Muriel, Leandro Pablo Nahuel

Aplicación de herramientas genómicas para el ensamblaje de cloroplastos de cannabis, su estudio filogenómico, y la evaluación de su potencial utilización en estudios de interés médico-científico

2023

Instituto de Ciencias de la Salud

Carrera: Bioquímica

Instituto de Ciencias de la Salud

Bioquímica

Trabajo final:

"Aplicación de herramientas genómicas para el ensamblaje de cloroplastos de cannabis, su estudio filogenómico, y la evaluación de su potencial utilización en estudios de interés médico-científico."

Estudiante: Muriel, Leandro Pablo Nahuel

Director: Mc Carthy, Andrés N.

Agradecimientos

Fueron unos años muy lindos vividos en esta hermosa facultad, nunca me imagine llegar a la etapa donde estoy ahora. Todo fue posible por mi familia, pareja, amigos, mascotas que te ayudaban a relajarte en cada noche de estudio y profesores.

Agradezco a mi mamá que me cuido y ayudo durante toda la carrera, siempre me tenia lista la comida al llegar, me daba la fuerza para seguir y su alegría cada vez que aprobaba una materia era mucho mas que la mía. Gracias a mis hermanos por ayudarme siempre en todo cuando lo necesitaba.

Algo muy especial que encontré en la facultad fue a mi pareja que la quiero mucho, siempre ayudándome en todo lo que podía, me daba ánimos, me motivaba a estudiar y mucho más.

Gracias a mis amigos, por no enojarse cuando no podía juntarme con ellos, en cambio, me daban apoyo y ánimo.

Agradezco a mi profesor de tesis por estar siempre en contacto, ayudarme con muy buena voluntad.

También quiero hacer una mención especial a los profesores que he tenido, sin sus enseñanzas no hubiera logrado estar donde estoy ahora. Y agradezco profundamente al coordinador de la carrera por siempre darme la oportunidad de seguir con ella, aunque me inscribiera tarde y me presentara fuera de termino, siempre me tuvo paciencia.

En la vida con esfuerzo y dedicación todo se puede. Gracias por todo a todos.

Resumen

En varias partes del mundo se sigue teniendo el estigma de que el Cannabis es una droga ilícita y peligrosa. Se ha estudiado su uso a lo largo de los años, en diversos campos. Después del descubrimiento de los receptores cannabinoides y el estudio de la estructura y componentes del cannabis, surgió el interés del ámbito científico para medir sus capacidades medicinales. Entre estas aplicaciones contamos, por ejemplo, con tratamientos para trastornos de ansiedad, inhibición del crecimiento celular en cáncer de cuello uterino, y su utilización como antimicrobiano. Por ello, es importante poder distinguir entre diferentes cepas de Cannabis sativa, debido a que pueden contar con diversas características, como tener mayor potencial medicinal que recreativo o viceversa. Para este fin se analizará el cloroplasto de esta planta, cuya estructura y genoma están altamente conservados con una ligera variabilidad entre especies vegetales. Un método para lograr una identificación presuntiva entre las diferentes variedades, es el análisis de genes específicos (rbcL y matK) del ADN del cloroplasto. Es por ello que en este trabajo se utilizaron 2 programas bioinformáticos (Getorganelle y NOVOPlasty) para ensamblar cloroplastos a partir de 12 secuencias de genoma completo de Cannabis sativa, con el fin de realizar un análisis comparativo entre ambos programas, como así también entre los genes antes mencionados. Los resultados revelan que NOVOPlasty demostró una tasa de éxito superior en el ensamblaje en comparación con Getorganelle, logrando ensamblar 11 de los 12 cloroplastos, mientras que el segundo solo alcanzó 10. Cabe destacar que el único cloroplasto no ensamblado por NOVOPlasty, fue ensamblado correctamente por Getorganelle. Con la utilización de Bandage para una visualización preliminar del ensamblaje, se vio que un cloroplasto contaba con una cantidad de bases mayor que el resto, los que dieron iguales entre sí. Con los cloroplastos ensamblados correctamente se anotaron, y localizaron los genes rbcL y matK. El gen rbcL se mantuvo igual en las 12 secuencias, en cambio, el gen matK dio diferente en el cloroplasto con el mayor número de bases mencionado anteriormente.

Índice

Abreviaturas

INTA: Instituto Nacional de Tecnología Agropecuaria

CONICET: Consejo Nacional de Investigaciones Científicas y Técnicas

INASE: Instituto Nacional de Semillas (INASE)

CB1: Receptor cannabinoide de tipo 1

CB2: Receptor cannabinoide de tipo 2

THC: Δ-9 tetrahidrocannabinol

CBD: Cannabidiol

LSC: Región larga de copia única

SSC: Región corta de copia única

IR: Región repetida inversa

Rubisco: Ribulosa bifosfato carboxilasa

matK: MaturaseK

rbcL: Gen de la subunidad grande de la ribulosa bifosfato carboxilasa

BLAST: Herramienta local básica de búsqueda de alineación

PCR: Reacción de la cadena de la polimerasa

ddNTPs: Dideoxinucleotidos

ADNds: ADN de doble cadena

ADNss: ADN de simple cadena

PE: Lecturas pareadas

ZMW: Guías de onda de modo cero

NCBI: Centro Nacional para la Información Biotecnológica

NIH: Institutos Nacionales de Salud

OGDWAW: Dibujo del genoma del organelo

1- **Introducción**

1.1-**Canabis sativa**

El cannabis se divide en tres especies *sativa*, *indica* y ruderalis, pero existe un gran debate al respecto, y algunos los consideran subespecies de la misma especie parental. Cannabis sativa es una planta herbácea de floración anual perteneciente a la familia Cannabaceae originaria de Asia. Esta muy extendido debido a su adaptación geo climática y ecosistémica, llegando a alturas entre 1 a 6 metros (ilustración 1). (2)

Ilustración 1- Cannabis Sativa. Fuente: Biblioteca de botánica y herbarios de la universidad de Harvard.

1.1.1-**Historia**

La planta ha sido utilizada durante siglos por diversas culturas, aprovechando sus amplias propiedades. Desde los inicios de la civilización humana registrada, se ha cultivado y empleado en la confección de diversos productos textiles, como fibras vegetales para tejidos, pasta de celulosa para la fabricación de papel, cuerdas y sogas, entre otras aplicaciones. Existen registros del uso en la medicina tradicional que se remontan a mediados del año 2700 a.C. (1). El primer uso medicinal del cannabis se lo asocia al emperador chino Shen Nung en el 2700 a. C. En la cultura asiria (1800 a. C.) se empleó con fines medicinales para el dolor, la epilepsia, la neuralgia y la pediculosis, y en el Antiguo Egipto (1700 a. C.) se utilizó como tratamiento ocular. (23)

Se han encontrado informes en papiros egipcios, entre nativos africanos y en la medicina popular sudamericana que destacan su aplicación como antiséptico en el tratamiento de

heridas, hinchazones y en diversas enfermedades como el ántrax, la sepsis, la disentería y la malaria. (23)

A pesar de sus históricos beneficios medicinales, sus efectos psicoactivos la convirtieron en una sustancia prohibida y fue incluida en la lista de sustancias ilegales de la Farmacopea Británica en 1932 y de la Farmacopea Estadounidense en 1941. (30)

1.1.2- **Situación en Argentina**

En Argentina, el cannabis medicinal está regulado por la Ley 27.350, que establece las bases para la investigación médica y científica del uso medicinal de la planta y sus derivados. Esta legislación ha dado lugar al Programa Nacional para el Estudio y la Investigación del uso medicinal de la planta de cannabis, con una serie de objetivos bien definidos:

- Promover acciones orientadas a garantizar el derecho a la salud.
- Promover medidas de concientización para la población en general.
- Establecer guías de asistencia, tratamiento y accesibilidad.
- Garantizar el acceso gratuito al aceite de cáñamo y demás derivados del cannabis a los pacientes que se inscriban en el Programa.
- Desarrollar alternativas terapéuticas a problemas de salud que no tratan los tratamientos médicos convencionales.
- Investigar los fines terapéuticos y científicos de la planta de cannabis y sus derivados.
- Conocer los efectos secundarios del uso medicinal de la planta de cannabis y sus derivados y establecer las limitaciones para su uso.
- Fomentar la participación de pacientes y sus familiares para que aporten su experiencia, vivencias y métodos usados.
- Brindar asesoramiento, cobertura y seguimiento del tratamiento a los pacientes que participen del programa.
- Contribuir a la capacitación continua de profesionales de la salud en el uso medicinal de la planta de cannabis y sus derivados.

En el ámbito del cultivo el Instituto Nacional de Tecnología Agropecuaria (INTA) y el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) están autorizados para llevar a cabo investigaciones médicas y científicas mediante el cultivo

de cannabis. Asimismo, el Instituto Nacional de Semillas (INASE) regula las condiciones de producción, difusión, manejo y acondicionamiento. (3)

1.1.3-**Compuestos**

Cannabis sativa contiene más de 100 compuestos químicos llamados Cannabinoides. (1). Estos compuestos actúan en los receptores cannabinoides, los cuales están implicados en una amplia variedad de procesos fisiológicos, como en la modulación de la liberación de neurotransmisores, la regulación de la percepción del dolor, y las funciones cardiovasculares, gastrointestinales y hepáticas. (31)

1.1.3.1-**Receptores y principales cannabinoides**

Los dos principales receptores que componen el sistema cannabinoide son el CB1 y CB2. Estos son proteínas transmembrana, capaces de transmitir una señal extracelular al interior de la célula. Los receptores CB1 son metabotrópicos, se encuentran con mayor abundancia en el cerebro, pero también están en el hígado, los pulmones, la musculatura lisa, el tracto gastrointestinal, las células pancreáticas ß, el endotelio vascular, los órganos reproductivos y en el sistema inmunológico (ilustración 2). Por otro lado, los receptores CB2 se encuentran ubicados en las células del sistema inmunitario, sistema nervioso central, las fibras nerviosas de la piel y en los queratinocitos, en las células óseas, hepáticas y pancreáticas (ilustración 2). (31)

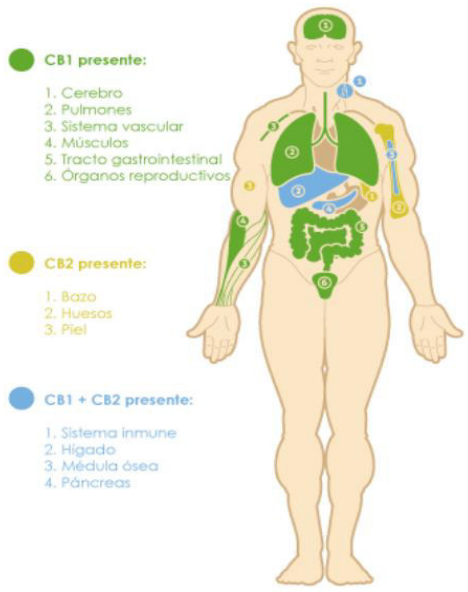

Ilustración 2- Distribución de receptores CB. Fuente: Fundación canna

El cannabinoide más estudiado es el THC, responsable de las propiedades psicoactivas. El otro cannabinoide principal es el CBD, que carece de propiedades psicoactivas, pero se cree que tiene propiedades terapéuticas. En general, el THC se encuentra en mayor concentración que el CBD. Actualmente, el CBD es el foco de una intensa investigación debido a su potencial en varias áreas terapéuticas. (2)

Desde la visualización estructural y la síntesis del THC e CBD en los años 1960, junto al reconocimiento del papel del sistema cannabinoide en la salud y la enfermedad, el interés científico en el Cannabis ha aumentado. (23)

La potencia de los efectos depende de las condiciones de crecimiento, las características genéticas del cannabis, la proporción de THC con respecto a otros cannabinoides y la parte de la planta utilizada. Las plantas de cannabis pueden cultivarse selectivamente para maximizar el contenido de THC o CBD. (24)

1.1.4-**Uso medicinal**

Aun se investiga el uso del CBD para fines médicos. Asimismo, numerosos hallazgos prometedores han evidenciado su contribución positiva a la salud.

1.1.4.1-**Apoyo en la superación de adicciones**

Las adicciones por el uso de sustancias son condiciones crónicas recurrentes, y el riesgo de recaída persiste por muchas razones, como la abstinencia, la susceptibilidad al estrés, la ansiedad elevada y el control deficiente de los impulsos.

