Universidad Autónoma de Baja California

Facultad de Ingeniería, Arquitectura y Diseño

Maestría y Doctorado en Ciencias e Ingeniería

“Ensamble y anotación del genoma del cloroplasto de la microalga *Dunaliella salina* aislada de Noruega e identificación de secuencias promotoras *in silico*”

Tesis para obtener el grado de:

Maestro en ingeniería

PRESENTA

León Jerónimo Murillo Acevedo

Director de tesis:

Dr. Dante Alberto Magdaleno Moncayo

Ensenada, Baja California, junio de 2025.

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

FACULTAD DE INGENIERÍA, ARQUITECTURA Y DISEÑO

MAESTRÍA Y DOCTORADO EN CIENCIAS E INGENIERÍA

ENSAMBLE Y ANOTACIÓN DEL GENOMA DEL CLOROPLASTO DE LA MICROALGA *DUNALIELLA SALINA* AISLADA DE NORUEGA E IDENTIFICACIÓN DE SECUENCIAS PROMOTORAS *IN SILICO*

TESIS

Para obtener el grado de maestría en ingeniería, presenta:

León Jerónimo Murillo Acevedo

Aprobada por:

Dr. Dante Alberto Magdaleno Moncayo
Director de tesis

Dr. Julio Enrique Valencia Suárez
Codirector de tesis

Dra. Haydee López Rodríguez
Miembro del comité

Dr. Rubén César Villarreal Sánchez
Miembro del comité

Dr. Priscy Alfredo Luque Morales
Miembro del comité

Ensenada Baja California, México, 16 de junio de 2025

Resumen de la tesis de León Jerónimo Murillo Acevedo, presentada como requisito para la obtención del grado de MAESTRO EN INGENIERÍA, Ensenada B.C a junio de 2025.

"Ensamble y anotación del genoma del cloroplasto de la microalga *Dunaliella salina* aislada de Noruega e identificación de secuencias promotoras *in silico*"

Aprobado por:

Dr. Dante Alberto Magdaleno Moncayo

Director de tesis

Resumen:

En el presente estudio se logró ensamblar y anotar el genoma del cloroplasto de la microalga *Dunaliella salina* cepa noruega de manera exitosa, presentando una estructura cuadripartida típica de cloroplastos, con una longitud de 241,000 pb. Dentro del genoma se encontraron 98 genes distintos, 66 codificantes, 3 rRNA y 29 tRNA. A pesar de la reducción de tamaño en comparación con los genomas del cloroplasto de *Dunaliella salina* cepa SQ y CCAP19/18, el genoma de la cepa noruega tiene 27.8% de regiones intergénicas que pueden ser utilizadas para insertar un casete de expresión de proteínas recombinantes. Además, se identificaron y caracterizaron 51 promotores endógenos putativos, los cuales contienen secuencias inducibles a la presencia de sal, luz, bajas temperaturas y baja concentración de CO_2.

Palabras claves: *Dunaliella salina;* cloroplasto; ensamble de genoma; anotación de genoma; promotores endógenos; análisis *in silico*.

Dedicatoria

Quiero dedicar este trabajo principalmente a mis padres, quienes, como en todas las etapas de mi vida, me apoyaron durante estos dos años de maestría.

A mi papá, Luis Enrique Murillo Moreno, por trabajar todo el año, guardando sus vacaciones, por si "surgía algo" y necesitaba que vinieran a Ensenada desde La Paz. Por facilitarme un vehículo para moverme con mayor facilidad de la escuela a la casa, y por hacer todo lo posible para que pudiera concentrarme únicamente en estudiar, sin mayores preocupaciones.

A mi mamá, Laura Esperanza Acevedo Acevedo, que iba y venía por temporadas desde La Paz, adaptando su vida para brindarme su compañía, su apoyo y su cariño en los momentos que más la necesité.

A ambos, por su paciencia, esfuerzo y amor constante, por ser mis pilares, sin los cuales este logro no habría sido posible.

A mi pareja, Angélica Luchi Hernández, por darme todo su amor, ser un lugar de refugio emocional y animarme en los momentos donde más lo necesite. También por los regaños constantes por el uso excesivo de muletillas mientras exponía (se agradece).

A Jacinto M, por hacerme compañía en mis jornadas de estudio.

A todos mis profesores, amigos y compañeros que me acompañaron durante cada etapa de mi formación escolar.

Agradecimientos

Agradezco al Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT) por el apoyo brindado durante la primera etapa de mis estudios de maestría. Asimismo, extiendo mi agradecimiento a la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), que a partir del 1 de enero de 2025 sustituyó a CONAHCYT como parte de una reforma institucional, por continuar con el respaldo que fue fundamental para la culminación exitosa de este proyecto académico.

A mi director de tesis, el Dr. Dante Alberto Magdaleno Moncayo, por brindarme las herramientas técnicas y conceptuales necesarias para la realización de este trabajo, por facilitarme el equipo de cómputo requerido para llevarlo a cabo, y por su constante orientación y valiosa guía a lo largo de los dos años que duró este posgrado.

A mi codirector de tesis, el Dr. Julio Enrique Valencia Suárez, por sus observaciones, sugerencias y por proporcionarme las secuencias utilizadas en este proyecto.

A mi sinodal, la Dra. Haydee López Rodríguez, quien, durante del posgrado, escuchó en diversas ocasiones mis presentaciones, brindándome orientaciones claves para el adecuado enfoque del proyecto.

A mis sinodales, el Dr. Rubén César Villarreal Sánchez y el Dr. Priscy Alfredo Luque Morales por sus retroalimentaciones durante los avances de este trabajo, las cuales contribuyeron al fortalecimiento de mis resultados y al enriquecimiento académico de este proyecto.

A la Universidad Autónoma de Baja California y su programa de posgrado MYDCI, por brindarme un espacio y las condiciones necesarias para llevar a cabo mis estudios de maestría.

A mis compañeros de maestría y doctorado que conocí en los cubículos de posgrado, por ser una guía constante y una grata compañía durante el desarrollo de este trabajo.

Índice

1 Introducción 12
2 Marco teórico 21
2.1 El cloroplasto: origen, evolución y diversidad 21
2.1.1 ¿Qué es el cloroplasto? 21
2.1.2 Origen del cloroplasto 21
2.1.3 Evolución del cloroplasto 22
2.1.4 Genoma de cloroplasto (plastoma) 24
2.1.5 Linaje verde 24
2.1.6 Género *Dunaliella* 26
2.2 Sistema de expresión 27
2.2.1 Transcripción en cloroplastos 27
2.2.2 RNA polimerasa y factores de transcripción 27
2.2.3 Promotores 29
2.2.4 Edición postranscripcional 32
2.2.5 Traducción en cloroplastos 35
2.2.6 Edición postraduccional 35
2.3 Herramientas de secuenciación 35
2.3.1 Generaciones en herramientas de secuenciación 36
2.3.2 Herramientas Next-Generation Sequencing (NGS) 36
2.3.3 Secuenciación illumina 38
2.4 Algoritmos de ensamble 39
2.4.1 Algoritmo Greedy (codicioso) 39
2.4.2 Algoritmo Overlap-Layout-Consensus (OLC) 39
2.4.3 Grafos de Bruijn 40
3 Antecedentes 41
4 Problema de investigación 46
5 Hipótesis 47
6 Justificación 47
7 Objetivos de investigación 47
7.1 Objetivo general 47

7.2 Objetivos específicos 47
8 Metodología 48
8.1 Limpieza y filtrado: 48
8.2 Alineación de lecturas cortas 48
8.3 Ensamble de novo: 49
8.4 Ensamble por referencia: 50
8.5 Anotación del genoma: 51
8.6 Análisis de sintenia 52
8.7 Predicción de regiones promotoras: 53
9 Resultados y discusión: 55
9.1 Secuencias sin procesar: 55
9.1.1 Basic Statistics 56
9.1.2 Per base sequence quality 57
9.1.3 Per tile sequence quality 58
9.1.4 Per sequence quality scores 60
9.1.5 Per base sequence content 61
9.1.6 Per sequence GC content 62
9.1.7 Per base N content 63
9.1.8 Sequence length distribution 64
9.1.9 Sequence duplication levels 65
9.1.10 Overrepresented sequences 66
9.1.11 Adapter Content 66
9.2 Secuencias limpias: 67
9.2.1 Basic Statistics 67
9.2.2 Per base sequence quality 68
9.2.3 Per tile sequence quality 69
9.2.4 Per sequence quality scores 70
9.2.5 Per base sequence content 71
9.2.6 Per sequence GC content 72
9.2.7 Per base N content 73
9.2.8 Sequence lenght distribution 74
9.2.9 Sequence duplication levels 75

9.2.10 Overrepresented sequences 76
9.2.11 Adapter content 76
9.3 Ensamble de novo 77
9.4 Ensamble por referencia: 82
9.5 Llamado de variantes 85
9.6 Anotación del genoma 87
9.7 Análisis de sintenia 92
9.8 Búsqueda de promotores 94
10 Conclusión 99
11 Referencias 100

Lista de figuras

Figura 1. Mercado global de proteínas recombinantes 14
Figura 2. Esquema de decisión para la selección de sistema de expresión de proteínas recombinantes 15
Figura 3. Vista esquemática de la evolución de los cloroplastos en la historia de las eucariotas 23
Figura 4. Reconstrucción consenso de las relaciones entre algas verdes, basada en datos moleculares 25
Figura 5. Representación esquemática de los mecanismos de control de la expresión de genes en el cloroplasto de plantas 32
Figura 6. Mecanismo de acción de proteínas penta, tetra y octotricopéptidos 33
Figura 7. Esquema general de secuenciación illumina 38
Figura 8 Diagrama de venn que representa el enfoque de anotación génica que se utilizará en esta tesis 52
Figura 9. Diagrama de flujo de la metodología 54
Figura 10. Resumen de módulos evaluados por fastqc 55
Figura 11: Tabla generada por fastqc que contiene datos generales de las secuencias evaluadas 56
Figura 12: Gráfico de calidad promedio por posición de base en las lecturas de secuenciación 57
Figura 13: Desviación de la calidad media de cada baldosa (tile) en función de la posición de las lecturas 58
Figura 14: Grafico de puntuaciones de calidad por secuencia 60
Figura 15: Grafico de contenido de bases por posición en el total de las lecturas 61
Figura 16: Grafico de contenido de gc en toda la longitud de cada secuencia 62
Figura 17: Grafico de contenido de n en todas las bases 63
Figura 18: Distribución de la longitud de secuencias 64
Figura 19: Niveles de duplicación de secuencias 65
Figura 20: Tabla de secuencias sobrerrepresentadas; en esta biblioteca de datos no hay secuencias sobrerrepresentadas 66
Figura 21: Grafico de contenidos de adaptadores 66
Figura 22: Resumen general de calidad de las secuencias después del filtrado. 67
Figura 23: Gráfico de calidad promedio por posición de base en las lecturas de secuenciación después del filtrado 68
Figura 24: Desviación de la calidad media de cada baldosa (tile) en función de la posición de las lecturas después del filtrado. 69
Figura 25: Grafico de puntuaciones de calidad por secuencia después del filtrado. 70
Figura 26: Grafico de contenido de bases por posición en el total de las lecturas después del filtrado 71
Figura 27: Grafico de contenido de gc en toda la longitud de cada secuencia después del filtrado 72
Figura 28: Grafico de contenido de n en todas las bases después del filtrado 73

Figura 29: Distribución de la longitud de secuencias después del filtrado. 74
Figura 30: Niveles de duplicación de secuencias después del filtrado. 75
Figura 31: Tabla de secuencias sobrerrepresentadas después del filtrado. 76
Figura 32: Grafico de contenidos de adaptadores después del filtrado. 76
Figura 33: Gráfico “per base sequence content” con sesgo (lado izquierdo) y sin sesgo (lado derecho) 78
Figura 34: Tabla generada por fastqc donde se resumen los datos generales de calidad y características de las secuencias provenientes de *Dunaliella salina* cepa noruega, después de ser limpiadas 80
Figura 35: Imagen de los grafos generados por los ensambladores contra la vista típica de un cloroplasto con una estructura quadripartida. A) estructura quadripartida típica de cloroplasto, b) grafo generado por spades, c) grafo generado por velvet. 81
Figura 36. Representación gráfica de la alineación de las lecturas contra los genomas de referencia sq y ccap19/18 con histograma de cobertura. 83
Figura 37: Captura de pantalla del programa de visualizador de alineaciones tablet, en la posición correspondiente al único snp detectado entre las lecturas del plastoma de *Dunaliella salina* cepa noruega contra el plastoma de la cepa sq. 86
Figura 38: Captura de pantalla del alineador tablet, visualizando la alineación de lecturas wgs de *Dunaliella salina* cepa noruega contra el plastoma de referencia sq, en la región correspondiente al extremo del gen petg que se perdió por tener una profundidad <10x. 89
Figura 39: Imagen de la secuencia del gen rps9 antes del snp (imagen a) con su respetiva búsqueda en ncbi e imagen de la secuencia del gen rps9 después de aplicar el snp y su respectiva búsqueda en ncbi. 90
Figura 40: Arquitectura del plastoma de *Dunaliella salina* cepa noruega. 91
Figura 41: Análisis de sintenia entre genomas de cloroplasto de cepas de *Dunaliella salina*. 92
Figura 42: Análisis de sintenia del genoma del cloroplasto entre cepas de *Dunaliella salina* ccap19/18 93
Figura 43: Análisis de sintenia del genoma del cloroplasto entre cepas de *Dunaliella salina* sq y noruega 94

Lista de tablas

Tabla 1. Metabolitos primarios de alto valor en microalgas 13
Tabla 2. Comparación de distintos sistemas de expresión de proteínas recombinantes..... 18
Tabla 3. Comparación de la ingeniería del genoma nuclear y del cloroplasto en microalgas para la producción de proteínas recombinantes. 20
Tabla 4. Características generales compartidas entre cianobacterias y cloroplastos utilizadas como evidencia base de la teoría endosimbiótica 22
Tabla 5. Comparación de maquinaria utilizada en la transcripción dentro del cloroplasto entre *Embryophytes* y *Chlamydomonadales* 29
Tabla 6. Con regiones consenso -35 y -10 de genes de cloroplasto en algas *Chlamydomonadales* 31
Tabla 7. Mecanismos de edición postranscripcionales entre *Chlamydomonadales* y *embryophytes* 34
Tabla 8. Generaciones en herramientas de secuenciación. 36
Tabla 9. Información acerca de distintas herramientas ngs 37
Tabla 11. Proteínas recombinantes producidas en el cloroplasto de la microalga modelo *Chlamydomonas reinhardtii* 43
Tabla 12. Proteínas recombinantes producidas en el cloroplasto de microalgas chlamydomonadales no modelo 44
Tabla 13. Comparación de los plastomas pertenecientes a *Dunaliella salina cp sq y ccap19/18* 45
Tabla 14. Comparación de alienaciones generadas con secuencias wgs de *Dunaliella salina* cepa noruega 77
Tabla 15. Tabla comparativa de la cantidad y porcentaje de lecturas alineadas de *Dunaliella salina* cepa noruega wgs al cloroplasto, utilizando distintos alineadores. 79
Tabla 16. Comparación de características de los ensambles generados con spades y velvet. 81
Tabla 17. Variantes genómicas típicas 85
Tabla 18: Variantes detectadas con bcftools sobre la alineación de lecturas wgs de *Dunaliella salina* cepa noruega contra el plastoma de referencia *Dunaliella salina* cepa sq 86
Tabla 19. Promotores pep putativos dependientes de -35 y -10, con elementos cis-reguladores inducibles 96
Tabla 20. Promotores pep putativos dependientes solo de -10, con elementos cis-reguladores inducibles 98

1 Introducción

Todos los organismos en el planeta tierra se pueden clasificar por sus mecanismos para la obtención de materias prima y energía, existen organismos heterótrofos que se alimentan de una fuente externa de compuestos orgánicos (carbohidratos, lípidos y proteínas), a su vez existen organismos autótrofos que se alimentan con el CO_2 del ambiente para usarlo como molécula orgánica en la construcción de tejidos. Entre los organismos autótrofos existen los organismos quimioautótrofos y los fotoautótrofos, los cuales fabrican moléculas orgánicas con la energía obtenida a partir de moléculas inorgánicas (como lo son el amoniaco, sulfuro de hidrogeno o nitritos) y de la radiación emitida por el sol, respectivamente. Los fotoautótrofos incluyen plantas, algas eucariotas, varios protistas flagelados y miembros de varios grupos de procariotas, siendo los responsables de alimentar las actividades de la mayoría de los organismos en la tierra [1].

Como se mencionó anteriormente, las algas eucariotas, son un grupo organismos fotoautótrofos, estos son capaces de vivir en todos los hábitats acuáticos (océanos, ríos, estanques y aguas residuales), además, cuentan con la capacidad de adaptarse a una variedad de condiciones ambientales (temperatura, salinidad, pH, intensidad de la luz, etc) [2], [3]. Tienen un papel vital en los ecosistemas acuáticos, siendo la base de la cadena alimenticia para todos los organismos acuáticos, además, ayudan a mitigar la eutrofización, mejoran la acidificación de los océanos, proveen hábitat y protegen las costas [4]. Pueden ser clasificadas en función de su tamaño, siendo las macroalgas organismos multicelulares visibles a simple vista mientras que las microalgas son organismos unicelulares no visibles a simple vista [5].

Macroalgas y microalgas no solo difieren en tamaño y en la presencia de estructuras diferenciadas, sino también en su composición, de forma general, las microalgas contienen mayores concentraciones de lípidos y proteínas que las macroalgas, mientras que las macroalgas tienen mayores concentraciones de carbohidratos [6].

Las microalgas tienen múltiples aplicaciones, entre las que se encuentran su uso directo como suplementos alimenticios, uso de sus componentes para cosméticos y medicamentos, uso

directo como herramientas para biorremediación, fuentes de enzimas de uso industrial, entre otras [7], [8].

La diversidad de estas aplicaciones se debe a su composición química, ya que las microalgas producen importantes niveles de metabolitos primarios (tabla 1) y metabolitos secundarios de alto valor, estos últimos están asociados con beneficios a la salud, gracias a que se generan como un mecanismo de defensa en contra del estrés ambiental, actuando como antioxidantes, antimicrobianos, antitumorales, anticancerígenos y antiinflamatorios [9].

Tabla 1. Metabolitos primarios de alto valor en microalgas

Metabolito	Descripción
Ácidos grasos poliinsaturados (PUFA)	Los ácidos grasos poliinsaturados no pueden ser producidos por el cuerpo humano, de ahí la importancia de obtenerlos a través del consumo de alimentos. Se dividen en dos grupos, omega-3 y omega-6. El valor de las microalgas para la salud se puede dirigir en parte a su composición PUFA, que se ha demostrado que promueve la salud del sistema nervioso, sistema cardiovascular y protegen contra la diabetes.
Polisacáridos	En microalgas los polisacáridos funcionan como agentes protectores, moléculas estructurales y reservas de energía, y se dividen en pectinas, proteínas de glicol y polisacáridos sulfatados (SPS). Entre los polisacáridos el más ampliamente informado es el grupo sulfatado que cuenta con beneficios antiinflamatorios.
Vitaminas	Son elementos indispensables para el correcto desarrollo humano y solo pueden obtenerse mediante la ingesta. Las microalgas son una excelente fuente potencial de vitaminas. Aunque no produzcan vitamina A, pueden acumular sus precursores como los carotenos.
Péptidos	Son aminoácidos de cadena corta (20 a 50 unidades), el 50% del mercado global de proteínas y péptidos proviene de plantas terrestres, pero puede ser reemplazado por proteínas de microalgas e insectos.

Nota. Tabla elaborada con información extraída de [9].

Además de ser una fuente de productos de alto valor, las microalgas presentan ventajas al ser utilizadas como plataforma de expresión de proteínas recombinantes [10], Este campo representa un mercado en crecimiento, habiendo generado cantidades aproximadas de 3.02 mil millones de dólares en 2024, con una tasa de crecimiento anual proyectada en 10.2% de 2025 a 2030 (Figura 1) [11].

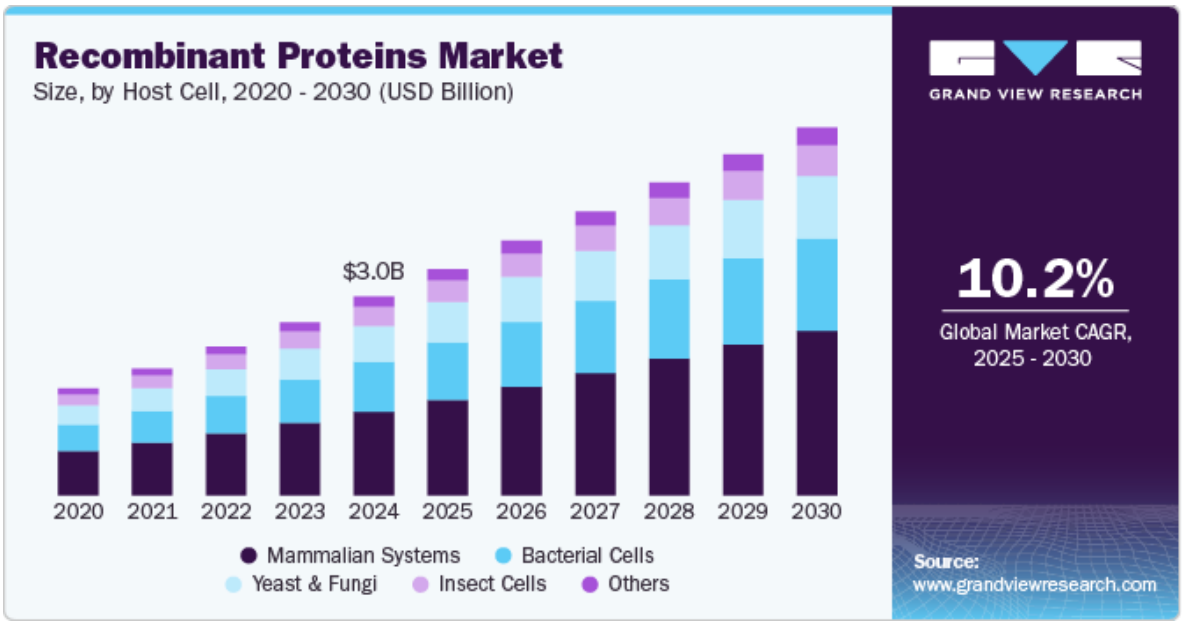

Figura 1. Mercado global de proteínas recombinantes, extraída de [11]

Tradicionalmente la expresión de proteínas recombinantes se realiza en bacterias, levaduras, hongos, líneas celulares de mamíferos y en plantas, cada uno con sus respectivas ventajas y desventajas [12]. El equipo de Schütz et al., en el 2023 generó una guía concisa para determinar la mejor plataforma de expresión dependiendo de cada caso en específico, está guía se generó con los datos obtenidos a través de una encuesta aplicada a más de 60 especialistas en la producción de proteínas, pertenecientes a la asociación para la producción y purificación de proteínas en Europa, resumiéndolo en 4 preguntas (Figura 2), de esta guía se destaca como en función de las características de la proteína y de las características de la producción, se determina el sistema de expresión de proteínas recombinantes más conveniente, sin embargo, aunque sea un reporte de hace 2 años, no se toman en cuenta las microalgas como plataforma de expresión.

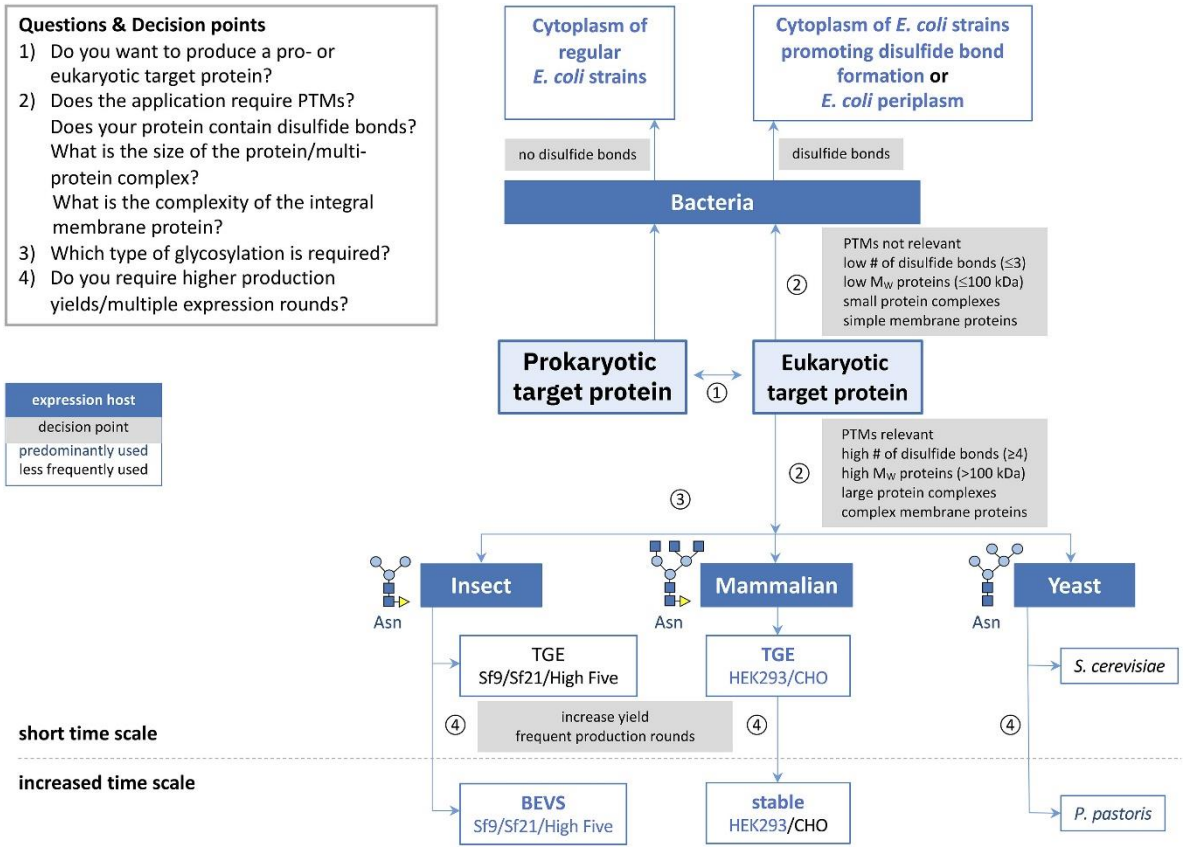

Figura 2. Esquema de decisión para la selección de sistema de expresión de proteínas recombinantes, extraído de [13]

Según el reporte de mercado generado por Grand View Research en 2024, la mayor parte del mercado de proteínas recombinantes es abarcado por las proteínas terapéuticas, dominando un 34.2% en 2024, esto se debe a la prevalencia de enfermedades como diabetes, cáncer, anemia, desordenes infecciosos, entre otras, siendo este tipo de proteínas un tratamiento muy efectivo en contra de ellas. La federación internacional de diabetes, informa que para 2030 habrá más de 643 millones de diabéticos en todo el mundo, aumentando su esta cifra a 783 millones para 2045, por lo que se espera que la demanda de este tipo de proteínas crezca, generando una continua búsqueda por mejoras o alternativas que cumplan con estas necesidades.

En función de la proteína y la metodología de producción deseada se determina el tipo de plataforma de expresión. Para la producción de proteínas eucariotas (siendo el mayor tipo de

proteína utilizada en el la industria farmacéutica) se necesitan sistemas con la capacidad de realizar modificaciones post-traduccionales (PTMs, Post-Translational modification), siendo las dos más importantes la glicosilación, cuando un glicano (cadena de azúcares) se une al grupo amino de un residuo de asparagina en la proteína (N-glicosilación) o se une a residuos de serina o treonina (O-glicosilación) (ambos presente solo en eucariotas) y los puentes disulfuro. Estas modificaciones son necesarias ya que logran que la proteína se pliegue adecuadamente, incrementando su estabilidad y permitiendo su función [14].

Para proteínas que no requieren modificaciones postraduccionales y poseen una cantidad limitada de puentes disulfuro (alrededor de 3 o menos), es posible utilizar sistemas de expresión bacterianos. Por otro lado, para expresar proteínas con modificaciones postraduccionales no es suficiente con que la plataforma sea del tipo eucariota, en su lugar, es necesario verificar si el tipo de modificación postraduccional sea lo más similar al necesario para el correcto funcionamiento de la proteína a producir. Por ejemplo, en el caso de producir proteínas terapéuticas para humanos, las células de insectos, aunque producen la N-glicosilación, sus glicanos difieren en composición con los generados por humanos, por otra parte, las levaduras son capaces de realizar N- y O-glicosilación, pero con glicanos del tipo hipermanosa (con muchas manosas) lo que puede generar respuestas inmunes, por último, las células de mamíferos (grupo al que pertenecemos) generan sus glicanos de la forma más similar a nosotros, por lo que para proteínas donde la glicosilación sea crítica, este sistema de expresión se vuelve el mejor candidato [15]. Esta idea coincide con el reporte de mercado generado por Grand View Research en 2024, donde la mayor parte de los ingresos del mercado de proteínas recombinantes a nivel global es producida en células de mamíferos, con un 42.1% del mercado global en 2024, esto es gracias a la similitud de sus modificaciones postraduccionales con las presentes en humanos.

La principal línea celular de mamíferos para la producción de proteínas recombinantes son las células de ovario de hámster chino (CHO), sin embargo, aunque sean muy utilizadas, aún tienen bastantes limitaciones, como la baja productividad, restricciones en el crecimiento de producción, inestabilidad en la expresión y baja resistencia al estrés del cultivo, altos costos

de expresión, sensibilidad a la contaminación por priones y virus humanos, además de un proceso de purificación largo y laborioso [16].

Una alternativa para la producción de proteínas recombinantes con modificaciones postraduccionales se ve en las algas eucariotas, muchas especies de este grupo producen proteínas con modificaciones mínimas en el patrón universal de glicosilación, presente en proteínas humanas, lo que facilita la glicoingeniería (modificación de las vías nativas de la glicosilación para imitar las rutas humanas) en microalgas en comparación con otras plataformas de expresión de bajo costo más establecidas, como las levaduras [17]. Además, algunas microalgas son consideradas como organismos GRAS (Generally Recognized As Safe), el cual es un estatus dado por la FDA (Food and Drug Administration) a cualquier sustancia o químico, incluyendo organismos enteros, que son considerados seguros para consumo humano [18].

Las microalgas se colocan como una plataforma de expresión de proteínas recombinantes emergente, con un amplio abanico de ventajas, empezando por el hecho de que crean su biomasa a partir del dióxido de carbono del ambiente y la luz solar, reduciendo los costos de producción en comparación con los sistemas de expresión heterotróficos [19]. Además, la producción de proteínas recombinantes en microalgas puede ser acompañada con la mitigación de CO_2, ser utilizada en el tratamiento de aguas residuales y/o biorremediación [20].

En comparación con bacterias, las microalgas eucariotas pueden producir proteínas más complejas, es decir, que necesiten mecanismos de edición postraduccional más extensos. Por otra parte, aunque comparta el mismo sistema de expresión que las levaduras (eucariota), poseen una capacidad más desarrollada para modificar proteínas. En comparación con las plantas terrestres, las microalgas tienen tiempo de cultivo más corto y menos restrictivo a las temporadas del año. Comparadas con células de mamífero, las microalgas producen proteínas a un costo menor y pueden ser cultivados a gran escala [21]. Además, algunas microalgas están clasificadas como organismos GRAS, es decir, pueden ser consumidas directamente, eliminando un proceso adicional de purificación, reduciendo costos [22].

En este sentido, las microalgas ofrecen un abanico de ventajas en el ámbito de producción de proteínas recombinantes, en comparación con los demás sistemas disponibles (tabla 2).

Tabla 2. Comparación de distintos sistemas de expresión de proteínas recombinantes.

Parámetro	Microalgas	Plantas	Bacterias	Levaduras	Células de mamífero
Costo operación	Bajo	Bajo	Bajo	Medio	Muy alto
Coto de escalamiento	Medio	Muy bajo	Alto	Alto	Muy alto
Velocidad de crecimiento	Rápida	Lento	Rápido	Rápido	Lento
Plegamiento de proteína	Alto	Alto	Bajo	Medio	Medio
Rendimiento de proteína	Alto	Alto	Alto	Medio	Medio
Seguridad	Alta	Alta	Baja	Desconocida	Bajo
Tamaño de gen	Sin límite	Sin limite	Desconocido	Desconocido	Limitado
Sensibilidad al estrés cortante	Bajo	N/A	Medio	Medio	Alto

Tabla elaborada complementando tablas de [21] y [23], este último utilizando información de [12].

La microalga más estudiada para la expresión de proteínas recombinantes es *Chlamydomonas reinhardtii*, siendo considerada un organismo modelo para la ingeniería genética (similar a *Escherichia coli* en bacterias), siendo utilizada en la expresión de más de 100 proteínas recombinantes [24].

Otra microalga qué ganando interés como sistema de expresión de proteínas recombinantes es *Dunaliella salina.* Gracias a su estrecha relación evolutiva con *Chlamydomonas reinhardtii*, ambas perteneciendo al grupo de *Chlamydomonadales* [25], muchos estudios realizados en este organismo modelo pueden servir como referencia para investigaciones en *Dunaliella salina*.

De entre todas las especies del género *Dunaliella*, *Dunaliella salina* destaca por ser la más halotolerante. Se puede encontrar en una amplia gama de hábitats marinos, como los océanos, los lagos de salmuera, marismas y lagunas saladas cerca del mar, sobre todo en los cuerpos de agua que contienen más de 2 M de cloruro de sodio y altos niveles de magnesio. Esta especie es capaz de adaptarse a altas salinidades y tolerar un amplio rango de condiciones ambientales, incluyendo altas intensidades luminosas y temperaturas, así como deficiencias de nitrógeno y/o fósforo [26].

