.6659	0.0833	0.12509
CCAP_19/18 vs CONC-001	0.6659	0.0833	0.12509
Promedio	0.4922	0.0610	0.12034

Tabla 19. Análisis dn/ds del gen nad5 de las cepas CCAP 19/18, CCM-UDEC 001, SQ y GN.

Comparación Gen nad5	ds	dn	dn/ds
SQ vs GN	0.2571	0.0131	0.05095
SQ vs CCAP_19/18	0.2608	0.0131	0.05023
SQ vs CONC-001	0.4099	0.0407	0.09929
GN vs CCAP_19/18	0.0028	0.0015	0.53571
GN vs CONC-001	0.4941	0.0448	0.09067
CCAP_19/18 vs CONC-001	0.5017	0.0453	0.09029
Promedio	0.3211	0.0264	0.15286

Tabla 20. Análisis dn/ds del gen nad6 de las cepas CCAP 19/18, CCM-UDEC 001, SQ y GN.

Comparación Gen nad6	ds	dn	dn/ds
SQ vs GN	0.6052	0.1284	0.21216
SQ vs CCAP_19/18	0.6503	0.1284	0.19745
SQ vs CONC-001	0.5912	0.1291	0.21837
GN vs CCAP_19/18	0.0929	0.0027	0.02906
GN vs CONC-001	0.3899	0.0849	0.21775
CCAP_19/18 vs CONC-001	0.4321	0.0804	0.18607
Promedio	0.4603	0.0923	0.17681

Los valores del radio dn/ds nos proporciona información referente a los mecanismos evolutivos de selección de las mutaciones que sufren los organismos, un radio dn/ds = 1, indica selección neutral, dn/ds < 1, indica selección negativa y un radio dn/ds > 1, indica selección positiva. Los resultados de las tablas 14, 15, 16, 17, 18, 19 y 20, muestran para cada uno de los genes codificantes de las cepas de *Dunaliella salina*, un radio dn/ds < 1, por lo tanto, las mutaciones que se generan en las regiones codificantes son seleccionadas de forma negativa.

Las regiones no codificantes de las cepas de *Dunaliella salina* no se lograron alinear, por lo tanto, no se obtuvo el valor del radio dn/ds. Las regiones no codificantes no se pudieron alinear debido a la gran diferencia en secuencia y tamaño que presentan estas regiones entre las cepas de *Dunaliella salina* reportadas hasta la fecha. El no lograr alinear las secuencias no codificantes, puede ser debido a que estas zonas han acumulado un gran número de mutaciones en el tiempo, ya que no están expuestas a la selección natural como las regiones codificantes.

Con los resultados obtenidos en este trabajo podemos decir que la gran variabilidad de arquitecturas genómicas de las mitocondrias de las cepas de *Dunaliella salina*, se deben a factores no adaptativos (Smith *et al.,* 2010), ya que las regiones no codificantes y el orden génico de estas en el genoma mitocondrial, son las principales regiones de variación entre las cepas aisladas de distintas regiones geográficas (CCAP 19/18 de Australia, CCM-UDEC 001 del desierto de atacama en Chile y SQ de San Quintín B.C México y GN Guerrero Negro B.C México), pero el número de genes entre cepas se mantiene constante (12 genes) y la identidad a nivel de nucleótidos y aminoácidos que se presentan entre las cepas nos indica una selección negativa, por eso los porcentajes de identidad son altos.

VI. Conclusiones

El genoma nuclear de *Dunaliella salina* SQ contiene una gran cantidad de secuencias repetitivas como la cepa CCAP 19/18, por tal motivo los ensambles *de novo* no fueron exitosos.

Los genomas mitocondriales de las cepas de *Dunaliella salina* SQ y GN contienen los 12 genes que se encuentran presentes en las cepas CCAP 19/18 y CONC-001, 7 codificantes para proteínas, 3 genes de RNA de transferencia y 2 genes de RNA ribosomal.

Los genomas mitocondriales de las cepas de *Dunaliella salina* varían en tamaño, debido a la diferencia en la longitud de las regiones intergénicas.