Estudios en modelos animales han revelado que la administración transdérmica de CBD atenúa la abstinencia y el estrés, sin provocar tolerancia ni efectos sedantes. Después de la finalización del tratamiento, la necesidad del consumo permaneció atenuada hasta 5 meses, a pesar de que el nivel de CBD en plasma y cerebro permaneció detectable solamente 3 días. (4)

1.1.4.2 **Tratamiento de trastornos de ansiedad**

El miedo y la ansiedad son respuestas adaptativas para enfrentar amenazas y para la supervivencia, pero el exceso de estas puede ser perjudicial para la salud. El CBD ha sido comprobado como un posible tratamiento ansiolítico, respaldado por estudios en humanos. En dosis orales que oscilan entre 300 y 600 mg, se observa una reducción en la ansiedad inducida experimentalmente en controles sanos, y de manera significativa, en pacientes con trastorno de ansiedad social (TAE). (5)

1.1.4.3 **Inhibición del crecimiento celular en cáncer de cuello uterino**

El cáncer de cuello uterino representa una gran problemática de salud a nivel mundial, reportando más de medio millón de casos nuevos al año. Estudios recientes han destacado la capacidad del CBD para detener la proliferación celular en líneas celulares de cáncer de cuello uterino. Estos resultados sugieren la posibilidad de que el CBD sea una herramienta valiosa en la búsqueda de terapias efectivas para combatir este tipo de cáncer (8)

1.1.4.4 **Propiedades antibacterianas**

La resistencia bacteriana a la terapia antimicrobiana es una preocupación a nivel mundial. Las enfermedades infecciosas son la segunda causa de muerte en todo el mundo, resultando el fallecimiento de 17 millones de personas cada año a causa de infecciones bacterianas. La resistencia de múltiples bacterias a una o más clases de antibióticos ha limitado las opciones para combatir este problema, principalmente debido al bajo número de nuevos antibióticos desarrollados y aprobados en las últimas décadas. (26)

La utilización de Cannabis sativa y sus compuestos contra bacterias patógenas, no había llamado la atención por parte de la comunidad científica en comparación con otras aplicaciones médicas. Sin embargo, recientemente, se han publicado nuevos estudios relacionados con los efectos antimicrobianos de los cannabinoides, lo que posiciona a los extractos de Cannabis como potenciales agentes antimicrobianos. (26)

Los alcaloides, flavonoides, péptidos, taninos y fenoles son conocidos por sus propiedades antibacterianas y muchos de ellos se encuentran en Cannabis sativa.

Se demostró que los cannabinoides mejoran la actividad antimicrobiana de los antibióticos convencionales contra bacterias resistentes. (27)

Un ejemplo de esto es como CBD mejoró significativamente el efecto antibacteriano de la eritromicina y la rifampicina contra E. coli VCS257, además de potenciar el efecto antibacteriano de la kanamicina contra S. aureus subsp. aureus Rosenbach. (25)

1.2- **El cloroplasto**

1.2.1 - **Descripción y estructura**

Los cloroplastos son organelos esenciales presentes en las células vegetales, desempeñan un papel fundamental en procesos energéticos, tales como la fotosíntesis, donde la energía lumínica se convierte en energía química, procesos de transporte de electrones y respiración. (28)

La mayoría de las células vegetales contienen entre 400 y 1.600 cloroplastos. La mayor concentración se encuentra en las hojas, debido a su gran importancia en la fotosíntesis. La estructura de los cloroplastos es clave para su funcionalidad (ilustración 3). Están rodeados por una doble membrana, siendo la externa permeable a pequeñas moléculas orgánica, mientras que la membrana interna no lo es, contando con proteínas transportadoras para este fin. (47)

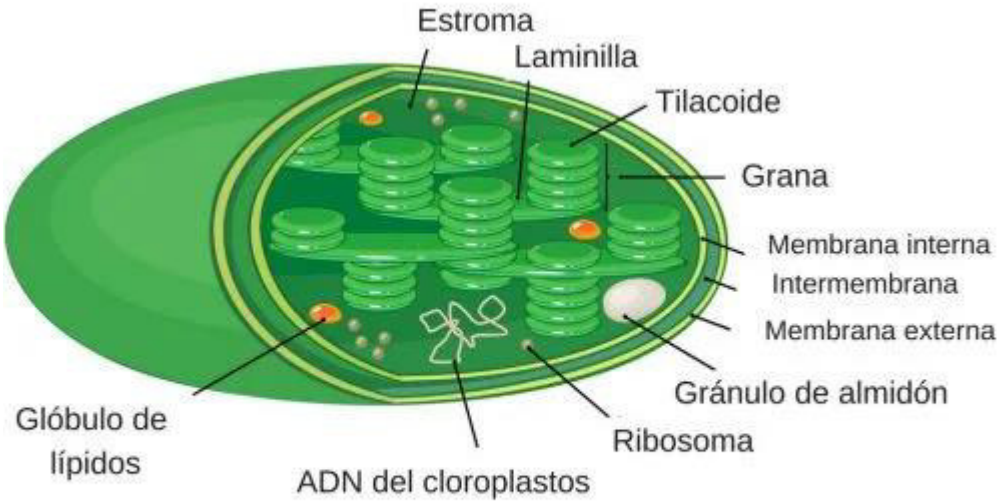

Ilustración 3- Estructura del cloroplasto. Fuente: Atlas de histología vegetal

El estroma es el compartimento vital donde se alberga el genoma y sirve como espacio para la transcripción y traducción. La síntesis de proteínas es llevada a cabo mediante el ribosoma 70 S, similares a los presentes en las bacterias. (47)

1.2.2 - **El genoma del cloroplasto**

El cloroplasto contiene su propio material genético circular, bicatenario, independiente del nuclear, con una replicación semiautónoma durante la división celular. Su genoma se compone de cuatro regiones bien definidas: una región larga de copia única (LSC), una región corta de copia única (SSC) y dos regiones repetidas inversas (IR) (ilustración 4).

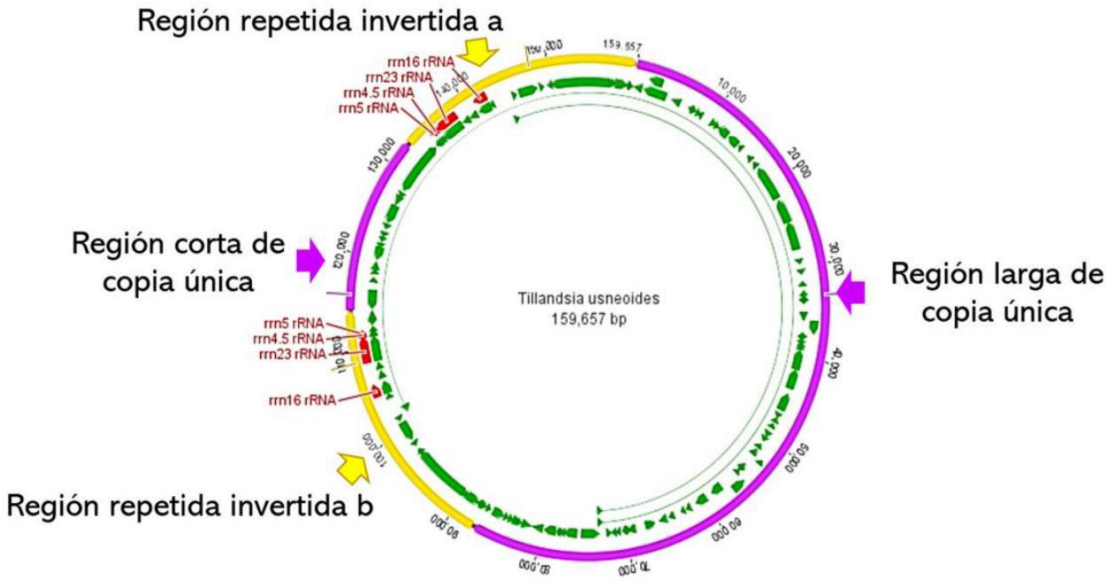

Ilustración 4- Genoma de Cloroplasto. Fuente: El otro genoma de las plantas de Espinosa Barrera

La región LSC es la más extensa, y en ella alberga genes que codifican para la Rubisco, responsable de la fijación de CO2 en la fotosíntesis. También se encuentra la ATP sintasa, necesaria para la producción de energía y algunos componentes del fotosistema II. Por otro lado, el SSC es la región más pequeña del genoma, contiene múltiples secuencias que codifican componentes de NADH deshidrogenasa, necesaria para el acarreo de electrones durante la fotosíntesis. Las IR son dos secuencias que flanquean a SCC y LSC, caracterizándose por tener la misma información, y contienen una alta cantidad de secuencias de ADN repetidas, destacándose la presencia de los genes ribosomales. (28)
El genoma del cloroplasto este compuesto por alrededor de unos 150 genes, incluyendo aproximadamente 80 proteínas, 30 ARNt y 4 ARNr. Estos genes se agrupan en operones, similar a los genomas bacterianos, expresando ARN mensajero policistrónico que posteriormente es clivado. (19)

Este genoma se encuentra altamente conservado y su estructura (tamaño, contenido y orden de genes) evoluciona mucho más lento que el genoma nuclear e incluso el mitocondrial. Esta elevada conservación podría deberse a la organización de los genes en operones, como ocurre en los genomas de las cianobacterias, algas verdes y plantas terrestres. (46)

En la célula, la cantidad de copias de la secuencia de DNA del cloroplasto supera al DNA nuclear, reflejando la mayor cantidad relativa del plasmido. Debido a su naturaleza no recombinante, mayor proporción en las células, herencia materna y tasas de mutación bajas, gran parte de la investigación genómica se ha focalizado en el genoma del cloroplasto.

1.2.3-**Teoría endosimbiótica**

El origen de los cloroplastos se explica mediante la teoría endosimbiótica.

Esta sostiene que los cloroplastos derivan de una cianobacteria fotosintética ancestral de vida libre que fue fagocitada por una célula eucariótica no fotosintética, estableciendo una relación de endosimbiosis. La asociación permitió que estos eucariotas adquirieran la capacidad de realizar la fotosíntesis, convirtiéndose finalmente en los cloroplastos.

A medida que la célula hospedadora y las cianobacterias simbiontes evolucionaron de manera conjunta, estas últimas perdieron su autonomía. El genoma del cloroplasto está compuesto por aproximadamente 150 genes, considerablemente inferior al de las cianobacterias actuales, que poseen entre 2.000 a 3000 genes. (14)

Esta discrepancia pone en manifiesto una dependencia creciente del cloroplasto hacia el hospedador, poniendo en evidencia el traspaso de los genes del cloroplasto al núcleo. La mayoría de las proteínas que desempeñan funciones en el organelo son codificadas por genes del genoma nuclear, sintetizadas en el citoplasma y transportadas posteriormente al cloroplasto. En cuanto a los genes que se han conservado en el cloroplasto, mayormente están relacionados con la expresión de sus propios genes y con procesos fotometabólicos. (14)

1.2.4- **Código de barras**

Por las características del cloroplasto, ha sido el foco de estudio en análisis evolutivos, análisis filogenómicos y estudios de diversidad entre especies. Una herramienta eficaz y ampliamente utilizada que permite la identificación rápida y precisa de plantas, análogamente a un código de barras para las especies vegetales.

Estos “códigos de barras” vegetales implican el uso estandarizado de una o varias regiones de ADN con el objetivo de proporcionar una identificación de especies rápida, precisa y automatizable. (29)

En el 2009, el Grupo de Trabajo de Plantas del Consorcio para el Código de Barras de la Vida (CBOL) publicó un metaanálisis de la eficiencia de los principales marcadores de códigos de barras, recomendando dos regiones del cloroplasto matK y rbcL. (22)

El gen rbcL codifica la subunidad más grande de la rubisco, una enzima fundamental para la adaptación de las plantas a las variaciones en las concentraciones de CO2. Este gen fue propuesto como un fragmento de código de barras por su característica de universalidad en los cloroplastos, su fácil amplificación y comparabilidad. Su tasa de evolución es baja complicando la diferenciación entre individuos de la misma especie, pero presenta resultados favorables a nivel de género. (33)

El gen *matK* está ampliamente presente en las plantas, localizándose en el intrón *trnK,* que codifica una proteína involucrada en el empalme de intrones del Grupo II. La secuencia *matK* es uno de los genes de más rápida evolución, con una tasa 2 a 3 veces mayor que rbcL, teniendo una buena tasa de éxito en la diferenciación entre individuos de la misma especie. (33)

1.3-**Importancia y aplicación del estudio genómico**

La identificación genética de diversas cepas de cannabis sativa es fundamental para el control de calidad en un mercado en crecimiento no regulado, intereses industriales y para un uso medicinal eficaz y especifico. (10)

Se han desarrollado cepas con bajo contenido en cannabinoides psicoactivos y alto contenido en CBD con un potencial uso terapéutico. Además, se han desarrollado otras cepas con diferentes relaciones de THC/CBD, según el uso y características deseadas. (40)

El análisis comparativo de las secuencias rbcL y matK entre diferentes cloroplastos daría las pautas para identificar la planta y ubicarla correctamente en el árbol genealógico correspondiente. Además, permitiría potencialmente distinguir el tipo de cannabis y prever las posibles características potenciales que podría presentar.

2- **Objetivos**

- Adquirir conocimiento en la búsqueda de genomas secuenciados, y aprender a utilizar los programas de ensamblaje.
- Realizar una comparación del flujo de trabajo y de los resultados generados por los programas de ensamblaje NOVOPlasty y GetOrganelle.
- Realizar anotaciones de los cloroplastos ensamblados, y localizar los genes rbcL y matK.
- Analizar posibles diferencias en la secuencia de nucleótidos de los genes rbcL y matK.

3- **Materiales y métodos**

3.1-**Secuenciación de ADN**

La determinación del orden de los ácidos nucleicos es un componente fundamental en diversos campos de investigación. A lo largo de los últimos cincuenta años, numerosos investigadores se han dedicado a desarrollar técnicas y tecnologías para facilitar esta tarea. En este periodo de tiempo, la tecnología experimento grandes avances, transitando desde la secuenciación corta y lenta de oligonucleótidos hasta la secuenciación rápida del genoma completo. (35)

En la actualidad conviven 3 generaciones de secuenciación de DNA: primera generación, segunda generación y tercera generación.