Dunaliella salina es capaz de acumular grandes cantidades de β-Caroteno y glicerol. Siendo explotada comercialmente con fines de uso farmacéuticos, cosméticos y agroalimentarios. Además de ser una fuente de grandes cantidades de triglicéridos [27]. Esta microalga es reconocida como un organismo GRAS, por lo que puede ser utilizada como una vía de administración oral de proteínas [28].

En microalgas, existen 3 genomas disponibles para la producción de proteínas recombinantes, el nuclear, mitocondrial y cloroplástico. De estos 3 genomas, los más utilizados para expresión de proteínas recombinantes son el nuclear y cloroplástico [29], cada una de sus ventajas y desventajas vienen en la tabla 3.

Tabla 3. Comparación de la ingeniería del genoma nuclear y del cloroplasto en microalgas para la producción de proteínas recombinantes.		
	Ingeniería de genoma en núcleo	**Ingeniería de genoma en cloroplasto**
Método de transformación	Electroporación, biolística, agitación con perlas de vidrio, bigotes de silicio	Biolístico.
Integración de genes	Unión final no homóloga	Recombinación homologa
Niveles de acumulación de proteínas	Valores máximos de 0.25% de la proteína soluble total.	Alto: 1-21% de la proteína soluble total
Modificaciones post-traduccionales	Formación de enlaces disulfuro, fosforilación, glicosilación	Formación de enlaces disulfuro, fosforilación
Localización de proteínas	Pueden dirigirse a varios lugares: citoplasma, núcleo, cloroplasto, ER, mitocondrias, secreción	Permanecen en el cloroplasto
Maquinaria de expresión de genes	Eucariota	Procariota
Expresión de genes inducible	Nutriente, químico, fisiológico	Luz inducible
Tabla extraída de [30].		

Para poder utilizar el cloroplasto *Dunaliella salina* cepa noruega como plataforma de expresión de proteínas recombinantes, es fundamental caracterizar su genoma. Para ello, es necesario secuenciar, ensamblar y anotar el genoma del cloroplasto, esto con el fin de identificar las regiones para la inserción de genes de interés. Además, para que la producción sea inducible, es necesario identificar y caracterizar los promotores endógenos del cloroplasto. En este contexto, el presente trabajo tiene como objetivo proporcionar las herramientas necesarias para la producción de proteínas recombinantes, ensamblando y

anotando el genoma del cloroplasto, así como identificando los promotores endógenos y sus mecanismos de inducción.

2 Marco teórico

2.1 El cloroplasto: origen, evolución y diversidad

2.1.1 ¿Qué es el cloroplasto?

El cloroplasto es un organelo presente en las células eucariotas fotosintéticas con dimensiones aproximadas de 2-4 µm de ancho y de 5-10 µm de largo. La cubierta exterior cuenta con dos membranas, la membrana de envoltura externa y la membrana de envoltura interna separadas por un estrecho espacio (la membrana externa derivaría de la vesícula que envolvió originalmente a la cianobacteria). La envoltura externa cuenta con canales relativamente grandes, donde existe cierta selectividad hacia diversos solutos, confiriéndole la característica de no ser libremente permeable. Por otro lado, la membrana interna es altamente permeable, movilizando sustancias a través de ella gracias a una variedad de transportadores. Dentro del cloroplasto existe una tercera membrana interna distribuida en forma de una red independiente de cientos de estructuras saculares (sacos membranosos) llamadas "tilacoides". El conjunto de tilacoides dispuestos en pilas interconectadas se denomina "grana". La mayoría de la maquinaria utilizada en la fotosíntesis está integrada en la membrana de los tilacoides. Por otra parte, la región externa a los tilacoides e interior a la envoltura de doble capa se denomina "estroma", la cual contiene enzimas responsables de la síntesis de carbohidratos [31].

2.1.2 Origen del cloroplasto

En 1905 se propuso una teoría por Mereschkowsky que busca explicar el nacimiento del cloroplasto. La teoría es llamada "Teoría endosimbiótica", en ella se propone el origen de este organelo a partir de una antigua célula cianobacteriana, la cual fue fagocitada por una célula eucariota primitiva. Dicha célula, en lugar de ingerir y utilizar los compuestos orgánicos propios de la cianobacteria, empezó a aprovechar los beneficios que le podía ofrecer la misma, formando una estrecha relación simbiótica. De esta relación simbiótica obtuvieron ventajas que les permitieron una mayor supervivencia al ambiente, formando con el paso del tiempo al cloroplasto [32].

Durante la primera mitad del siglo XX esta teoría no tuvo fundamento, sin embargo, con el tiempo surgieron descubrimientos que constituyeron la base de la teoría endosimbiótica del origen del cloroplasto (Tabla 4).

Tabla 4. Características generales compartidas entre cianobacterias y cloroplastos utilizadas como evidencia base de la teoría endosimbiótica

	Cianobacteria	Cloroplasto
Fotosíntesis oxigenada	Si	Si
Número de fotosistemas	2	2
Clorofila como pigmento esencial en la fotosíntesis	Si	Si
Presencia de membranas tilacoides	Si	Si
Genes implicados en la fotosíntesis	Si	Si
Presencia de ADN circular	Si	Si
Maquinaria de replicación de ADN	Si	Si
ARN polimerasa bacteriana	Si	Si
Ribosomas 70s (a diferencia de 80s presente en eucariotas)	Si	Si

Tabla elaborada con información de [32].

2.1.3 Evolución del cloroplasto

Del primer evento endosimbiótico ocurrido hace aproximadamente entre 1 y 1.5 millones de años, resultó la formación del cloroplasto, lo que dio lugar a tres linajes distintos: las *Glaucófitas*, caracterizadas por la presencia de una pared celular de peptidoglicano entre las membranas del cloroplasto, una característica típica de las células cianobacterianas. *Rhodophyta* (algas rojas), notablemente diversas, con especies que van desde células pequeñas no flageladas hasta macroalgas marinas. Se caracterizan por tener cloroplastos con clorofila a y ficoeritrina, este último siendo un pigmento accesorio que captura la luz en un rango de espectro poco absorbido por la clorofila, particularmente la luz verde. Por último, *Chlorophyta* (algas verdes), son un grupo extremadamente diverso y abundante, se encuentran en una variedad de entornos marinos y de agua dulce, así como en algunas áreas terrestres secas. Este linaje es el precursor de las plantas terrestres (*Embriophyta*). Los

cloroplastos de las algas verdes contienen clorofila a y b, y se distinguen por su capacidad para almacenar almidón dentro del cloroplasto. El conjunto de *Chlorophytas* y sus descendientes *Embriophytas* se denomina *Viridiplantae* (linaje verde). A partir de estos tres linajes distintos, la fotosíntesis se propagó a diversos protistas eucariotas mediante nuevos eventos endosimbióticos, que involucraron la captura de algas verdes o rojas por protistas no fotosintéticos (Figura 3) [33].

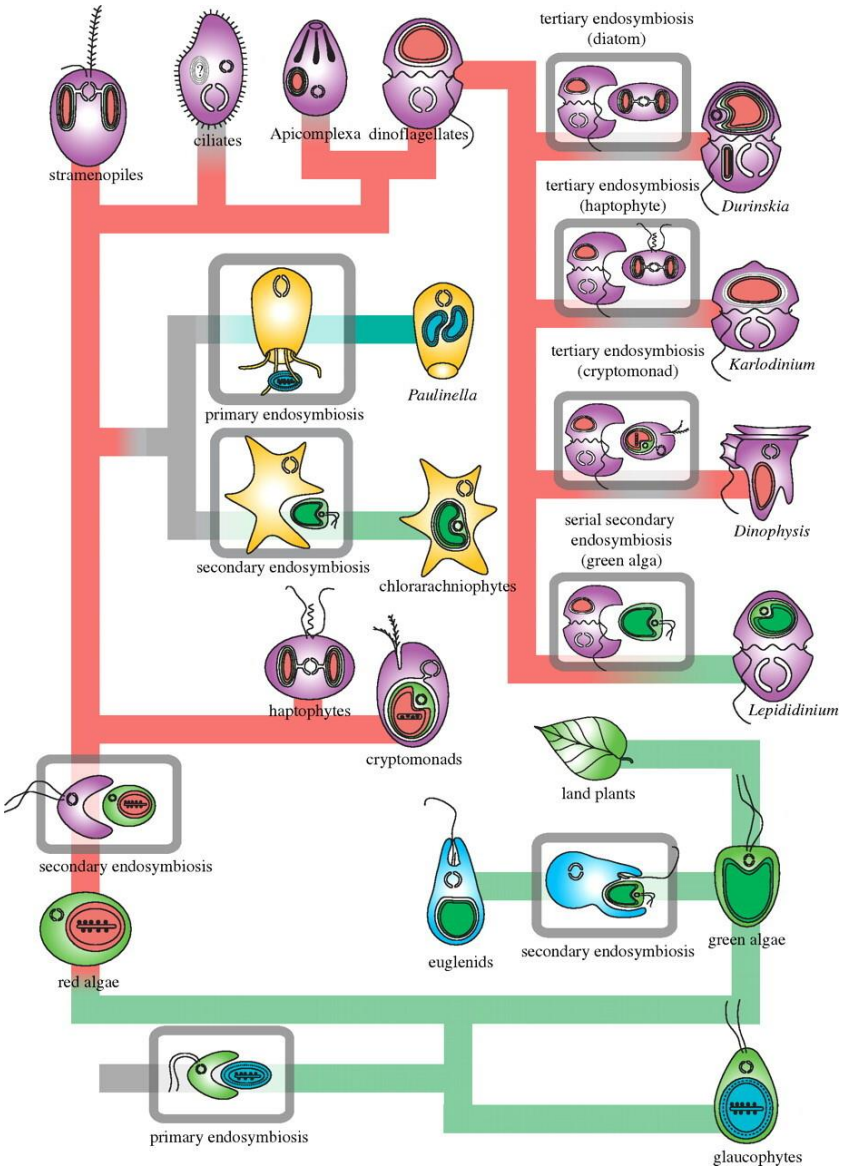

Figura 3. Vista esquemática de la evolución de los cloroplastos en la historia de las eucariotas, extraída de [33]

2.1.4 Genoma de cloroplasto (plastoma)

El tamaño del genoma del cloroplasto por lo general varía entre 120,000 y 160,000 bases nucleotídicas. Los genes dentro del cloroplasto codifican para la maquinaria fotosintética, transcripción, traducción y replicación de su propio ADN, el genoma del cloroplasto está constituido de 4 regiones delimitadas, la región larga de copia única o LSC (Large Single Copy Region), una región corta de copia única o SSC (Small Single Copy region) y dos regiones repetidas inversas (IR). Dentro de estas regiones podemos encontrar aproximadamente 130 genes, de los cuales corresponden aproximadamente a 80 proteínas, 30 ARNt y 4 ARNr, así como intrones y espaciadores intergénicos [34].

2.1.5 Linaje verde

El linaje verde está conformado por las algas verdes (*Chlorophytas*) y *Streptophytas*, a su vez, *Streptophyta* incluye *Charophytes* (un conjunto parafilético de algas de agua dulce) y plantas terrestres (*Embriophyta*). Por otra parte, *Chlorophyt*a está dividido en *Prasinophytes* (predominantemente plancton marino), que dieron lugar al núcleo de los Chlorophytes, el cual se divide en *Ulvophyceae*, *Trebouxiophyceae*,y *Chlorophyceae* (Figura 4); el núcleo de los *Chlorophytes* se caracteriza por un nuevo modo de división celular mediado por un ficoplasto, que posteriormente se perdió en *Ulvophyceae* [35].

Chlorophyceae es un grupo diverso que incluye unicélulas no móviles, móviles, colonias, filamentos ramificados y no ramificados. Dentro de esta clase existen diversas formas de reproducirse, incluyendo varios modos sexuales y asexuales. Esta clase es caracterizada por mitosis cerrada durante la división celular, citocinesis mediada por ficoplastos y diversas configuraciones del aparato flagelar de células móviles. Dicho grupo se divide en cinco clados principales, *Chlamydomonadales* y *Sphaeropleales*, compartiendo un ancestro en común y *Chaetopeltidales*, *Chaetophorales* y *Odeogoniales* compartiendo otro ancestro en común; Siendo *Dunaliella salina* miembro de los *Chlamydomonadales*; Los *Chlamydomonadales* se caracterizan por tener estructuras biflagelares (importantes para la movilidad) orientadas en sentido horario, sin embargo, en *Chlamydomonadales* de cuatro flagelos la orientación puede variar, comprenden entornos de agua dulce y terrestres, además

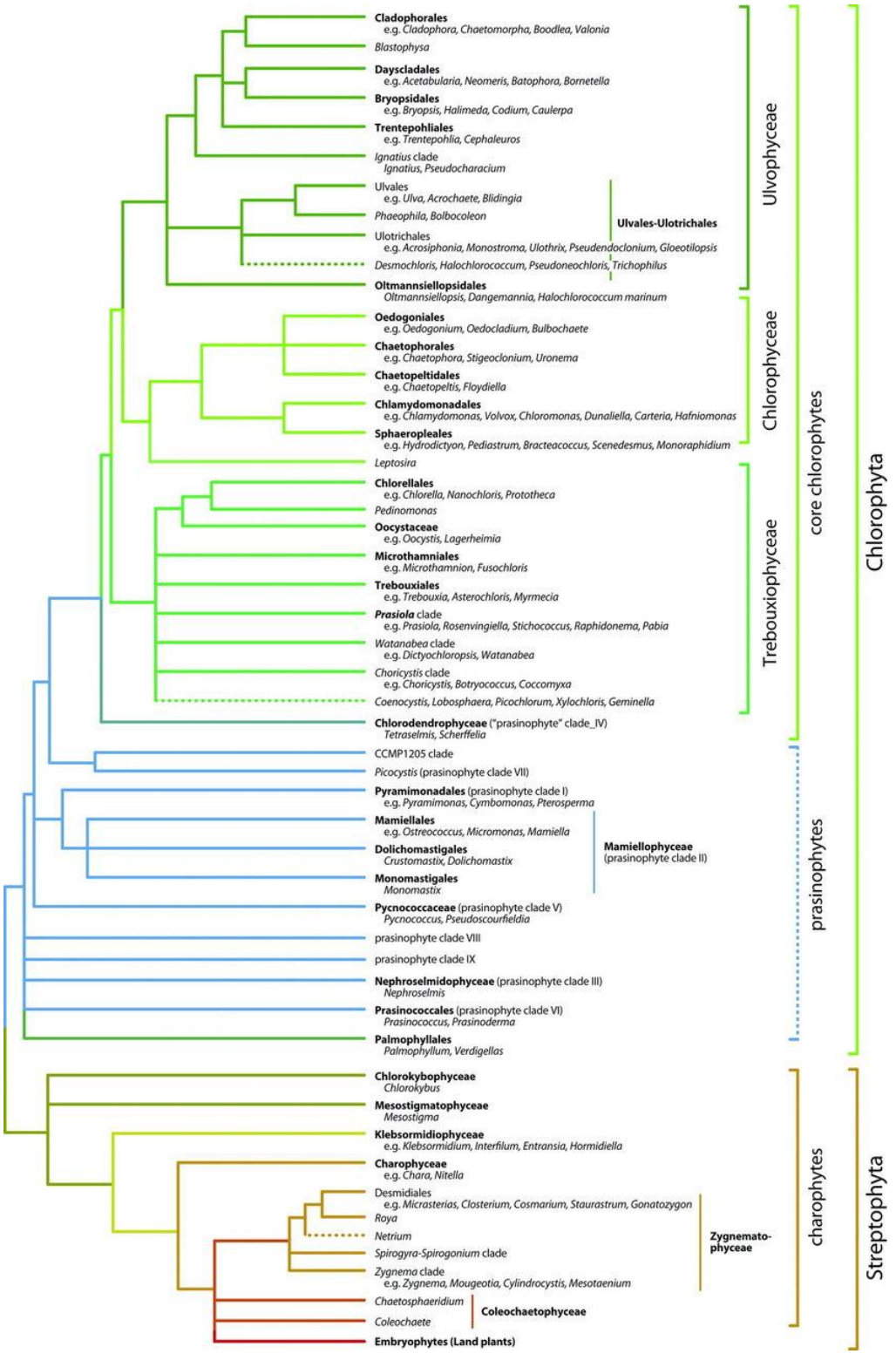

Figura 4. Reconstrucción consenso de las relaciones entre algas verdes, basada en datos moleculares; extraída de [35]

de varias especies psicofílicas (prosperan en ambientes fríos) y halotolerantes (capaces de tolerar condiciones de alta salinidad, como es el caso de *Dunaliella salina*) [36].

2.1.6 Género *Dunaliella*

El género *Dunaliella* consiste en algas verdes unicelulares, responsables de gran parte de la producción primaria en ambientes hipersalinos. El primer reporte de especies de este género fue publicado en 1838 por Michel Felix Dunal, en estanques de evaporación salina en el sur de Francia, *Dunaliella* fue nombrada en nombre de su descubridor (Michel Dunal) por Teodoresco en 1905 [37].

Las especies mejor conocidas del género *Dunaliella* son *Dunaliella salina, Dunaliella trertiolecta, Dunaliella primolecta, Dunaliella viridis, Dunaliella parva, Dunaliella acidophyla, Dunaliella media* y *Dunaliella bioculata.;* Las células de *Dunaliella* carecen de una pared celular rígida, en su lugar está encerrada por una membrana de plasma delgada y elástica, cubierta por una capa superficial de mucosidad, la capa delgada permite respuestas rápidas a los cambios osmóticos extracelulares [38]. Diferentes especies de *Dunaliella* pueden acumular cantidades significativas de valiosos químicos finos como carotenoides, glicerol, lípidos, vitaminas, minerales y proteínas [39].

Se ha reportado que *Dunaliella salina* muestra altas tasas de crecimiento en rangos de concentración de NaCl que van desde el 1 M a 3 M, obteniendo la mejor tasa de crecimiento a 2 M [40]. Para darnos una idea de lo grande que es esta tolerancia, la microalga *Chlamydomonas reinhardtii* entra en choque osmótico cuando la concentración de su medio sobrepasa 0.2 M de NaCl [41]. Aunque el rango óptimo de *Dunaliella salina* va de 1 a 3 M de NaCl, se ha reportado que es capaz de sobrevivir a ambientes con concentraciones de sal mayores a 4.5 M [42], esto gracias a la capacidad que tiene de acumular grandes cantidades de glicerol (se ha reportado que sus concentraciones de glicerol llegan hasta el 50% de su peso seco) que utiliza como osmólitos compatibles para contrarrestar la alta presión osmótica extracelular [43].

Dunaliella salina se caracteriza por producir altas concentraciones de betacaroteno bajo condiciones estresantes, acumulándose dentro de glóbulos lipídicos intracelulares en más del 10% del peso seco de la microalga [44].

2.2 **Sistema de expresión**

Desde la presentación del dogma central de la biología por Francis Crick en 1957 donde establecía la idea de que la síntesis proteica se basa en los ácidos nucleicos, la búsqueda del entendimiento de la maquinaria molecular de la transcripción y su regulación han sido un objeto importante de estudio [45].

2.2.1 Transcripción en cloroplastos

Un paso importante en el flujo de información genética es la transcripción. Durante la transcripción una hebra de DNA sirve como plantilla para la síntesis de una hebra de RNA, este proceso se logra gracias a la intervención de una enzima llamada RNA polimerasa, la cual lee cada base nitrogenada del DNA molde y coloca, una a la vez, su base complementaria en una hebra de RNA. El primer paso en la transcripción es la unión del RNA polimerasa a la plantilla de DNA, la enzima logra unirse gracias a la ayuda de proteínas adicionales llamadas factores de transcripción, los cuales les permite reconocer secuencias específicas de nucleótidos dentro de una secuencia larga de DNA.

Aunque el cloroplasto surgió de la unión endosimbiótica de una cianobacteria procariota y un organismo eucariota primitivo, muchos de los mecanismos originales en las cianobacterias fueron modificándose a lo largo del tiempo, entre ellas la transcripción.

Al igual que en bacterias, la transcripción en el cloroplasto se lleva a cabo en operones, denominados operones plástidos, los cuales también generan transcritos policistrónicos [46].

2.2.2 RNA polimerasa y factores de transcripción

La enzima más importante en el proceso de transcripción genómica es la RNA polimerasa, la cual fue descubierta en 1959 por Samuel Weiss y Leonard Gladstone mientras investigaban la fracción nuclear de células de hígado de rata trayendo consigo múltiples investigaciones

alrededor de ella [47]. En los últimos 30 años del siglo XX se desarrollaron grandes avances en el entendimiento del proceso de transcripción, como lo son la composición de las subunidades, propiedades bioquímicas y factores accesorios de la RNA polimerasa. Uno de los más importantes fue descubrir que la RNA polimerasa se une a sus factores de transcripción (σ en bacterias; TFIIA, B, D, E, F, y H en eucariotas), ayudando a reconocer y unirse a regiones reguladoras dentro del DNA, permitiendo la transcripción. Las regiones a las que se unen las RNA polimerasas con ayuda de los factores de transcripción se conocen como promotores [48].

El cloroplasto conservó y adquirió en el tiempo sus propias enzimas encargadas de la transcripción de los genes plástidos, las principales RNA polimerasas dentro del cloroplasto son: polimerasas codificadas en plástidos de origen bacteriano (PEP, por sus siglas en ingles "Plastid-encoded Polymerase") y Polimerasas codificadas nuclearmente de tipo fago (NEP, por sus siglas en ingles "Nuclear-encoded polymerase") [49].

Los genes involucrados en la formación de la PEP se encuentran dentro del plastoma, estos genes son rpoA, rpoB, rpoC1 y rpoC2, que codifican 4 diferentes subunidades, las subunidades α, β, β' y β'', esta RNA polimerasa es capaz de detectar subunidades sigma 70 (σ^{70}) presentes en *E.coli* [50]. En el cloroplasto de *Arabidopsis thaliana* se conocen 6 subunidades sigma que se unen a la PEP para iniciar la transcripción, sin embargo, en algas Chlamydomonadales, específicamente en *Chlamydomonas reinhardtii* solo se ha encontrado uno [51], al igual que este ejemplo, existen otras diferencias en el proceso de transcripción dentro del cloroplasto de *Embryophytes* y *Chlamydomonadales*, siendo uno de los más destacables la falta de NEP en *Chlamydomonadales* [52]. La tabla 5 compara las RNAp y las subunidades de *Embryophytes* y *Chlamydomonadales*.

Tabla 5. Comparación de maquinaria utilizada en la transcripción dentro del cloroplasto entre *Embryophytes* y *Chlamydomonadales*		
	Embryophytes (plantas terrestres)	*Chlamydomonadales*
PEP (plastid encoded polymerase) de origen bacteriano	Si	Si
Factores sigma	6	1
NEP (Nuclear encoded polymerase)	Si	No
Subunidad	1	---
Tabla elaborada con información extraída de: [49], [51], [52].		

2.2.3 Promotores

La mayoría de los estudios en promotores de cloroplasto van enfocados a plantas terrestres, por lo que muchos de los resultados no pueden ser aplicados a este trabajo, ya que, como se mencionó en el apartado de RNA polimerasa, las microalgas *Chlamydomonadales* solo tienen un tipo de RNA polimerasa, de origen bacteriano [52], por lo que resultados dirigidos a la NEP no aplican debido su ausencia en este tipo de algas. Al estar tan estrechamente relacionados, muchos de los estudios en microalgas como *Chlamydomonas reinhardtii*, sirven como una guía para estudios aplicados a *Dunaliella salina.*

Los promotores reconocidos por la PEP comparten similitud con los promotores bacterianos σ70, ambos contienen los elementos -35 y -10, localizados a 35 pb aguas arriba y a 10 pb aguas arriba del sitio de inicio de la transcripción respectivamente [53]. Por esta razón las regiones consenso de los elementos -35 y -10 70 de bacterias se han utilizado para la búsqueda de promotores putativos [54].

En 1992 se hizo un estudio en el cloroplasto de la microalga *Chlamydomonas reinhardtii*, en el cual se analizó la estructura de los promotores de los genes atpB y 16S rRNA i*n vivo*. Para ello construyeron vectores de transformación que contenía el promotor del gen atpB o el promotor del gen 16S rRNA junto con el gen uidA de *Escherichia coli* (proteína -glucuronidase, GUS) como reportero. La actividad de los promotores se midió con la

abundancia de transcritos de GUS. Utilizando exonucleasas eliminaron gradualmente bases en dirección 5’ -> 3’ del promotor, midiendo la actividad en cada una de las modificaciones. Se encontraron dos estructuras distintas de promotores, en el promotor del gen 16s rRNA se necesita la presencia de ambos elementos -35 y -10 para que ocurra la transcripción. El otro tipo de estructura está presente en el promotor del gen atpB, la eliminación del elemento -35 no tenía efecto en la transcripción, empezando a mostrar un declive en la cantidad de transcritos cuando la eliminación de bases pasaba las 22 pb aguas arriba del sitio de inicio de la transcripción, mostrando incluso un 3% de transcritos (siendo 100% la cantidad de transcritos sin eliminar bases del extremo 5’) cuando el promotor se extiende hasta el -10. En el tipo de promotores que comparten estructura con el promotor del gen atpB, se notó que la región consenso del elemento -10 pasó de ser un hexámero a un octámero [55].

Algunas de las regiones consenso -35/-10 de promotores identificadas en el cloroplasto de especies *Chlamydomonadales* se resumen en la tabla 6. A partir de estas secuencias se hará la búsqueda *in silico* de promotores putativos en el cloroplasto de *Dunaliella salina* cepa noruega.

Tabla 6. Con regiones consenso -35 y -10 de genes de cloroplasto en algas *Chlamydomonadales*

Título del articulo	Especie	Gen	Consenso -35	Consenso -10
Two types of chloroplast gene promoters in Chlamydomonas reinhardtii	*Chlamydomonas reinhardtii*	rrn16	TTGACA	TAAATT
		atpB	No obligatorio	TATAATAT
Sequence of the Chloroplast DNA Region of Chlumydomonas reinhardii Containing the Gene of the Large Subunit of Ribulose Bisphosphate Carboxylase and Parts of its Flanking Genes		rbcL	TTTACA	TATAAT
Analysis of heterologous regulatory and coding regions in algal chloroplasts		psbA	TTGATT	TATAAT
	Haematococcus pluvialis	psbA	TTCAGA	ATAAGT
	Dunaliella salina	psbA	TTGATA	TATTAT

Tabla elaborada con información de [55], [56], [57].

2.2.4 Edición postranscripcional

Durante la evolución el genoma del cloroplasto abandonó en gran medida la regulación de la expresión génica procariota basada en la transcripción para adaptar mecanismos postranscripcionales más elaborados, generando a su vez una estrecha coordinación espaciotemporal de la síntesis e importación de proteínas desde el núcleo, generando una red de factores codificado por el núcleo que se unen a regiones 3' y 5' no traducidas (UTR) de los ARNm, por lo tanto, el cloroplasto representa un sistema hibrido con características eucariotas como la existencia de intrones y la presencia de un extenso sistema de procesamiento de ARNm, lo que la hace capaz de soportar el plegamiento nativo de proteínas recombinantes complejas [58].

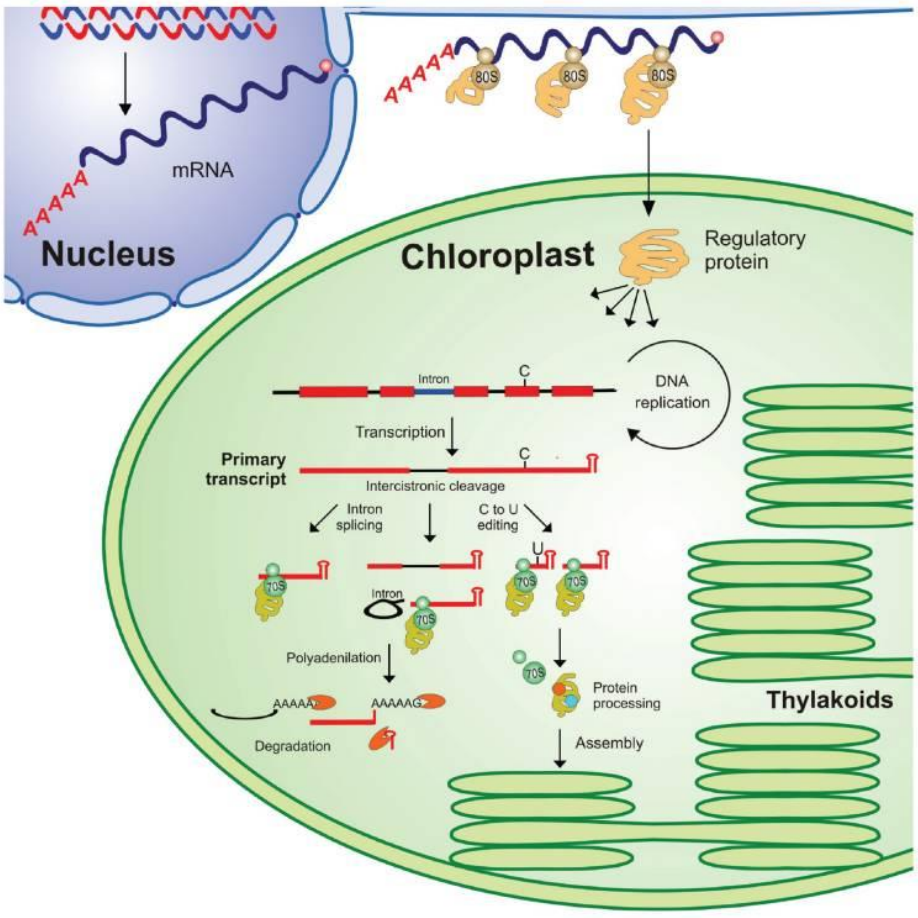

Figura 5. Representación esquemática de los mecanismos de control de la expresión de genes en el cloroplasto de plantas, extraída de [59]

Las cadenas de RNA generados en la transcripción plastida son policistrónicos, lo que significa que tienen más de un gen, sin embargo, este transcrito pasa por distintos procesos postranscripcionales antes de llegar a ser traducido. Primero que nada, el transcrito policistrónico se separara en transcritos monocistrónicos, esto se logra gracias a distintas proteínas llamadas penta, tetra y octotricopéptidos, estas proteínas son producidas nuclearmente y viajan al interior del cloroplasto (Figura 5) [59]. Una vez dentro del cloroplasto las proteínas penta, tetra y octotricopéptidos regulan la actividad de las ribonucleasas, esto lo logran uniéndose a los espacios intergénicos del transcrito, evitando así la escisión del gen (Figura 6) [60].

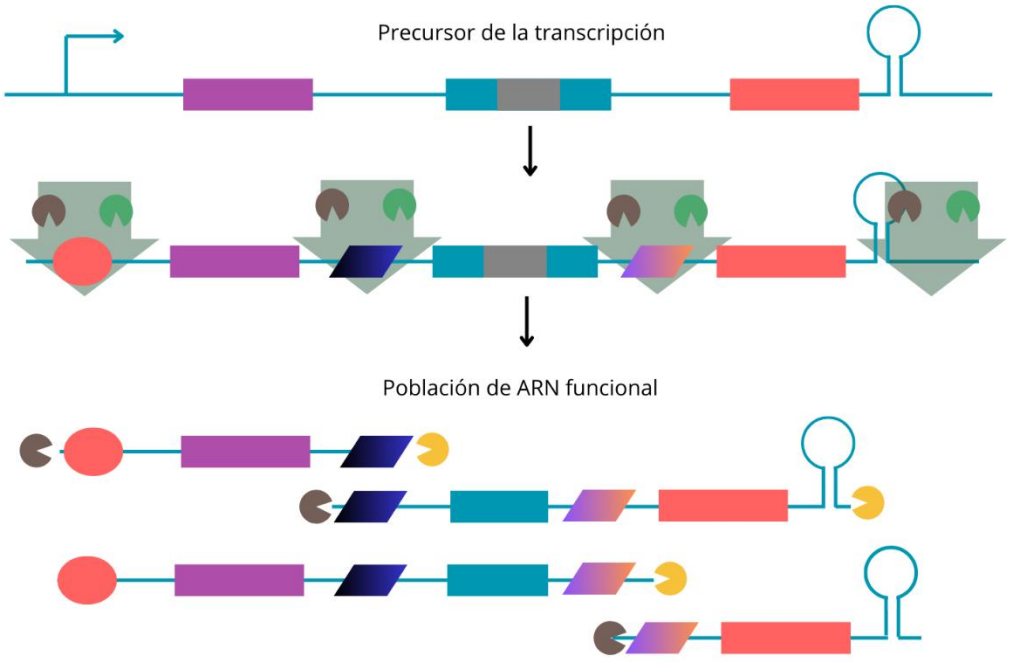

Figura 6. Mecanismo de acción de proteínas penta, tetra y octotricopéptidos, imagen extraída y editada de [60]

Además de la actividad de los penta, tetra y octotricopéptidos, existen distintos mecanismos de edición postranscripcionales. La tabla 7 resume algunos de ellos separados entre *Chlamydomonadales* y *Embryophytes*.