La variación en los genomas mitocondriales de las cepas de *Dunaliella salina* es causada por factores no-adaptativos.

La identificación de cepas de *Dunaliella salina* debe ser por medio de alineamientos de genomas mitocondriales, debido a la variación en arquitectura genómica entre cepas.

Referencias

Al-Hasan, R.H., Ghannoum, M.A., Sallal, A.K., Abu-Elteen, K.H. y Radwan, S.S. 1987. Correlative changes of growth, pigmentation and lipid composition of *Dunaliella salina* in response to halostress. *Microbiology*, *133*(9), pp.2607-2616.

Alkayal, F., Albion, R.L., Tillett, R.L., Hathwaik, L.T., Lemos, M.S. y Cushman, J.C. 2010. Expressed sequence tag (EST) profiling in hyper saline shocked *Dunaliella salina* reveals high expression of protein synthetic apparatus components. *Plant science*, *179*(5), pp.437-449.

Andrews, S. 2010. FastQC: a quality control tool for high throughput sequence data.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M. y Brzezinski, M.A. 2004. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science*, *306*(5693), pp.79-86.

Arredondo, B. O. y Voltolina, D. 2007. Métodos y herramientas analíticas en la evaluación de la biomasa microalgal. CIBNOR pp. 49-50.

Ben-Amotz, A. y Avron, M. 1983. On the factors which determine massive β-carotene accumulation in the halotolerant alga *Dunaliella bardawil. Plant Physiology*, *72*(3), pp.593-597.

Ben-Amotz, A., Gressel, J. y Avron, M. 1987. Massive accumulation of phytoene induced by norflurazon in Dunaliella bardawil (Chlorophyceae) prevents recovery from photoinhibition. *Journal of phycology*, *23*(1), pp.176-181.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R. y Boutell, J.M. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *nature*, *456*(7218), p.53.

Blanc, G., Duncan, G., Agarkova, I., Borodovsky, M., Gurnon, J., Kuo, A., Lindquist, E., Lucas, S., Pangilinan, J., Polle, J. y Salamov, A. 2010. The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *The Plant Cell*, *22*(9), pp.2943-2955.

Bouvier-Navé, P., Benveniste, P., Oelkers, P., Sturley, S.L. y Schaller, H. 2000. Expression in yeast and tobacco of plant cDNAs encoding acyl CoA: diacylglycerol acyltransferase. *The FEBS Journal*, *267*(1), pp.85-96.

Chikhi, R. y Medvedev, P. 2013. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, *30*(1), pp.31-37.

Chikhi, R. y Rizk, G. 2013. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithms for Molecular Biology*, *8*(1), p.22.

Darling, A.C., Mau, B., Blattner, F.R. y Perna, N.T. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, *14*(7), pp.1394-1403.

Davis, M.S., Solbiati, J. y Cronan, J.E, 2000. Overproduction of acetyl-CoA carboxylase activity increases the rate of fatty acid biosynthesis in Escherichia coli. *Journal of Biological Chemistry*, *275*(37), pp.28593-28598.

Dehesh, K., Tai, H., Edwards, P., Byrne, J. y Jaworski, J.G. 2001. Overexpression of 3-ketoacyl-acyl-carrier protein synthase IIIs in plants reduces the rate of lipid synthesis. *Plant physiology*, *125*(2), pp.1103-1114.

Del Vasto, M., Figueroa-Martinez, F., Featherston, J., Gonzalez, M.A., Reyes-Prieto, A., Durand, P.M. y Smith, D.R. 2015. Massive and widespread organelle genomic expansion in the green algal genus Dunaliella. *Genome biology and evolution*, *7*(3), pp.656-663.

Deth, S.K, 1999. *Antimicrobial compounds from marine cyanobacteria with special reference to the bioactivity of a purified compound from Oscillatoria laete-virens BDU 20801* (Doctoral dissertation, Ph D thesis, Bharathidasan University, Thiruchirappalli, India).