3.1-**Método de primera generación**

En 1977, se desarrolló el método de Sanger, considerado de primera generación. Este método utiliza el enfoque de secuenciación por síntesis de una cadena de ADN marcada radiactivamente, complementaria a una cadena molde, mediante la utilización de la técnica de terminación de cadena dideoxi. La metodología, como se observa en la ilustración 5, se basa en la utilización de 4 tubos. En cada uno de ellos, ocurre una reacción de PCR para cada tipo diferente de nucleótido marcado. Para la realización de estas reacciones se requiere dNTPs, ddNTPs y los primers complementarios a la cadena molde, que sirven como iniciadores para la polimerasa termorresistente. La polimerasa va amplificando y uniendo dNTPs hasta la incorporación de forma aleatoria de un ddNTPs, lo que finaliza la amplificación. Este proceso se realiza simultáneamente en los 4 tubos, generando fragmentos de diferentes tamaños. (46)

La lectura se lleva a cabo mediante electroforesis en gel de agarosa, donde las muestras son separadas en distintas columnas. Los fragmentos de menor tamaño se desplazan a mayor velocidad que los fragmentos de mayor de tamaño, ubicándose en la parte inferior del gel. Finalmente, se arma la secuencia siguiendo el orden de abajo hacia arriba del gel, donde cada marca corresponde a un nucleótido. (46)

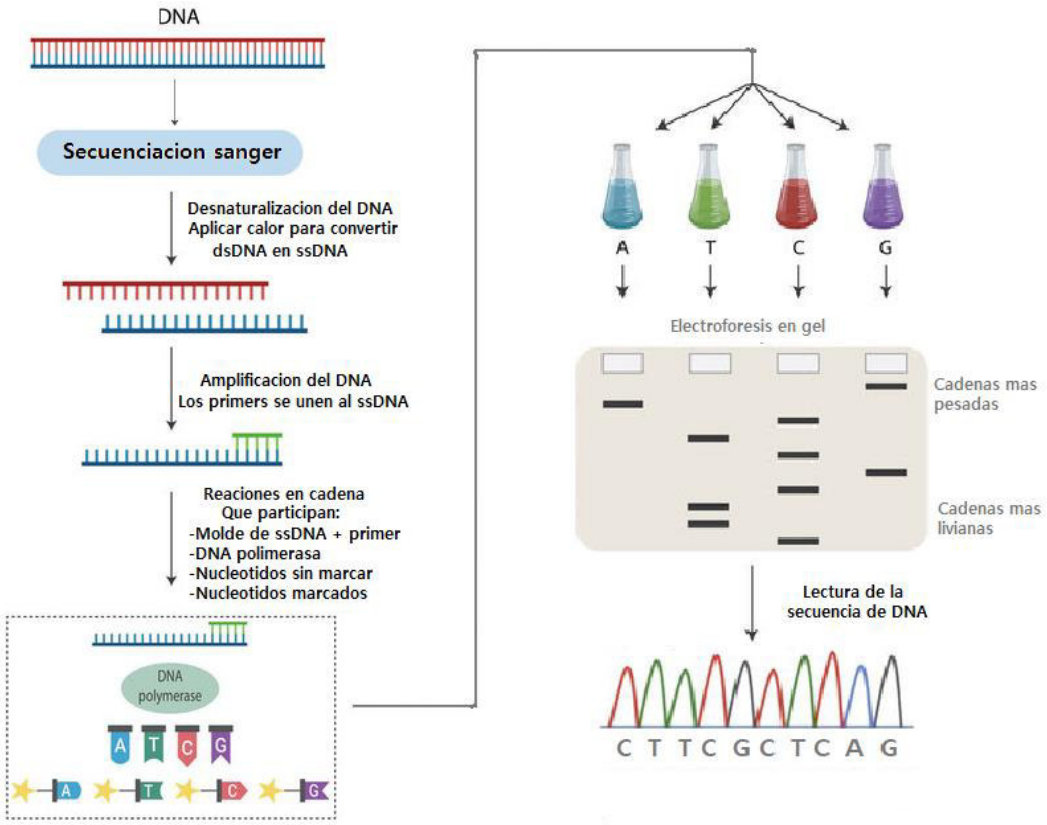

Ilustración 5- Pasos del método de Sanger. Fuente: Filogenomica principios, oportunidades y dificultades de la filogenia

El método experimento mejoras, las cuales incluyeron la sustitución de los ddNTPs marcados con radioactividad por una detección basada en fluorometría. Esta modificación posibilita que la reacción ocurra en un solo recipiente en lugar de cuatro. Además, se implementó la electroforesis capilar para aumentar la velocidad del proceso. Ambas mejoras contribuyeron al desarrollo de máquinas de secuenciación de ADN cada vez más automatizadas. (35)

3.2-**Método de segunda generación**

3.2.1-**Cobertura y profundidad**

Existen dos conceptos fundamentales para entender el proceso y los resultados de las pruebas basadas en tecnologías de segunda generación: cobertura y profundidad.

La cobertura se refiere al porcentaje de bases del genoma de referencia que están siendo secuenciada. Por otro lado, la profundidad representa el número promedio de veces que cada base es secuenciada en los fragmentos de ADN (ilustración 6). (34)

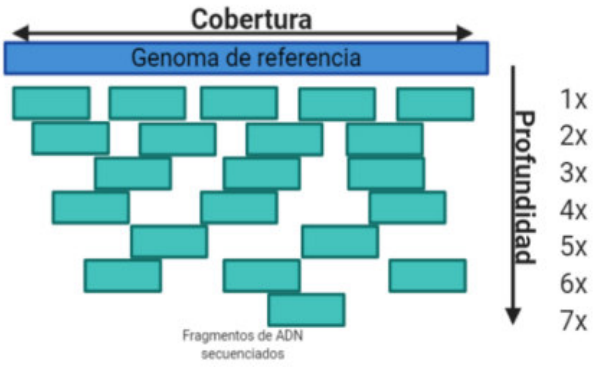

Ilustración 6- Concepto cobertura y profundidad. Fuente: Secuenciación de nueva generación (NGS): Presente y futuro en la práctica clínica.

3.2.2-**Metodología Illumina**

Illumina es la compañía líder a nivel mundial en el desarrollo de tecnologías para la secuenciación masiva. Esta tecnología se basa en la secuenciación por síntesis, está ampliamente adoptada, y es responsable de generar más del 90% de los datos de secuenciación del mundo. El flujo de trabajo consta de 3 pasos básicos. (50)

3.2.2.1 **Preparación de la Biblioteca**

La biblioteca de secuenciación se prepara mediante la fragmentación aleatoria de la muestra de DNA o cDNA con ondas sonoras o enzimas en fragmentos de entre 200 a 800 pares de bases. Posteriormente, se realiza una ligación de adaptadores en ambos extremos 5' y 3', formando así la biblioteca de secuenciación (ilustración 7). Estos adaptadores son secuencias conocidas, y son esenciales para la amplificación e identificación posterior. (50)

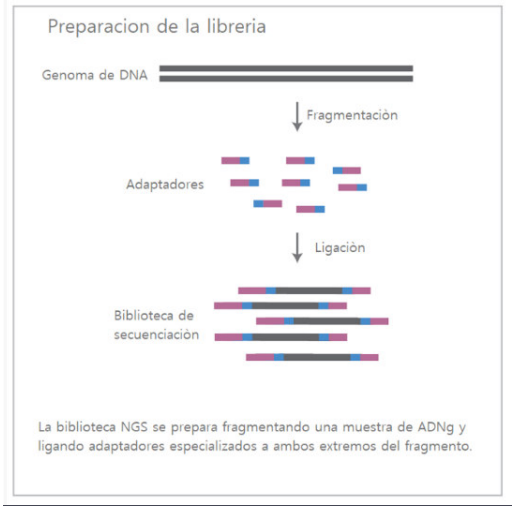

Ilustración 7- Preparación de la librería. Fuente: Principios básicos de la secuenciación de nueva generación.

3.2.2.2 **Generación de Clústers**

La biblioteca se carga en una celda de flujo donde los fragmentos son capturados por una capa de oligonucleótidos unidos a la superficie, ya que estos son complementarios a los adaptadores de la biblioteca (ilustración 8-A). La polimerasa inicia el proceso de amplificación de la hebra de ADN, sintetizando la hebra complementaria. Luego, se produce la desnaturalización del ADNds y se elimina por lavado el ADN original (ilustración 8-B). (50)

La hebra se pliega sobre sí misma, y el adaptador libre se ancla a su respectiva secuencia complementaria de oligonucleótido, adopta así una forma de puente. La polimerasa sintetiza la hebra complementaria al puente, el cual posteriormente se desnaturaliza. Se obtienen así dos hebras ADNss (ilustración 8-C). Este proceso se repite una gran cantidad de veces, ocurriendo simultáneamente en millones de grupos, lo que da como resultado amplificaciones clonales de los fragmentos (ilustración 8-D). Después de la amplificación, los ADNds se desnaturalizan y se lava una de las cadenas, en este ejemplo la cadena R, mientras la F forma los clusters. Se bloquean los oligonucleótidos libres para evitar que sean un iniciador de polimerización. Y se prosigue a la secuenciación (ilustración 8-E). (50)

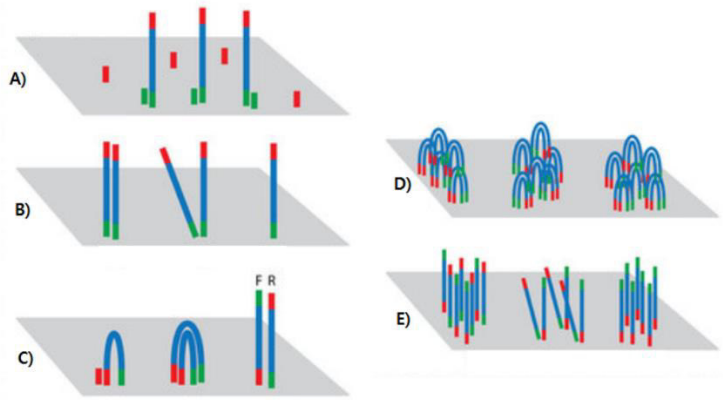

Ilustración 8- Amplificación método illumina. Fuente: Filogenomica: principios, oportunidades y dificultades de la filogenia

3.2.2.3-**Secuenciación**

Illumina utiliza un método patentado basado en nucleótidos con una modificación química, conocida como terminadores reversibles, que evita la unión de más de un nucleótido marcado en cada sitio de reacción. Cada vez que una base se adhiere, emite una señal de fluorescencia propia que permite su identificación. El marcador debe ser removido antes de la incorporación del siguiente nucleótido para evitar que las señales se solapen. Este paso se repite simultáneamente en todas las hebras y clústers de forma paralela, hasta completar la secuenciación de la primera cadena.

Luego se retira el bloqueo de los oligonucleótidos, se forman nuevamente los clusters y, en esta ocasión, se elimina la cadena F. Se procede a secuenciar la cadena R.

Finalmente, los datos se exportan a un archivo de salida en formato PE para su futuro uso (ilustración 9). (34)

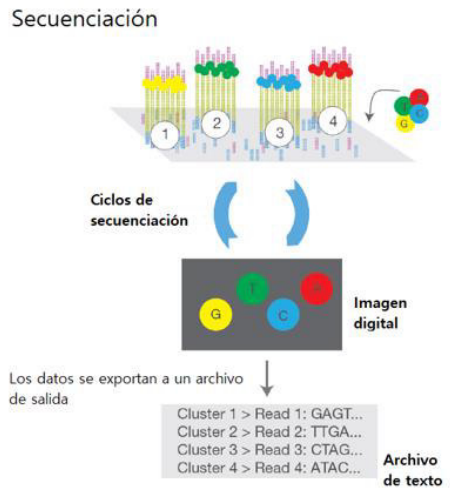

Ilustración 9. Secuenciación. Fuente: Principios básicos de la secuenciación de nueva generación

3.2.2.4- **Secuenciación de Lecturas Pareadas**

Un avance importante en la tecnología de secuenciación ocurrió con el desarrollo de la secuenciación de lecturas pareadas (PE). Implica secuenciar ambos extremos de los fragmentos de ADN de una biblioteca y alinear las lecturas como pares de lecturas. Además de producir el doble de la cantidad de lecturas en el mismo tiempo y esfuerzo, las PE permiten una mayor precisión y la capacidad de detectar inserciones o deleciones, lo que no es posible con datos de lecturas individuales. (45)

3.3 - **Métodos de tercera generación**

Las cortas longitudes de lectura del método Illumina hacen que sea poco adecuado para abordar algunos problemas biológicos particulares, como el ensamblaje y determinación de regiones genómicas complejas.

Las tecnologías de tercera generación utilizan fragmentos de gran tamaño (>10 Kb). Aunque esta técnica pueda ser rápida y permitir lecturas de fragmentos más extensos, presenta una alta tasa de error (>5%), lo que puede afectar el análisis. (37)

3.3.1 **Metodología SMRT de Pacific Biosciences (PacBio)**

A diferencia de Illumina, la secuenciación de una sola molécula en tiempo real (SMRT) es un método que no requiere una pausa entre los pasos de lectura. Utiliza longitudes más extensas de DNA y no requiere una amplificación clonal.

En primer lugar, se crea una librería SMRTbell, que implica la fragmentación del DNA de interés y unión de adaptadores a ambos extremos para formar el SMRTbell (Ilustracion 10). A continuación, se agrega la DNA polimerasa junto al primer complementario a los adaptadores. (37)

Ilustración 10- SMRTbell. Fuente: Introducción a SMRTbell preparación temprana.