Tabla 7. Mecanismos de edición postranscripcionales entre *Chlamydomonadales* y *Embryophytes*				
Mecanismo	Embryophytes (plantas terrestres)	Chlamydomonadales	Descripción	Referencias
Proteínas PPR, TPR y OPR	Presente	Presente	Las proteínas PPR, TPR y OPR se unen a las regiones intergénicas de los transcritos de ARN para regular la actividad de las ribonucleasas, transformando los transcritos policistrónicos en monocistrónicos	[49], [60], [61]
Transformación de Citidina a Uridina	Presente	No presente	Se transforma el segundo nucleósido de cada codón de los transcritos de RNA, mostrando una tendencia de pasar de Citidina a Uridina. Los cambios más comunes en aminoácidos son de Pro a Leu, Ser a Leu y Ser a Phe.	[62], [63], [64]
Poly(A)	Presente	Presente	Adición de una cola poly(A) (muchas adeninas) en el extremo 3' UTR, la cual participa en la degradación eficiente de los transcritos de RNAm	[65], [66]
Poly(G)	Presente	Presente	Adición de una secuencia poly(G) (muchas guaninas) en el extremo 5' UTR, la cual impide el paso de las exorribonucleasas a lo largo de las moléculas de RNAm, lo que produce transcritos con mayor tiempo de vida.	[65], [67]
Stem loop	Presente	Presente	La mayoría de los extremos 3' UTR tienen una repetición invertida (IR) que puede doblarse para formar una estructura de bucle, este proceso postranscripcional interviene en la maduración y la estabilidad de ARN	[68]

2.2.5 Traducción en cloroplastos

La traducción en el cloroplasto es realizada en ribosomas de tipo 70s, al igual que en sus antecesoras, las cianobacterias. Los transcritos de ARN mensajero (ARNm) ya procesados (monocistronicos) llegan a los ribosomas, donde son leídos en tripletes de nucleótidos llamados codones [60]. El cloroplasto utiliza la tabla de traducción de codones número 11, reportada en NCBI.

2.2.6 Edición postraduccional

Existen notables diferencias ente cianobacterias y cloroplastos, las cuales le otorgan ventajas a este último como sistema de expresión de proteínas recombinantes. Una de estas ventajas se encuentra en la maquinaria de traducción del cloroplasto, el cual está equipado con proteínas chaperonas, como las disulfuro-isomerasas [69] y las peptidilprolil-isomerasas [70], que facilitan el correcto plegamiento de las proteínas. Gracias a estas chaperonas, el cloroplasto es capaz de producir proteínas recombinantes complejas, ya que asisten en el plegamiento de proteínas en cada etapa de su ciclo de vida.

En el genoma de *Chlamydomonas reinhardtii*, una microalga perteneciente a las algas verdes *chlamydomonadales* (al igual que *Dunaliella salina*), se han identificado genes que codifican miembros de las principales familias de chaperonas: Hsp100/Clp, Hsp90, Hsp70, Hsp60 y las pequeñas proteínas de choque térmico (sHsp), así como sus co-chaperonas GrpE y Cpn10/20 [71].

2.3 Herramientas de secuenciación

El código genético, común a todos los organismos, está conformado por una secuencia específica de bases nitrogenadas (adenina, timina, guanina y citosina), cuya disposición determina la síntesis de proteínas con funciones esenciales para la vida. Para modificar genéticamente un organismo, es fundamental conocer la secuencia exacta de estas bases en su genoma. Por ello, resulta indispensable comprender los métodos y tecnologías disponibles para la secuenciación genómica. Tener una visión clara de este panorama permite seleccionar la herramienta más adecuada, considerando sus ventajas y limitaciones según el objetivo planteado.

2.3.1 Generaciones en herramientas de secuenciación

De forma general, las tecnologías de secuenciación se dividen en generaciones. La tabla 8 aborda cada generación, destacando las características que las definen y que herramientas pertenecen a ella. A partir de la segunda generación también se les conoce como tecnologías Next-Generation Sequencing (NGS).

Tabla 8. Generaciones en herramientas de secuenciación.

	Característica principal	Principales herramientas
Primera	Son las más antiguas, producen lecturas cortas con bajos rendimientos, pero alta precisión. A menudo se utilizan para validar resultados obtenidos por herramientas de generaciones posteriores	Sanger, Maxam- Gilbert
Segunda	Generan grandes volúmenes de información en menor tiempo, esto se logra gracias a la secuenciación simultanea de miles de millones de fragmentos cortos de ADN.	Illumina, Ion torrentm 454 pyrosequencing, SOLiD.
Tercera	Capacidad de secuenciar fragmentos de ADN más largos, sin embargo, son más caras y manejan tasas de error más altas.	PacBio Single-molecule real-time sequencing (SMRT) technology, Nanopore DNA sequencing
Tabla elaborada con información de: [72]		

2.3.2 Herramientas Next-Generation Sequencing (NGS)

A los equipos desarrollados a partir de la segunda generación, se les conoce a como Next-Generation Sequencing (NGS), estas herramientas combinan las ventajas de diferentes técnicas químicas de secuenciación con diversas matrices en las cuales el proceso ocurre de forma paralela [73]. Todas las plataformas NGS se caracterizan por generar de manera simultánea millones de fragmentos de ADN [74]. Algunas de las herramientas NGS más utilizadas se encuentran en la tabla 9.

Tabla 9. Información acerca de distintas herramientas NGS					
	454 pyrosequencing	Ion Torrent	Illumina	PacBio Single-molecule real-time sequencing (SMRT) technology	Nanopore DNA sequencing
Uso	Secuenciación de lecturas cortas	Secuenciación de lecturas cortas	Secuenciación de lecturas cortas	Secuenciación de lecturas largas	Secuenciación de lecturas largas
Tecnología de secuenciación	Secuenciación por síntesis	Secuenciación por síntesis	Secuenciación por síntesis	Secuenciación por síntesis	Secuenciación por detección de impedancia eléctrica
Tipo de amplificación	Emulsión PCR	Emulsión PCR	PCR de puente	Sin PCR	Sin PCR
Principio	Detección de pirofosfato liberado durante la incorporación de nucleótidos	Principio de secuenciación de semiconductores de iones que detecta los iones de H+ generados durante la incorporación de nucleótidos	Secuenciación en fase solida en superficie inmovilizada aprovechando la formación de matrices clonales utilizando tecnología de terminador reversible patentada para una secuenciación rápida y precisa a gran escala.	Utiliza un tipo de celda especial (SMRT) que alberga numerosos pozos pequeños conocidos como zero-mode waveguides, las moléculas de ADN individual se inmovilizan dentro de estas celdas, emitiendo luz al tiempo que una polimerasa incorpora un nucleótido en tiempo real	Se basa en la linealización del ADN o ARN y su capacidad de moverse a través de un poro biológico llamado "nanopores" que tiene 8 nanómetros de ancho. La movilidad electroforética genera una señal de corriente que será utilizada para realizar la secuenciación
Longitud de lecturas	400 - 1000	200 - 400	36 - 300	10,000 – 25,000	10,000 – 30,000
Limitaciones	Puede contener errores de secuenciación de eliminación e inserción debido a la determinación ineficiente de la longitud del homopolímero.	Cuando se secuencian secuencias de homopolímeros, puede provocar una pérdida de intensidad de la señal	En caso de una sobrecarga de la muestra, la secuenciación puede resultar en una superposición de señales, aumentando la tasa de error hasta el 1%.	Costo superior en comparación con otras plataformas de secuenciación	La tasa de error puede dispararse arriba del 15%, comparado con secuenciadores de lecturas cortas tiene una menor precisión
Tabla extraída y editada de: [72]					

2.3.3 Secuenciación illumina

Para la generación de los datos utilizados en esta tesis se utilizó la secuenciación de Illumina, de forma general, la secuenciación de Illumina se puede dividir en cuatro pasos principales (figura 7) [75].

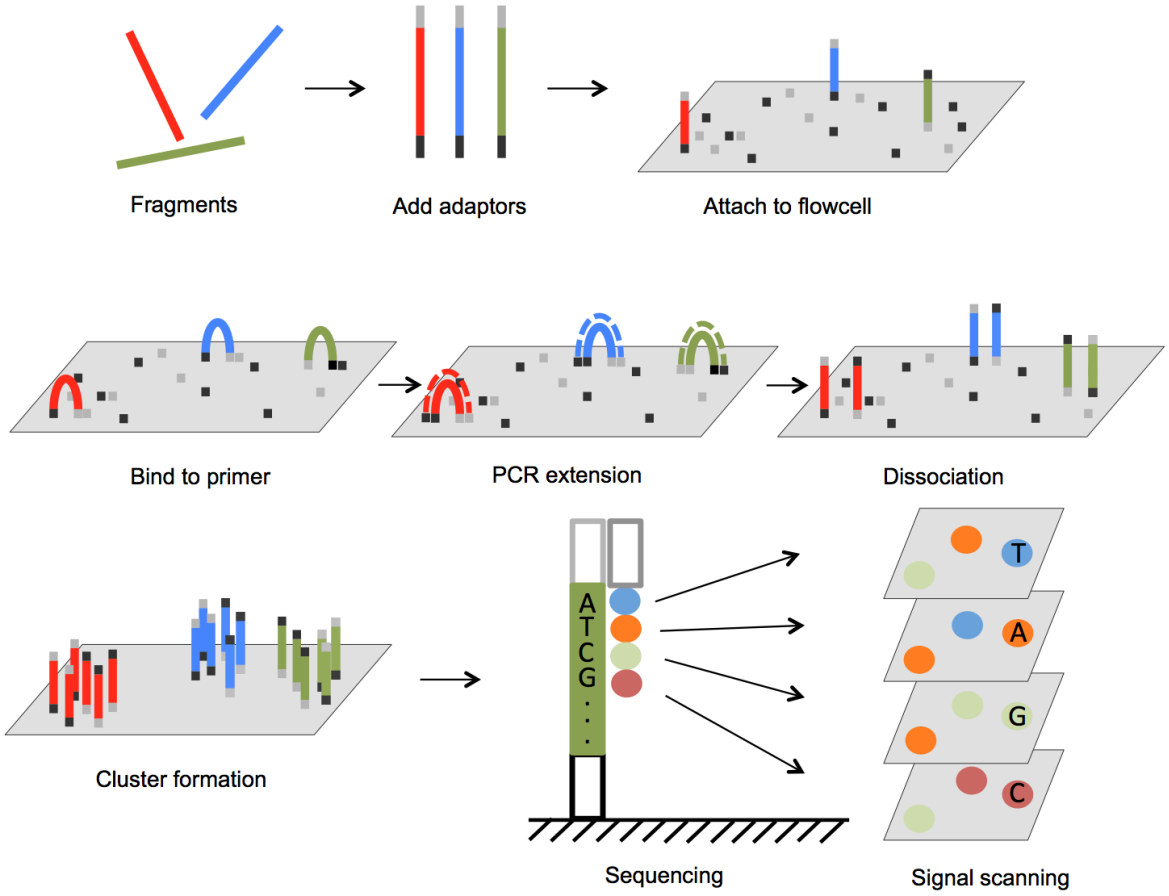

Figura 7. Esquema general de secuenciación illumina, extraído de [75]

Primero, se prepara la librería: después de extraer y fragmentar el ADN, se añaden adaptadores específicos a cada extremo de los fragmentos, facilitando su unión a la flow cell, la identificación de las muestras y la unión de la ADN polimerasa. En el segundo paso, los fragmentos con adaptadores se unen a la flow cell, que está preparada con oligonucleótidos complementarios. El tercer paso es la amplificación en puente: Los fragmentos de ADN se

doblan y se hibridan con oligonucleótidos cercanos en la flow cell. La ADN polimerasa sintetiza copias complementarias, creando múltiples copias (clústeres) de cada fragmento.
Finalmente, se realiza la secuenciación por síntesis: Nucleótidos marcados con fluoróforos se incorporan a la cadena de ADN en crecimiento. Un láser ilumina los clústeres, detectando la fluorescencia emitida por los nucleótidos incorporados. La secuencia de ADN se determina a partir de estas señales de fluorescencia [76].

2.4 **Algoritmos de ensamble**

Las herramientas de secuenciación otorgan como resultados archivos con muchas secuencias cortas, estas secuencias cortas no pueden ser utilizadas sin ayuda de un algoritmo de ensamble que se encargue de juntar todos estos fragmentos. Existen distintas metodologías que no permiten ensamblar estas lecturas.

2.4.1 Algoritmo Greedy (codicioso)

El algoritmo Greedy (codicioso) es una técnica sencilla para ensamblar secuencias de ADN que construye una solución paso a paso, eligiendo en cada paso la opción más prometedora. Este método funciona calculando las distancias entre pares de lecturas para encontrar las superposiciones más largas, agrupando las lecturas que tienen la mayor superposición y ensamblándolas en contigs más largos. Aunque es rápido y eficiente en términos de tiempo de ejecución, el algoritmo Greedy no garantiza alcanzar el óptimo global y puede tener problemas con regiones repetitivas. A pesar de estas limitaciones, ha demostrado ser útil en ciertos contextos, como en el ensamblaje de genomas virales grandes con el ensamblador S-aligner, que aplica una estrategia Greedy modificada para mejorar el rendimiento y la cobertura del genoma [77].

2.4.2 Algoritmo Overlap-Layout-Consensus (OLC)

El método Overlap-Layout-Consensus (OLC) es una estrategia más robusta para el ensamblaje de genomas, que maneja mejor la complejidad de los datos de secuenciación al identificar superposiciones significativas entre pares de lecturas. Este enfoque organiza las lecturas en un grafo de superposición y busca una ruta consistente que recorra todas las lecturas, construyendo finalmente una secuencia de consenso a partir de las rutas

identificadas mediante alineación múltiple de secuencias. Aunque el algoritmo OLC puede generar ensamblajes de alta calidad y maneja bien los datos con errores y regiones repetitivas, requiere mucho tiempo y memoria, especialmente con grandes conjuntos de datos de secuenciación [78], [79].

2.4.3 Grafos de Bruijn

Los grafos de Bruijn son una técnica ampliamente utilizada para el ensamblaje de genomas, especialmente eficaz en la era de la secuenciación de próxima generación debido a su eficiencia. Este método divide las secuencias en fragmentos cortos llamados k-mers, que se utilizan como nodos en un grafo, donde las superposiciones entre k-mers definen las aristas. Existen dos tipos principales de ensambladores de Bruijn: los ensambladores eulerianos, que encuentran un camino que recorre cada arista exactamente una vez, y los ensambladores hamiltoneanos, que encuentran un camino que visita cada nodo exactamente una vez. Los grafos de Bruijn son muy eficientes para manejar grandes volúmenes de datos y reducen significativamente la complejidad del problema de ensamblaje. Sin embargo, pueden generar caminos ambiguos debido a regiones repetitivas en el genoma y requieren un manejo cuidadoso de la longitud de los k-mers [80].

3 Antecedentes

El ensamble de genomas es un proceso bioinformático que permite la reconstrucción de una secuencia genómica completa, a través del alineamiento y unión de lecturas generadas del material genético, a través de tecnologías de secuenciación [81]. Ese proceso ha permitido avances en el estudio y análisis de los genomas, incluido el de los cloroplastos.

El primer genoma de cloroplasto secuenciado y reportado fue el de *Marchantia polymorpha* en agosto de 1986 [82]. Seguido del genoma de cloroplasto de *Nicotiana tabacum* en septiembre de 1986 [83]. Estos primeros avances abrieron el camino para utilizar al cloroplasto como recurso biotecnológico, entre ellos su uso como plataforma de expresión de proteínas recombinantes. En la actualidad el cloroplasto de *Marchantia polymorpha* está emergiendo como una plataforma de expresión [84], [85], mientras que el cloroplasto *Nicotiana tabacumt* está fuertemente establecido desde hace tiempo [86], [87], siendo utilizado para la producción de enzimas y proteínas terapéuticas, entre otras. [88], [89].

Actualmente, NCBI (Centro Nacional de Información Biotecnológica) están subidos 12,988 genomas de cloroplastos pertenecientes al reino viridiplantae. El genoma de cloroplasto más largo registrado en NCBI es del alga verde perteneciente al orden de *Chlamydomonadales*, *Haematococcus lacustri* con una longitud de 1352306 pb y más del 90% de DNA no codificante [90], mientras que el más pequeño pertenece a la planta terrestre *Asarum minus*, con una longitud de 15553 pb, teniendo una forma linear (en lugar de una circular que es lo más común visto) perteneciente al orden de *Piperales* [91].

Dentro de las microalgas *Chlamydomonadales* (orden al que pertenece *Dunaliella salina*), hay 17 genomas de cloroplasto reportados (Tabla 10), siendo el más largo el de *Haematococcus lacustri,* y el más pequeño el de *Chlorogonium euchlorum,* midiendo 22988 pb.

Tabla 10. Genomas de cloroplastos de algas *Chlamydomonadales* reportadas en NCBI				
Nombre científico	**Modificador**	**Topología**	**Tamaño (kpb)**	**Numero de genes**
Chlamydomonas reinhardtii	-	Circular	203.83	109
Pleurastrum terricola	-	Circular	195.08	119
Dunaliella salina	CCAP 19/18 (strain)	Circular	269.04	122
Gonium pectorale	K3-F3-4 (isolate)	Circular	222.58	108
Oogamochlamys gigantea	-	Circular	254.08	119
Hafniomonas laevis	-	Circular	263.33	119
Characiochloris acuminata	-	Circular	197.18	108
Carteria cerasiformis	-	Circular	318.97	113
Phacotus lenticularis	-	Circular	203.37	113
Haematococcus lacustris	UTEX 2505 (strain)	Circular	1,352.31	143
Chlorococcum tatrense	-	Circular	242.17	104
Spermatozopsis similis	-	Circular	134.87	103
Chlorogonium euchlorum	SAG 12-2a (strain)	Circular	22.99	11
Chlamydomonas chlamydogama	11-48b (strain)	Circular	198.35	108
Hyalomonas oviformis	SAG62-27 (strain)	Circular	131.94	66
Edaphochlamys debaryana	-	Circular	227.46	92
Hydrocytium acuminatum	SAG 40.91 (strain)	Circular	450.75	152
Tabla elaborada con información recolectada del NCBI				

De entre los genomas de cloroplasto de algas *Chlamydomonadales* registrados en NCBI, solo unos pocos han sido utilizados para la expresión de proteínas recombinantes, el más utilizado

es el de la microalga modelo *Chlamydomonas reinhardtii*, algunos ejemplos del uso de su cloroplasto para expresar proteínas recombinantes se encuentran en la tabla 11.

Tabla 11. Proteínas recombinantes producidas en el cloroplasto de la microalga modelo *Chlamydomonas reinhardtii*

Producto expresado	Categoría/aplicación	Promotor	Referencia
Isoforma 121 del factor de crecimiento endotelial vascular humano (VEGF)	Tratamiento del enfisema pulmonar.	PpsbA	[92]
Hormona del crecimiento humano	Deficiencia en la hormona del crecimiento humano	PpsbA	[93]
Alérgeno principal del polen de abedul Bet v 1	Inmunoterapia con alérgenos (ITA) para el tratamiento de enfermedades alérgicas	PpsaA	[94]
Interleucina-29 humana (IL-29)	Tratamiento contra enfermedades autoinmunes	PpsaA	[95]
Polyethylene terephthalate hydrolases (PETases)	Degradación enzimática del PET (polietilentereftalato)	PpsaA	[96]

Además de los ejemplos mostrados en la tabla 11, *Chlamydomonas reinhardtii* tiene más casos de éxitos demostrados en la expresión de proteínas recombinantes. Aparte de esta microalga, dentro del grupo de *Chlamydomonadales*, solo se ha utilizado el genoma del cloroplasto de 3 organismos más para la expresión de proteínas recombinantes, estos genomas son de la microalga *Haematococcus lacustris* (*pluvialis), Dunaliella tertiolecta y Dunaliella salina* (la microalga de este estudio), algunos de los ejemplos de su uso están en la tabla 12.

En el caso de la transformación de *Dunaliella tertiolecta*, utilizaron el genoma del cloroplasto de *Chlamydomonas reinhartii* para el diseño del casete de expresión, logrando expresar las proteínas de interés. Este experimento demuestra cierto grado de conservación del genoma de este organelo dentro del orden *Chlamydomonadales* [97]. Por otro lado, la transformación de *Dunaliella salina* fue lograda con un casete de expresión construido originalmente para arroz, lo que nos indica que entre clases alejadas como las algas *Chlamydomonadales* y las

plantas con flores *Magnoliopsida* existen regiones del cloroplasto que están sometidas a una fuerte presión selectiva [98].

Tabla 12. Proteínas recombinantes producidas en el cloroplasto de microalgas *Chlamydomonadales* no modelo

Nombre	Producto expresado	Categoría/aplicación	Promotor	Referencia
Haematococcus lacustris (*pluvialis)*	Péptido antimicrobial piscidin-4 (ant1)	Marcador de selección de transformación	PpsbA	[99]
Haematococcus lacustris (*pluvialis)*	Astaxanthin	Enfermedades inflamatorias crónicas, antienvejecimiento, cánceres, síndrome metabólico, enfermedades neurodegenerativas, protección contra el daño por luz ultravioleta.	PpsbA	[100]
Dunaliella tertiolecta	Xilanasa, α-galactosidasa, fitasa, fosfato anhidrolasa, y β-mananasa	Suplementos en alimentos de animales	PpsbD	[97]
Dunaliella salina	EGFP	Marcador de selección	Prrn16	[98]

Sin embargo, aunque existan partes del genoma del cloroplasto que están fuertemente conservadas entre los integrantes del reino *Viridiplantae*, incluso dentro del orden de los *Chlamydomonadales* (tabla 4) existen diferencias entre longitudes y contenido de genes. Estudiar estas regiones nos puede proporcionar información acerca de su función reguladora, variabilidad genética y evolución del genoma, lo que se traduce en mejoras biotecnológicas y un mejor entendimiento en filogenética.

Actualmente se encuentran secuenciados los genomas del cloroplasto de dos cepas de *Dunaliella salina*, la cepa CCAP19/18 y la cepa SQ [101], [102]. La tabla 13 muestra sus diferencias.

Tabla 13. Comparación de los plastomas pertenecientes a *Dunaliella salina cp SQ y CCAP19/18*

Cepa	Longitud	%CG	SSC	LSC	IR	GENES
CCAP 19/18	269,044 pb	32,1%	112,900 pb	127,300 pb	14,400 pb	102 genes 66 codificantes 3ARNr 28ARNt
SQ	243,635 pb	29,73%	101,527 pb	107,815 pb	17,145 pb	98 genes 66 codificantes 3ARNr 29 ARNt

Tabla elaborada con información de [101], [102].

Estudiar el genoma del cloroplasto de otra cepa de *Dunaliella salina*, nos ofrece el potencial de profundizar en la biotecnología y la filogenia de esta microalga. Comparándola con los genomas de cloroplasto ya reportados de otras cepas, se pueden descubrir nuevas adaptaciones y marcadores evolutivos, lo que tiene aplicaciones en el estudio de la evolución de las algas. Además, conocer información acerca de sus regiones reguladores puede abrir las puertas para nuevas herramientas para la ingeniería de cloroplastos.

4 Problema de investigación

La industria de la producción de proteínas recombinantes está en constante expansión debido a su importancia en áreas como la medicina, la biotecnología y la agricultura. Sin embargo, enfrenta desafíos significativos, especialmente relacionados con las condiciones específicas necesarias para la producción eficiente de dichas proteínas. Los microorganismos tradicionalmente empleados en la producción de proteínas recombinantes presentan limitaciones en cuanto a estabilidad y rendimiento en entornos con alta salinidad.

Estos desafíos se deben principalmente a que las condiciones salinas extremas pueden afectar negativamente la funcionalidad y viabilidad de los microorganismos convencionales, comprometiendo así la eficiencia de la producción de proteínas recombinantes. En consecuencia, existe una necesidad urgente de encontrar y desarrollar conocimientos sobre organismos alternativos que puedan soportar y prosperar en estos ambientes adversos, manteniendo altos niveles de producción y estabilidad.

El genoma del cloroplasto de *Dunaliella salina* aislada de Noruega ya ha sido secuenciado mediante la tecnología Illumina MiSeq, generando un total de 15,742,069 secuencias. Sin embargo, la secuenciación genómica produce fragmentos cortos de ADN, en este caso de 300 pares de bases cada secuencia, por lo que se requiere de un procesamiento bioinformático exhaustivo para ensamblar así un genoma completo y funcional. Además, el genoma del cloroplasto contiene regiones repetitivas, que añaden un nivel significativo de complejidad al proceso de ensamblaje y análisis. Este proceso incluye varias etapas críticas, como lo son el limpiado de secuencias, el ensamblaje *de novo* y por referencia, la anotación del genoma y la identificación de promotores.

Estas etapas son indispensables para convertir los datos de secuenciación en información utilizable. La falta de un ensamblaje y anotación precisos del genoma del cloroplasto puede limitar la capacidad de *Dunaliella salina* aislada de Noruega para ser utilizada de manera eficiente en la producción de proteínas recombinantes. Es así que, sin una identificación adecuada de las regiones promotoras, el potencial de esta microalga como plataforma de expresión en condiciones salinas no puede ser plenamente realizado.

5 Hipótesis

Al ensamblar el genoma del cloroplasto de *Dunaliella salina* aislada de Noruega, se logrará identificar secuencias promotoras de la transcripción por medio de análisis *in silico*.

6 Justificación

Dunaliella salina es una microalga versátil, adaptada a ambientes extremos, especialmente salinos, siendo un foco de estudio en biotecnología. Actualmente, los genomas del cloroplasto de las cepas SQ y CCAP19/18 han sido secuenciados y ensamblados, proporcionando una base sólida para la investigación. Sin embargo, la diversidad genética dentro de la especie y las posibles variaciones genéticas entre diferentes cepas sugieren la necesidad de explorar y ensamblar el genoma del cloroplasto de otras cepas de *Dunaliella salina*. Ensamblar el genoma del cloroplasto de otra cepa permitiría entender mejor la adaptación de *Dunaliella salina* a diversos entornos extremo. Comparar los genomas de cloroplastos de diferentes cepas de *Dunaliella salina* mejorará nuestra comprensión de su evolución y diversidad genética, con potenciales aplicaciones en biotecnología y producción sostenible de bioproductos.

7 Objetivos de investigación

7.1 **Objetivo general**

Ensamblar el genoma del cloroplasto de *Dunaliella salina* obtenida de Noruega e identificar elementos cis promotores *in silico*.

7.2 **Objetivos específicos**

- Ensamblar el genoma del cloroplasto de *Dunaliella salina* procedente de Noruega con SPADes y Bowtie2.
- Anotar el genoma del cloroplasto de *Dunaliella.salina* procedente de Noruega GeSeq
- Analizar la sintenia del genoma del cloroplasto de *Dunaliella salina* procedente de Noruega con MAUVE.
- Predecir de regiones promotoras en el genoma del cloroplasto de *Dunaliella salina* procedente de Noruega.

- Analizar y seleccionar de regiones promotoras candidatas dentro del genoma del cloroplasto de *Dunaliella salina* con newPLACE.

8 Metodología

8.1 **Limpieza y filtrado:**

Durante la secuenciación de illumina, es necesario agregar secuencias adaptadoras a los fragmentos de ADN que se van a secuenciar, de esta manera se vuelve posible la generación de clúster y posteriormente el llamado de bases [103], sin embargo, una vez completada la secuenciación, las secuencias adaptadoras se conservan en los datos generados, además, por la naturaleza de la tecnología de secuenciación existen posibles escenarios que generan errores que afectarán el tratamiento de los datos, estos escenarios involucran: Decaimiento de la señal, desincronización entre la cámara y la luz emitida por los fluoróforos, y fluoróforos sin remover, generando llamados de base con alta probabilidad de no ser correctos (Quality Score bajo) [104]. Es por estas razones que, para poder trabajar con la información obtenida de la secuenciación, es necesario procesarla por herramientas de recorte. Para este trabajo se utilizará la herramienta trimmomatic, la cual es un recortador flexible (permite ajustar los parámetros de recorte) compatible con secuencias fastq de pair-end (lecturas complementarias de la hebra foward y reverse), capaz de eliminar secuencias adaptadoras con dos enfoques, el modo palíndromo, que se aprovecha de las secuencias foward y reverse para detectar con mayor precisión los adaptadores a los extremos de las secuencias, y el modo simple, que funciona deslizando las secuencias adaptadoras a lo largo de las lecturas hasta encontrar coincidencias y recortarlas. De igual forma esta herramienta permite la eliminación de secuencias de baja calidad, lo que nos permitirá contar con secuencias sin adaptadores y con puntajes altos, dando mayor veracidad a los resultados de los siguientes pasos de la metodología [105].

8.2 **Alineación de lecturas cortas**

Cuando se extrae una muestra de ADN de un organismo, es común que contenga material genético de toda la célula. En el caso de células vegetales, la muestra puede contener información del núcleo, cloroplasto y mitocondria [106].

Un paso previo pertinente antes de ensamblar un genoma consiste en extraer las lecturas pertenecientes al tipo de material genético a necesitar, en este caso, al trabajar con el material

genético del cloroplasto, es ventajoso alinear el total de lecturas una vez limpias y filtradas (sin errores de secuenciación, secuencias de baja calidad, adaptadores ni barcodes), contra el genoma de cloroplasto de una especie cercana evolutivamente [107].

Con el fin de separar las lecturas se alinearán las lecturas limpias contra el genoma del cloroplasto de *Dunaliella salina* cepa SQ y cepa CCAP19/18. Para este paso se utilizará el alineador de lecturas cortas bowtie2 [108].

8.3 **Ensamble de novo:**

Una vez separadas las secuencias limpias y seleccionadas las lecturas del cloroplasto, solo quedan lecturas de DNA listos para ser procesados (el tamaño varía dependiendo de la herramienta de secuenciación utilizada). En el caso de esta tesis, se secuenció el genoma utilizando illumina Miseq, la cual da lecturas de 300 pb de longitud [109]. El genoma de un cloroplasto tiene longitudes aproximadas de 120,000 pb a 160,000 pb [110]. Aún después de la limpieza y filtrado de secuencias es posible que las lecturas aún contengan errores en el ||llamado de cada base, es por esto que, para tener la certeza probabilística de que el ensamble y el llamado de bases fue correcto, es deseable tener una profundidad de 100x. Siguiendo el estimado para determinar el número de bases necesarias para cumplir con la profundidad deseada se tomará en cuenta la siguiente fórmula disponible en [111]:

$$C = \frac{L * N}{G}$$

Donde:

- C = Cobertura (Profundidad)
- G = Largo del genoma haploide
- L = Largo de la lectura
- N = Número de lecturas

Con el fin de obtener el número de lecturas necesarias para tal profundidad despejaremos N, quedando:

$$N = \frac{G * C}{L}$$

$$N = \frac{160000 * 100}{300} = 53{,}334\ lecturas\ minimas\ aproximadas$$

Con un aproximado del mínimo de lecturas necesarias para un ensamble de genoma de cloroplasto de 53,334 lecturas (recordando que un volumen mayor de lecturas da resultados más precisos), se utilizará la herramienta SPAdes para el ensamble *de novo.* Esta herramienta basa su metodología en grafos de Bruijn Eulerianos (utiliza rutas que recorran todas las aristas una sola vez). SPAdes es un ensamblador especializado en secuenciación single-cell (células individuales), por lo que fue pensado para trabajar con datos complejos y ruidosos que resultan de la secuenciación de una célula individual, como lo son los sesgos obtenidos durante la amplificación y secuenciación al partir de poco material genético (el de una sola célula). SPAdes utiliza grafos de Bruijn emparejados (una especie de grafico de Bruijn de doble capa). La capa interna se construye utilizado k-mers individuales extraídos directamente de las lecturas de ADN, mientras que la capa externa se realiza utilizado k-mers emparejados (provenientes de lecturas pair-end) que están separados por una cierta distancia (insert size) en la secuencia original de ADN, se aprovechan los k-mers con mayor insert size para resolver repeticiones en el genoma y para construir un esqueleto (scaffolding) que conecte los contigs en una secuencia más larga y continua [112].

8.4 **Ensamble por referencia:**

La mayoría de los ensambladores de secuencias cortas ensamblan los genomas a nivel de contigs. En el caso de SPAdes, este va un paso más adelante, con la integración del proceso del scaffolding (proceso en el que se unen los contigs en secuencias más largas), aun así es muy probable que el genoma quede abierto (queden fragmentos largos que no estén conectados entre sí), esto puede deberse a regiones repetitivas, a contaminación con material genético de otras especies o contaminación con fragmentos del genoma nuclear o mitocondrial de la misma especie que contenga secuencias muy similares a las del cloroplasto, metiendo ruido suficiente para imposibilitar el cerrado del genoma.

Independientemente de la causa, se puede utilizar el genoma de una especie estrechamente relacionada para realizar el scaffolding y terminar de cerrar el genoma. Esto se logra mapeando los contigs previamente generados por el ensamblador sobre el genoma de una especie estrechamente relacionada, utilizándolo como referencia en el acomodo de los

contigs [113]. Se utilizarán los genomas de *Dunaliella salina* cepa noruega y *Dunaliella salina* CCAP19/18 para organizar los contigs.

En caso de no conseguir cerrarse el genoma de cloroplasto utilizando la estrategia previamente explicada, se recurrirá a generar una secuencia consenso a partir de la alineación de lecturas cortas contra el genoma de cloroplasto de una especie relacionada, se optará por el genoma que presente un mayor porcentaje de cobertura y profundidad, permitiendo reconstruir la secuencia genómica mediante el consenso de múltiples lecturas. Esta aproximación se realizará con bcftools [114] .

8.5 **Anotación del genoma:**

Una vez ensamblado y cerrado el genoma del cloroplasto de *Dunaliella salina* cepa noruega, es necesario interpretar la secuencia de sus bases nitrogenadas (A, T, G, C), y adjuntarle información biológica mediante el análisis de sus partes.

Para la predicción de genes en un genoma existen 3 enfoques principales, el intrínseco, extrínseco y el combinado, el intrínseco centrado únicamente en la información que se puede obtener de la secuencia por sí misma, la forma extrínseca utiliza transcritos de RNA y/o polipéptidos como información, en este enfoque proteínas secuenciadas de otras especies proporcionan una buena indicación sobre la presencia y ubicación de los genes. El enfoque combinado es el más utilizado de todos, integran las ventajas de los ambos modos [115]. Sin embargo, para este proyecto no se cuentan con transcritos de RNA, por lo que se optará por utilizar un enfoque intrínseco, además de secuencias de especies relacionadas.