Dismukes, G.C., Carrieri, D., Bennette, N., Ananyev, G.M. y Posewitz, M.C. 2008. Aquatic phototrophs: efficient alternatives to land-based crops for biofuels. *Current opinion in biotechnology*, *19*(3), pp.235-240.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. y Bibillo, A. 2009. Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), pp.133-138.

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), p.1433.

Gouveia, L. 2011. Microalgae as a Feedstock for Biofuels. In *Microalgae as a Feedstock for Biofuels* (pp. 1-69). Springer Berlin Heidelberg.

Griffiths, M.J. y Harrison, S.T. 2009. Lipid productivity as a key characteristic for choosing algal species for biodiesel production. *Journal of Applied Phycology*, *21*(5), pp.493-507.

Gurevich, A., Saveliev, V., Vyahhi, N. y Tesler, G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), pp.1072-1075.

Hosseini Tafreshi, A. y Shariati, M. 2009. Dunaliella biotechnology: methods and applications. *Journal of applied microbiology*, *107*(1), pp.14-35.

Hahn, C., Bachmann, L. y Chevreux, B. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic acids research*, *41*(13), pp.e129-e129.

Jako, C., Kumar, A., Wei, Y., Zou, J., Barton, D.L., Giblin, E.M., Covello, P.S. y Taylor, D.C. 2001. Seed-specific over-expression of an Arabidopsis cDNA encoding a diacylglycerol acyltransferase enhances seed oil content and seed weight. *Plant physiology*, *126*(2), pp.861-874.

Kenrick, P. y Crane, P.R., 1997. *The origin and early diversification of land plants. A cladistic study* (Vol. 560). Smithsonian Institute Press Washington DC.: A cladistic study. Smithsonian Institute Press.

Kreimer, G.,2009. The green algal eyespot apparatus: a primordial visual system and more?. *Current genetics*, *55*(1), pp.19-43.

Langmead, B. y Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), pp.357-359.

Leliaert, F., Smith, D.R., Moreau, H., Herron, M.D., Verbruggen, H., Delwiche, C.F. y De Clerck, O. 2012. Phylogeny and molecular evolution of the green algae. *Critical Reviews in Plant Sciences*, *31*(1), pp.1-46.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y. y Zhang, Z. 2010. The sequence and de novo assembly of the giant panda genome. *Nature*, *463*(7279), p.311.

Lin, H., Castro, N.M., Bennett, G.N. y San, K.Y. 2006. Acetyl-CoA synthetase overexpression in Escherichia coli demonstrates more efficient acetate assimilation and lower acetate accumulation: a potential tool in metabolic engineering. *Applied microbiology and biotechnology*, *71*(6), pp.870-874.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z. y Dewell, S.B. 2006. Corrigendum: Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *441*(7089), pp.120-121.

Mata, T.M., Martins, A.A. y Caetano, N.S. 2010. Microalgae for biodiesel production and other applications: a review. *Renewable and sustainable energy reviews*, *14*(1), pp.217-232.

Mendoza, H., De la Jara, A., Freijanes, K., Carmona, L., Ramos, A.A., de Sousa Duarte, V., Varela, S. y Carlos, J. 2008. Characterization of *Dunaliella salina* strains by flow cytometry: a new approach to select carotenoid hyperproducing strains. *Electronic Journal of Biotechnology*, *11*(4), pp.5-6.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Maréchal-Drouard, L. y Marshall, W.F. 2007. The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, *318*(5848), pp.245-250.

Metzker, M.L. 2010. Sequencing technologies--the next generation. *Nature reviews. Genetics*, *11*(1), p.31.

Nozaki, H., Takano, H., Misumi, O., Terasawa, K., Matsuzaki, M., Maruyama, S., Nishida, K., Yagisawa, F., Yoshida, Y., Fujiwara, T. y Takio, S. 2007. A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga Cyanidioschyzon merolae. *BMC biology*, *5*(1), p.28.

Odjadjare, E.C., Mutanda, T. y Olaniran, A.O. 2017. Potential biotechnological application of microalgae: a critical review. *Critical reviews in biotechnology*, *37*(1), pp.37-52.