En cada ZMW (pozos microscópicos), se inmoviliza una única DNA polimerasa en la parte inferior junto a su SMRTbell (ilustración 10). Se le agrega los cuatro nucleótidos marcados con diferentes indicadores de fluorescencia, para obtener espectros de emisión diferenciados. (37)

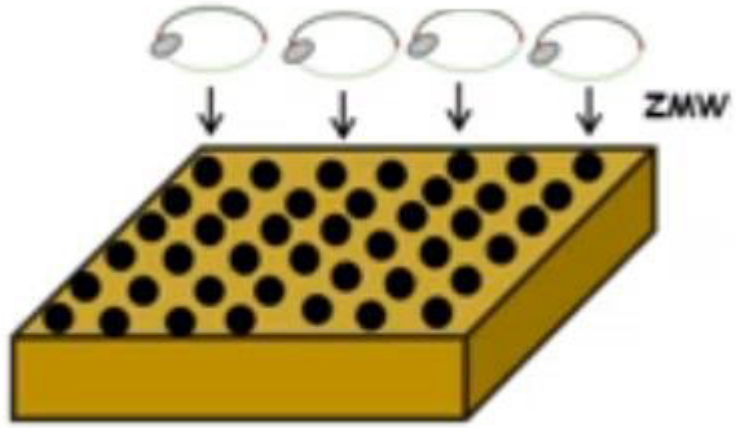

Ilustración 11- Placa con ZMW. Fuente: Secuenciadores de ADN miniaturizados PacBio.

A medida que la polimerasa agrega bases, se produce una excitación y emisión diferente para cada tipo de nucleótido (ilustración 12). Posteriormente, se elimina el marcador y se incorpora otra base, repitiendo el ciclo hasta la finalización de la secuenciación. (37)

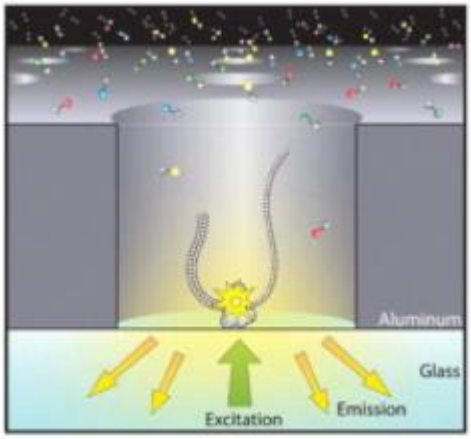

Ilustración 12- Excitación y emisión de señal. Fuente: Secuenciadores de ADN miniaturizados PacBio

3.4- **Estrategia Simultánea**

Las fortalezas y debilidades de la secuenciación illumina y SMRT son complementarias, lo que llevo al desarrollo de una estrategia de secuenciación simultánea integrando ambas técnicas. Estos enfoques a menudo implican el uso de datos de lectura cortas de alto rendimiento y alta precisión de illumina, para corregir errores en las lecturas largas y reducir la cantidad de datos de secuencia de lectura larga que son costosas.

Las lecturas largas de SMRT pueden proporcionar alineamientos y detecciones confiables de variantes genómicas, mientras que las lecturas cortas refinan las alineaciones y ensamblajes. (37)

3.5-**Fuente de datos de secuenciación**

El Centro Nacional para la Información Biotecnológica (NCBI) es parte de la Biblioteca Nacional de Medicina de Estados Unidos, una rama de los Institutos Nacionales de Salud (NIH). Almacena y constantemente actualiza la información relacionada con las secuencias genómicas, además cuenta con un índice de artículos científicos referentes a biomedicina, biotecnología, bioquímica, etc. (39)

En la actualidad cuenta con más de 25.000.000 secuencias genómicas cargadas, donde más de 19.000.000 son de lectura pareada. La mayoría de las secuenciaciones almacenadas son realizadas mediante la metodología de Illumina.

3.6-**Ensamblaje del genoma**

El proceso de descifrar la secuencia genómica a partir de fragmentos de ADN, junto con alguna información adicional disponible como referencia, se denomina ensamblaje de genoma. Este procedimiento es fundamental para comprender la organización y estructura genética de un organismo, ya sea una bacteria, planta, animal o humano.

Las estrategias para el ensamblaje se pueden dividir en dos categorías: ensamblaje por comparación, que utiliza un genoma como referencia; y ensamblaje *de novo*, en el cual se utiliza solo la información obtenida de la secuenciación para reconstruir el genoma en cuestión, sin conocimiento previo de la organización del mismo. (41)

El ensamblaje genómico enfrenta desafíos, como la presencia de secuencias repetitivas, variaciones genéticas y errores en los datos de secuenciación. Estos factores pueden complicar la tarea de reconstruir de manera precisa y completa el genoma de un organismo. (38)

Hay dos rutas generales para el ensamblaje de un genoma. Una consiste en usar una ruta personalizada, requiriendo la combinación de varios programas bioinformáticos de forma individual. Estos programas cumplen funciones como la de filtrar datos, generar alineamientos, ensamblar secuencias, realizar búsquedas de coincidencia, anotación de genes, etc. La segunda forma es utilizar ensambladores especializados. Estos

ensambladores agrupan un conjunto de programas bioinformáticos para un uso directo y rápido. (41)

Existe una gran cantidad de ensambladores, los que se van a utilizar en este trabajo son los dos de código libre más utilizadas actualmente en la literatura científica especializada: GetOrganelle y NOVOPlasty.

3.6.1-**Getorganelle**

GetOrganelle es un conjunto de herramientas de última generación para ensamblar con precisión genomas de organelos a partir de datos de secuenciación del genoma completo (WGS). Este conjunto integra Bowtie2, BLAST y SPAdes. Además, trabaja en conjunto con el programa Bandage para una visualización preliminar del gráfico de ensamblaje. (7)

3.6.1.1-**Bowtie2**

Bowtie2 es una herramienta utilizada para alinear las lecturas de secuenciación con secuencias de referencia, a este procedimiento se lo denomina mapeado. Admite tanta alineación de lectura simple y de lecturas pareadas. Genera un archivo de alineamiento en un formato SAM compatible con una gran cantidad de otras herramientas bioinformáticas. Es a menudo el primer paso en la iniciación de los proyectos de ensamblaje y en los procesos de genómica comparativa. (36)

3.6.1.2-**BLAST**

BLAST es una herramienta de búsqueda de similitud que compara secuencias de nucleótidos o proteínas con bases de datos de secuencias caracterizadas, y calcula la importancia estadística de las coincidencias, para así poder inferir su función. (49)

3.6.1.3-**SPAdes**

SPAdes es una popular herramienta de ensamblaje de código abierto utilizada para la reconstrucción del genoma de novo a partir de lecturas de secuenciación de WGS. Admite la utilización de lecturas pareadas como de lecturas simples. (9)

3.6.1.4-**Bandage**

Bandage es una herramienta bioinformática para la visualización de gráficos de ensamblaje y para obtener información básica estructural. Trabaja en conjunto con Getorganelle, ya que este genera un gráfico de ensamblaje al terminar el proceso, para una visualización preliminar. (43)

3.6.2-**Flujo de trabajo de Getorganelle**

3.6.2.1-**Proceso estándar**

El proceso estándar inicia con la ejecución de una instrucción que activa el conjunto de herramientas las cuales incluyen Bowtie2, BLAST y SPAdes.

En primera instancia, se utiliza Bowtie2 para asociar las secuencias totales con una base de datos semilla (referencia). Esta semilla puede consistir en el genoma completo de un organelo de referencia o fragmentos del mismo. Las secuencias asociadas con las semillas, se ensamblan de manera básica y se tratan como una nueva semilla, actuando como "cebos" para el reclutamiento (pesca) de lecturas. Este proceso se repite de forma iterativa el número solicitado de veces necesario para cada proyecto. Estas iteraciones son fundamentales, ya que refuerzan la información con más secuencias, para un correcto ensamblaje futuro (ilustración 13, flecha 2). (7)

A continuación, se empleará SPAdes, que utiliza las lecturas mapeadas por bowtie2 para generar el ensamble de novo del cloroplasto. Este ensamblaje comprenderá tanto las regiones correctamente ensambladas del organelo, como otras regiones no deseadas, que incluyen el DNA nuclear y mitocondrial (ilustración 13, flecha 3).

Posteriormente, se empleará BLAST, el cual utiliza una base de datos de genes de cloroplastos para realizar una búsqueda de similitud en las secuencias ensambladas. Las secuencias que muestran similitud significativa se conservan, mientras que aquellas que no están relacionados se eliminan (ilustración 13, flecha 4). Generando una anotación básica que incluye el número de bases en cada región del cloroplasto (SCL, SSC y IR). Luego se procederá con el análisis del resultado del ensamblaje (ilustración 13, flecha 5). Dentro de los resultados, Getorganelle proporciona un archivo para la visualización preliminar del grafico de ensamblaje con el programa Bandage. En caso de un correcto ensamblaje el procedimiento concluye, en cambio, ante un ensamblaje incorrecto, se procede a utilizar otra ruta de Getorganelle. (7)

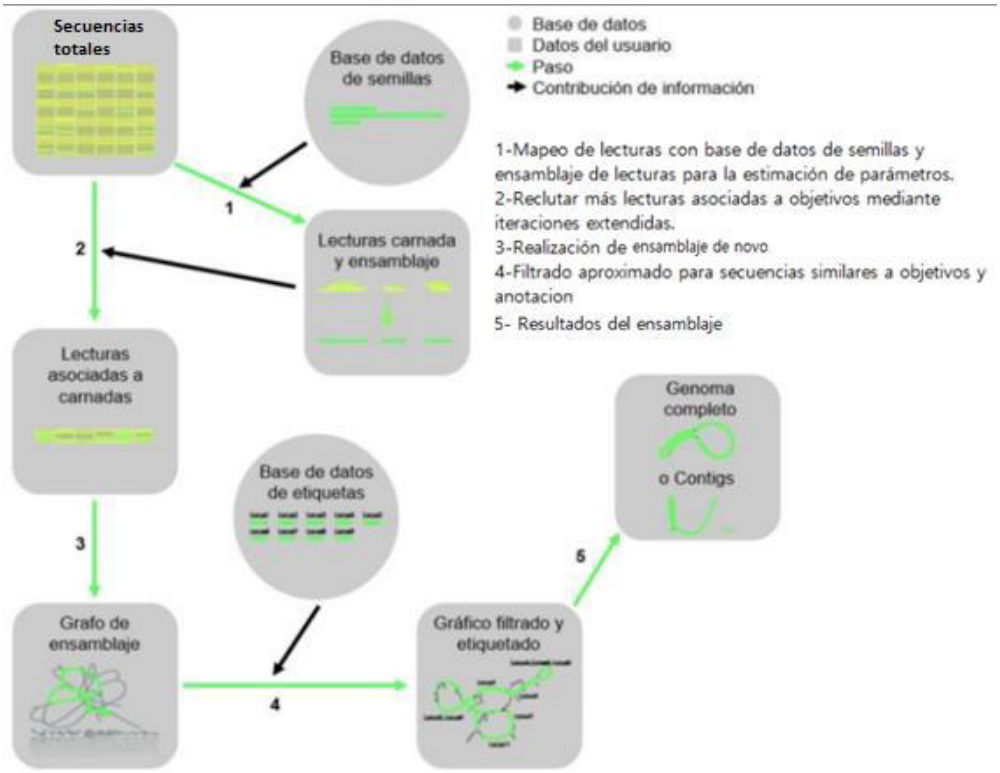

Ilustración 13- Flujo de trabajo básico de Getorganelle. Fuente: Getorganelle a fast and versatile toolkit

3.6.2.2-**Segunda ruta**

La segunda ruta consta de dos partes:

- La primera parte comienza a partir del ensamblaje crudo realizado por SPAdes (ilustración 14, flecha 1). Se utiliza un nuevo comando con la función de extraer el genoma del cloroplasto del gráfico de ensamblaje, utilizando un genoma de referencia proporcionado por el programa, al mismo tiempo eliminando los genomas restantes, y anotando de forma básica con BLAST (ilustración 14, flecha 2). Si el ensamblaje es satisfactorio, el procedimiento concluye, por el contrario, si no es correcto, se prosigue con la segunda parte (ilustración 14, flecha 4). (7)
- La segunda parte se lleva a cabo mediante un nuevo comando, el cual utiliza el resultado anterior para unir los fragmentos de secuencias, usando de referencia un genoma de cloroplasto aportado por el usuario. (Ilustración 14, flecha 5). El resultado puede ser satisfactorio o insatisfactorio. (7)