Para la anotación del genoma de cloroplasto de *Dunaliella salina* cepa noruega se utilizarán dos herramientas especializadas en anotación del cloroplasto, las cuales se complementan entre sí. La primera llamada Geseq es una herramienta en línea especializada en la anotación precisa de cloroplastos, cuenta con una selección de genomas anotados de NCBI curados manualmente para utilizarlos como referencia en la anotación del genoma objetivo, además, ofrece la posibilidad de subir un genoma fuera de su catálogo como genoma de referencia. El programa toma las anotaciones de los genomas de referencia seleccionados y busca alineaciones con BLAT, con el fin de identificar esas mismas anotaciones en nuestro genoma objetivo [116]. Para apoyar la anotación generada con Geseq, también se anotará la secuencia generada en la metodología anterior con la herramienta MFannot, la cual, a diferencia de

Geseq que se aprovecha del uso de BLAT para buscar similitud de secuencias con genes conocidos, un enfoque que puede llegar a ser insuficiente. MFannot utiliza HMMs (Hiden Markov Models) para predecir genes, partiendo de una anotación conceptual, donde se buscan marcos de lectura abiertos y a partir de ellos se hace la búsqueda de genes conocidos [117].

De esta manera, se busca complementar ambas anotaciones, tal como se observa en la figura 8, con el fin de reducir el trabajo de curación manual. Los genes con discrepancias entre ambas anotaciones se confirmarán con blast [118].

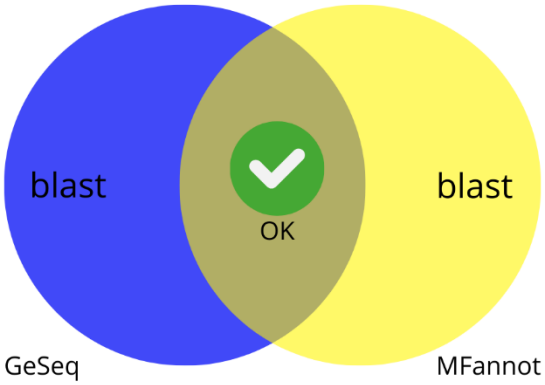

Figura 8 Diagrama de Venn que representa el enfoque de anotación génica que se utilizará en esta tesis. Los genes con anotaciones coincidentes serán tomadas como verdaderas. Aquellos genes cuya anotación presente discrepancia se confirmarán con blast.

8.6 **Análisis de sintenia**

Al ensamblar y anotar un genoma nuevo generalmente se le realizan estudios de genómica comparativa. En este sentido, un análisis de sintenia proporciona un marco en el que se identifica conservación de genes homólogos y el orden genético entre los genomas de diferentes especies. En general es un proceso de filtrado y organización de todas las similitudes locales entre secuencias del genoma en una imagen global [119].

Para fines de esta tesis, se realizará la sintenia con los genomas de cloroplasto de *Dunaliella salina* cepa CCAP19/18 y cepa SQ. Esta comparación nos permitirá inferir la relación evolutiva de *Dunaliella salina* entre sus distintas cepas en función del genoma sus cloroplastos.

Para esta tesis se utilizará la herramienta MAUVE, un sistema diseñado para construir alineamientos múltiples de genomas en presencia de eventos evolutivos de gran escala, como lo son reordenamientos e inversiones. Mauve permite detectar bloques conservados, conocidos como "Locally Collinear Blocks (LCBs)", entre distintos genomas, proporcionando una base sólida para la inferencia de relaciones evolutivas [120].

8.7 **Predicción de regiones promotoras:**

Para la predicción de promotores putativos en el cloroplasto de *Dunaliella salina* cepa noruega, es necesario conocer su posición filogenética. Como está descrito en el marco teórico, las algas *Chlamydomonadales,* orden al que pertenece *Dunaliella salina,* carecen de la ARN polimerasa dependiente de fago (NEP), por lo que la transcripción en estos cloroplastos depende únicamente de la ARNp de origen bacteriano (PEP).

La búsqueda de promotores se centrará únicamente en los promotores de la ARNp de origen bacteriano (PEP). La metodología utilizará regiones -35 y -10 de promotores para la PEP previamente caracterizados. La tabla 6 resume algunas regiones consenso caracterizadas en el cloroplasto de algas *chlamydomonadales.*

Con estas regiones, se generará un script en Python, el cual trabajará con el archivo de anotación del cloroplasto GenBank (.gb) del genoma del cloroplasto de *Dunaliella salina* cepa noruega. El script extraerá las regiones 1000 pb aguas arriba del codón de inicio de todos los genes (sabiendo que el promotor se encuentra aguas arriba de los genes), y buscará las regiones -35 y -10, a distancia variables entre sí de 12 pb a 25 pb [121].

Finalmente, las secuencias que presenten los motivos conservados serán analizadas mediante la base de datos newPLACE, con el objetivo de identificar elementos cis-reguladores relacionados con diversos estímulos, como la luz, temperatura, estrés biótico, etc. Lo cual permitirá completar la caracterización funcional de los promotores putativos identificados [122].

La figura 9 muestra un diagrama de flujo de la metodología utilizada en esta tesis.

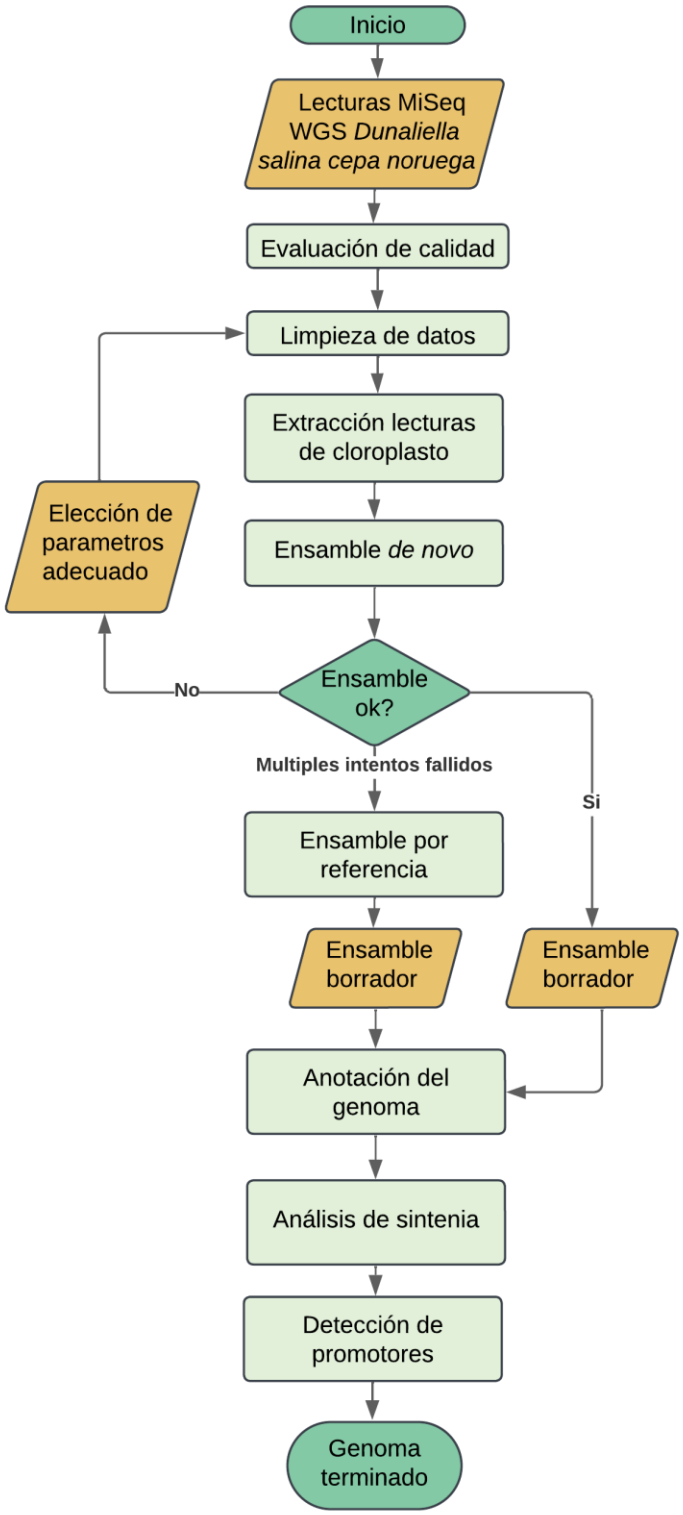

Figura 9. Diagrama de flujo de la metodología

9 Resultados y discusión:

9.1 Secuencias sin procesar:

Las herramientas de secuenciación, como la utilizada en esta tesis (illumina Miseq), generan grandes cantidades de datos que necesitan ser analizados para evaluar y mejorar la calidad de las secuencias obtenidas.

Para la evaluación de la calidad de los datos generados por secuenciación de nueva generación (NGS) se utilizó FastQC, una herramienta ampliamente utilizada gracias a sus gráficos interactivos, los cuales facilitan la interpretación de los resultados [123].

Las figuras que se muestran en la sección 9.1 serán las correspondientes a los datos de secuenciación R1 (hebra forward) sin procesar, omitiendo las secuencias correspondientes a R2 (hebra reverse) con el propósito de evitar extender el documento.

En la figura 10 se muestra un resumen de los módulos evaluados, para cada uno el programa presenta un símbolo, el cual varía entre una marca de verificación en color verde, que corresponde a resultados dentro de los rangos considerados normales por FASTQC, un signo de exclamación naranja correspondiente a resultados ligeramente anormales y una equis roja correspondiente a resultados totalmente anormales. En esta imagen se pueden observar resultados ligeramente anormales en los módulos “Per tile sequence quality” y en “Per sequence GC content”, mientras que se pueden ver resultados muy totalmente anormales en los módulos “Per base sequence quality”, “Per base sequence content” y “Adapter content”.

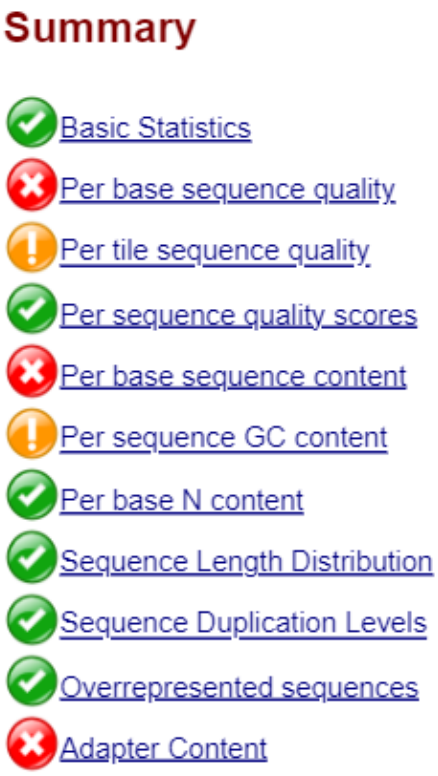

Figura 10: Resumen de módulos evaluados por FASTQC

9.1.1 Basic Statistics

Measure	Value
Filename	AD_DUNALIELLA_TAAGGCGA-CTCTCTAT_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	15742069
Total Bases	4.7 Gbp
Sequences flagged as poor quality	0
Sequence length	300
%GC	47

Figura 11: Tabla generada por FASTQC que contiene datos generales de las secuencias evaluadas

La figura 11 muestra datos generales del archivo fastq, estos datos incluyen: nombre del archivo (filename); tipo de archivo (file type): indica si el archivo contiene llamadas de base o datos de la colorimetría de la secuenciación (en algunas tecnologías de secuenciación los datos iniciales no se reportan como bases directamente, si no como colores que representan las bases), en este caso los datos se encuentran reportados como bases, por lo que se puede omitir el paso de convertir los colores a sus bases correspondientes; Codificación (Encoding): Menciona que tipo de codificación ASCII de valores de calidad se encontró en el archivo, nuestro archivo FASTQC dice “Sanger/illumina 1.9” lo cual indica que el archivo FASTQ usa codificación de calidad Phred+33 (tipo de codificación inicialmente utilizado en secuenciación de Sanger y posteriormente adoptado en Illumina a partir de la versión 1.8); Total de secuencias (Total sequence): Muestra el recuento total del número de secuencias contenidas en el archivo FASTQ, en el caso de nuestro archivo contamos con 15742069 secuencias sin procesar; Total de bases (Total Base): Menciona el número total de bases nucleotídicas en el total de secuencias dentro del archivo FASTQ, en nuestro caso contamos con 4.7 Gpb en total; Secuencias marcadas como de baja calidad (Sequence flagged as por quality): es una métrica que indica el número de secuencias etiquetadas como de baja calidad por la plataforma de secuenciación, sin embargo aunque en nuestro caso marca 0, es necesario revisar otros módulos (como “Per base sequence quality”) para determinar si existen o no existen secuencias de mala calidad; Longitud de secuencias (Sequence length):

proporciona la longitud de las secuencias desde la más corta a la más larga, en nuestro caso, todas las secuencias tienen la misma longitud correspondiente a 300 pares de bases); %GC: proporciona el porcentaje general de guaninas y citosinas en todas las bases de todas las secuencias, en nuestro caso el 47% corresponde a GC.

9.1.2 Per base sequence quality

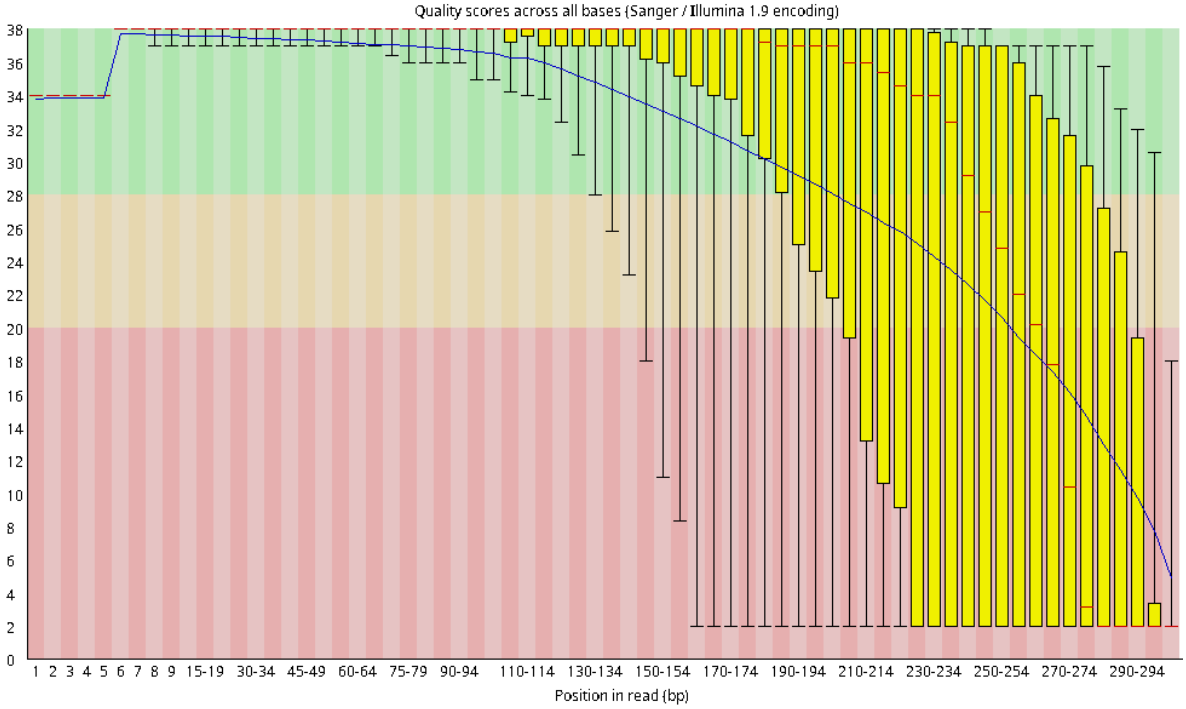

Figura 12: Gráfico de calidad promedio por posición de base en las lecturas de secuenciación

La figura 12 muestra una gráfica de calidad de las lecturas crudas. El eje X corresponde a la posición nucleotídica de las lecturas, el eje Y correspondiente al quality score (Q) (medida numérica que indica la probabilidad de que cada llamado de base sea correcto), entre mayor sea Q, mejor es la calidad del llamado de base. Por cada posición se dibuja un gráfico de tipo BoxWhisker, los elementos del grafico son los siguientes: la linea central roja significa el valor medio (mediana); la caja amarilla representa el rango intercuartil de 25% a 75%; Los bigotes superiores e inferiores representan el 10% y 90% de los puntos; la linea azul representa la calidad media (promedio). El fondo divide el eje Y en 3 secciones representadas con colores: verde: llamados de bases muy buenos; naranja: llamados de bases con calidad razonable; rojo: llamados de base pobres.

Al final del gráfico de calidad existe una caída general de la calidad de los llamados de base (extremo 3'). Es común ver una disminución en la calidad al final de las secuencias de la mayoría de las NGS, esto se debe al degradamiento de la maquinaria implicada en la secuenciación a medida que avanza la ejecución, provocando en la mayoría de los casos que Q caiga al área naranja [123]. En nuestra grafica viene adjunto el símbolo fracaso o failure (equis roja), FASTQC muestra este símbolo cuando el cuartil inferior para cualquier base es inferior a 5, o si la mediana de cualquier base baja de 20, en nuestro caso se cumplen las dos condiciones. Lo que sugiere este módulo de FASTQC es la necesidad de limpiar las secuencias, removiendo las que tengan menor calidad ya que pueden meter ruido en futuros pasos de la metodología.

9.1.3 Per tile sequence quality

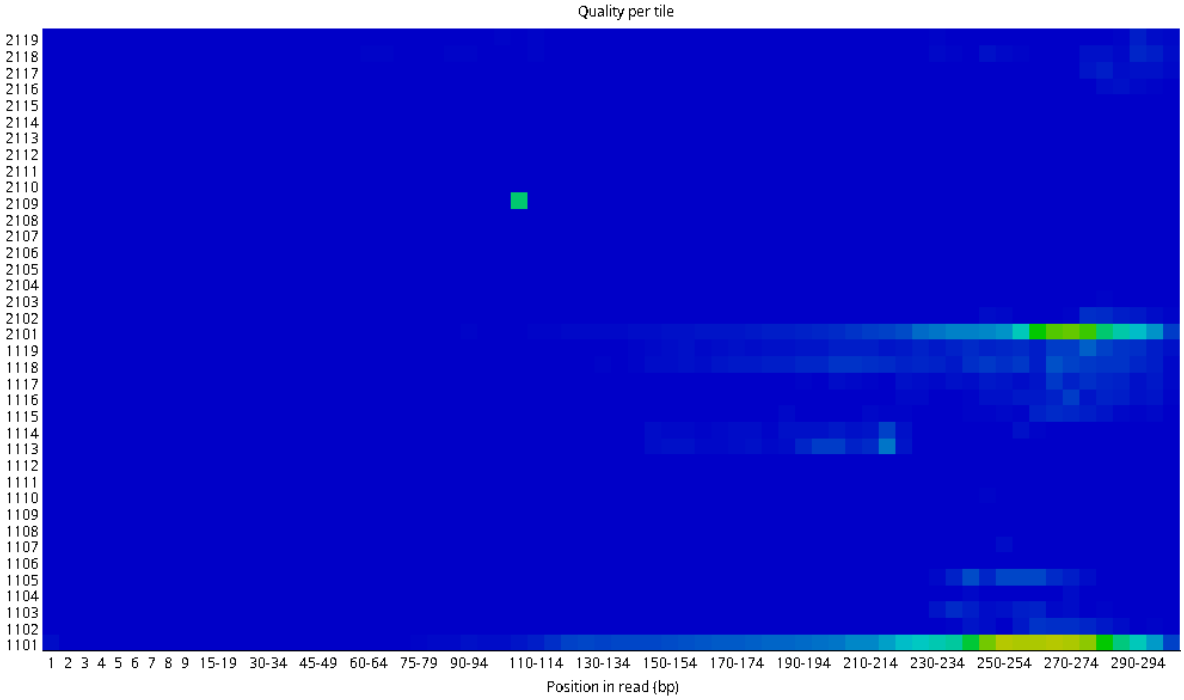

Figura 13: Desviación de la calidad media de cada baldosa (tile) en función de la posición de las lecturas

La presencia del grafico 13 entre los módulos de FASTQC significa que se está analizando una biblioteca de illumina que aún conserva sus identificadores de secuencias originales (barcodes) [123].

En la figura 13 los colores del grafico están de escala fría a caliente: los colores fríos aparecen en posiciones en las que la calidad está en o por encima del promedio de esa base durante la secuenciación, los colores muy calientes significan que una baldosa tenía una calidad menor a las demás baldosas en esa base.

En el eje X se encuentra la posición en la lectura, mientras que en el eje Y se ve las baldosas (tiles), en nuestro caso se ve que inicia desde la numeración 1101, esto no significa que se haya secuenciado en mil ciento uno baldosas, este número es un código que otorga información sobre los carriles, sectores y baldosas de la secuenciación, el número 1101 significa que se utilizó el carril 1, del sector 1, en la baldosa 01, de igual forma, cuando cambia el numero de 1101 a 2101 significa que se cambió de carril, siendo carril 2, del sector 1, en la baldosa 01 [124]. Ahora bien, cada una de estas baldosas tiene dentro un clúster generado por la amplificación en puente de illumina, estos clústeres están conformados por fragmentos de DNA que otorgarán lecturas de 300 pb, es por esto que a cada baldosa le corresponde 300 llamados de bases. En nuestra figura se logra ver un cambio de color en el carril 1, del sector 1, en la baldosa 01 (1101), alrededor de la base 114 incrementando en la 214, este cambio de color de frío a cálido demuestra una reducción en la calidad para la misma base en distintas baldosas. De igual manera en el carril 2, sector 1, baldosa 01 (2101) se logra ver el cambio de frío a cálido, mostrando el mismo comportamiento que en 1101, de igual forma en el carril 2, sector 1, baldosa 09 se logra ver un error puntual alrededor de la base 100. Estos errores pueden deberse a problemas transitorios como burbujas que pasan por la celda de flujo, o problemas más permanentes como manchas en la celda de flujo o escombros dentro del carril de cada celda, a juzgar por la rápida solución del problema (tan solo unas cuantas bases más adelante regresa a los colores fríos parecería una burbuja dentro de la celda de flujo. Junto a esta gráfica viene adjunto un símbolo de resultados ligeramente inusuales (naranja), lo que significa que una baldosa muestra una puntuación Phred menor en dos que las otras baldosas para la misma base nitrogenada. En este caso, aunque aparezca una advertencia, las áreas donde se encuentran estas bases de mala calidad están localizadas y no son muy extensas, por lo que, al ser la mayoría de mis baldosas de color azul, se puede pasar un poco por desapercibida esta grafica.

9.1.4 Per sequence quality scores

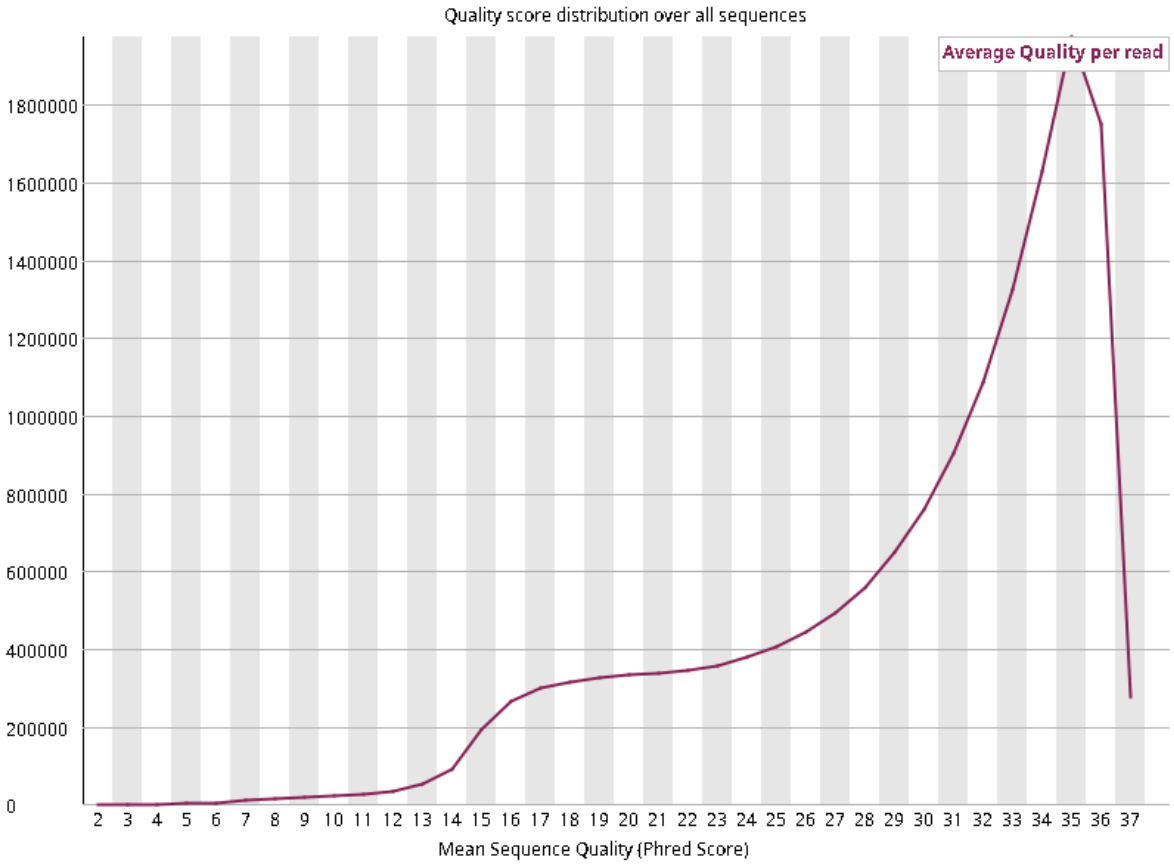

Figura 14: Grafico de puntuaciones de calidad por secuencia

En la figura 14 el eje X corresponde a la calidad media de las secuencias, mientras que el eje Y representa la cantidad de lecturas que presentan dicho valor. Se observa el pico más pronunciado (mayor cantidad) alrededor del valor medio de calidad de 35, el cual se considera alto. Esto sugiere que, tras una etapa de filtrado, no se perderán cantidades significativas de lecturas.

9.1.5 Per base sequence content

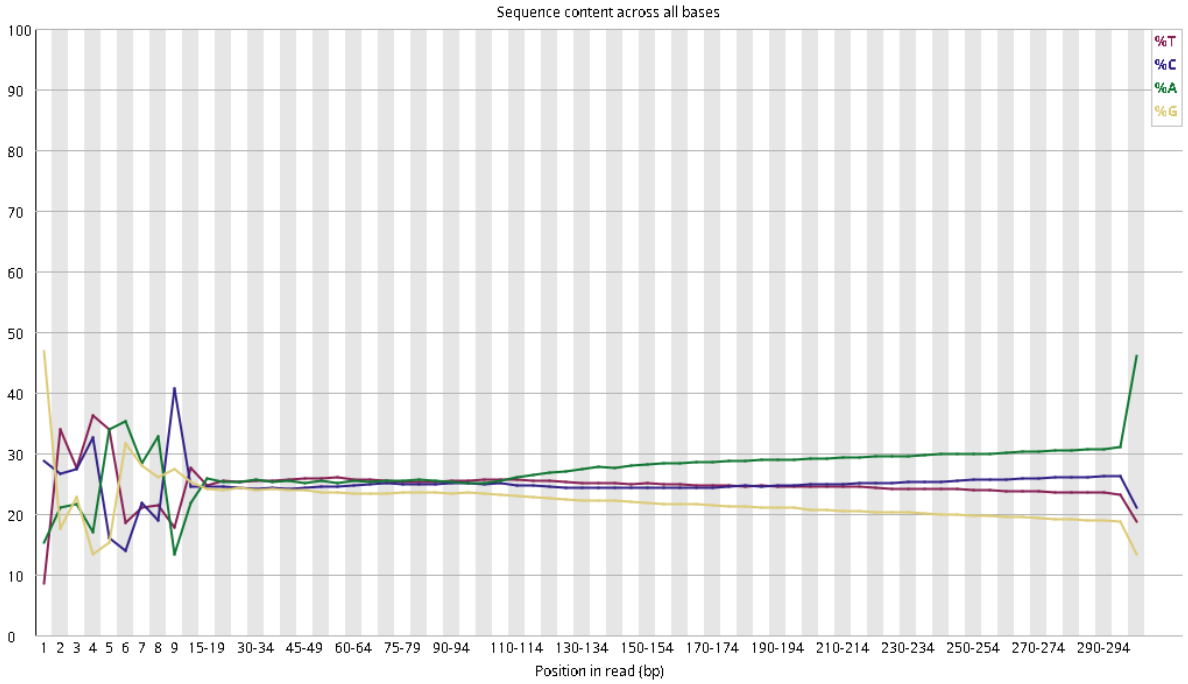

Figura 15: Grafico de contenido de bases por posición en el total de las lecturas

En la figura 15, el eje X representa la posición en la lectura, mientras que el eje Y él % de cada base nucleotídica, las líneas de colores representan cada uno de los porcentajes de bases nitrogenadas, siendo: rojo: timina; azul: citosina; verde: adenina; amarillo: guanina.

Se puede observar un sesgo en las primeras 15 y en las últimas 6 bases de las secuencias, lo cual hace que se dispare una alerta de resultados muy inusuales, este failure aparece cuando hay una diferencia entre A y T o G y C superior a 20% en cualquier posición. Existen múltiples escenarios que pueden generan una advertencia de este tipo, uno de ellos serían las secuencias sobrerrepresentadas, como lo son adaptadores. Otra posible causa es la biblioteca de composición sesgada, por ejemplo, tratar la biblioteca con bisulfito de sodio que transforma las citosinas en timinas, sin embargo, no parece ser el caso en esta secuencia porque en el sesgo hay mayor cantidad de citosinas que de adeninas en varias posiciones. Otras causas pueden involucrar presencia de contaminantes en la muestra secuenciada (material genético del núcleo, mitocondria o de otra especie). En este punto de los resultados es imposible inferir cual es el causante de este sesgo, sin embargo, esta incógnita se buscará responder durante el desarrollo de la metodología.

9.1.6 Per sequence GC content

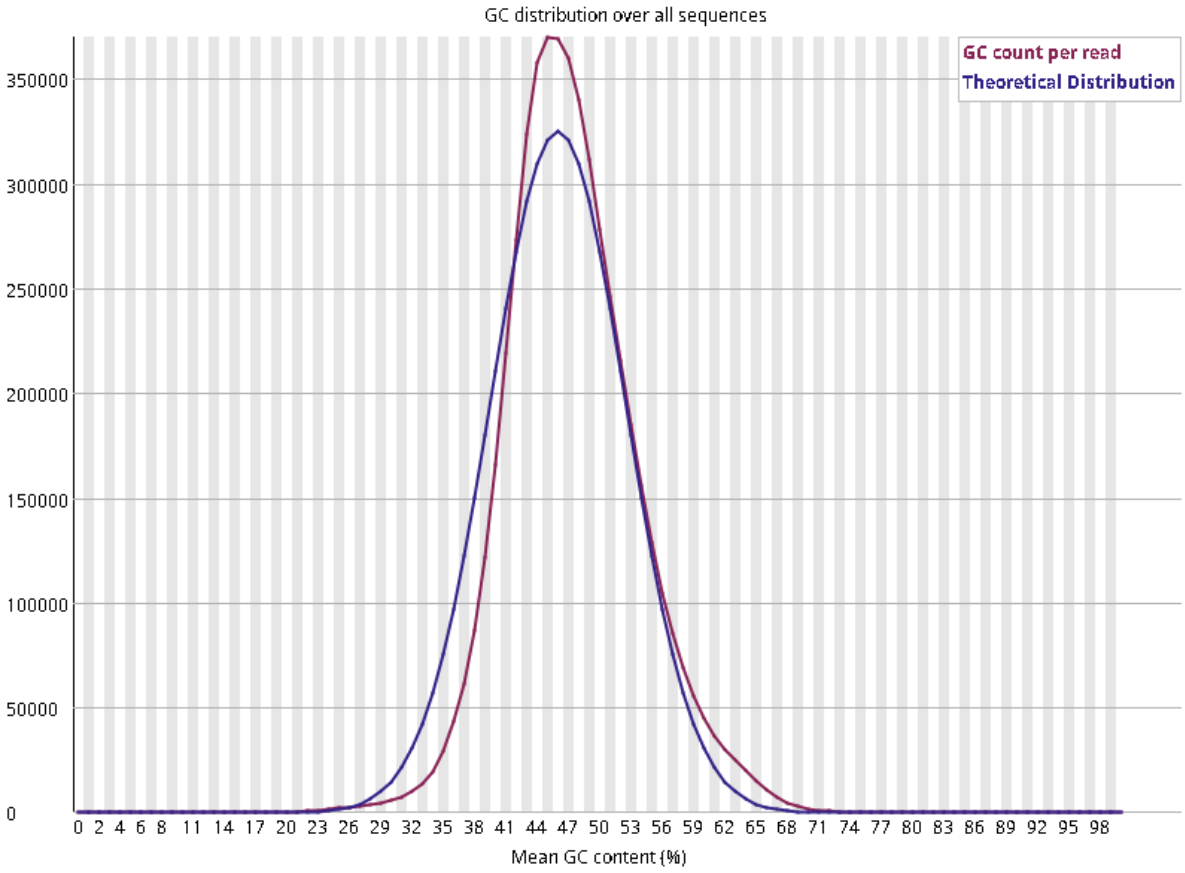

Figura 16: Grafico de contenido de GC en toda la longitud de cada secuencia

En la figura 16 el eje X significa el %GC, mientras que el Y significa el número de secuencias con cada %GC. La línea roja representa el %GC por lectura, mientras que la línea azul representa una distribución teórica del %GC basada en una distribución normal.