Oren, A. 2005. A hundred years of Dunaliella research: 1905–2005. *Saline systems*, *1*(1), p.2.

Radakovits, R., Jinkerson, R.E., Fuerstenberg, S.I., Tae, H., Settlage, R.E., Boore, J.L. y Posewitz, M.C. 2012. Draft genome sequence and genetic transformation of the oleaginous alga Nannochloropis gaditana. *Nature communications*, *3*, p.686.

Ramos, A., Coesel, S., Marques, A., Rodrigues, M., Baumgartner, A., Noronha, J., Rauter, A., Brenig, B. y Varela, J. 2008. Isolation and characterization of a stress-inducible *Dunaliella salina* Lcy-β gene encoding a functional lycopene β-cyclase. *Applied microbiology and biotechnology*, *79*(5), p.819.

Rangasamy, D. y Ratledge, C. 2000. Genetic enhancement of fatty acid synthesis by targeting rat liver ATP: citrate lyase into plastids of tobacco. *Plant physiology*, *122*(4), pp.1231-1238.

Ratledge, C. 2002. Regulation of lipid accumulation in oleaginous micro-organisms. Biochemical Society Transactions 30 (6), 1047–1050.

Rismani-Yazdi, H., Haznedaroglu, B.Z., Bibby, K. y Peccia, J. 2011. Transcriptome sequencing and annotation of the microalgae *Dunaliella tertiolecta*: pathway description and gene discovery for production of next-generation biofuels. *BMC genomics*, *12*(1), p.148.

Rittmann, B.E. 2008. Opportunities for renewable bioenergy using microorganisms. *Biotechnology and bioengineering*, *100*(2), pp.203-212.

Simpson, J.T. y Durbin, R. 2010. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics*, *26*(12), pp. i367-i373.

Simpson, J.T. y Durbin, R. 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome research*, *22*(3), pp.549-556.

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), p.e31468

Smith, D.R., Lee, R.W., Cushman, J.C., Magnuson, J.K., Tran, D. y Polle, J.E. 2010. The *Dunaliella salina* organelle genomes: large sequences, inflated with intronic and intergenic DNA. *BMC plant biology*, *10*(1), p.83.

Subrahmanyam, S. y Cronan, J.E. 1998. Overproduction of a Functional Fatty Acid Biosynthetic Enzyme Blocks Fatty Acid Synthesis in Escherichia coli. *Journal of bacteriology*, *180*(17), pp.4596-4602.

Takagi, M. y Yoshida, T. 2006. Effect of salt concentration on intracellular accumulation of lipids and triacylglyceride in marine microalgae Dunaliella cells. *Journal of bioscience and bioengineering*, *101*(3), pp.223-226.

Verwoert, I.I., van der Linden, K.H., Walsh, M.C., Nijkamp, H.J.J. y Stuitje, A.R. 1995. Modification of Brassica napus seed oil by expression of the Escherichia coli fabH gene, encoding 3-ketoacyl-acyl carrier protein synthase III. *Plant molecular biology*, *27*(5), pp.875-886.

Yan, Y., Zhu, Y.H., Jiang, J.G. y Song, D.L. 2005. Cloning and sequence analysis of the phytoene synthase gene from a unicellular chlorophyte, *Dunaliella salina*. *Journal of agricultural and food chemistry*, *53*(5), pp.1466-1469.

Zhao, R., Cao, Y., Xu, H., Lv, L., Qiao, D. y Cao, Y. 2011. Analysis of expressed sequence tags from the green alga *Dunaliella salina* (Chlorophyta). *Journal of phycology*, *47*(6), pp.1454-1460.

Zou, J., Katavic, V., Giblin, E.M., Barton, D.L., MacKenzie, S.L., Keller, W.A., Hu, X. y Taylor, D.C. 1997. Modification of seed oil content and acyl composition in the Brassicaceae by expression of a yeast sn-2 acyltransferase gene. *The Plant Cell*, *9*(6), pp.909-923.