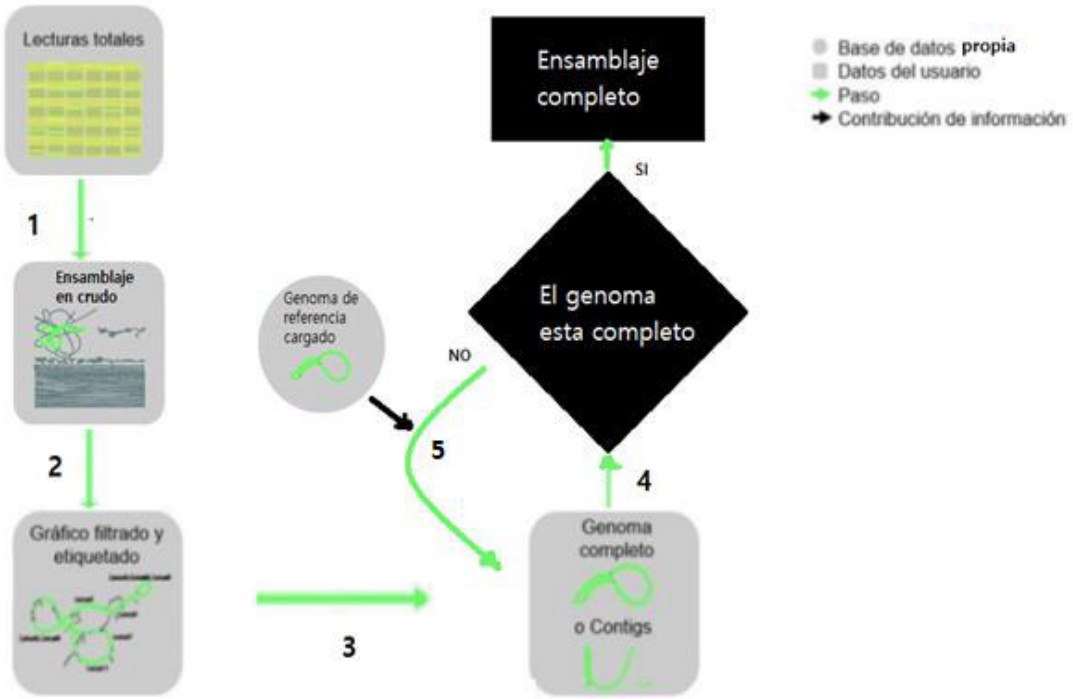

Ilustración 14- Flujo de trabajo secundario. Fuente: Getorganelle a fast and versatile toolkit

3.6.2-**NOVOPlasty**

NOVOPlasty es una herramienta de ensamblaje que se basa en la utilización inicial de una semilla (genoma de referencia). La semilla puede ser una secuencia de lectura, un gen conservado o incluso un genoma completo de un organelo de una especie distante. Previamente a la utilización del ensamblador, el archivo WGS debe ser pretratado, ya que el ensamblador no lo puede utilizar en su estado completo. (15)

A diferencia de Getorganalle, que posee el conjunto de herramientas necesarias para el ensamblaje incorporado, Novoplasty no cuenta con un conjunto integrado y tampoco trabaja en conjunto con Bandage. Es necesario armar un flujo de trabajo específico; para este trabajo se utilizarán Bowtie2 y Samtools. (15)

3.6.2.1-**Samtools**

Samtools es un conjunto de herramientas diseñada para interactuar con los datos de secuenciación. Ofrece funcionalidades que permiten la lectura, escritura, edición, indexación y visualización de las secuencias. (6)

3.6.3-**Flujo de trabajo de Novoplasty**

Antes de utilizar Novoplasty, el objetivo es obtener las secuencias del organelo en formato de secuencia pareada (PE) a partir del conjunto de lecturas de genoma completo. Se emplea Bowtie2 para mapear las lecturas utilizando una semilla con una relación cercana al organelo a ensamblar.

Posteriormente se utilizará Samtools en una serie de procesos, que abarcan desde la limpieza del mapeado para eliminar las secuencias no relacionadas con la semilla de referencia, hasta la preparación de las lecturas pareadas del genoma del cloroplasto para su uso en NOVOPlasty.

Después del pretratamiento, se avanza hacia el ensamblaje. Inicialmente, cada lectura se almacena en tablas hash (memorias) con una identificación única, permitiendo un acceso rápido a las mismas (ilustración 15, A).

El proceso de ensamblaje no se inicia directamente con la secuencia semilla. En cambio, esta semilla recluta una secuencia altamente afín dentro de todas las lecturas. La lectura reclutada se puede considerar como la primera secuencia ensamblada, y a partir de ella, el ensamblaje se extiende de manera iterativa en ambas direcciones. Para una mejor comprensión, en la ilustración 15 se representa en un solo sentido. (15)

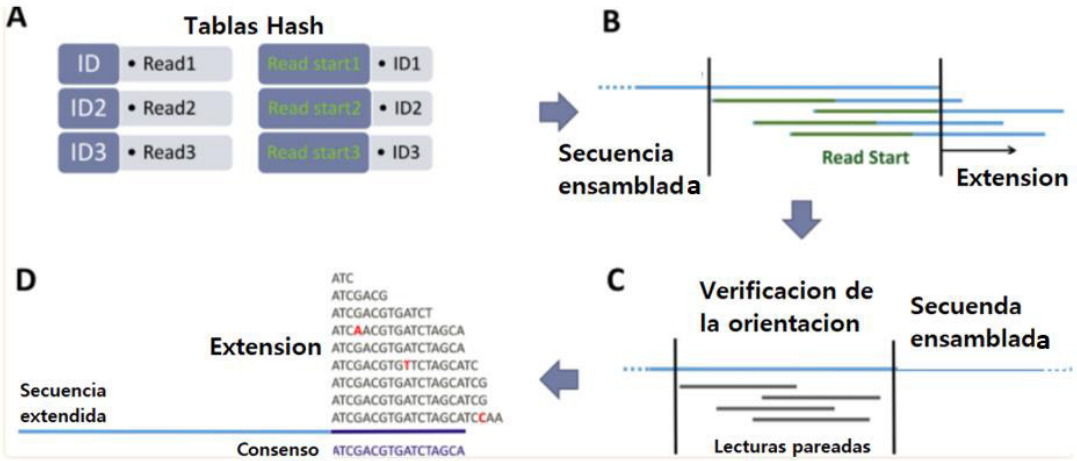

Ilustración 15- Flujo de trabajo NOVOPlasty. Fuente: Novoplasty orgánulos a partir de datos de genoma completo

Se agrupan las lecturas que son relativamente similares a la secuencia reclutada, y se almacena la información de la secuencia de inicio de lectura en tablas hash para su posterior utilización (Read Start) (Ilustración 15 B).

Antes de proceder con la nueva extensión, se verifica la correcta posición del ensamblaje anterior utilizando las lecturas pareadas. (Ilustración 15 C)

Una vez realizada la verificación y en caso de ser exitosa, se continua con la extensión. Se reclutan nuevamente las secuencias similares y se organizan rápidamente según sus inicios de lectura guardados. Éstas se alinean y, si se observa alguna diferencia en algún nucleótido, se resuelve mediante consenso entre las lecturas superpuestas (Ilustración 15 D). En caso de que haya la misma cantidad de diferencias, se generaran dos ensamblajes, uno para cada opción de consenso. A diferencia de la mayoría de los ensambladores, NOVOPlasty no busca ensamblar cada lectura individual; en cambio, extiende el ensamblaje hasta que se forme el genoma circular. La circularización ocurre cuando ambos extremos se superponen en al menos 200 pares de bases. Si se detecta una región repetitiva, la circularización se pospondrá hasta que el conjunto salga de dicha región. (15)

3.7- **Anotación genómica**

La anotación de un genoma se refiere a la tarea de identificar y asignar funciones a los distintos elementos presentes en la secuencia genética de un organismo.

Este proceso puede dividirse conceptualmente en dos etapas: anotación estructural y anotación funcional. La anotación estructural se centra en encontrar genes y sitios con relevancia biológica, determinando sus posiciones en el genoma y su estructura, como regiones intrónicas y sitios de splicing. En contra-parte, la anotación funcional busca determinar las funciones y toda la información biológica relevante de cada uno de los sitios anotados en la primera etapa.

En este trabajo, se emplearon las herramientas GeSeq para la anotación, y OGDRAW para la visualización de los resultados. Ambas herramientas están disponibles en el servidor de Chlorobox.

3.7.1-**GeSeq**

GeSeq es una herramienta de anotación online diseñada para la anotación de secuencias de organelos de plantas. Sin embargo, con un conjunto de referencia apropiado, también puede anotar genomas de mitocondrias, como la de los mamíferos. Utiliza los resultados de los ensamblajes y métodos de anotación basados en homología, donde compara las secuencias ensambladas con un genoma de cloroplasto de referencia proporcionado por la biblioteca del NCBI o cargado por el mismo usuario. Estos métodos parten del supuesto

de que secuencias muy similares deben desempeñar las mismas funciones en especies distintas, ya que provienen de un ancestro común y han sido sometidas a los mismos procesos evolutivos. La efectividad de la anotación aumenta a medida que el genoma de referencia tenga una estrecha relación con el cloroplasto ensamblado. (44)

3.7.2-**OGDRAW**

Es una herramienta ampliamente utilizada para dibujar mapas de gráficos de genomas de organelos. Antes de su creación los mapas a menudo se dibujaban manualmente, carecían de un diseño homogéneo y eran inconsistentes en la visualización de las características. (42)

3.8- **Alineación de secuencias**

Obtener información de la similitud o diferencia entre dos secuencias es importante para inferir relaciones estructurales, funcionales o evolutivas. En este contexto, se emplea la herramienta de alineación online VectorBuilder, que permite la comparación entre dos secuencias, ya sea a nivel de ADN o proteínas, además de brindar información de la presencia de gaps. (51)

4-**Flujo de trabajo y resultados**

En primer lugar, se seleccionaron de manera aleatoria 12 secuenciaciones de genoma completo (WGS) de Cannabis Sativa de la biblioteca del NCBI, que cumplan con las siguientes características:

- Estar secuenciado con el método de ilumina.
- Seguir una estrategia de genoma completo
- El diseño debe ser de lecturas pareadas.

En la tabla 1 se muestra el número de serie del NCBI y el tamaño de cada archivo.

Número de serie NCBI	Tamaño del archivo
SRR20856300	4,2GB
SRR20856307	2,2GB
SRR24187772	2,7GB
SRR24187773	2,4GB
SRR24187777	3Gb
SRR24187779	2,9GB
SRR24187783	2,5GB
SRR24187784	2,5GB
SRR20856281	3,5GB
SRR20856286	2,6GB
SRR20856296	1,3GB
SRR20856258	1.6Gb

Tabla 1- Numero se serie y tamaño de las secuencias WGS

En todos los casos se procedió al ensamblaje del cloroplasto, primero con Getorganelle y después con NOVOPlasty. A los 12 WGS se les aplico el método estándar de Getorganelle, y se verifico su eventual correcto ensamblaje evaluando el grafico preliminar mediante Bandage. De todos los ensamblajes realizados con el primer método de Getorganelle, solamente tres no dieron resultados satisfactorios, por lo que para estos casos se procedió al segundo método disponible para este ensamblador.

Un ensamblaje correcto de cloroplasto se observa en la ilustración 16-A, correspondiente al genoma SRR20856300, el cual forma una estructura de dos bucles unidos. Por otro

lado, para el genoma SRR20856258 en la ilustración 16-B, se observa una considerable cantidad de secuencias separadas, indicando un ensamblaje insatisfactorio. Todos los ensamblajes correctos de los cloroplastos comparten la misma estructura de dos bucles unidos. Se ilustra una como ejemplo.

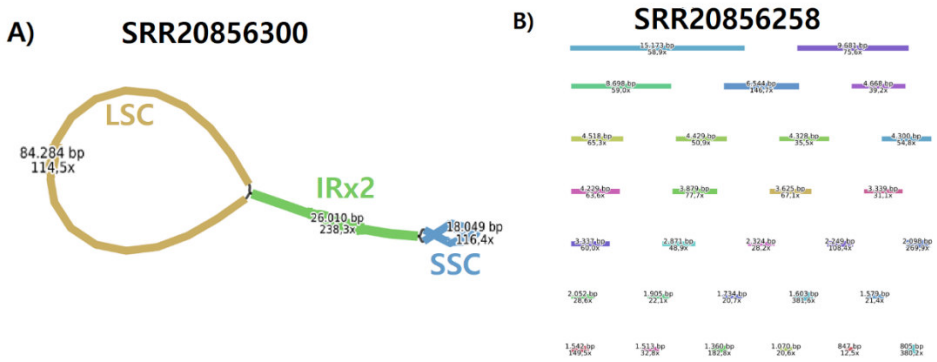

Ilustración 16- Visualización del ensamblaje preliminar utilizando Bandage. A) Ensamblaje correcto B) Ensamblaje incorrecto

A partir de los ensamblajes crudos realizados por SPAdes sobre estos tres genomas incorrectamente ensamblados por la primera estrategia de Getorganelle, se llevó a cabo la primera parte del segundo método, que consistió en la extracción del genoma del cloroplasto del gráfico de ensamblaje. Se prosiguió a su visualización preliminar con Bandage, obteniendo resultados insatisfactorios en los tres casos.

Se realizó la segunda parte del segundo método, con el uso del resultado previo y un cloroplasto de cannabis sativa de referencia obtenido del NCBI, el cual sirvió como guía para las uniones de las secuencias.

En la ilustración 17 se puede observar el genoma SRR20856286, desde el ensamblaje en crudo hasta la conclusión del segundo método, para el cual Getorganelle no logró un ensamblaje satisfactorio.

En el genoma SRR20856296, también se observó que el ensamblaje no fue satisfactorio; sin embargo, a diferencia del caso anterior, se pudieron visualizar fragmentos más extendidos de secuencia de ADN, tal como se muestra en la ilustración 18.

Por último, como se observa en la ilustración 19, para el genoma SRR20856258 se alcanzó un ensamblaje satisfactorio al visualizar la estructura de dos bucles unidos. Es decir, se observa el genoma del cloroplasto. Tambien se encuentran restos de fragmentos de ADN que pueden ser eliminados.