El grafico de contenido de GC por secuencia (figura 15) viene adjunto con una alerta de resultados ligeramente inusuales o warning, esta señal aparece cuando la suma de las desviaciones de la distribución normal representa más del 15% de las lecturas. La aparición de esta alerta puede ser resultado de un contaminante especifico (como adaptadores) o la contaminación con material genético del núcleo o de la mitocondria, al igual que la aparición de material genético de una especie diferente. La causa de esto lo averiguaremos con el desarrollo de la metodología.

9.1.7 Per base N content

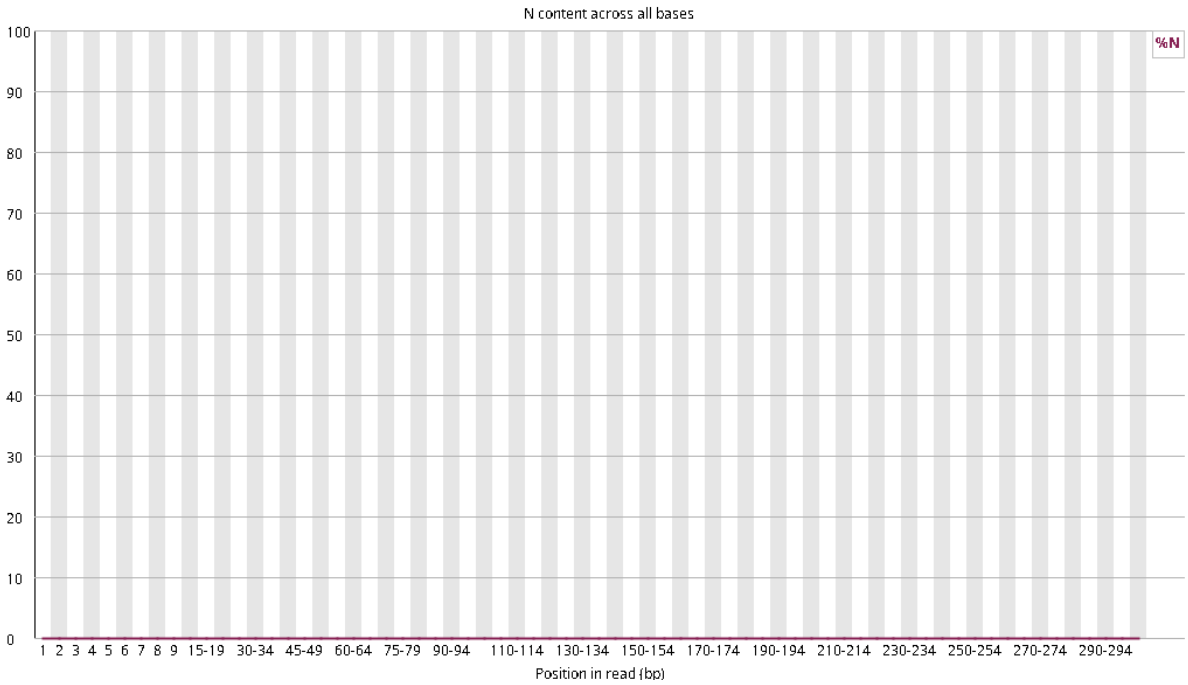

Figura 17: Grafico de contenido de N en todas las bases

En la figura 17 el eje X significa la posición en la lectura y mientras que el eje Y significa el porcentaje de presencia de N.

Cuando un secuenciador no puede realizar el llamado de base de manera precisa, en lugar de asignar un nucleótido específico en las secuencias generadas, se coloca una "N". En el contexto de los datos que se analizarán en esta tesis, no se observa la presencia de "N". Esto sugiere que todos los llamados de base se llevaron a cabo correctamente.

9.1.8 Sequence length distribution

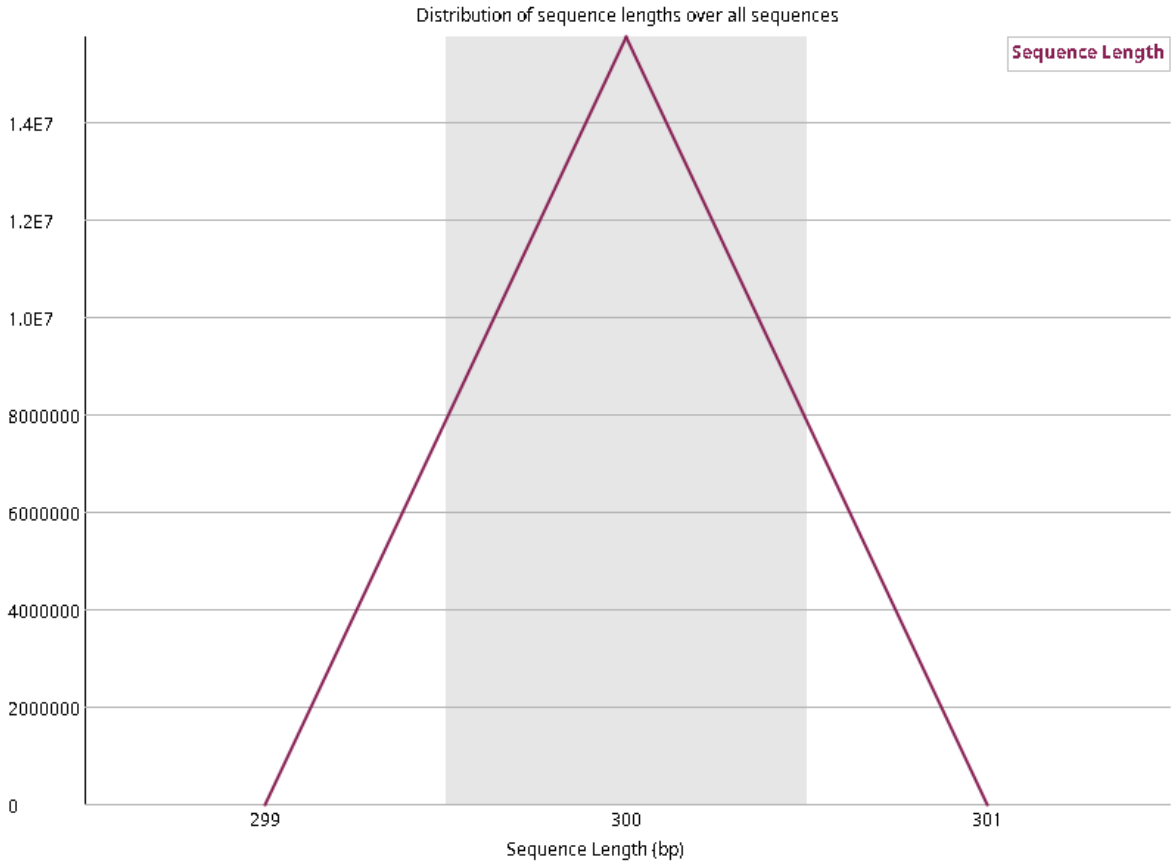

Figura 18: Distribución de la longitud de secuencias

En la figura 18 el eje X significa la longitud de las secuencias, mientras que el eje Y es la cantidad de secuencias con este tamaño.

Algunos secuenciadores de alto rendimiento generan fragmentos de secuencias de longitud uniforme, sin embargo, otros pueden generar lecturas con longitudes muy variables. Este módulo nos muestra que la longitud de las secuencias generadas por illumina Miseq son uniformes, coincidiendo en 300 pb.

9.1.9 Sequence duplication levels

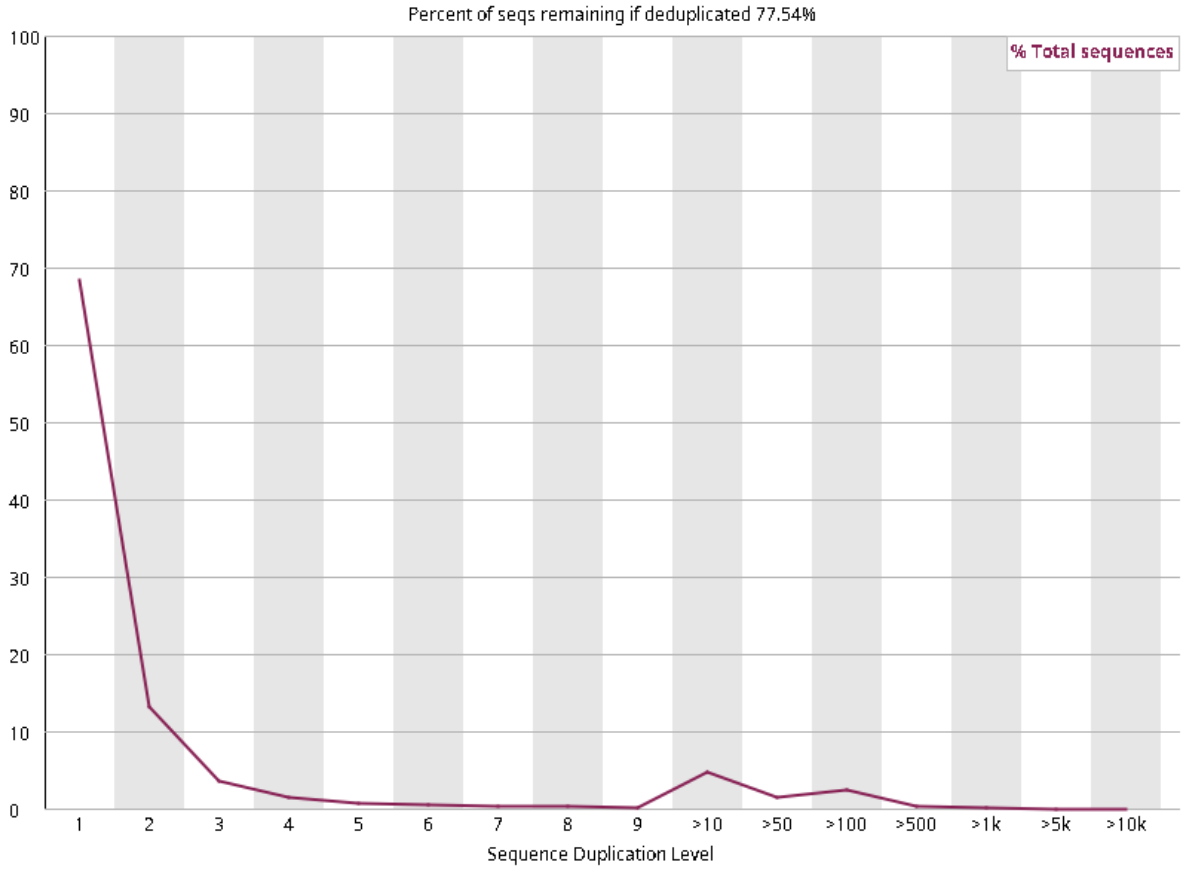

Figura 19: niveles de duplicación de secuencias

En la figura 19, el eje X significa el nivel de duplicación de secuencias, mientras que el eje Y significa el porcentaje de secuencias que representa ese nivel de duplicación.

En una secuencia diversa la mayoría de fragmentos de DNA se secuenciarán una sola vez, un nivel alto de duplicación indica un sesgo de enriquecimiento; el gráfico de este módulo se genera utilizando las primeras 100,000 secuencias de cada archivo, de esta manera se evita sobrecargar el equipo computacional, siendo 100,000 secuencias un buen representante de la población total de secuencias. Con los datos utilizados en esta tesis se puede observar que no hay gran cantidad de datos duplicados, lo que indica que nuestra biblioteca de datos es abundantemente diversa.

9.1.10 Overrepresented sequences

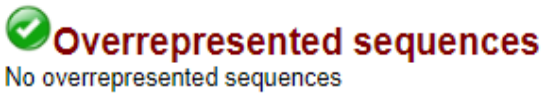

Figura 20: Tabla de secuencias sobrerrepresentadas; en esta biblioteca de datos no hay secuencias sobrerrepresentadas

Generalmente este módulo muestra una tabla con las secuencias sobrerrepresentadas, mostrando la secuencia, numero de repeticiones, porcentaje que representan del total y la posible fuente de esta sobrerrepresentación, sin embargo, para estos datos no se generó ninguna tabla, ya que como muestra FASQC, no se cuenta con secuencias sobrerrepresentadas.

9.1.11 Adapter Content

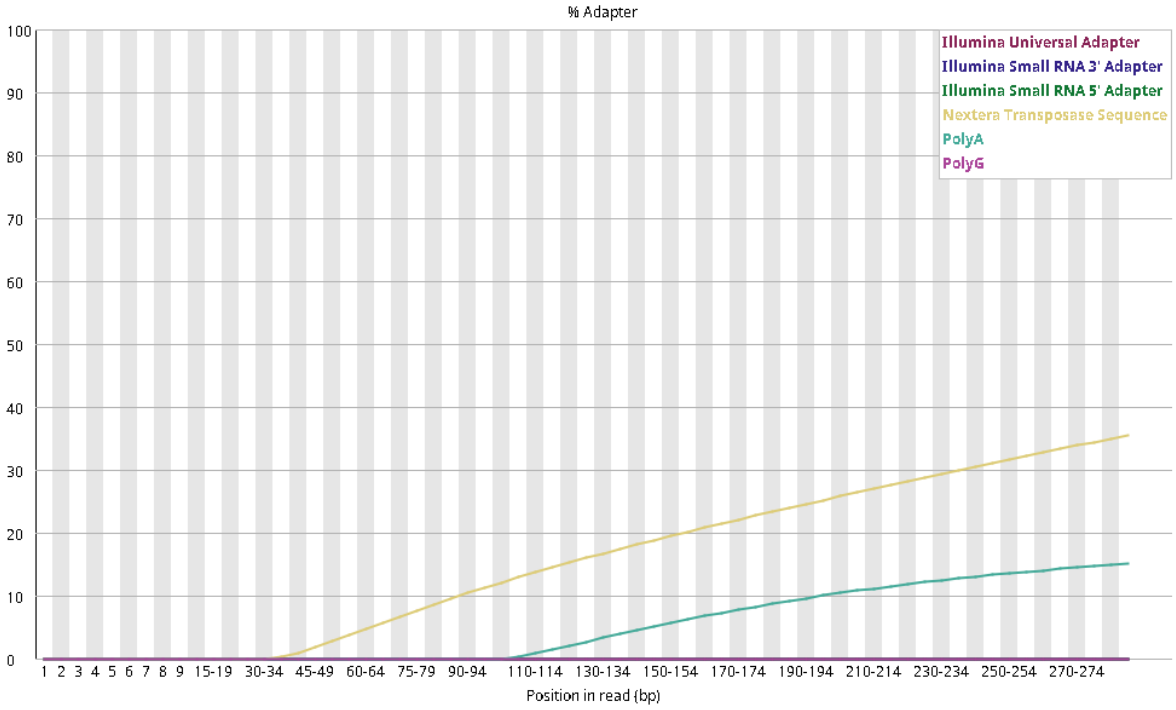

Figura 21: Grafico de contenidos de adaptadores

En la figura 21 el eje X significa la posición en la lectura y el eje Y significa el porcentaje de representación de estas secuencias adaptadoras correspondiente a cada adaptador.

Las secuencias adaptadoras son un tipo de secuencia que muy probablemente este presente en un conjunto de datos sin procesar, por lo mismo, es necesario identificar si nuestras

secuencias tienen una cantidad significativa de adaptadores, con el fin de evaluar si es necesario recortar los adaptadores. Este módulo nos muestra que nuestras secuencias aún contienen adaptadores, específicamente los Nextera Transposase Sequence y PolyA. Por lo que será necesario un recorte de adaptadores para poder continuar con los siguientes pasos de la metodología.

9.2 Secuencias limpias:

Tras analizar los gráficos de calidad de las secuencias del genoma de *Dunaliella salina* cepa noruega con FASTQC, se determinó necesaria una etapa de limpieza de datos, con el fin de eliminar errores intrínsecos de la secuenciación que podrían estropear los análisis siguientes.

Tras repetir múltiples veces los pasos de filtrado y alienación buscando la mejor combinación de parámetros, se obtuvieron secuencias limpias que balancean la calidad con la cantidad de lecturas. Las gráficas generadas por FASTQC a partir de las lecturas limpias son las siguientes:

9.2.1 Basic Statistics

Measure	Value
Filename	AD_DUNALIELLA_TAAGGCGA-CTCTCTAT.foward.trimmed.paired_R1_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	11307005
Total Bases	2.5 Gbp
Sequences flagged as poor quality	0
Sequence length	100-300
%GC	48

Figura 22: Resumen general de calidad de las secuencias después del filtrado.

De forma general se puede observar una reducción en la cantidad de secuencias, pasando de 15,742,069 secuencias con los datos crudos a 11,307,005 secuencias con los datos ya filtrados, además de esto se observa que la longitud en las secuencias pasó de mantener todas con una longitud de 300 bases en las lecturas crudas a variar entre un rango de 100 a 300 bases en las secuencias ya limpias.

9.2.2 Per base sequence quality

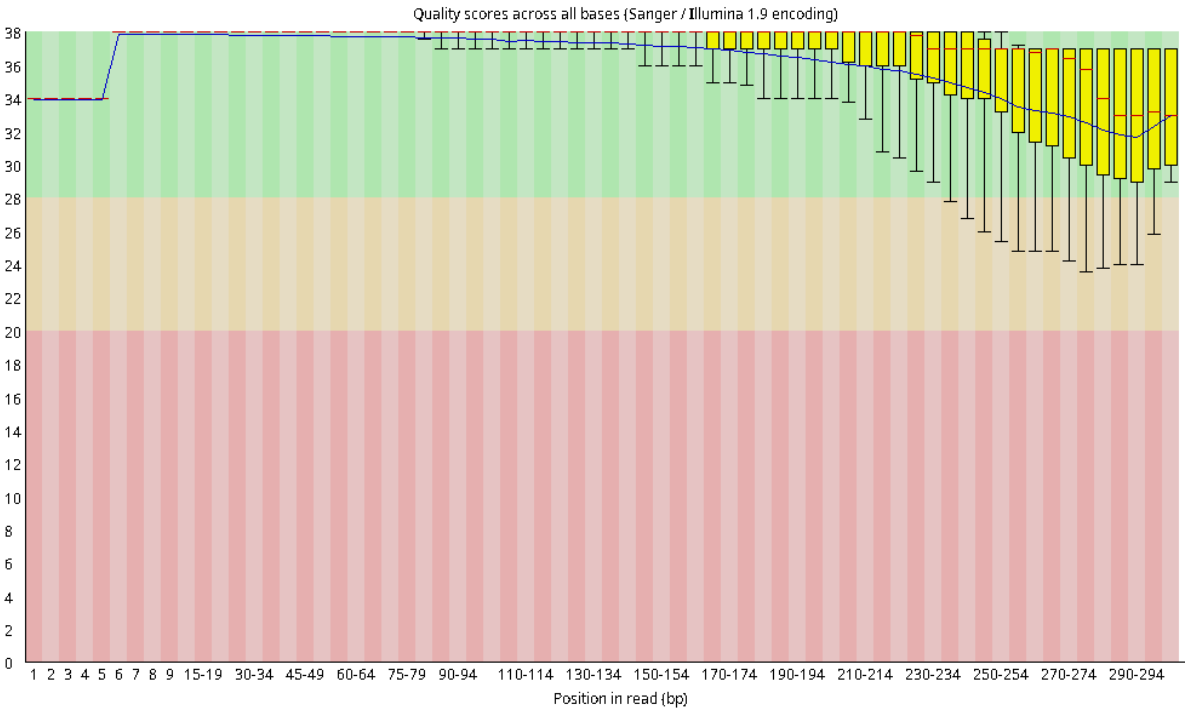

Figura 23: Gráfico de calidad promedio por posición de base en las lecturas de secuenciación después del filtrado.

En comparación con la figura 12, se puede observar un aumento de la calidad en cada posición de la figura 23, manteniendo un promedio de calidad por arriba de 30, el cual es el valor considerado estándar para secuencias de buena calidad, ya que representa un error en la secuenciación por cada 1000 bases [125].

9.2.3 Per tile sequence quality

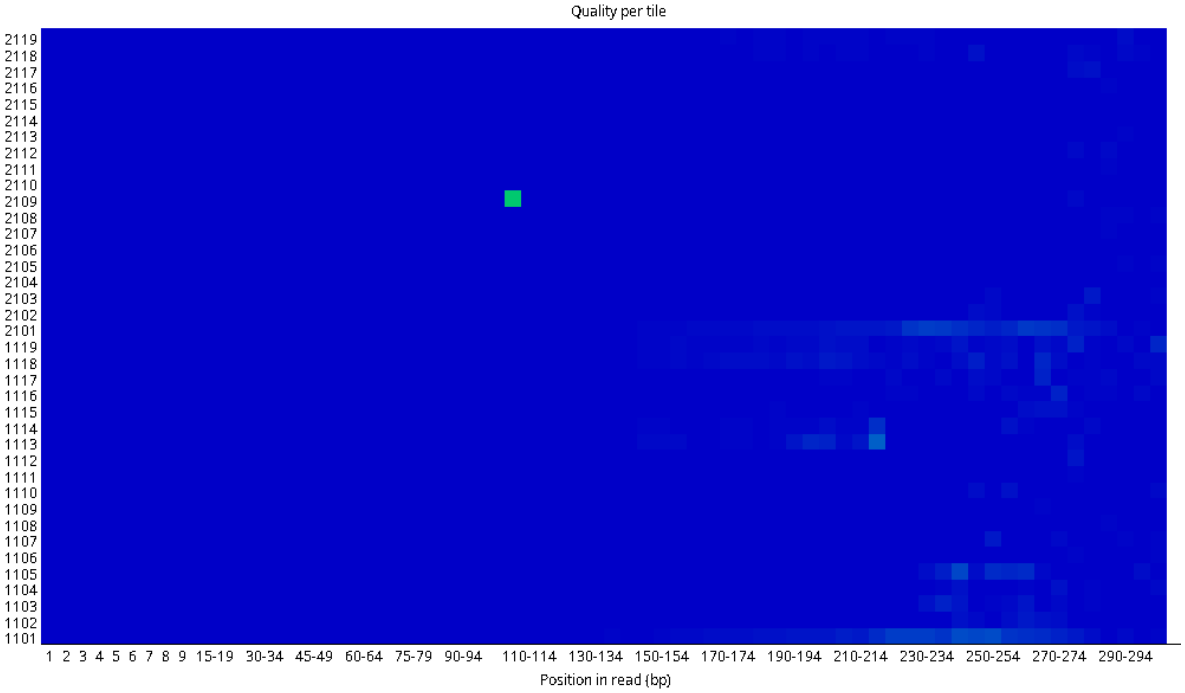

Figura 24: Desviación de la calidad media de cada baldosa (tile) en función de la posición de las lecturas después del filtrado.

Comparando la gráfica 24 con la gráfica 13 generada antes de la limpieza, se observa una homogenización de la calidad en las baldosas, ya que los colores más cercanos al rojo desaparecieron tras eliminar las lecturas con baja calidad, lo que indica que todas las baldosas mantenían valores de calidad cercanos.

9.2.4 Per sequence quality scores

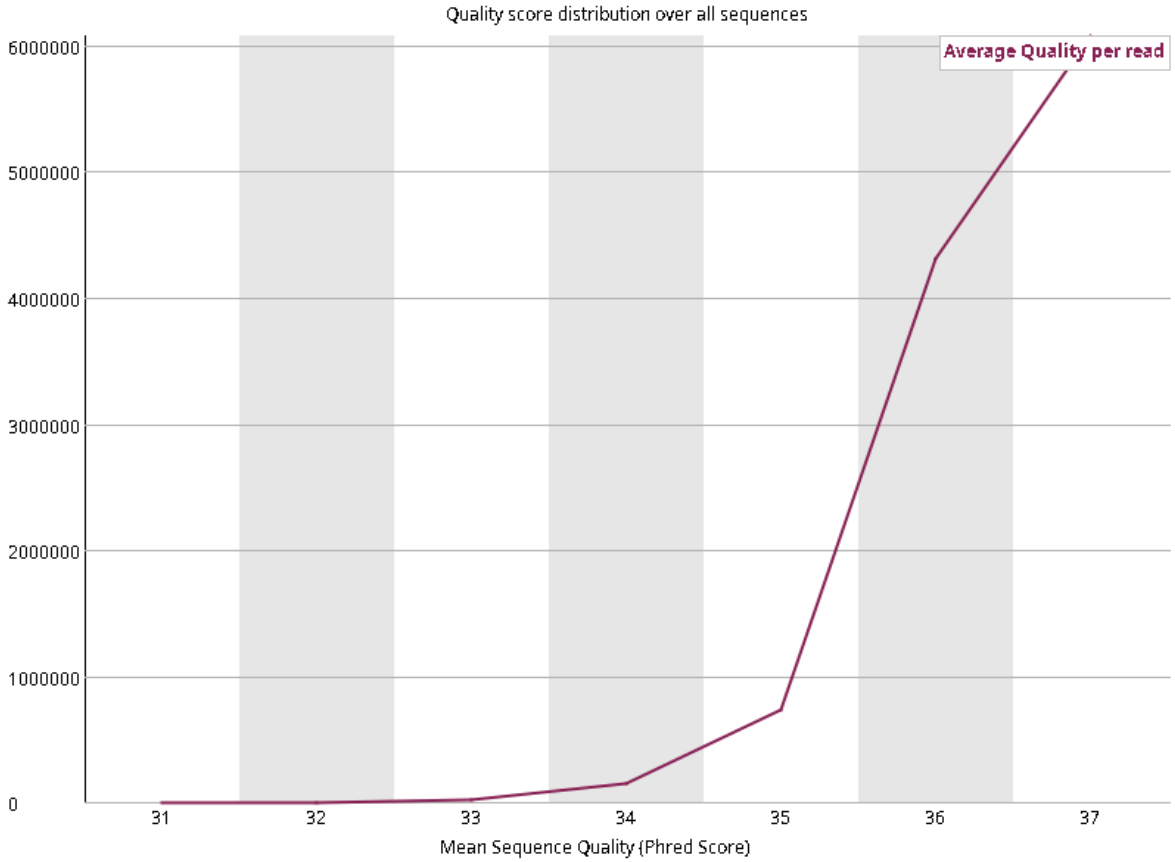

Figura 25: Grafico de puntuaciones de calidad por secuencia después del filtrado.

A diferencia de lo observado en la figura 14, todas las secuencias se encuentran sobre un quality score mayor a 31 después del filtrado, lo cual supera el umbral de calidad para considerarlas secuencias de alta calidad, por lo que este es un buen indicador para ya trabajar con estas secuencias.

9.2.5 Per base sequence content

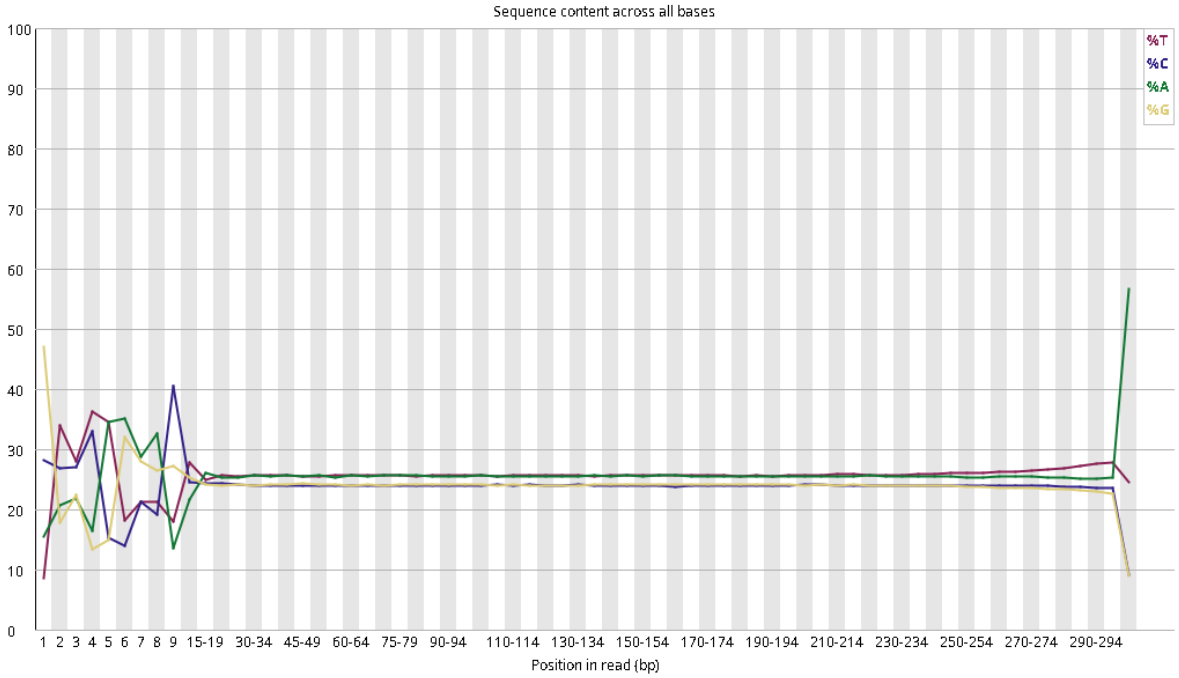

Figura 26: Grafico de contenido de bases por posición en el total de las lecturas después del filtrado

Al igual que en la figura 15, en la figura 26 se observa un sesgo en las primeras 15 y en los últimos 6 pares de bases, al mismo tiempo, en las demás posiciones se ve una regularización de las líneas, quedando más paralelas unas de otras después del filtrado. Sin embargo, llama la atención este sesgo que sigue sin desaparecer, el cual podría ocasionar errores en futuros pasos.

9.2.6 Per sequence GC content

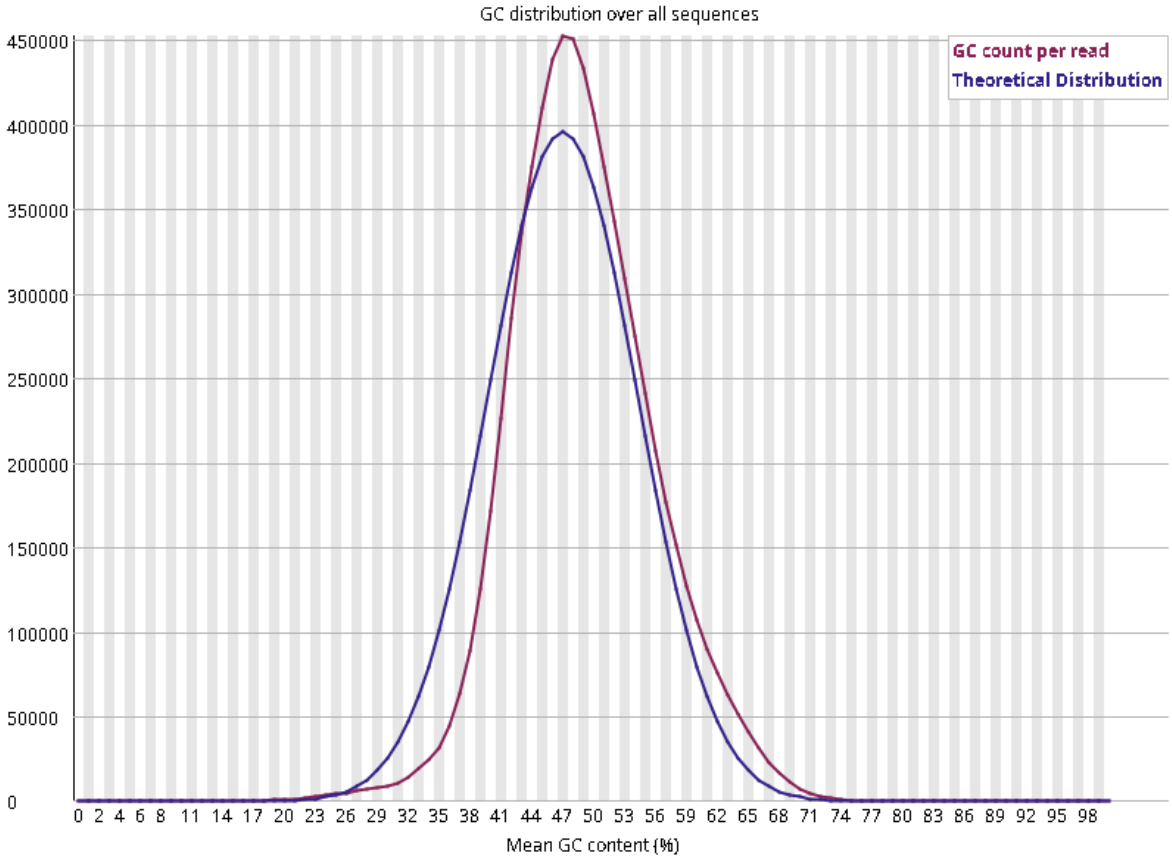

Figura 27: Grafico de contenido de GC en toda la longitud de cada secuencia después del filtrado

Al igual que en la figura 16, en la figura 27 se puede observar que distribución real de %GC es mayor que la distribución teórica de este porcentaje, lo cual es de esperar al trabajar con secuencias de un organismo extremófilo.