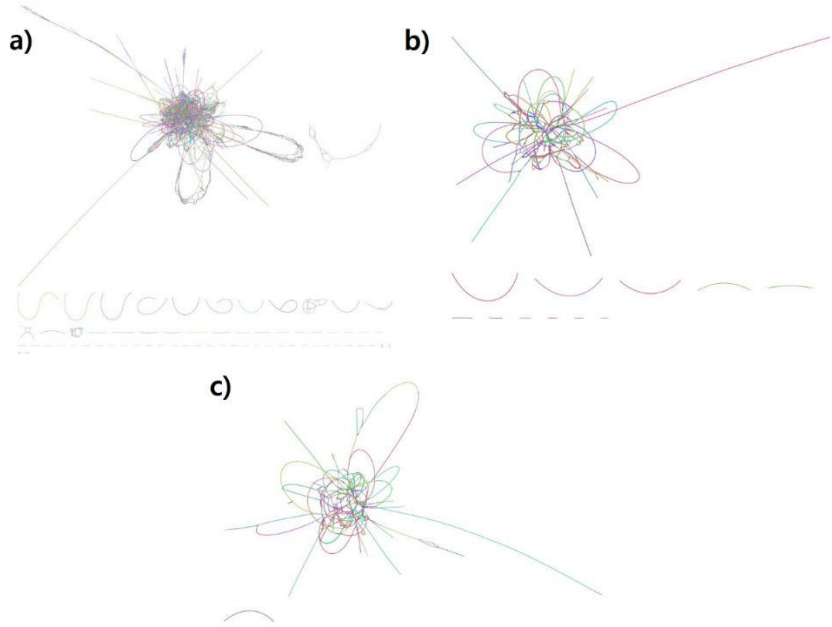

Ilustración 17- Utilización del segundo método en SRR20855686- A) Ensamblaje Crudo. B) Luego de la primera parte del método. C) Luego de la segunda parte del método

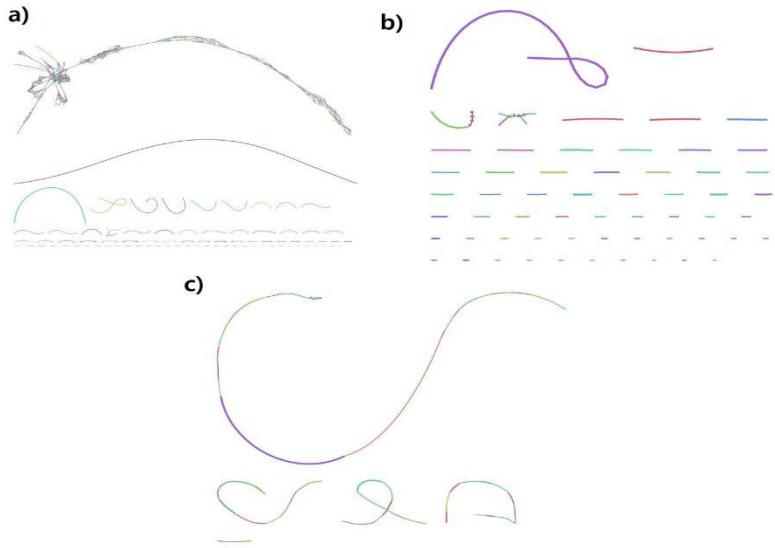

Ilustración 18- Utilización del segundo método en SRR20856296. A) Ensamblaje Crudo. B) Luego de la primera parte del método. C) Luego de la segunda parte del método

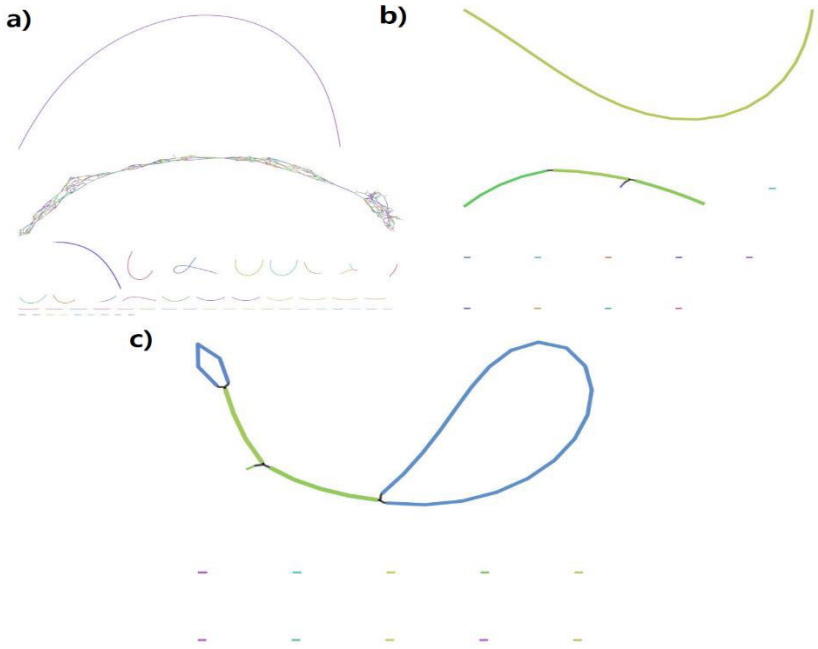

Ilustración 17- Utilización del segundo método en SRR20856258. A) Ensamblaje Crudo. B) Luego de la primera parte del método. C) Luego de la segunda parte del método

De cada grafico preliminar donde el ensamblaje se realizó correctamente, se recopilaron datos sobre el tamaño de los pares de bases de cada fragmento del cloroplasto (tabla 3). Se observa una diferencia en el tamaño del cloroplasto SRR20856300 en sus regiones LSC y SSC en comparación con los demás.

Paralelamente, se procedió a realizar los ensamblajes de los 12 cloroplastos de las muestras mediante el flujo de trabajo personalizado de NOVOPlasty (ilustración 20).

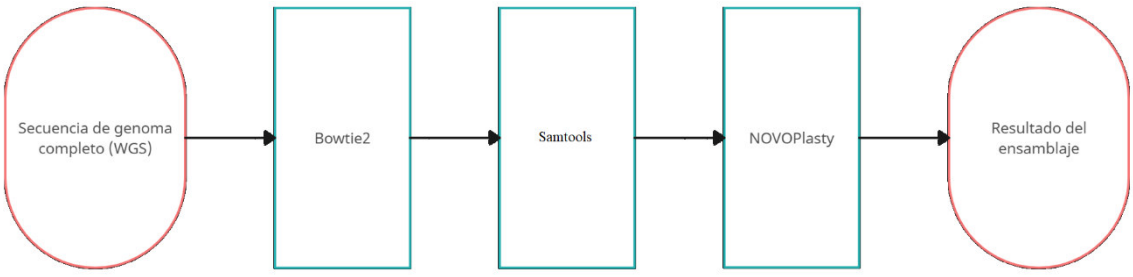

Ilustración 20- Flujo de trabajo NOVOPlasty

El proceso comenzó con el uso de Bowtie2, que mapeo los fragmentos de secuencia del genoma completo con un cloroplasto de cannabis sativa de referencia descargado del banco de datos del NCBI. Al finalizar el proceso, se obtuvo un archivo en formato SAM que contenía tanto las secuencias mapeadas como el resto del genoma de la célula.

Posteriormente, se utilizó Samtools para llevar a cabo varios procedimientos, iniciando con la eliminación de las secuencias no mapeadas y preparando correctamente el archivo para su utilización por NOVOPlasty.

Antes de proseguir con el ensamblaje, se indicó el cloroplasto de referencia a utilizar y el tamaño de cada fragmento secuenciado, esta información ayuda a mejorar el proceso. Se realizó el ensamblaje del cloroplasto de los doce genomas.

Los resultados de todos los ensamblajes de cloroplastos para cada WGS se presentan en la tabla 2 y el porcentaje en la ilustración 21.

Número de serie NCBI	Ensamblaje			
	Getorganelle			NOVOPlasty
	Getotganelle básico	Segundo método - primera parte	Segundo método - segunda parte	
SRR20856300	Logrado	-	-	Logrado
SRR20856307	Logrado	-	-	Logrado
SRR24187772	Logrado	-	-	Logrado
SRR24187773	Logrado	-	-	Logrado
SRR24187777	Logrado	-	-	Logrado
SRR24187779	Logrado	-	-	Logrado
SRR24187783	Logrado	-	-	Logrado
SRR24187784	Logrado	-	-	Logrado
SRR20856281	Logrado	-	-	Logrado
SRR20856286	Fallo	Fallo	Fallo	Logrado
SRR20856296	Fallo	Fallo	Fallo	Logrado
SRR20856258	Fallo	Fallo	Logrado	Fallo

Tabla 2- Resultados de los ensamblajes realizados por Getorganelle y NOVOPlasty

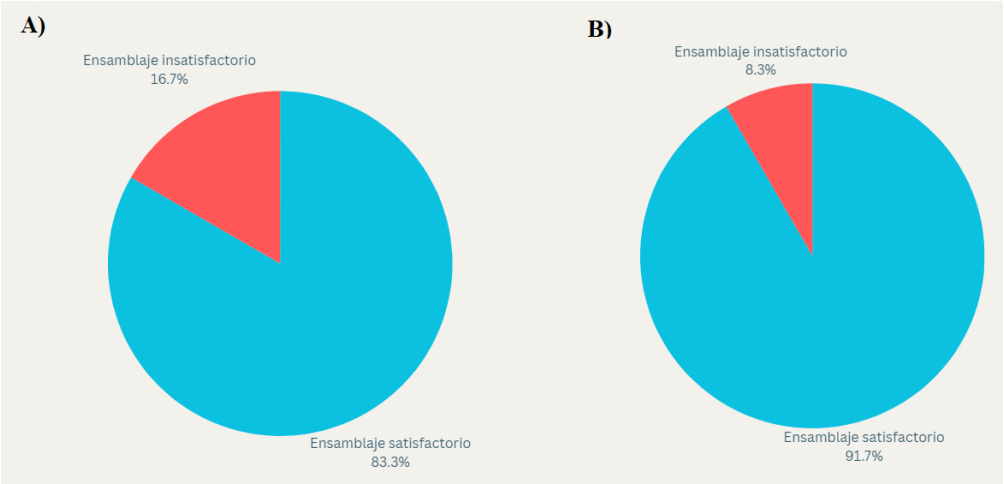

Ilustración 21- Porcentaje de ensamblaje de los 12 WGS. A) Getorganelle B) NOVOPlasty

Número de serie ncbi	Tamaño de bases estimado con Bandage		
	Región larga de copia única (LSC)	Región corta de copia única (SSC)	Región repetida invertida x2 (IR)
SRR20856300	84284	18049	26010
SRR20856307	84277	18031	26010
SRR24187772	84277	18031	26010
SRR24187773	84277	18031	26010
SRR24187777	84277	18031	26010
SRR24187779	84227	18031	26010
SRR24187783	84277	18031	26010
SRR24187784	84277	18031	26010
SRR20856281	84277	18031	26010
SRR20856286	-	-	-
SRR20856296	-	-	-
SRR20856258	84277	18031	26010

Tabla 3- Numero de bases estimado de cada parte del Cloroplasto de los WGS utilizados.

Se realizó la anotación de los cloroplastos ensamblados correctamente con la herramienta GeSeq y se visualizó con OGDRAW. En la ilustración 22, generada del ensamblaje de SRR20856300, se pueden observar todos los genes del cloroplasto, y la estructura cuatripartida del mismo. Es importante destacar que este gráfico se mantiene igual tanto para el ensamblaje realizado con Getorganelle como para el realizado por NOVOPlasty.

En todas las anotaciones de las secuencias ensambladas para todos los cloroplastos estudiados se observaron las mismas estructuras y distribución de genes, por este motivo se ilustra solamente uno como ejemplo.

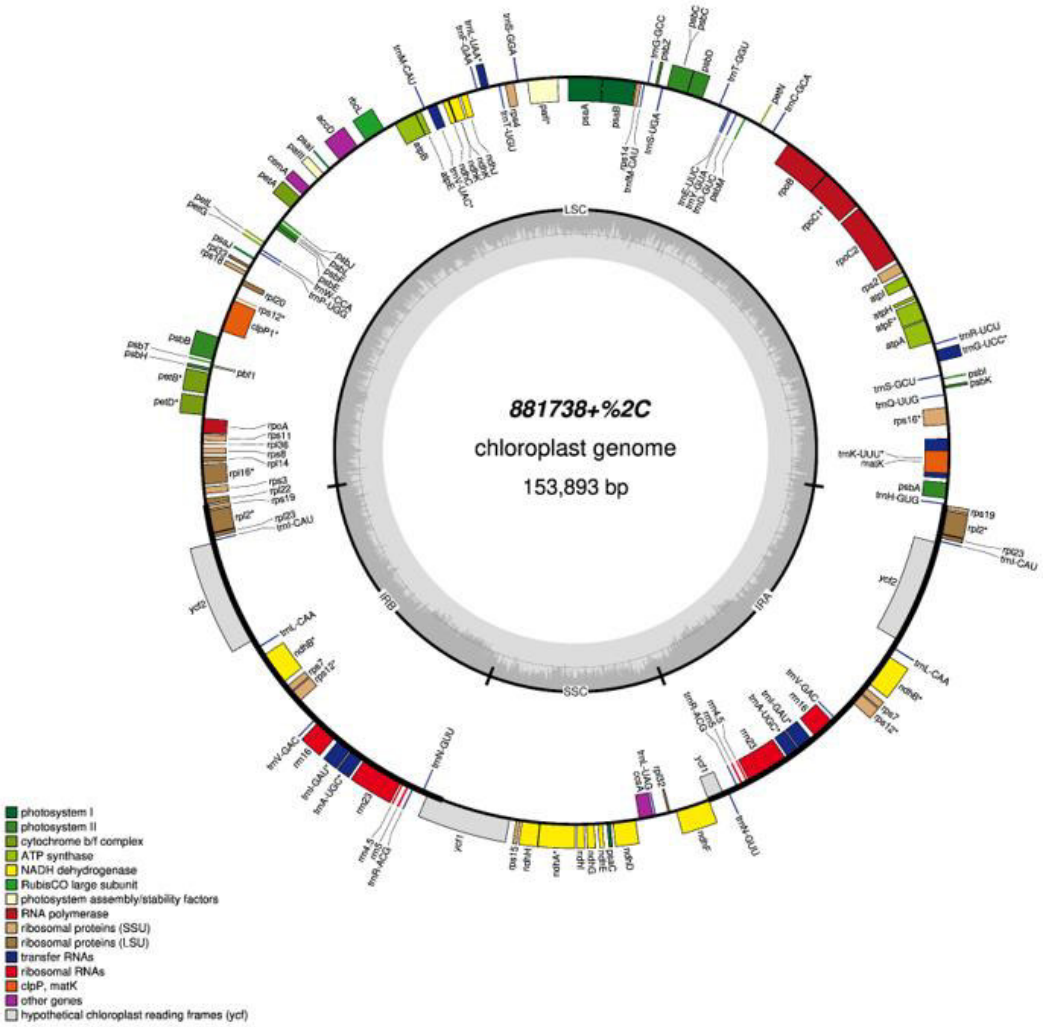

Ilustración 22- Mapa del genoma completo del cloroplasto de cannabis sativa SRR20856300. Los genes representados en el interior del círculo se transcriben en sentido de las agujas del reloj, mientras que los representados en el exterior se transcriben en sentido contrario. En el centro del círculo se representa la variación en el contenido en G+C del genoma.

En las anotaciones, se pueden identificar los genes utilizados en los estudios filogenomicos, como el gen rbcL (ilustracion 23) y el gen matK (ilustracion 24).

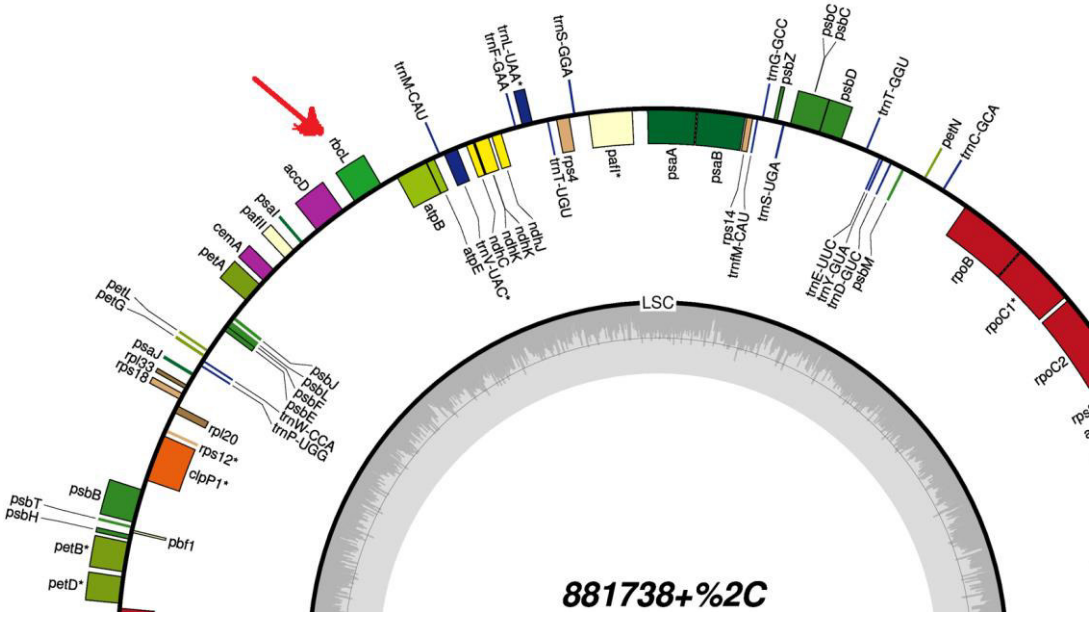

Ilustración 23- Localización del gen rbcL en la anotación del genoma

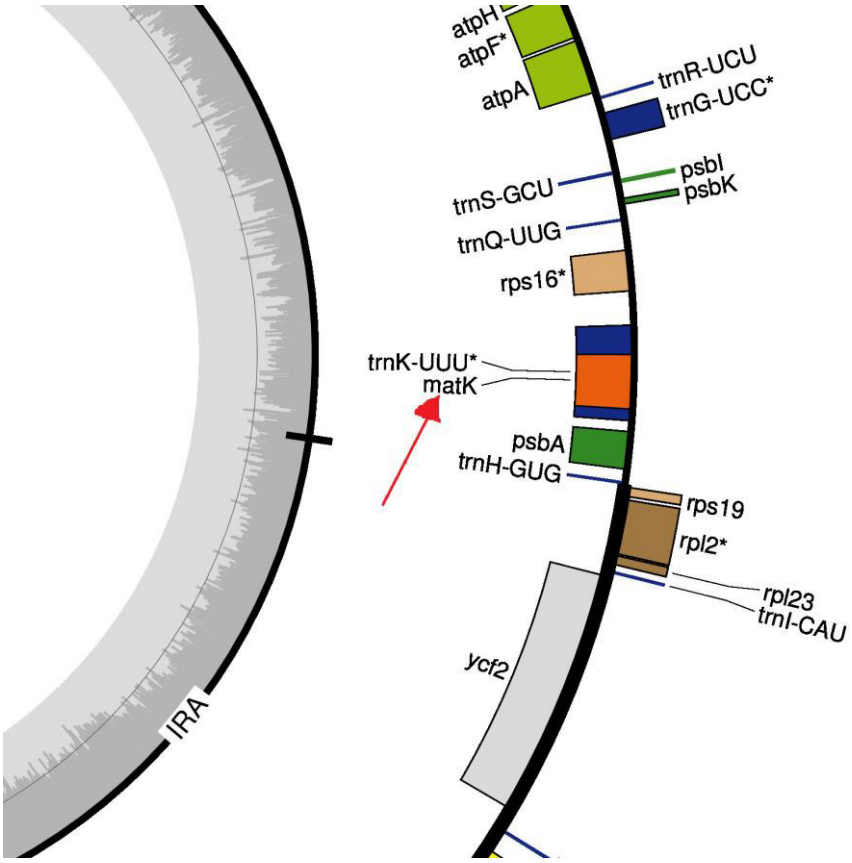

Ilustración 24- Localización del gen matK en la anotación del genoma

La herramienta GeSeq proporciona la oportunidad de visualizar la secuencia de nucleótidos utilizada para la anotación de cada gen. Aprovechando esta capacidad, buscamos y obtenemos las secuencias de los genes matK y rbcL de cada uno del cloroplasto, con el fin de realizar un análisis comparativo.

Se inicio la comparación de las secuencias entre el cloroplasto SRR24187772 y SRR20856307, realizando una alineación con la página VectorBuilder para determinar la similitud entre los genes matK, encontrándose idénticos. (ilustración 27). En cambio, en la comparación de las secuencias entre los cloroplastos SRR20856300 y SRR24187772, se observó una diferencia en el gen matK del cloroplasto SRR20856300. (ilustración 28).

Se realizo el mismo procedimiento para el gen rbcL de los tres cloroplastos, resultando ser todas las alineaciones iguales (ilustración 29 y 30).

De los doce cloroplastos analizados, se constató que en once presentaban secuencias de nucleótidos idénticas para el gen matK, y todas iguales para gen rbcL. Para facilitar la exposición, se optó por ilustrar la comparación de tres secuencias de cloroplastos (Ilustración 25 y 26): el diferente en su secuencia matK y otros dos idénticos, abarcando todas las posibilidades comparativas.

Ilustración 25- Secuencia del gen matK. A) SRR20856300. B) SRR24187772. C) SRR20856307.

Ilustración 26- Secuencia del gen rbcL A) SRR20856300 B) SRR24187772 C) SRR20856307

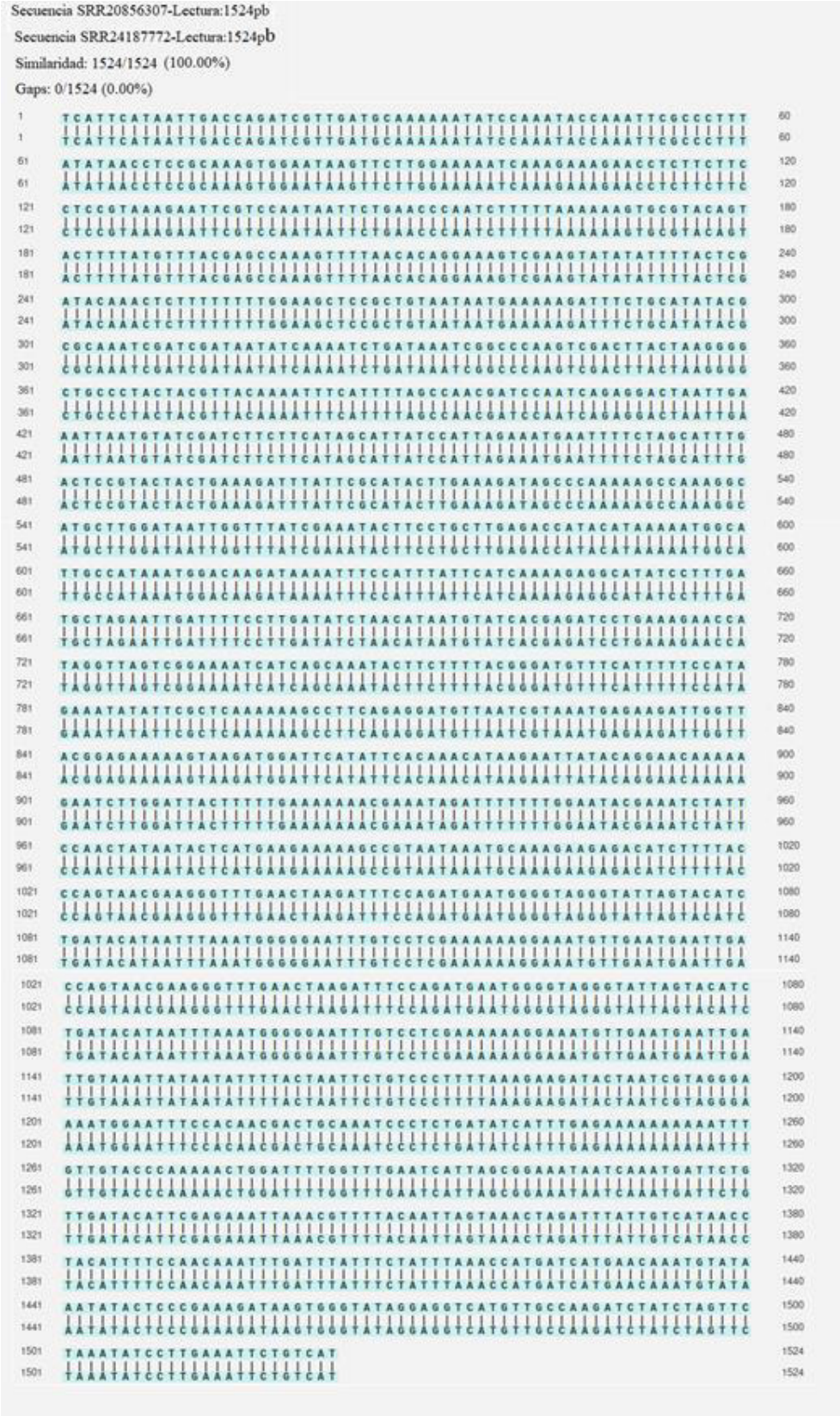

Ilustración 27- Comparación del gen matK del cloroplasto SRR20856307 con el cloroplasto SRR24187772