9.2.7 Per base N content

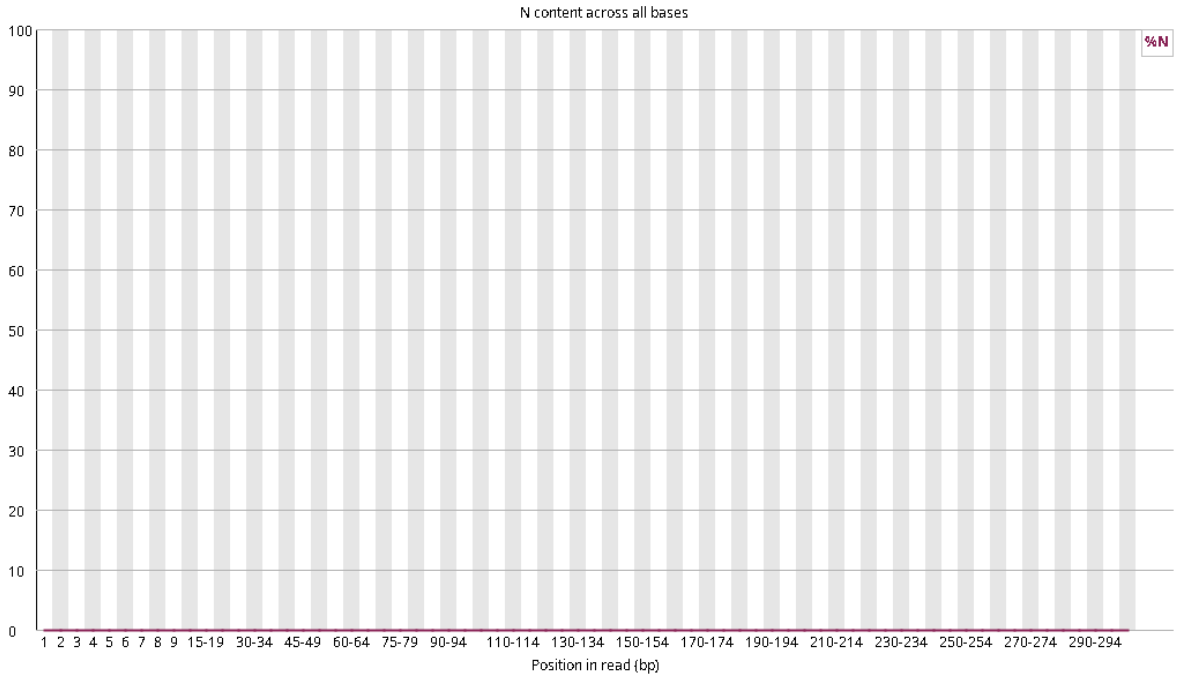

Figura 28: Grafico de contenido de N en todas las bases después del filtrado

Al igual que en la figura 17, no se detectaron presencia de “N”s en la figura 28.

9.2.8 Sequence lenght distribution

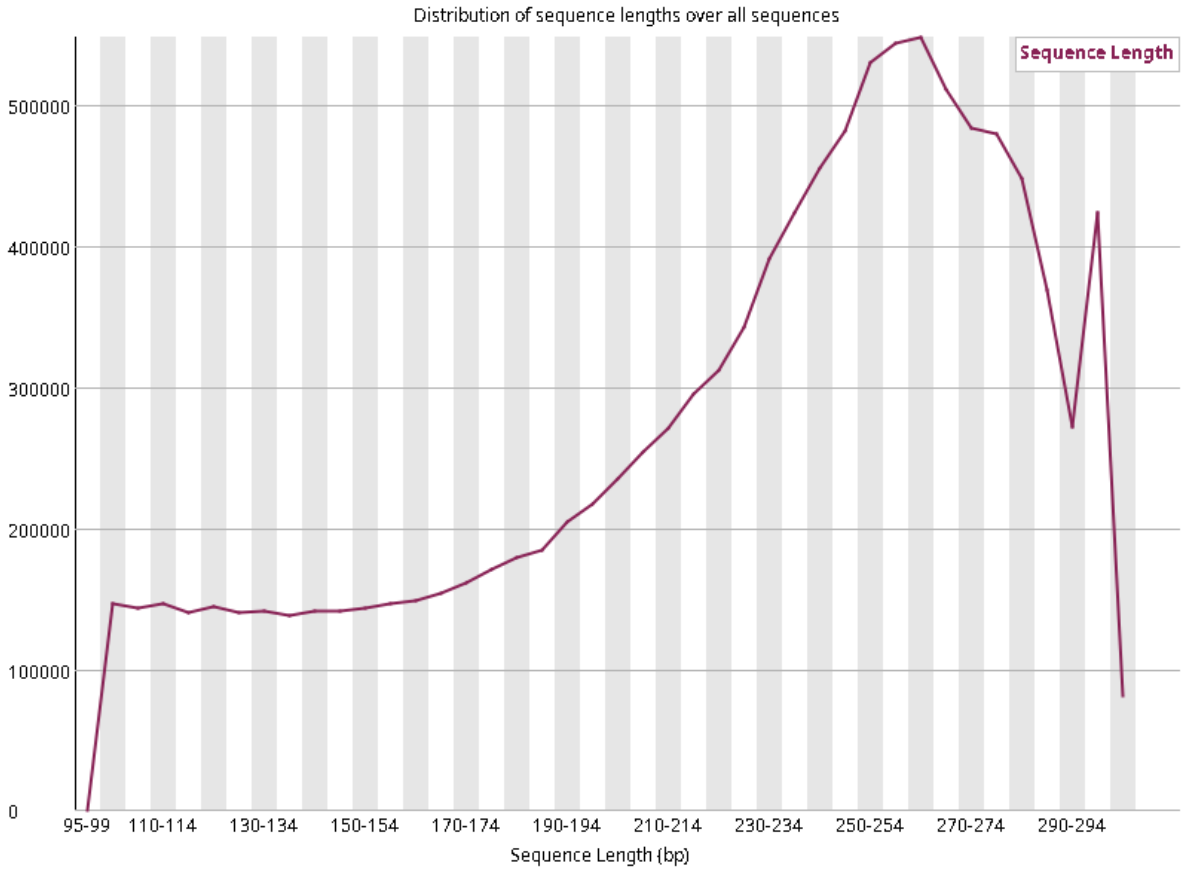

Figura 29: Distribución de la longitud de secuencias después del filtrado.

A diferencia de lo visto en la figura 18, en la figura 29 se eliminaron secuencias adaptadoras y regiones de mala calidad de las lecturas, generando una variedad de longitudes de lecturas que van desde 100 bases hasta 300 bases, con una mayor cantidad de lecturas con una longitud entre las 250 y 254 bases.

9.2.9 Sequence duplication levels

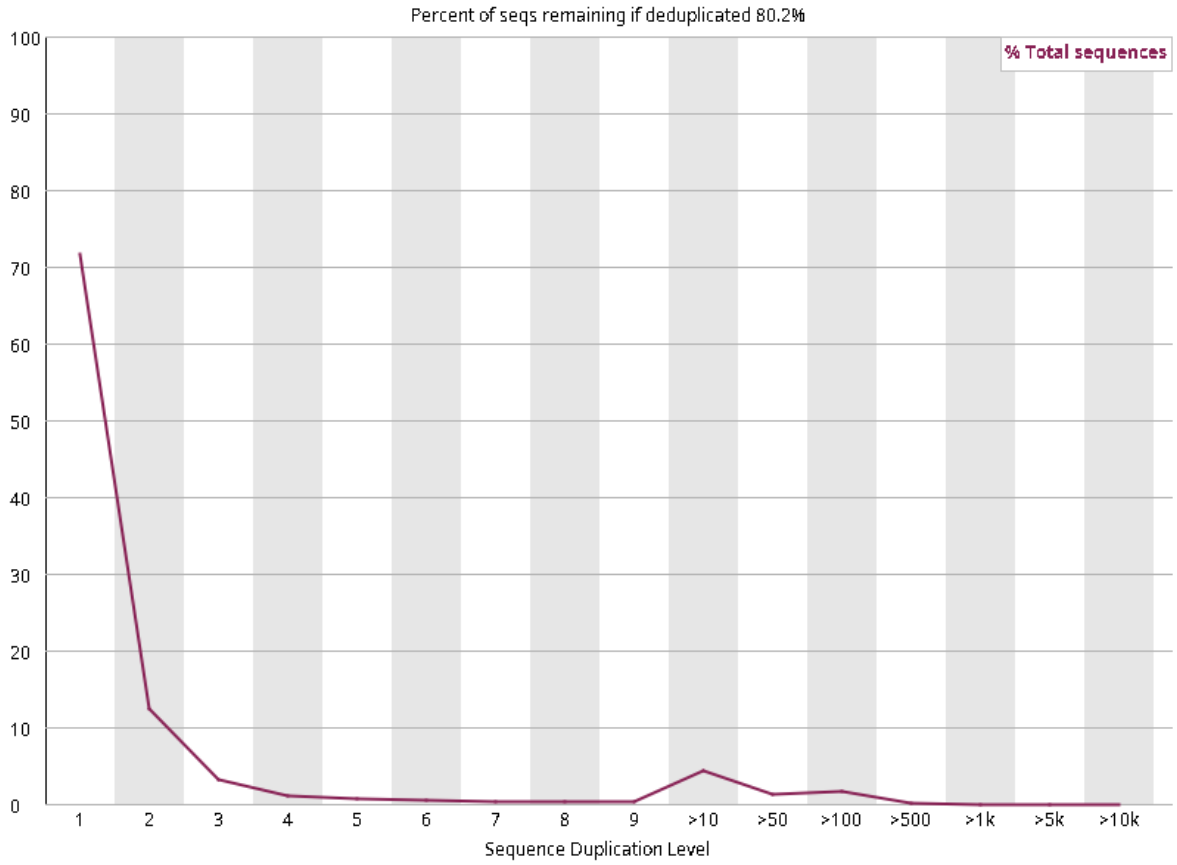

Figura 30: Niveles de duplicación de secuencias después del filtrado.

Los niveles de duplicación de secuencias en la figura 30 se mantuvieron similares a los vistos antes del filtrado en la figura 19, conservando la diversidad de datos en nuestras secuencias.

9.2.10 Overrepresented sequences

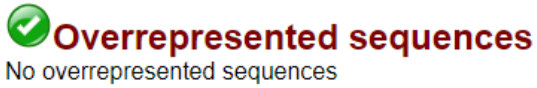

Figura 31: Tabla de secuencias sobrerrepresentadas después del filtrado.

Al igual que lo visto en la figura 20, en la figura 31 no se encuentran secuencias sobrerrepresentadas, lo que nos confirma una vez más que la biblioteca de datos es abundantemente diversa.

9.2.11 Adapter content

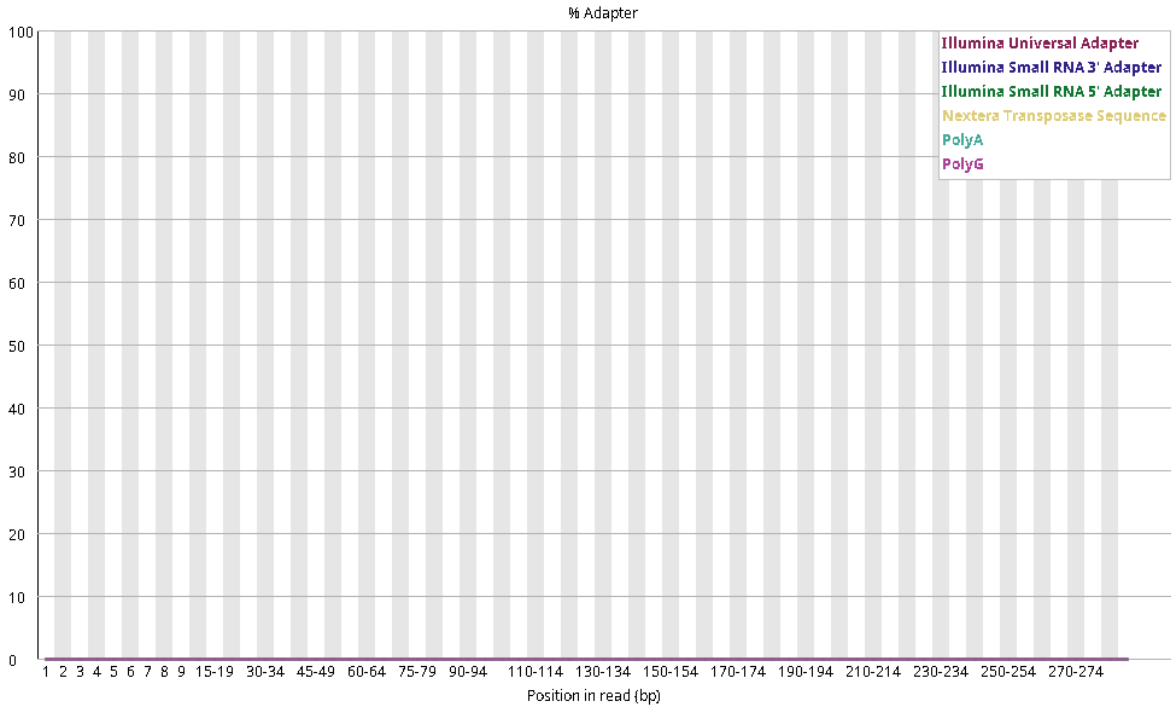

Figura 32: Grafico de contenidos de adaptadores después del filtrado.

A diferencia de lo observado antes de la limpieza y filtrado en la figura 21, los adaptadores en las lecturas han desaparecido en la figura 32, otorgando un conjunto de datos limpios que nos permite continuar con los siguientes pasos de la metodología.

9.3 Ensamble de novo

Continuando con el ensamble *de novo* a partir de las lecturas limpias, es necesario extraer las lecturas del cloroplasto de todo el conjunto de secuenciación WGS alineándolo contra la referencia de un genoma de cloroplasto evolutivamente cercano.

Para *Dunaliella salina* cepa noruega se utilizaron como referencia dos plastomas de *Dunaliella salina* disponibles en NCBI, el primero correspondiente a la cepa CCAP19/18 extraída de la laguna Hutt en Australia Occidental (Numero de acceso en RefSeq: NC_016732.1) y el segundo obtenido de la cepa SQ recolectada en San Quintín, Baja California, México (Número de acceso GenBank KX530454.1).

El programa utilizado para realizar las alineaciones fue bowtie2, las características de cada una de las alineaciones se resumen en la tabla 14:

Tabla 14. Comparación de alienaciones generadas con secuencias WGS de *Dunaliella salina* cepa noruega

	CCAP19/18	**SQ**
Tamaño referencia	269,044	243,635
Número de lecturas	22,614,010	22,614,010
Lecturas mapeadas	78,802 / 0.35%	173,286 / 0.77%
Lecturas no mapeadas	22,535,208 / 99.65%	22,440,724 / 99.23%
Lecturas mapeadas en pares	65,466 / 0.29%	173,084 / 0.77%
Lecturas alineadas en singular (singletons)	13,336 / 0.06%	202 / 0%
Porcentaje cubierto	56.21%	99.96%
Profundidad de cobertura	61.1481	146.7702

En un conjunto de datos WGS de células de planta es común encontrar de un 4.5 – 12 % de lecturas provenientes del cloroplasto (dependiendo del método de conservación de la muestra) [126]. Del total de nuestras secuencias se ve un porcentaje más bajo de lecturas alienadas que lo reportado anteriormente, siendo de 0.35% y 0.77% contra los genomas de las cepas CCAP19/18 y SQ respectivamente, por lo que, sumado a lo reportado en la figura

26 (per base sequence después del filtrado), es posible que el sesgo reportado en las primeras 16 y ultimas 9 bases cause problemas en el puntaje de alineación de cada secuencia, ocasionando que se disminuya el número de lecturas alineadas.

Para corroborar si este sesgo no ocasiona problemas en la extracción de secuencias del cloroplasto, se decidió utilizar dos distintos alineadores para realinear las lecturas:

- Bowtie2
- Burrows-Wheeler Aligner (bwa) [127]

Estos alineadores se utilizaron con dos conjuntos de datos, los datos como salieron de la limpieza y otro conjunto donde se removieron los sesgos presentes al inicio y al final de las secuencias. Para la eliminación de estos sesgos se utilizaron las funciones crop y headcrop de trimmomatic que cortan las lecturas desde sus extremos (figura 33).

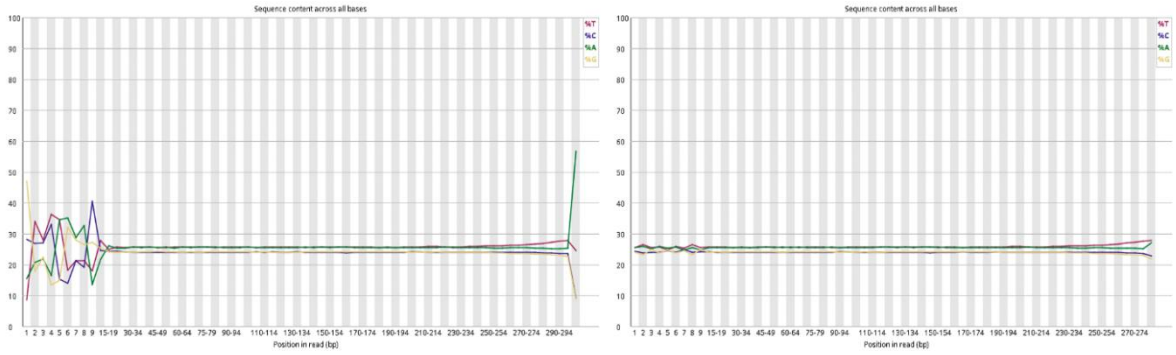

Figura 33: Gráfico “Per base sequence content” con sesgo (lado izquierdo) y sin sesgo (lado derecho)

Los resultados de los realineamientos se ven en la tabla 15, donde se comparan los dos alineadores con distintos modos de trabajo y distintos parámetros de alineación.

Tabla 15. Tabla comparativa de la cantidad y porcentaje de lecturas alineadas de *Dunaliella salina* cepa noruega WGS al cloroplasto, utilizando distintos alineadores.			
Con sesgo			
Alineador	Modo	Parámetros	Lecturas alineadas
Bowtie2	End to end	Default	173084 / 0.77%
Bowtie2	local	Very sensitive local	174140 / 0.77%
bwa	MEM (similar a local)	Default	187378 / 0,.83%
Sin sesgo			
Alineador	Modo	Parámetros	Lecturas alineadas
Bowtie2	End to end	Very sensitive	173100 / 0.77%
Bowtie2	End to end	Very sensitive -N 1 -L 15	173058 / 0.77%
Bowtie2	Local	Very sensitive local	174126 / 0.77%
Bowtie2	Local	Very sensitive local -N 1 -L 15	175016 / 0.77%
bwa	MEM (similar a local)	Default	185816 / 0.82%

Al hacer esta comparativa, se detectó que no existe una variación significativa entre el número de lecturas alineadas en función del sesgo detectado, aun así, se detectó una variación mínima al cambiar de alineador, sin embargo, no es para nada cercano a lo esperado. Aun así, podemos decir que no existe problema con la secuenciación ni con el proceso de alineación.

Esta baja cantidad de lecturas alineadas se puede explicar revisando el proceso de preparación de una muestra de ADN de cloroplasto (cpDNA) para secuenciar. Por lo general, al secuenciar lecturas de cpDNA, las muestras pasan por un proceso de aislamiento y enriquecimiento, donde se separa el material genético del material nuclear y mitocondrial, estas técnicas incluyen un proceso de separación del cloroplasto de otros organelos, lisis de cloroplasto y purificación de cpDNA [128], y se realizan en medios de alta fuerza iónica [129]. Sin un paso de enriquecimiento, las lecturas provenientes del cloroplasto se verán "diluidas" en comparación con las lecturas que provienen del núcleo y la mitocondria.

En la actualidad aún no se ha logrado resolver la arquitectura del genoma nuclear de *Dunaliella salina*, sin embargo, se estima un tamaño aproximado de 300 Mpb para *Dunaliella salina* cepa CCAP19/18 [102], lo cual considerado los plastomas *Dunaliella salina* ya reportados para las cepas CCAP19/18 y cepa SQ, con tamaños de 269 kpb y 243 kpb

respectivamente y recordando que existen múltiples copias del plastoma dentro de un cloroplasto hace sentido el porcentaje tan pequeño de lecturas provenientes de este organelo en nuestros datos WGS.

Debido a que se tenía un mayor progreso con los datos alineados con bowtie2 modo –end-to-end con parámetros default, se decidió mantener el trabajo con estos resultados.

A partir del archivo de alineación, se extrajeron las lecturas correspondientes al cloroplasto, esto se logró utilizando la herramienta samtools, generando dos archivos fastq, uno para las lecturas R1 y otro para las lecturas R2 (figura 34).

Measure	Value
Filename	SAMPLE_R1_2.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	86542
Total Bases	19.5 Mbp
Sequences flagged as poor quality	0
Sequence length	100-300
%GC	32

Figura 34: Tabla generada por FASTQC donde se resumen los datos generales de calidad y características de las secuencias provenientes de *Dunaliella salina* cepa noruega, después de ser limpiadas

Una profundidad de 50x es considerada adecuada para realizar un ensamble *de novo* de manera eficiente [130], [131]. Con las lecturas alineadas contra la referencia SQ se alcanza una cobertura de 147x, lo cual es más que óptimo para realizar un ensamble *de novo.*

Los programas elegidos para realizar el ensamble de novo con las secuencias del cloroplasto fueron SPADES y Velvet [132].

Se generó un script en bash que automatiza el uso de SPADES con distintos valores de kmers. Los valores utilizados con este script varían entre 21 y 99 incrementándose en intervalos de 2 unidades entre cada valor. Dando un total de 40 ensambles de los cuales se revisó cada uno,

separando los que presentan valores de N50 más altos para utilizarlos en conjunto en un ensamble posterior con el mismo programa. Los valores de kmer seleccionados fueron 47, 49, 51, 53, 55, 57, 59, 63, 65, 69, 71, 77 y 97. Por otro lado, con velvet se utilizó el script "VelvetOptimiser", una herramienta diseñada para automatizar y optimizar el proceso de ensamble con esta herramienta, el script explora múltiples ensambles generados con diferentes valores de kmers dentro de un rango definido por el usuario, buscando el valor que optimice la calidad del ensamblaje, utilizando la métrica N50 para definir su calidad. Las características de los ensambles se ven en la tabla 16, mientras que los grafos resultantes se ven en la figura 35.

Tabla 16. Comparación de características de los ensambles generados con spades y velvet.

Ensamblador	SPADES	Velvet
Número de nodos	12	23
Longitud total	227,046 pb	225,138 pb
N50	34,555 pb	17,056 pb
Nodo más corto	98 pb	467 pb
Nodo más largo	53,540 pb	30,534 pb
Profundidad promedio	67.3x	68.8x

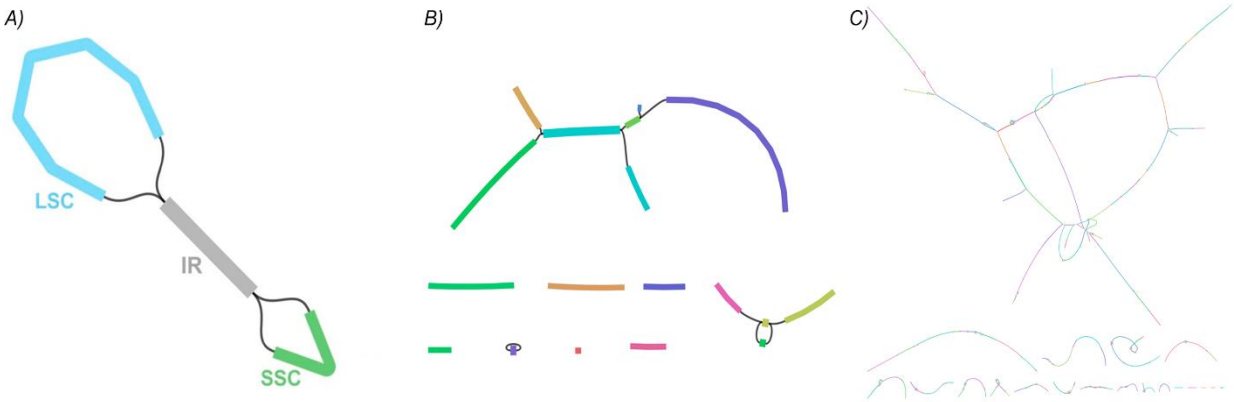

Figura 35: Imagen de los grafos generados por los ensambladores contra la vista típica de un cloroplasto con una estructura quadripartida. A) Estructura quadripartida típica de cloroplasto, B) grafo generado por SPADES, C) grafo generado por velvet.

Como se puede observar en los gráficos de ensamble del cloroplasto de *Dunaliella salina*, ni el ensamble con SPADES (Figura 35B) ni el ensamble con velvet (figura 35C) logran cerrarse ni alcanzar la estructura cuadripartida típica (figura 35A), la cual si se encuentra

presente en los plastomas de las cepas CCAP19/18 y la cepa SQ. Para buscar explicar este problema debemos de revisar el origen del cloroplasto, recordemos que este organelo proviene de una cianobacteria ancestral, la cual a lo largo de los años fue mudando genes al núcleo y a la mitocondria, esta transferencia horizontal de genes dio como resultado fragmentos de ADN plastídico presentes en el genoma nuclear (NUPTs, nuclear plastid DNA) y mitocondrial (MTPTs, Mitochondrial Plastid DNA), los cuales tienen una alta similitud con el ADN del cloroplasto [133].

El bajo porcentaje de lecturas de cloroplasto presentes en el conjunto de datos WGS de *Dunaliella salina* cepa noruega ocasiona que, al momento de hacer el ensamble, las lecturas derivadas de los fragmentos NUPTs y MTPTs generen más ruido en comparación con un conjunto de datos donde se enriquecieron las lecturas de cloroplasto correctamente, imposibilitando el cierre correcto del genoma. Estos fragmentos se unen al conjunto de lecturas debido a su alta similitud con lecturas del cloroplasto [134]. Como se menciona en el artículo de Takamatsu et al (2018) “Una proporción alta de cpDNA es fundamental para reducir las lecturas mal alineadas en el plastoma, como las NUPTs o MTPTs,”. Por lo que podemos suponer que es debido a esta contaminación de ADN nuclear y ADN mitocondrial que el ensamble *de novo* fracasó.

9.4 Ensamble por referencia:

Aunque no se haya podido completar el ensamble del genoma del cloroplasto de *Dunaliella salina* cepa noruega con una estrategia *de novo*, aún se puede obtener por medio de un ensamble por referencia. El ensamble por referencia consiste en alinear las lecturas sin artefactos de secuenciación contra un genoma muy cercano evolutivamente, como lo hicimos al momento de extraer las lecturas del cloroplasto del conjunto de lecturas WGS.

Utilizando los archivos generados con el programa bowtie2, se decidió utilizar el plastoma de la cepa SQ para el ensamble, ya que presenta un mayor porcentaje de cobertura, casi el 100%, siendo de 99.96%, contra los 56.21% obtenidos al alinear las lecturas contra el plastoma CCAP19/18, lo que nos indica una mayor similitud entre los plastomas de la cepa noruega y la cepa SQ en comparación con la cepa australiana.

En la figura 35, se logra observar de manera gráfica la cobertura de las lecturas a través de los genomas de referencia, en las figuras 35A y 35B se representa mediante una línea roja la profundidad (número de veces en que una base del genoma de referencia está cubierta por las lecturas alineadas) en función del genoma de referencia,

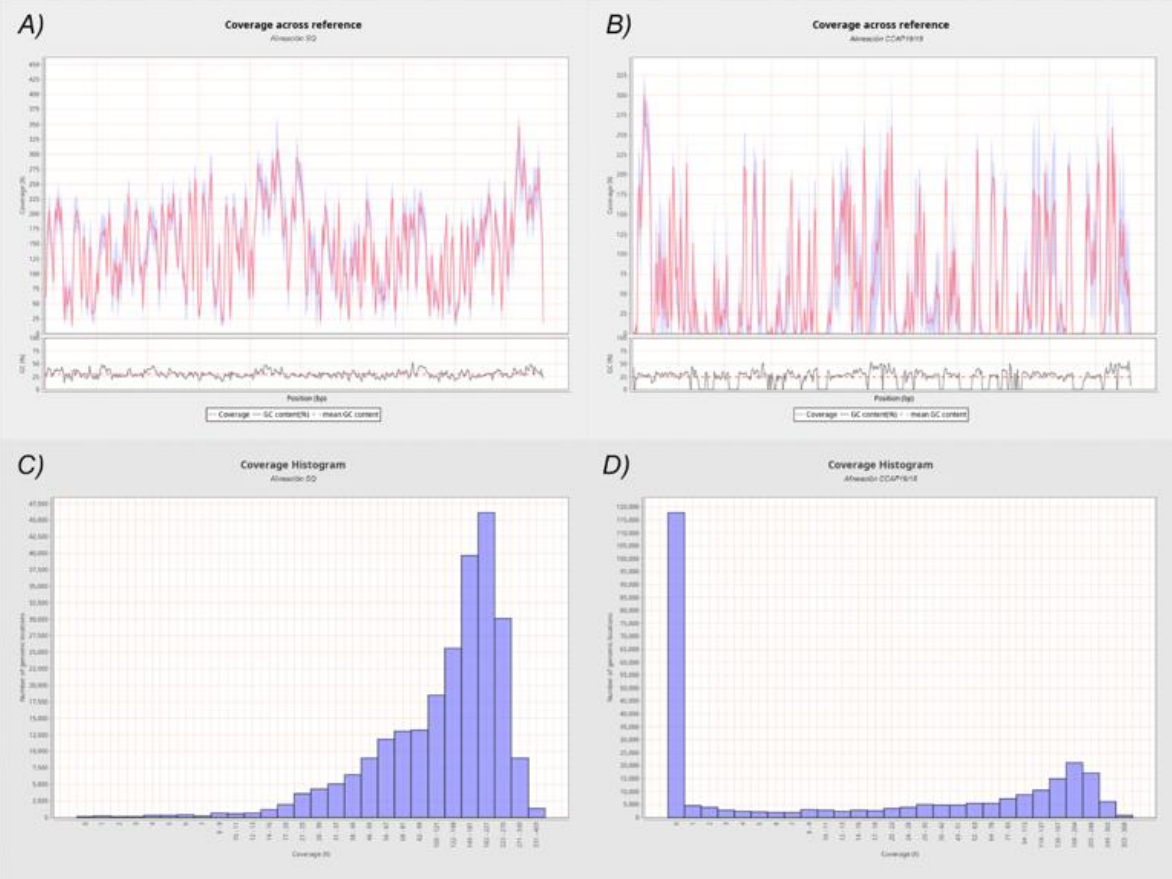

Figura 36. Representación gráfica de la alineación de las lecturas contra los genomas de referencia SQ y CCAP19/18 con histograma de cobertura, generadas con qualimap [142].

siendo 35A el gráfico para el genoma de referencia SQ y 35B el gráfico para el genoma CCAP19/18. Como se puede observar, para el genoma SQ existen pocas regiones donde la profundidad llegará a alcanzar valores por debajo de 20x, a diferencia de la alineación contra el genoma CCAP19/18, donde a simple vista se pueden observar varias regiones sin profundidad, es decir, que no están presentes en el plastoma de *Dunaliella salina* cepa noruega. En las figuras 35C y 35D se representa la cobertura de las alineaciones con histogramas, siendo la figura 35C el histograma de cobertura de la cepa SQ y la figura 35D el histograma de cobertura con la cepa CCAP19/18, en estos histogramas se representan la

cantidad de posiciones en el genoma de referencia (eje Y) que pertenecen a un grupo con una determinada cobertura (eje X), de esta manera se puede apreciar de mejor manera la mayor cantidad de regiones sin cobertura en la alineación contra la cepa CCAP19/18, superando las 115,000 pb sin cobertura, a diferencia de la alineación contra la cepa SQ, que no superan ni las 1000 pb sin cobertura, lo que nos refuerza la idea de trabajar con el plastoma de la cepa SQ para el ensamble por referencia del plastoma de la cepa noruega.

Para un ensamble por referencia, la profundidad recomendable es de 10x [111]. Con esto en cuenta, se procedió a detectar las regiones que presentaran una profundidad menor a la recomendable para ambas alineaciones. La identificación de estas regiones se logró utilizando la herramienta bedtools [135] para generar un archivo de cobertura, en conjunto del comando awk que ayudó a filtrar las regiones que tuvieran una cobertura menor a 10, después se unieron las regiones con baja profundidad que fueran adyacentes utilizando la función "merge" de bedtools, por último, con el comando getfasta se obtuvieron las secuencias correspondientes a las regiones sin cobertura. Una vez con las secuencias extraídas de las regiones sin cobertura en ambas alineaciones, se hicieron nuevas alienaciones, pero ahora entre estas secuencias, con el fin de detectar regiones conservadas entre la cepa SQ y la cepa CCAP19/18. Para buscar estas coincidencias se utilizaron las 3 variaciones de blast disponibles (blast, megablast y discontinuous megablast), siendo disccontinuous megablast el que presentó más coincidencias (incluyendo las obtenidas con blast y megablast).

Las regiones que coincidieron se almacenaron en un archivo fasta, sobre el cual se volvieron a mapear todas las lecturas limpias WGS de *Dunaliella salina* cepa noruega, sin embargo, ninguna lectura se alineó a estas regiones, por lo que podemos decir que, aunque estén presentes en el cloroplasto de *Dunaliella salina cepa* SQ y cepa CCAP19/18, no están presentes en la cepa noruega.

Una vez confirmado que las regiones sin cobertura no están presentes en el cloroplasto de *Dunaliella salina* cepa noruega se puede continuar con el siguiente paso de la metodología.

9.5 Llamado de variantes

El llamado de variantes es el proceso mediante el cual se identifican y catalogan las diferencias observadas entre las lecturas secuenciadas contra un genoma de referencia. Para poder realizarlo se necesita seguir una serie de pasos que consisten en la alineación de lecturas contra el genoma de referencia, paso que quedó completado en el punto anterior. A partir de aquí se puede determinar las variantes de las alineaciones. Para esto se utilizaron dos programas, bcftools y freebayes [136].

Existen distintas clasificaciones para las variantes que pueden encontrarse, Istvan en su libro "The The Biostar Handbook: 2nd Edition" (2020) menciona las siguientes (tabla 17):

Tabla 17. Variantes genómicas típicas

Nombre	Acrónimo	Definición
Polimorfismo de un solo nucleótido (Single Nucleotide Polymorphism)	SNP	Un cambio en una sola base
Inserción o eliminación	INDEL	Una sola base añadida o eliminada
Variante de un solo nucleótido	SNV	Una variante de un solo nucleótido, puede ser un SPN o INDEL, pero el cambio debe de tener un solo par de base de longitud
Polimorfismo de múltiples nucleótidos (multi-nucleotide Polymorphism)	MNP	Variante de múltiples nucleótidos, varios SNPs consecutivos.
Variante de múltiples nucleótidos (multi-nucleotide variant)	MNV	Variante de múltiples nucleótidos, pueden ser múltiples SNPs o múltiples INDELs
Variación corta	-	Variaciones MNVs menores a 50 pb de longitud
Variación larga	-	Variaciones MNVs mayores a 50 pb de longitud

Tabla elaborada con información obtenida de [137]

En un primer llamado de variantes se utilizó bcftools sobre el archivo de alineación (.bam) correspondiente a la alineación contra el plastoma SQ con el programa bowtie2 y parámetros default. De este llamado de se detectaron únicamente 3 variantes, las cuales se ven reflejadas en la tabla 18:

Tabla 18: Variantes detectadas con bcftools sobre la alineación de lecturas WGS de *Dunaliella salina* cepa noruega contra el plastoma de referencia *Dunaliella salina* cepa SQ					
Cromosoma	#	Posición	Ref	Alt	QUAL
Plastoma	1	4	ACA	ACATATTAATCGCAGACA	9.478
Plastoma	2	36448	G	T	225.417
Plastoma	3	47892	T	A	3.7766

Sin embargo, las variantes 1 y 3 se encuentran en posiciones con una cobertura menor a 10x, además de contar con una calidad de variante muy baja. Por lo que, tras filtrar las variantes solo sobrevivió la numero 2, esta variante entra en la categoría de SNP, ya que cambia un nucleótido de la referencia, en este caso una guanina en la posición 36448, por un nuevo nucleótido, siendo este una timina (figura 37).

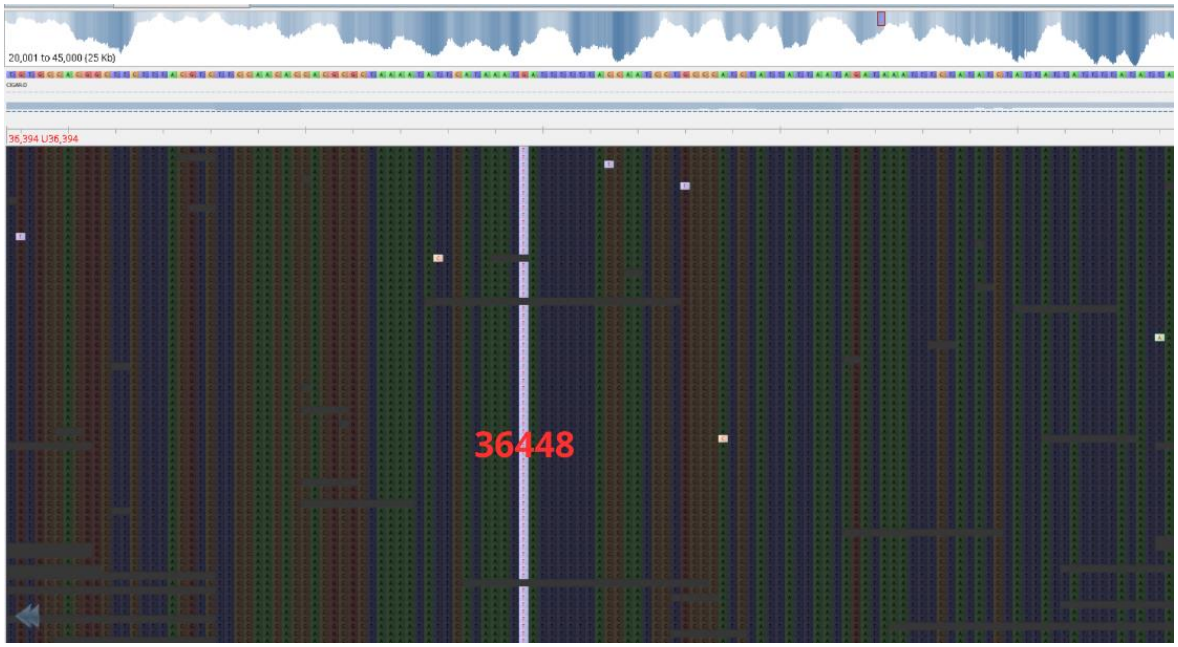

Figura 37: Captura de pantalla del programa de visualizador de alineaciones Tablet, en la posición correspondiente al único SNP detectado entre las lecturas del plastoma de *Dunaliella salina* cepa noruega contra el plastoma de la cepa SQ.

Esta cantidad de variantes puede considerarse muy baja en comparación con lo reportado en análisis similares [138], por lo que se recurrió a pasos anteriores de la metodología, específicamente dentro del ensambla *de novo*, al análisis donde se realizaron múltiples alineaciones para corroborar el impacto del sesgo presente en las primeras y últimas bases de las lecturas. Se utilizó la alineación en la que más lecturas WGS se mapearon contra el plastoma SQ, siendo este el obtenido con el programa bwa con lecturas con sesgo.

A este archivo con más lecturas alineadas se le realizó el llamado de variantes con bcftools y con freebayes, obteniendo exactamente el mismo SNP, por lo que podemos decir que solo existe esta variante entre el plastoma de *Dunaliella salina* cepa noruega y el plastoma de *Dunaliella salina* cepa SQ. Con el único SNP detectado se procedió a obtener una secuencia consenso utilizando la función "consensus" de bcftools. De esta secuencia consenso se eliminaron las regiones con baja cobertura detectadas anteriormente, obteniendo así el genoma del cloroplasto de *Dunaliella salina* cepa noruega.

9.6 Anotación del genoma

Para la anotación del genoma se utilizaron dos programas distintos, geseq el cual es una herramienta bioinformática especializada en la anotación funcional y estructural de cloroplastos, apoyándose de otras herramientas como blat, trna-scan y ARAGORN para ello. La razón de utilizar este anotador es porque permite al usuario introducir un genoma a partir del cual se va a determinar la anotación de los genes en una nueva secuencia, para este caso, al ser tan similares los genomas del cloroplasto de *Dunaliella salina* cepa noruega y SQ entre sí, es de esperar que contengan la misma cantidad de genes y la misma arquitectura, aun así, debido a que el anotador geseq se basa en alineaciones elaboradas con blat utilizando los genes ya anotados de la referencia como secuencia, puede llegar a pasar por alto características importantes de los genes (como codones de inicio y codones de terminación) siempre y cuando el resto de la secuencia genere el puntaje adecuado para determinar que la alineación es correcta. Previendo estos errores se decidió utilizar la herramienta de anotación MFannot, para complementar esta necesidad. Mfannot, al igual que GeSeq, es una herramienta bioinformática diseñada específicamente para la anotación automática de genomas organelares, como lo es el genoma del cloroplasto. La razón de complementar la anotación con esta herramienta es por la diferencia entre sus flujos de trabajo, Mfannot se

enfoca en la detección de open reading frames (ORFs) para identificar genes mediante la alineación de estos mismos contra bases de datos de generales de genes de cloroplasto. De esta manera, podemos decir que, si un gen tiene las mismas coordenadas en ambas anotaciones, esa anotación no necesita curarse.

Los genes que coincidieron en una primera instancia fueron 75. Además, hubo 21 genes donde falló la anotación realizada con mfannot, pero coincidía la anotación de GeSeq con lo anotado en el plastoma SQ, se detectaron otros 3 genes donde ninguna anotación coincidía entre sí. La curación manual de estos genes se realizó mediante búsquedas con blast en NCBI.

Los primeros 21 genes donde falló mfannot se pueden explicar con presencias de codones de stop en intrones, como lo fue en el caso del gen psaB, que tenía un codón de stop dentro de un intrón, lo que hacía que se acortará la anotación para este programa, ya que recordemos que trabaja basándose en la búsqueda de ORFs, sin embargo, no discrimina entre intrones y exones, razón por la cual en la mayoría de genes con intrones falló la anotación. Otro error común por el que falló mfannot es la detección de un codón de inicio exclusivo para plástidos (transl_table 11 de NCBI) antes o después del codón de inicio real del gen, modificando su longitud, estos errores se corroboraron con búsquedas en blast contra genes de cloroplasto de especies relacionadas evolutivamente.

Hubo otros 3 casos donde ninguna de las anotaciones coincidió entre sí ni con el plastoma de *Dunaliella salina* cepa SQ. El primero es para el gen atpA, la longitud variaba entre las 3 anotaciones debido a la eliminación de un pedazo de intrón al momento de generar la secuencia consenso (donde se eliminaron las regiones con una cobertura menor a 10x), cortando el gen, esta eliminación no causa problemas ya que recordemos que los intrones varían en tamaño y secuencia entre distintas especies, ya que al no expresarse no están sometidos a una presión selectiva, por lo que pueden acumular mutaciones neutrales, modificando su tamaño.

Para el caso de petG, al momento de eliminar las regiones con una cobertura menor a 10x, se eliminó parte del extremo final del gen, sin embargo, al realizar un blast con la secuencia resultante, se observó que ningún gen reportado compartía esta característica, por lo que, al revisar el archivo de alineación, se observó que el extremo del gen estaba representado por una profundidad de 8x (figura 38).

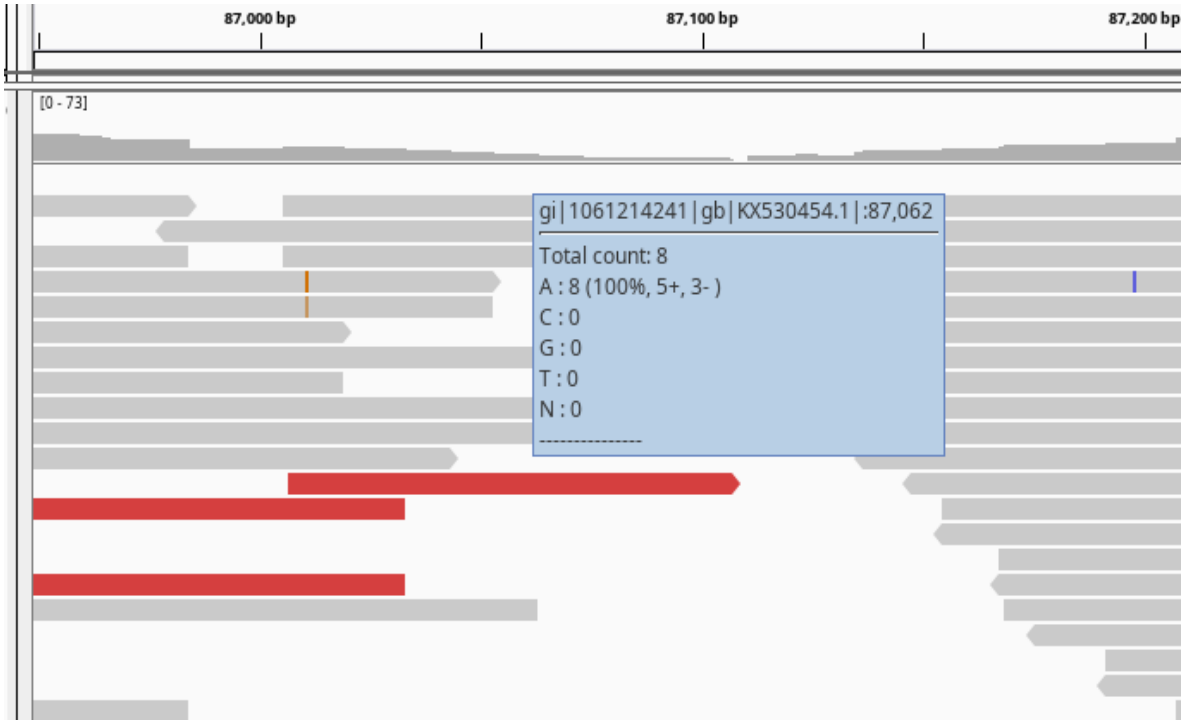

Figura 38: Captura de pantalla del alineador Tablet, visualizando la alineación de lecturas WGS de *Dunaliella salina* cepa noruega contra el plastoma de referencia SQ, en la región correspondiente al extremo del gen petG que se perdió por tener una profundidad <10x.

Debido a la conservación de ese extremo en los genes de especies relacionadas en NCBI, además de que en algunos trabajos se manejan profundidades mínimas de alineación de 4x [139], se decidió agregar la sección faltante de este gen.

Para el caso del SNP detectado en el llamado de variantes, este se encuentra en el la posición 36448 y corresponde a un cambio de guanina a timina, este cambio se encuentra dentro del gen rps9, pero este gen se encuentra en la hebra complementaria, recordemos que el SNP fue detectado en sentido 5' a 3', por lo que es necesario pasarlo a su hebra complementaria, quedando el cambio de citosina a guanina. El cambio se da dentro del codón TCA que traduce a serina, generando un nuevo codón TAA, el cual es un codón de parada, acortando su tamaño (Figura 39). Tras buscar la secuencia resultante en NCBI, se encontró que *Dunaliella*

tertiolecta también presenta este acortamiento en el gen rps9 en comparación con *Dunaliella salina* cepa SQ.

Figura 39: Imagen de la secuencia del gen rps9 antes del SNP (imagen A) con su respetiva búsqueda en NCBI e imagen de la secuencia del gen rps9 después de aplicar el SNP y su respectiva búsqueda en NCBI.

Una vez terminada la curación manual, se obtuvo el gráfico del plastoma de *Dunaliella salina* cepa noruega con OGDRAW [140] (figura 40):

Como resultado de la anotación se detectaron:

- 98 genes
    - 66 codificantes
    - 3 rRNA (por cada IR)
    - 29 tRNA

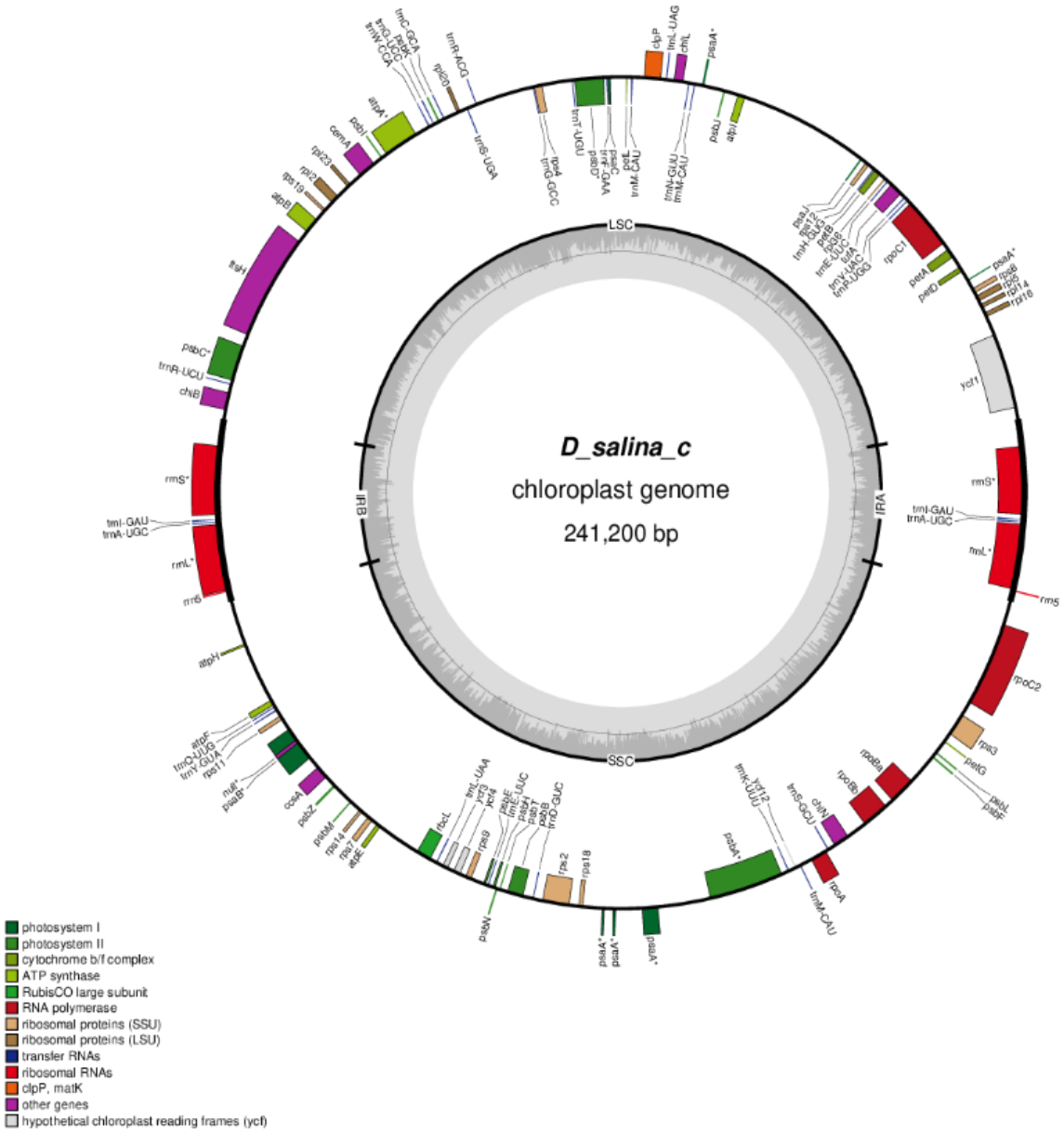

Figura 40: Arquitectura del plastoma de Dunaliella salina cepa noruega.

9.7 Análisis de sintenia

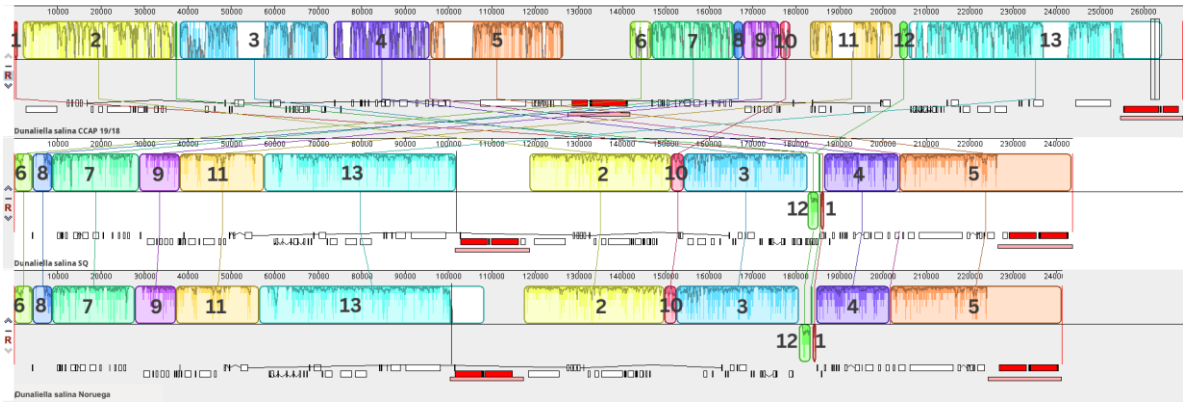

Figura 41: Análisis de sintenia entre genomas de cloroplasto de cepas de *Dunaliella salina*, generado con mauve, editada.

Se realizó un análisis de sintenia entre las los genomas de cloroplasto de *Dunaliella salina* cepa CCAP19/18 (Australia), cepa SQ (San Quintín, Baja California) y Noruega con el programa mauve (figura 41).

El análisis mostró un reordenamiento drástico de los bloques localmente colineales entre la cepa de australiana contra las cepas de San Quintín y Noruega por ejemplo, en el genoma de la cepa australiana se presenta un patrón de colores ordenados de izquierda a derecha con la siguiente secuencia: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13. Por otra parte, las cepas San Quintín y Noruega presentan una arquitectura idéntica entre sí: 6, 8, 7, 9, 11, 13, 2, 10, 3, 12, 1, 4, 5, pero muestran diferencias contra el genoma australiano, como el movimiento de los bloques 1, 8, 10 y 12. También se observa que los bloques de la cepa noruega son de menor tamaño en comparación con la cepa SQ, esto debido a las regiones que no presentaron la profundidad adecuada en el ensamble por referencia.

El reordenamiento de la cepa CCAP19/18 podría indicar que el genoma de su cloroplasto pudo haber sufrido algún tipo de lesión dañina, como rupturas de doble cadena (DSBs, Double-stranded breaks) o ruptura de un solo extremo DSBs (ocurre cuando la horquilla de replicación encuentra una ruptura en una sola hebra), ocasionado un proceso de reparación denominado recombinación homologa (proceso de reparación conservador basado en el intercambio de información genética entre dos secuencias de ADN cuasi-idénticas), que

puede traer como consecuencia reordenamientos en el genoma [141]. Sin embargo, es una hipótesis apresurada que necesita mayor información para ser respaldada.

Las gráficas dentro de los rectángulos de colores representan la identidad entre los bloques localmente colineales, cuando bajan significan baja identidad, entre más alta, más alta la identidad, en este sentido, los picos bajos representan regiones intergénicas (no conservadas), siendo la sepa australiana la que tiene mayores regiones intergénicas.

En esta comparación, se puede observar en el extremo derecho del bloque 13 una similitud que lleva a abarcar una de las regiones IR, compartida entre la cepa noruega y la cepa australiana, mientras que en la cepa SQ no se presenta, en un principio se pensó que esta diferencia se debía a la eliminación de las zonas que no presentaron la profundidad suficiente en el ensamble por referencia, sin embargo, al revisar el ensamble por referencia se encontró que esta zona presentaba una profundidad alta de manera continua, por lo que se generaron análisis de sintenia utilizando dos cepas a la vez.

El primer análisis se hizo entre la cepa australiana y la cepa noruega (figura 41), en esta imagen se buscó descartar que estás regiones compartidas por la cepa noruega y australiana no fueran causadas por el programa al querer procesar la sintenia de 3 genomas al mismo tiempo, en este sentido, al colocar solo estos dos genomas, se perdió la región que alcanzaba las zonas IR (recuadro rojo) por lo que se puede estimar que esta similitud ubicada en el bloque 13 fue generada por el programa al alinear 3 genomas.

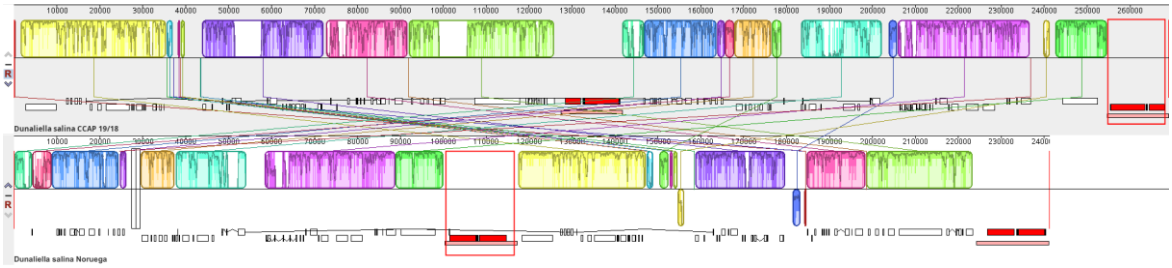

Figura 42: Análisis de sintenia del genoma del cloroplasto entre cepas de *Dunaliella salina* CCAP19/18 y Noruega con mauve, editada.

Por último, se hizo el análisis de sintenia entre la cepa noruega y la cepa SQ (figura 42), en este caso se nota la existencia de un solo bloque localmente colineal, con ciertas regiones con

el gráfico caído en la cepa SQ, esto se debe las regiones que se eliminaron en el ensamble por referencia de la cepa noruega durante el ensamble por referencia.

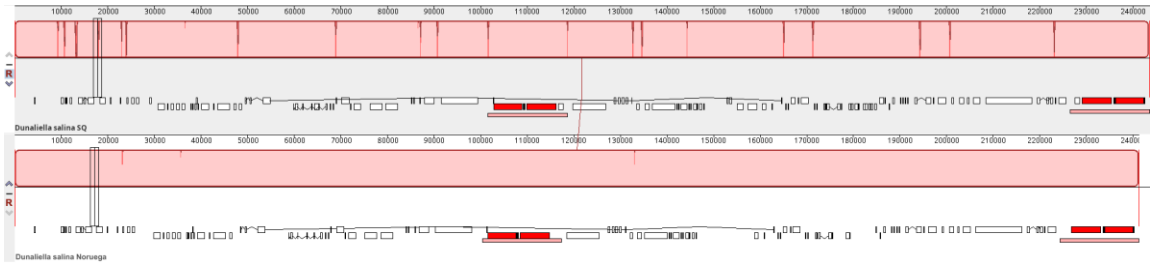

Figura 43: Análisis de sintenia del genoma del cloroplasto entre cepas de *Dunaliella salina* SQ y Noruega con mauve, editada.

9.8 Búsqueda de promotores

Para la predicción de promotores, se utilizó un script que busca motivos conservados de los promotores de PEP. Las regiones conceso utilizadas son las reportadas en la tabla 6 del marco teórico.

Los resultados de la búsqueda se dividieron en dos, tomando como base el artículo escrito por Klein en 1992 [55], la tabla 19 muestra los promotores que siguen la estructura bacteriana con regiones -35 y -10. Mientras que la tabla 20 muestra los promotores encontrados que solo tienen región -10.

Se encontraron 35 regiones aguas arriba de los genes que contienen la estructura bacteriana de promotores -10 y -35, sin embargo, en el caso de los genes adyacentes trnQ y trnY, en sus regiones aguas arriba se encontraron los mismos elementos -35 (TTTACA) y -10 (TATTAT) a 24 pb de distancia entre sí, lo que sugiere que estos genes están regulados por el mismo promotor. Aun así, quedan 34 promotores putativos con esta estructura. Por otra parte, se encontraron 17 promotores putativos que presentan solo la región -10, dando un total de 51 promotores putativos a analizar.

Debido a la falta de información de elementos cis-reguladores inducibles propios del cloroplasto, se caracterizaron los promotores putativos de cloroplasto utilizando elementos del núcleo tanto de *Embryophytes* como de la microalga modelo *Chlamydomonas reinhardtii.*

Los elementos cis-reguladores inducibles a buscar fueron GATABOX, EECCRCAH1, GT1GMSCAM4, CBFHV y LTRECOREATCOR15, los cuales son inducibles a luz, niveles bajos de CO_2, sal, deshidratación y baja temperatura respectivamente.

Se reconocieron múltiples elementos cis-funcionales inducibles en los promotores putativos del cloroplasto de *Dunaliella salina* cepa noruega, la mayoría con múltiples elementos reguladores, lo que sugiere una regulación modulable, sin embargo, se necesita experimentación para confirmar el hecho de que estos elementos son funcionales y no generados por casualidad o transferidos del núcleo al cloroplasto, pero sin función real.

Tabla 19. Promotores PEP putativos dependientes de -35 y -10, con elementos cis-reguladores inducibles			
ID Promotor	Motivo	Secuencia_Motivo	Inducibilidad
atpF	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
trnQ	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
rps11	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
rps14	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
rps9	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
	LTRECOREATCOR15	CCGAC	Baja temperatura
psbE	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
trnE	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
psbH	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
psbT	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
	LTRECOREATCOR15	CCGAC	Baja temperatura
psbB	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
rps2	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
psbA	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
chlN	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
psbF	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
petG	GATABOX	GATA	Luz
rrnS	GATABOX	GATA	Luz

	GT1GMSCAM4	GAAAAA	Sal
rpl16	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
	LTRECOREATCOR15	CCGAC	Baja temperatura
tufA	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
trnE	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
psbJ	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
trnM	GATABOX	GATA	Luz
trnN	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
trnL	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
clpP	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
	LTRECOREATCOR15	CCGAC	Baja temperatura
trnM	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
rps4	GATABOX	GATA	Luz
trnG	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
rpl20	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
psbI	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
atpB	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
ftsH	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
psbC	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
	LTRECOREATCOR15	CCGAC	Baja temperatura
rrnS	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
rrn5	GATABOX	GATA	Luz

Tabla 20. Promotores PEP putativos dependientes solo de -10, con elementos cis-reguladores inducibles			
ID Promotor	Motivo	Secuencia	Inducibilidad
ccsA	GT1GMSCAM4	GAAAAA	Sal
psbN	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
trnD	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
ycf12	GT1GMSCAM4	GAAAAA	Sal
trnM	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
rpoBb	GATABOX	GATA	Luz
ycf1	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
trnP	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
trnV	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
psaJ	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
chlL	GATABOX	GATA	Luz
psaC	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
trnF	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
psbD	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
trnT	GATABOX	GATA	Luz
	EECCRCAH1	GA[ATGC]TT[ATGC]C	Bajo CO_2
	GT1GMSCAM4	GAAAAA	Sal
trnC	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal
psbK	GATABOX	GATA	Luz
	GT1GMSCAM4	GAAAAA	Sal

10 Conclusión

Se ensambló el genoma del cloroplasto de la microalga *Dunaliella salina* cepa noruega, con una longitud de 241,200 pb, siendo el más pequeño en comparación con genomas de cloroplasto de *Dunaliella salina* cepa SQ y cepa CCAP19/18. Se logró la anotación del genoma, con un total de 98 genes distintos y mostrando un 27.8% de regiones intergénicas, lo cual representa un espacio potencial para la inserción de vectores de expresión.

Asimismo, se identificaron 51 promotores endógenos putativos que contienen motivos inducibles por sal, luz, bajas temperaturas y concentración reducida de CO_2. Estos resultados respaldan la hipótesis inicial sobre identificación de secuencias promotoras reguladoras de la transcripción que podrían utilizarse en metodologías de expresión génica regulada.

Por otro lado, se analizó la sintenia entre los genomas del cloroplasto de las distintas cepas de *Dunaliella salina.* Observándose que la organización genómica del cloroplasto de la cepa noruega guarda una alta similitud con la cepa SQ. Este hallazgo sugiere que el conjunto de promotores identificados podría ser funcional entre ambas, lo que amplía la aplicabilidad biotecnológica de los elementos descubiertos.

Finalmente, el script desarrollado para la detección automática de motivos conservados (-35 y -10) demostró ser eficaz, por lo que se abre la posibilidad de transferirlo a otros genomas cloroplásticos de microalgas del orden *Chlamydomonadales*, sentando las bases para su implementación en estudios comparativos.

11 Referencias

[1] J. Iwasa, *Karp Biología Celular y Molecular*, 8th ed., vol. 1. McGrawHill, 2018.

[2] M. I. Khan, J. H. Shin, and J. D. Kim, “The promising future of microalgae: Current status, challenges, and optimization of a sustainable and renewable industry for biofuels, feed, and other products,” Mar. 05, 2018, *BioMed Central Ltd.* doi: 10.1186/s12934-018-0879-x.

[3] M. D. H. da Rosa *et al.*, “Macroalgae and Microalgae Biomass as Feedstock for Products Applied to Bioenergy and Food Industry: A Brief Review,” Feb. 01, 2023, *MDPI*. doi: 10.3390/en16041820.

[4] J. Cai *et al.*, “Seaweeds and microalgae: an overview for unlocking their potential in global aquaculture development,” *FAO Fisheries and Aquaculture Circular*, no. 1229, 2021, doi: 10.4060/cb5670en.

[5] P. Carillo, L. F. Ciarmiello, P. Woodrow, G. Corrado, P. Chiaiese, and Y. Rouphael, “Enhancing sustainability by improving plant salt tolerance through macro-and micro-algal biostimulants,” Sep. 01, 2020, *MDPI AG*. doi: 10.3390/biology9090253.

[6] T. Cai, S. Y. Park, and Y. Li, “Nutrient recovery from wastewater streams by microalgae: Status and prospects,” *Renewable and Sustainable Energy Reviews*, vol. 19, pp. 360–369, 2013, doi: https://doi.org/10.1016/j.rser.2012.11.030.

[7] G. M. Vingiani, P. De Luca, A. Ianora, A. D. W. Dobson, and C. Lauritano, “Microalgal enzymes with biotechnological applications,” Aug. 05, 2019, *MDPI AG*. doi: 10.3390/md17080459.

[8] R. J. Lawton, A. J. Cole, D. A. Roberts, N. A. Paul, and R. de Nys, “The industrial ecology of freshwater macroalgae for biomass applications,” *Algal Res*, vol. 24, pp. 486–491, 2017, doi: https://doi.org/10.1016/j.algal.2016.08.019.

[9] J. Ampofo and Lord Abbey, “Microalgae: Bioactive Composition, Health Benefits, Safety and Prospects as Potential High-Value Ingredients for the Functional Food Industry,” Jun. 01, 2022, *MDPI*. doi: 10.3390/foods11121744.

[10] P. Barciela *et al.*, “Macroalgae as biofactories of metal nanoparticles; biosynthesis and food applications,” *Adv Colloid Interface Sci*, vol. 311, p. 102829, 2023, doi: https://doi.org/10.1016/j.cis.2022.102829.

[11] Grand View Research, “Global Recombinant Proteins Market Size & Outlook,” 2024. Accessed: Mar. 21, 2025. [Online]. Available: www.grandviewresearch.com/industry-analysis/recombinant-proteins-market-report

[12] N. Yan, C. Fan, Y. Chen, and Z. Hu, “The potential for microalgae as bioreactors to produce pharmaceuticals,” *Int J Mol Sci*, vol. 17, no. 6, Jun. 2016, doi: 10.3390/ijms17060962.

[13] A. Schütz *et al.*, “A concise guide to choosing suitable gene expression systems for recombinant protein production,” *STAR Protoc*, vol. 4, no. 4, p. 102572, 2023, doi: https://doi.org/10.1016/j.xpro.2023.102572.

[14] G. Walsh and R. Jefferis, “Post-translational modifications in the context of therapeutic proteins,” *Nat Biotechnol*, vol. 24, no. 10, pp. 1241–1252, 2006, doi: 10.1038/nbt1252.

[15] M. Smith *et al.*, “Producing recombinant proteins in Vibrio natriegens,” *Microb Cell Fact*, vol. 23, no. 1, Dec. 2024, doi: 10.1186/s12934-024-02455-5.

[16] Z.-M. Li, Z.-L. Fan, X.-Y. Wang, and T.-Y. Wang, “Factors Affecting the Expression of Recombinant Protein and Improvement Strategies in Chinese Hamster Ovary Cells,” *Front Bioeng Biotechnol*, vol. 10, 2022, [Online]. Available: https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2022.880155

[17] A. Banerjee and V. Ward, “Production of recombinant and therapeutic proteins in microalgae,” *Curr Opin Biotechnol*, vol. 78, p. 102784, 2022, doi: https://doi.org/10.1016/j.copbio.2022.102784.

[18] Y. Torres-Tiji, F. J. Fields, and S. P. Mayfield, “Microalgae as a future food source,” *Biotechnol Adv*, vol. 41, p. 107536, 2020, doi: https://doi.org/10.1016/j.biotechadv.2020.107536.

[19] L. Muñoz-Solórzano, K. Willis-Ureña, S. Valverde-Rojas, M. Jarquín-Cordero, and L. Barboza-Fallas, “Microalgae as expression systems for recombinant protein production,” *Revista Tecnología en Marcha*, Nov. 2024, doi: 10.18845/tm.v37i9.7608.

[20] E. Forján *et al.*, “Microalgae: Fast-Growth Sustainable Green Factories,” *Crit Rev Environ Sci Technol*, vol. 45, no. 16, pp. 1705–1755, Aug. 2015, doi: 10.1080/10643389.2014.966426.

[21] Z. Xie, J. He, S. Peng, X. Zhang, and W. Kong, “Biosynthesis of protein-based drugs using eukaryotic microalgae,” *Algal Res*, vol. 74, p. 103219, 2023, doi: https://doi.org/10.1016/j.algal.2023.103219.

[22] A. Castillo *et al.*, “The generally recognized as safe (GRAS) microalgae Haematococcus pluvialis (wet) as a multifunctional additive for coloring and improving the organoleptic and functional properties of foods,” *Food Funct*, vol. 14, no. 13, pp. 6023–6035, 2023, doi: 10.1039/D3FO01028G.

[23] C. Carreño-Campos, M. L. Villarreal, and A. Ortiz Caltempa, “El potencial del genoma del cloroplasto de Chlamydomonas reinhardtii para la producción de proteínas recombinantes,” 2021.

[24] K. Ma, L. Deng, H. Wu, and J. Fan, “Towards green biomanufacturing of high-value recombinant proteins using promising cell factory: Chlamydomonas reinhardtii chloroplast,” *Bioresour Bioprocess*, vol. 9, no. 1, p. 83, 2022, doi: 10.1186/s40643-022-00568-6.

[25] R. Kalra, S. Gaur, and M. Goel, “Chapter 13 - Harnessing the potential of microalgal species Dunaliella: A biofuel and biocommodities perspective,” in *Algal Biotechnology*, A. Ahmad, F. Banat, and H. Taher, Eds., Elsevier, 2022, pp. 259–279. doi: https://doi.org/10.1016/B978-0-323-90476-6.00008-X.

[26] M. Guevara *et al.*, “Influencia de la salinidad y la irradiancia sobre el crecimiento y composición bioquímica de una nueva cepa de Dunaliella salina, proveniente de las salinas de Araya, Venezuela,” *Saber*, vol. 28, pp. 494–501, 2016.

[27] G. de S. Celente, T. M. Rizzetti, Y. Sui, and R. de C. de S. Schneider, “Potential use of microalga Dunaliella salina for bioproducts with industrial relevance,” *Biomass Bioenergy*, vol. 167, p. 106647, 2022, doi: https://doi.org/10.1016/j.biombioe.2022.106647.

[28] I. Hyrslova *et al.*, “Functional Properties of Dunaliella salina and Its Positive Effect on Probiotics,” *Mar Drugs*, vol. 20, no. 12, Dec. 2022, doi: 10.3390/md20120781.

[29] Y. M. Kwon, K. W. Kim, T.-Y. Choi, S. Y. Kim, and J. Y. H. Kim, “Manipulation of the microalgal chloroplast by genetic engineering for biotechnological utilization as a green biofactory,” *World J Microbiol Biotechnol*, vol. 34, no. 12, p. 183, 2018, doi: 10.1007/s11274-018-2567-8.

[30] B. A. Rasala and S. P. Mayfield, “Photosynthetic biomanufacturing in green algae; production of recombinant proteins for industrial, nutritional, and medical uses,” *Photosynth Res*, vol. 123, no. 3, pp. 227–239, 2015, doi: 10.1007/s11120-014-9994-7.

[31] S. Freeman, K. Quillin, L. Allison, M. Black, G. Podgorski, and E. Taylor, *Fundamentos de biología*, 6th ed. Pearson, 2018.

[32] N. Sato, “Complex origins of chloroplast membranes with photosynthetic machineries: multiple transfers of genes from divergent organisms at different times or a single endosymbiotic event?,” *J Plant Res*, vol. 133, no. 1, pp. 15–33, 2020, doi: 10.1007/s10265-019-01157-z.

[33] P. J. Keeling, “The endosymbiotic origin, diversification and fate of plastids,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1541, pp. 729–748, Mar. 2010, doi: 10.1098/rstb.2009.0103.

[34] L. Espinosa-Barrera and E. Chávez-Sahagún, “El otro genoma de las plantas: los cloroplastos y su ADN,” *Herbario CICY*, vol. 11, pp. 201–206, 2019, [Online]. Available: http://www.cicy.mx/sitios/desde_herbario/

[35] F. Leliaert *et al.*, “Phylogeny and Molecular Evolution of the Green Algae,” *CRC Crit Rev Plant Sci*, vol. 31, no. 1, pp. 1–46, Jan. 2012, doi: 10.1080/07352689.2011.615705.

[36] K. Fučíková, M. Taylor, L. A. Lewis, B. K. Niece, A. S. Isaac, and N. Pietrasiak, “Johansenicoccus eremophilus gen. et sp. nov., a novel evolutionary lineage in Chlorophyceae with unusual genomic features,” *Plant Ecol Evol*, vol. 156, no. 3, pp. 311–325, Sep. 6AD, [Online]. Available: https://doi.org/10.5091/plecevo.105762

[37] A. Oren, “A hundred years of Dunaliella research: 1905-2005,” *Saline Syst*, vol. 1, p. 2, Aug. 2005, doi: 10.1186/1746-1448-1-2.

[38] J. Masojídek and G. Torzillo, “Mass Cultivation of Freshwater Microalgae,” in *Encyclopedia of Ecology*, S. E. Jørgensen and B. D. Fath, Eds., Oxford: Academic Press, 2008, pp. 2226–2235. doi: https://doi.org/10.1016/B978-008045405-4.00830-2.

[39] M. A. Borowitzka and C. J. Siva, “The taxonomy of the genus Dunaliella (Chlorophyta, Dunaliellales) with emphasis on the marine and halophilic species,” *J Appl Phycol*, vol. 19, no. 5, pp. 567–590, 2007, doi: 10.1007/s10811-007-9171-x.

[40] M. R. Hadi, M. Shariati, and S. Afsharzadeh, “Microalgal biotechnology: Carotenoid and glycerol production by the green algae Dunaliella isolated from the Gave-Khooni salt marsh, Iran,” *Biotechnology and Bioprocess Engineering*, vol. 13, no. 5, pp. 540–544, 2008, doi: 10.1007/s12257-007-0185-7.

[41] R. León and F. Galván, “Halotolerance studies onChlamydomonas reinhardtii: glycerol excretion by free and immobilized cells,” *J Appl Phycol*, vol. 6, no. 1, pp. 13–20, 1994, doi: 10.1007/BF02185898.

[42] P. Ramachandran, N. K. Pandey, R. M. Yadav, P. Suresh, A. Kumar, and R. Subramanyam, “Photosynthetic efficiency and transcriptome analysis of Dunaliella salina under hypersaline: a retrograde signaling mechanism in the chloroplast,” *Front Plant Sci*, vol. 14, 2023, doi: 10.3389/fpls.2023.1192258.

[43] A. Ben-Amotz, I. Sussman, and M. Avron, “Glycerol production by Dunaliella,” in *New Trends in Research and Utilization of Solar Energy through Biological Systems*, H. Mislin and R. Bachofen, Eds., Basel: Birkhäuser Basel, 1982, pp. 55–58. doi: 10.1007/978-3-0348-6305-6_12.

[44] A. Ben-Amotz, “Production of β-Carotene and Vitamins by the Halotolerant Alga Dunaliella,” in *Pharmaceutical and Bioactive Natural Products*, D. H. Attaway and O. R. Zaborsky, Eds., Boston, MA: Springer US, 1993, pp. 411–417. doi: 10.1007/978-1-4899-2391-2_11.

[45] M. Cobb, “60 years ago, Francis Crick changed the logic of biology,” *PLoS Biol*, vol. 15, no. 9, pp. e2003243-, Sep. 2017, [Online]. Available: https://doi.org/10.1371/journal.pbio.2003243

[46] K. S. Macedo-Osorio, A. Martínez-Antonio, and J. A. Badillo-Corona, “Pas de Trois: An Overview of Penta-, Tetra-, and Octo-Tricopeptide Repeat Proteins From Chlamydomonas reinhardtii and Their Role in Chloroplast Gene Expression,” *Front Plant Sci*, vol. Volume 12-2021, 2021, [Online]. Available: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2021.775366

[47] S. B. Weiss and L. Gladstone, “A MAMMALIAN SYSTEM FOR THE INCORPORATION OF CYTIDINE TRIPHOSPHATE INTO RIBONUCLEIC ACID1,” *J Am Chem Soc*, vol. 81, no. 15, pp. 4118–4119, Aug. 1959, doi: 10.1021/ja01524a087.

[48] R. Landick, “A Long Time in the Making—The Nobel Prize for RNA Polymerase,” *Cell*, vol. 127, no. 6, pp. 1087–1090, Dec. 2006, doi: 10.1016/j.cell.2006.11.036.

[49] J. Dobrogojski, M. Adamiec, and R. Luciński, “The chloroplast genome: a review,” Jun. 01, 2020, *Springer*. doi: 10.1007/s11738-020-03089-x.

[50] M. C. Little and R. B. Hallick, “Chloroplast rpoA, rpoB, and rpoC genes specify at least three components of a chloroplast DNA-dependent RNA polymerase active in tRNA and mRNA transcription.,” *J Biol Chem*, vol. 263, no. 28, pp. 14302–14307, 1988, doi: 10.1016/s0021-9258(18)68221-3.

[51] A.-V. Bohne, A. E. Vered, I. Andreas, W. Ae, and D. B. Stern, “Chlamydomonas reinhardtii encodes a single sigma 70-like factor which likely functions in chloroplast transcription,” *Springer*, vol. 49, no. CurrGenet, pp. 333–340, 2006, doi: 10.1007/s00294-006.

[52] T. Börner, A. Y. Aleynikova, Y. O. Zubo, and V. V. Kusnetsov, “Chloroplast RNA polymerases: Role in chloroplast biogenesis,” Dec. 18, 2015, *Elsevier B.V.* doi: 10.1016/j.bbabio.2015.02.004.

[53] Á. Vergara-Cruces *et al.*, “Structure of the plant plastid-encoded RNA polymerase,” *Cell*, vol. 187, no. 5, pp. 1145-1159.e21, 2024, doi: https://doi.org/10.1016/j.cell.2024.01.036.

[54] C. L. Gutiérrez, C. Muñoz, M. San Martín, J.-P. Cadoret, and V. Henríquez, “Chloroplast Dual Divergent Promoter Plasmid for Heterologous Protein Expression in Tetraselmis suecica (Chlorophyceae, Chlorodendrales),” *J Phycol*, vol. 56, no. 4, pp. 1066–1076, Aug. 2020, doi: https://doi.org/10.1111/jpy.13013.

[55] U. Klein, J. D. De Camp, and L. Bogorad, “Two types of chloroplast gene promoters in Chlamydomonas reinhardtii.,” *Proceedings of the National Academy of Sciences*, vol. 89, no. 8, pp. 3453–3457, Apr. 1992, doi: 10.1073/pnas.89.8.3453.

[56] J. A. Gimpel and S. P. Mayfield, “Analysis of heterologous regulatory and coding regions in algal chloroplasts,” *Appl Microbiol Biotechnol*, vol. 97, no. 10, pp. 4499–4510, 2013, doi: 10.1007/s00253-012-4580-4.

[57] M. Dron, M. Rahire, and J.-D. Rochaix, “Sequence of the chloroplast DNA region of Chlamydomonas reinhardii containing the gene of the large subunit of ribulose bisphosphate carboxylase and parts of its flanking genes,” *J Mol Biol*, vol. 162, no. 4, pp. 775–793, 1982, doi: https://doi.org/10.1016/0022-2836(82)90547-2.

[58] E. A. Cutolo, G. Mandalà, L. Dall’osto, and R. Bassi, “Harnessing the Algal Chloroplast for Heterologous Protein Production,” Mar. 30, 2022, *MDPI*. doi: https://doi.org/10.3390/microorganisms10040743.

[59] E. M. del Campo, “Post-transcriptional control of chloroplast gene expression,” *Gene Regul Syst Bio*, vol. 2009, no. 3, pp. 31–47, 2009, doi: 10.4137/GRSB.S2080.

[60] D. B. Stern, M. Goldschmidt-Clermont, and M. R. Hanson, “Chloroplast RNA metabolism,” *Annu Rev Plant Biol*, vol. 61, pp. 125–155, Jun. 2010, doi: 10.1146/annurev-arplant-042809-112242.

[61] L. Z. Miandoab, “Transcription Flexibility of Dunaliella Chloroplast Genome,” 2022. doi: 10.5772/intechopen.105125.

[62] R. Freyer, M.-C. Kiefer-Meyer, and H. Kössel, “Occurrence of plastid RNA editing in all major lineages of land plants,” *Proceedings of the National Academy of Sciences*, vol. 94, no. 12, pp. 6285–6290, Jun. 1997, doi: 10.1073/pnas.94.12.6285.

[63] A. B. Cahoon, J. A. Nauss, C. D. Stanley, and A. Qureshi, “Deep transcriptome sequencing of two green algae, Chara vulgaris and Chlamydomonas reinhardtii, provides no evidence of organellar RNA editing,” *Genes (Basel)*, vol. 8, no. 2, Feb. 2017, doi: 10.3390/genes8020080.

[64] L. T. Vu and T. Tsukahara, “C-to-U editing and site-directed RNA editing for the correction of genetic mutations,” *Biosci Trends*, vol. 11, no. 3, pp. 243–253, 2017, doi: 10.5582/bst.2017.01049.

[65] I. L. ANTHONISEN, M. L. SALVADOR, and U. W. E. KLEIN, “Specific sequence elements in the 5′ untranslated regions of rbcL and atpB gene mRNAs stabilize transcripts in the chloroplast of Chlamydomonas reinhardtii,” *RNA*, vol. 7, no. 7, pp. 1024–1033, 2001, doi: DOI: 10.1017/S1355838201001479.

[66] R. Hayes, J. Kudla, and W. Gruissem, “Degrading chloroplast mRNA: the role of polyadenylation,” *Trends Biochem Sci*, vol. 24, no. 5, pp. 199–202, 1999, doi: https://doi.org/10.1016/S0968-0004(99)01388-2.

[67] R. G. Drager, D. C. Higgs, K. L. Kindle, and D. B. Stern, “5′ to 3′ exoribonucleolytic activity is a normal component of chloroplast mRNA decay pathways,” *The Plant Journal*, vol. 19, no. 5, pp. 521–531, Sep. 1999, doi: https://doi.org/10.1046/j.1365-313X.1999.00546.x.

[68] Z. Zou, C. Eibl, and H.-U. Koop, “The stem-loop region of the tobacco psbA 5′UTR is an important determinant of mRNA stability and translation efficiency,” *Molecular Genetics and Genomics*, vol. 269, no. 3, pp. 340–349, 2003, doi: 10.1007/s00438-003-0842-2.

[69] J. Kim and S. P. Mayfield, “Protein Disulfide Isomerase as a Regulator of Chloroplast Translational Activation,” *Science (1979)*, vol. 278, no. 5345, pp. 1954–1957, Dec. 1997, doi: 10.1126/science.278.5345.1954.

[70] O. Vallon, “Chlamydomonas immunophilins and parvulins: survey and critical assessment of gene models,” *Eukaryot Cell*, vol. 4, no. 2, pp. 230–241, 2005.

[71] M. Schroda, “The Chlamydomonas genome reveals its secrets: chaperone genes and the potential roles of their gene products in the chloroplast,” *Photosynth Res*, vol. 82, no. 3, pp. 221–240, 2004, doi: 10.1007/s11120-004-2216-y.

[72] H. Satam *et al.*, “Next-Generation Sequencing Technology: Current Trends and Advancements,” *Biology (Basel)*, vol. 12, p. 997, Jul. 2023, doi: 10.3390/biology12070997.

[73] D. Qin, “Next-generation sequencing and its clinical application,” *Cancer Biol Med*, vol. 16, no. 1, pp. 4–10, 2019, doi: 10.20892/j.issn.2095-3941.2018.0055.

[74] S. Behjati and P. S. Tarpey, “What is next generation sequencing?,” *Arch Dis Child Educ Pract Ed*, vol. 98, no. 6, pp. 236–238, Dec. 2013, doi: 10.1136/archdischild-2013-304340.

[75] Y. Lu, Y. Shen, W. Warren, and R. Walter, “Next Generation Sequencing in Aquatic Models,” in *Next Generation Sequencing*, J. K. Kulski, Ed., Rijeka: IntechOpen, 2016, p. Ch. 2. doi: 10.5772/61657.

[76] Illumina, “An introduction to Next-Generation Sequencing Technology,” 2024. [Online]. Available: www.illumina.com/technology/next-generation-sequencing.html

[77] J. Bermúdez, “s-aligner: a greedy algorithm for non-greedy de novo genome assembly,” 2021, doi: 10.1101/2021.02.02.429443.

[78] H. E. L. Lischer and K. K. Shimizu, “Reference-guided de novo assembly approach improves genome reconstruction for related species,” *BMC Bioinformatics*, vol. 18, no. 1, Nov. 2017, doi: 10.1186/s12859-017-1911-6.

[79] F. Dida and G. Yi, “Empirical evaluation of methods for de novo genome assembly,” *PeerJ Comput Sci*, vol. 7, pp. 1–31, 2021, doi: 10.7717/PEERJ-CS.636.

[80] J. Il Sohn and J. W. Nam, “The present and future of de novo whole-genome assembly,” *Brief Bioinform*, vol. 19, no. 1, pp. 23–40, Jan. 2018, doi: 10.1093/bib/bbw096.

[81] M. K. Basantani, D. Gupta, R. Mehrotra, S. Mehrotra, S. Vaish, and A. Singh, “An update on bioinformatics resources for plant genomics research,” *Curr Plant Biol*, vol. 11–12, pp. 33–40, 2017, doi: https://doi.org/10.1016/j.cpb.2017.12.002.

[82] K. Ohyama *et al.*, “Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA,” *Nature*, vol. 322, no. 6079, pp. 572–574, 1986, doi: 10.1038/322572a0.

[83] K. Shinozaki *et al.*, “The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression,” *EMBO J*, vol. 5, no. 9, pp. 2043-2049–2049, Sep. 1986, doi: https://doi.org/10.1002/j.1460-2075.1986.tb04464.x.

[84] S. Chiyoda, K. T. Yamato, and T. Kohchi, “Plastid Transformation of Sporelings and Suspension-Cultured Cells from the Liverwort Marchantia polymorpha L.,” in *Chloroplast Biotechnology: Methods and Protocols*, P. Maliga, Ed., Totowa, NJ: Humana Press, 2014, pp. 439–447. doi: 10.1007/978-1-62703-995-6_30.

[85] C. R. Boehm, M. Ueda, Y. Nishimura, T. Shikanai, and J. Haseloff, “A Cyan Fluorescent Reporter Expressed from the Chloroplast Genome of Marchantia polymorpha,” *Plant Cell Physiol*, vol. 57, no. 2, pp. 291–299, Feb. 2016, doi: 10.1093/pcp/pcv160.

[86] P. Maliga and T. Tungsuchat-Huang, “Plastid Transformation in Nicotiana tabacum and Nicotiana sylvestris by Biolistic DNA Delivery to Leaves,” in *Chloroplast Biotechnology: Methods and Protocols*, P. Maliga, Ed., Totowa, NJ: Humana Press, 2014, pp. 147–163. doi: 10.1007/978-1-62703-995-6_8.

[87] T. Golds, P. Maliga, and H.-U. Koop, “Stable Plastid Transformation in PEG-treated Protoplasts of Nicotiana tabacum,” *Bio/Technology*, vol. 11, no. 1, pp. 95–97, 1993, doi: 10.1038/nbt0193-95.

[88] T. Ruhlman, R. Ahangari, A. Devine, M. Samsam, and H. Daniell, “Expression of cholera toxin B–proinsulin fusion protein in lettuce and tobacco chloroplasts – oral administration protects against development of insulitis in non-obese diabetic mice,” *Plant Biotechnol J*, vol. 5, no. 4, pp. 495–510, Jul. 2007, doi: https://doi.org/10.1111/j.1467-7652.2007.00259.x.

[89] R. Tamburino, D. Castiglia, L. Marcolongo, L. Sannino, E. Ionata, and N. Scotti, “Tobacco Plastid Transformation as Production Platform of Lytic Polysaccharide MonoOxygenase Auxiliary Enzymes,” *Int J Mol Sci*, vol. 24, no. 1, Jan. 2023, doi: 10.3390/ijms24010309.

[90] D. R. Smith, “Haematococcus lacustris: the makings of a giant-sized chloroplast genome,” *AoB Plants*, vol. 10, no. 5, p. ply058, Sep. 2018, doi: 10.1093/aobpla/ply058.

[91] B. T. Sinn, D. D. Sedmak, L. M. Kelly, and J. V Freudenstein, “Total duplication of the small single copy region in the angiosperm plastome: Rearrangement and inverted repeat instability in Asarum,” *Am J Bot*, vol. 105, no. 1, pp. 71–84, Jan. 2018, doi: https://doi.org/10.1002/ajb2.1001.

[92] B. A. Rasala *et al.*, “Production of therapeutic proteins in algae, analysis of expression of seven human proteins in the chloroplast of Chlamydomonas reinhardtii,” *Plant Biotechnol J*, vol. 8, no. 6, pp. 719–733, Aug. 2010, doi: https://doi.org/10.1111/j.1467-7652.2010.00503.x.

[93] T. Wannathong, J. C. Waterhouse, R. E. B. Young, C. K. Economou, and S. Purton, “New tools for chloroplast genetic engineering allow the synthesis of human growth hormone in the green alga Chlamydomonas reinhardtii,” *Appl Microbiol Biotechnol*, vol. 100, pp. 5467–5477, 2016.

[94] S. Hirschl *et al.*, “Expression and characterization of functional recombinant Bet v 1.0101 in the chloroplast of Chlamydomonas reinhardtii,” *Int Arch Allergy Immunol*, vol. 173, no. 1, pp. 44–50, 2017.

[95] M. Akram *et al.*, “Cloning and expression of an anti-cancerous cytokine: human IL-29 gene in Chlamydomonas reinhardtii,” *AMB Express*, vol. 13, no. 1, p. 23, 2023, doi: 10.1186/s13568-023-01530-1.

[96] G. Di Rocco *et al.*, “A PETase enzyme synthesised in the chloroplast of the microalga Chlamydomonas reinhardtii is active against post-consumer plastics,” *Sci Rep*, vol. 13, no. 1, p. 10028, 2023, doi: 10.1038/s41598-023-37227-5.

[97] D. R. Georgianna *et al.*, “Production of recombinant enzymes in the marine alga Dunaliella tertiolecta,” *Algal Res*, vol. 2, no. 1, pp. 2–9, 2013, doi: https://doi.org/10.1016/j.algal.2012.10.004.

[98] D. Li *et al.*, “Construction of rice site-specific chloroplast transformation vector and transient expression of EGFP gene in Dunaliella salina,” *J Biomed Nanotechnol*, vol. 7, no. 6, pp. 801–806, 2011.

[99] K. Wang *et al.*, “Chloroplast Genetic Engineering of a Unicellular Green Alga Haematococcus pluvialis with Expression of an Antimicrobial Peptide,” *Marine Biotechnology*, vol. 22, no. 4, pp. 572–580, 2020, doi: 10.1007/s10126-020-09978-z.

[100] J. I. Galarza, J. A. Gimpel, V. Rojas, B. O. Arredondo-Vega, and V. Henríquez, “Over-accumulation of astaxanthin in Haematococcus pluvialis through chloroplast genetic engineering,” *Algal Res*, vol. 31, pp. 291–297, 2018, doi: https://doi.org/10.1016/j.algal.2018.02.024.

[101] H. Lopez, D. Magdaleno, and J. Stephano, “The complete chloroplast genome of the green microalgae Dunaliella salina strain SQ,” *Mitochondrial DNA B Resour*, vol. 2, no. 1, pp. 225–226, Jan. 2017, doi: 10.1080/23802359.2017.1310610.

[102] D. R. Smith, R. W. Lee, J. C. Cushman, J. K. Magnuson, D. Tran, and J. E. W. Polle, “The Dunaliella salina organelle genomes: Large sequences, inflated with intronic and intergenic DNA,” *BMC Plant Biol*, vol. 10, May 2010, doi: 10.1186/1471-2229-10-83.

[103] B. E. Slatko, A. F. Gardner, and F. M. Ausubel, “Overview of Next-Generation Sequencing Technologies,” *Curr Protoc Mol Biol*, vol. 122, no. 1, Apr. 2018, doi: 10.1002/cpmb.59.

[104] Harvard Chan Bioinformatics Core (HBC), “Introduction to RNA-Seq using high-performance computing,” https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html.

[105] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimmer for Illumina sequence data,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.

[106] A. D. Twyford and R. W. Ness, “Strategies for complete plastid genome sequencing,” *Mol Ecol Resour*, vol. 17, no. 5, pp. 858–868, Sep. 2017, doi: https://doi.org/10.1111/1755-0998.12626.

[107] S. Izan, D. Esselink, R. G. F. Visser, M. J. M. Smulders, and T. Borm, “De Novo Assembly of Complete Chloroplast Genomes from Non-model Species Based on a K-mer Frequency-Based Selection of Chloroplast Reads from Total DNA Sequences,” *Front Plant Sci*, vol. Volume 8-2017, 2017, [Online]. Available: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2017.01271

[108] B. Langmead and S. Salzberg, “Fast gapped-read alignment with Bowtie 2.,” *Nat Methods*, vol. 9, pp. 357–359, Mar. 2012, doi: 10.1038/nmeth.1923.

[109] Illumina, “MiSeq Sequencing System specifications,” https://www.illumina.com/systems/sequencing-platforms/miseq/specifications.html.

[110] D. E. Fosket, “3 - The Size and Complexity of Plant Genomes,” in *Plant Growth and Development*, D. E. Fosket, Ed., Boston: Academic Press, 1994, pp. 79–152. doi: https://doi.org/10.1016/B978-0-12-262430-8.50007-3.

[111] Illumina, “Estimating Sequencing Coverage,” 2014. [Online]. Available: http://genome.ucsc.edu/ENCODE/protocols/dataStandards/

[112] A. Bankevich *et al.*, “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing,” *Journal of Computational Biology*, vol. 19, no. 5, pp. 455–477, Apr. 2012, doi: 10.1089/cmb.2012.0021.

[113] J. Luo *et al.*, “A comprehensive review of scaffolding methods in genome assembly,” *Brief Bioinform*, vol. 22, no. 5, p. bbab033, Sep. 2021, doi: 10.1093/bib/bbab033.

[114] P. Danecek *et al.*, “Twelve years of SAMtools and BCFtools,” *Gigascience*, vol. 10, no. 2, p. giab008, Feb. 2021, doi: 10.1093/gigascience/giab008.

[115] H. Lantz *et al.*, “Ten steps to get started in Genome Assembly and Annotation,” *F1000Res*, vol. 7, 2018, doi: 10.12688/f1000research.13598.1.

[116] M. Tillich *et al.*, “GeSeq – versatile and accurate annotation of organelle genomes,” *Nucleic Acids Res*, vol. 45, no. W1, pp. W6–W11, Jul. 2017, doi: 10.1093/nar/gkx391.

[117] B. Lang, N. Beck, S. Prince, M. Sarrasin, P. Rioux, and G. Burger, “Mitochondrial genome annotation with MFannot: a critical analysis of gene identification and gene model prediction,” *Front Plant Sci*, vol. 14, Jul. 2023, doi: 10.3389/fpls.2023.1222186.

[118] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *J Mol Biol*, vol. 215, no. 3, pp. 403–410, 1990, doi: https://doi.org/10.1016/S0022-2836(05)80360-2.

[119] D. Liu, M. Hunt, and I. J. Tsai, “Inferring synteny between genome assemblies: a systematic evaluation,” *BMC Bioinformatics*, vol. 19, no. 1, p. 26, 2018, doi: 10.1186/s12859-018-2026-4.

[120] A. C. E. Darling, B. Mau, F. R. Blattner, and N. T. Perna, “Mauve: multiple alignment of conserved genomic sequence with rearrangements,” *Genome Res*, vol. 14, no. 7, pp. 1394–1403, 2004.

[121] P. Zhelyazkova, C. M. Sharma, K. U. Förstner, K. Liere, J. Vogel, and T. Börner, “The Primary Transcriptome of Barley Chloroplasts: Numerous Noncoding RNAs and the Dominating Role of the Plastid-Encoded RNA Polymerase ,” *Plant Cell*, vol. 24, no. 1, pp. 123–136, Jan. 2012, doi: 10.1105/tpc.111.089441.

[122] K. Higo, Y. Ugawa, M. Iwamoto, and T. Korenaga, “Plant cis-acting regulatory DNA elements (PLACE) database: 1999,” *Nucleic Acids Res*, vol. 27, no. 1, pp. 297–300, Jan. 1999, doi: 10.1093/nar/27.1.297.

[123] S. Andrews, “FastQC,” Braham Bioinformatics .

[124] Illumina, “MiSeq System Guide,” 2019. [Online]. Available: www.illumina.com/company/legal.html.

[125] Illumina, “Quality Scores for Next-Generation Sequencing,” 2011. [Online]. Available: http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/

[126] F. T. Bakker *et al.*, “Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline,” *Biological Journal of the Linnean Society*, vol. 117, no. 1, pp. 33–43, 2016.

[127] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows–Wheeler transform,” *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, Jul. 2009, doi: 10.1093/bioinformatics/btp324.

[128] S. Sakaguchi *et al.*, “Application of a simplified method of chloroplast enrichment to small amounts of tissue for chloroplast genome sequencing,” *Appl Plant Sci*, vol. 5, no. 5, p. 1700002, May 2017, doi: https://doi.org/10.3732/apps.1700002.

[129] G. Bookjans, B. M. Stummann, and K. W. Henningsen, “Preparation of chloroplast DNA from pea plastids isolated in a medium of high ionic strength,” *Anal Biochem*, vol. 141, no. 1, pp. 244–247, 1984, doi: https://doi.org/10.1016/0003-2697(84)90452-4.

[130] A. Desai *et al.*, “Identification of Optimum Sequencing Depth Especially for De Novo Genome Assembly of Small Genomes Using Next Generation Sequencing Data,” *PLoS One*, vol. 8, no. 4, pp. e60204-, Apr. 2013, [Online]. Available: https://doi.org/10.1371/journal.pone.0060204

[131] Illumina, “Coverage depth recommendations,” Illumina. Accessed: Nov. 18, 2024. [Online]. Available: https://sapac.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html

[132] D. Zerbino and E. Birney, “Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs,” *Genome Res*, vol. 18, pp. 821–829, Jun. 2008, doi: 10.1101/gr.074492.107.

[133] S. Stegemann, M. Keuthe, S. Greiner, and R. Bock, “Horizontal transfer of chloroplast genomes between plant species,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 7, pp. 2434–2438, Feb. 2012, doi: 10.1073/pnas.1114076109.

[134] T. Takamatsu *et al.*, “Optimized Method of Extracting Rice Chloroplast DNA for High-Quality Plastome Resequencing and de Novo Assembly,” *Front Plant Sci*, vol. 9, 2018, [Online]. Available: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2018.00266

[135] A. R. Quinlan and I. M. Hall, “BEDTools: a flexible suite of utilities for comparing genomic features,” *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010, doi: 10.1093/bioinformatics/btq033.

[136] E. Garrison and G. Marth, “Haplotype-based variant detection from short-read sequencing,” *arXiv preprint arXiv:1207.3907*, 2012.

[137] A. Istvan, *Biostars Handbook (February 2020)*, 2nd ed. 2020.

[138] B. Nguyen *et al.*, "Comprehensive comparative analysis of chloroplast genomes from seven Panax species and development of an authentication system based on species-unique SNP markers," *J Ginseng Res*, vol. 44, Jun. 2018, doi: 10.1016/j.jgr.2018.06.003.

[139] Y. Jiang, Y. Jiang, S. Wang, Q. Zhang, and X. Ding, "Optimal sequencing depth design for whole genome re-sequencing in pigs," *BMC Bioinformatics*, vol. 20, no. 1, p. 556, 2019, doi: 10.1186/s12859-019-3164-z.

[140] S. Greiner, P. Lehwark, and R. Bock, "OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes," *Nucleic Acids Res*, vol. 47, no. W1, pp. W59–W64, Jul. 2019, doi: 10.1093/nar/gkz238.

[141] A. Maréchal and N. Brisson, "Recombination and the maintenance of plant organelle genome stability," *New Phytologist*, vol. 186, no. 2, pp. 299–317, Apr. 2010, doi: https://doi.org/10.1111/j.1469-8137.2010.03195.x.

[142] F. García-Alcalde *et al.*, "Qualimap: evaluating next-generation sequencing alignment data," *Bioinformatics*, vol. 28, no. 20, pp. 2678–2679, Oct. 2012, doi: 10.1093/bioinformatics/bts503.