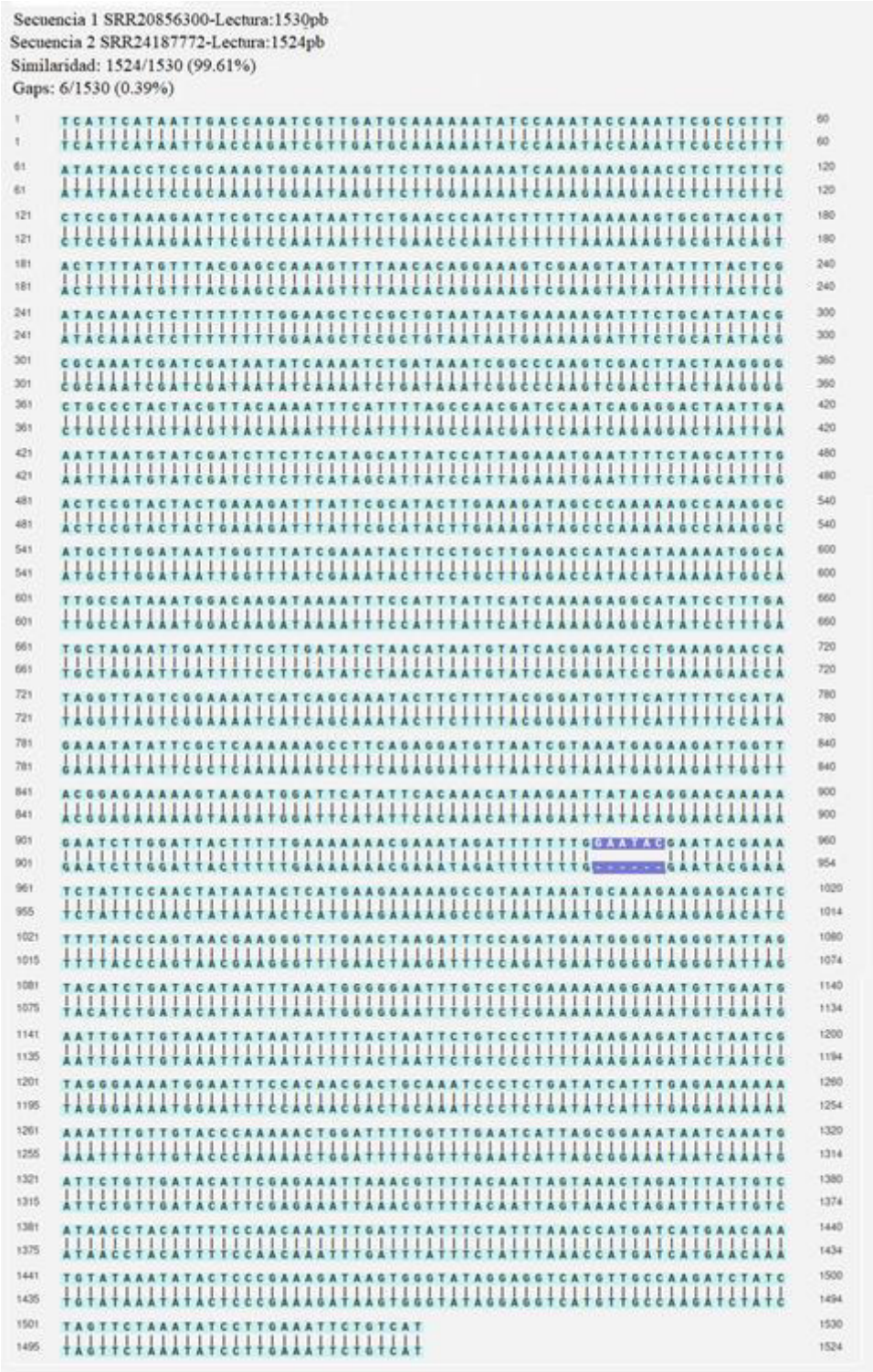

Ilustración 28- Comparación del gen matK del cloroplasto SRR208556300 con el cloroplasto SRR24187772

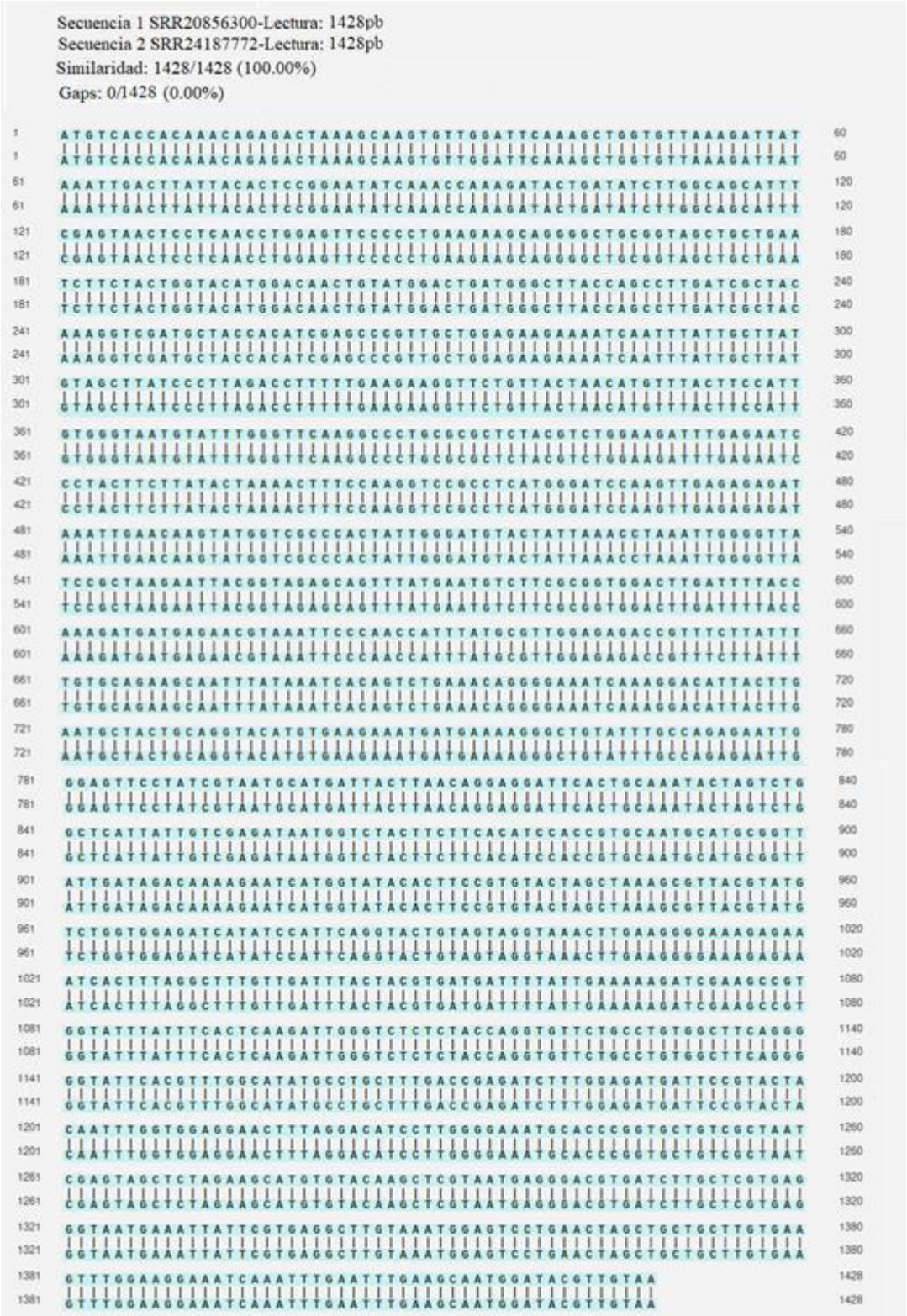

Ilustración 29- Comparación del gen rbcL del cloroplasto SRR20856307 con el cloroplasto SRR24187772

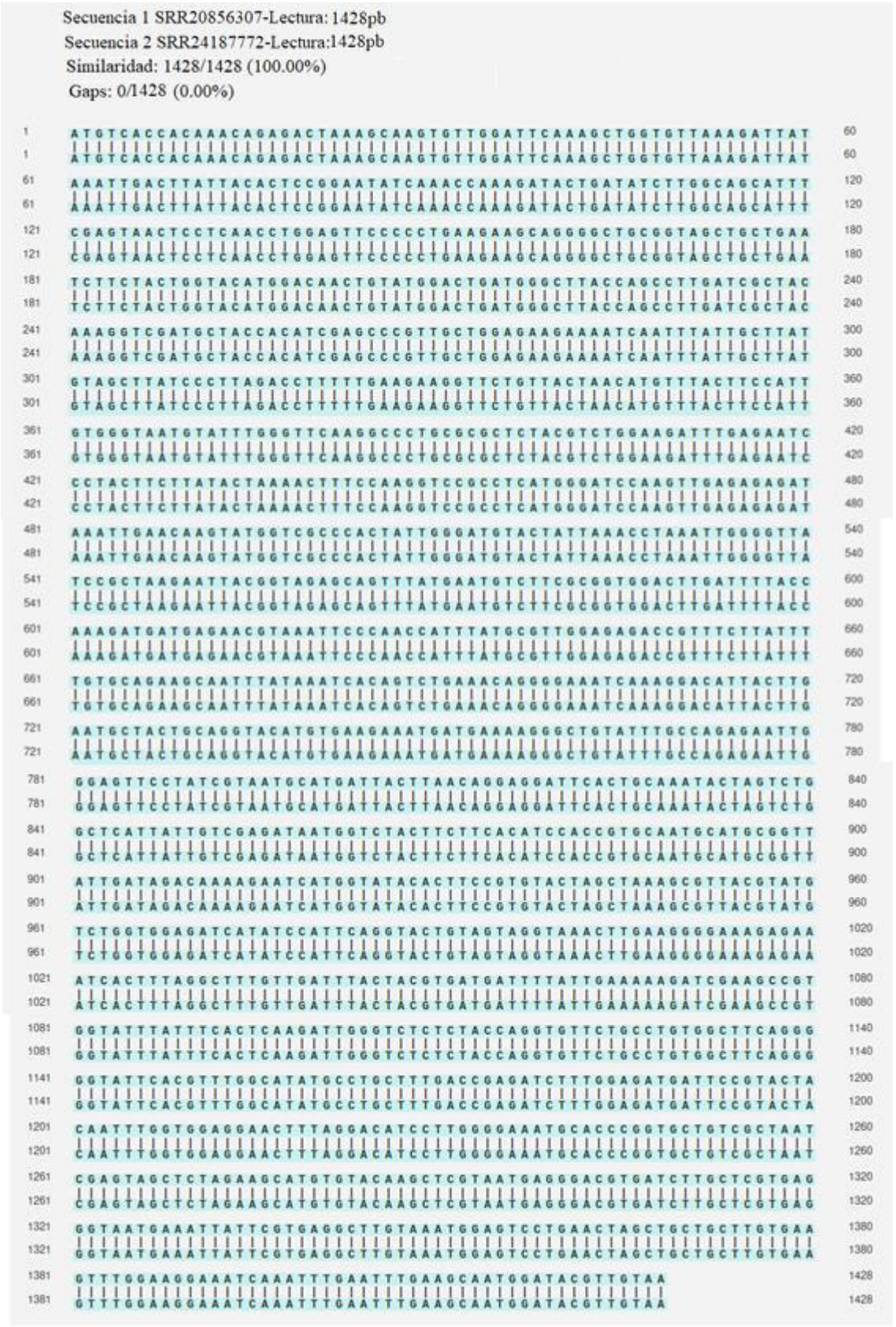

Ilustración 30- Comparación del gen rbcL del cloroplasto SRR208556300 con el cloroplasto SRR24187772

5-**Conclusión**

El conocimiento sobre el manejo de programas bioinformáticos de ensamblaje, anotación y alineamiento, es una herramienta de gran relevancia para el estudio de genomas.

Los resultados indican que NOVOPlasty tuvo, para la muestra reducida abordada en el presente estudio, una tasa mayor de éxito en el ensamblaje que Getorganelle, pudiendo ensamblar 11 de los 12 cloroplastos, mientras que Getorganelle solamente 10. Sin embargo, cabe destacar que el único cloroplasto no ensamblado por NOVOPlasty, fue ensamblado correctamente por Getorganelle. Podemos inferir por lo tanto que es necesario en la actualidad disponer del uso de los principales ensambladores disponibles a fin de contar con una mayor posibilidad de un proceso exitoso.

El cloroplasto del genoma de Cannabis Sativa SRR20856300, presenta un tamaño mayor en sus regiones LSC y SSC, además de contar con una variación en el gen matk en comparación con los otros cloroplastos. Esto podría ser un indicio de que esta cepa sea el pariente más lejano del subgrupo bajo estudio, aunque esta afirmación no es concluyente y requiere de un estudio filogenómico más detallado y extenso.

A partir de la experticia obtenida en el presente trabajo, se propone proyectar un subsiguiente estudio de la diferenciación del cannabis con el fin de realizar un posterior análisis filogenómico, con la expectativa de aportar a la eventual diferenciación entre cepas de cannabis con mayores propiedades de uso médico de otras con fines recreativos.

6-**Fuentes de información y referencia**

1) León Cam, Juan José. (2017). El aceite de Cannabis. *Revista de la Sociedad Química del Perú*, *83*(3), 261-263.
2) Atakan Z. (2012). Cannabis, a complex plant: different compounds and different effects on individuals. *Therapeutic advances in psychopharmacology*, *2*(6),
3) *Cannabis medicinal*. (2019, julio 12). Argentina.gob.ar.
4) Gonzalez-Cuevas, G., Martin-Fardon, R., Kerr, T. M., Stouffer, D. G., Parsons, L. H., Hammell, D. C., Banks, S. L., Stinchcomb, A. L., & Weiss, F. (2018). Unique treatment potential of cannabidiol for the prevention of relapse to drug use: preclinical proof of principle. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, *43*(10), 2036–2045.
5) Blessing, E. M., Steenkamp, M. M., Manzanares, J., & Marmar, C. R. (2015). Cannabidiol as a Potential Treatment for Anxiety Disorders. *Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics*, *12*(4), 825–836.
6) *Samtools*. (s/f). Htslib.org. Recuperado el 3 de diciembre de 2023, de https://www.htslib.org/
7) Jin, JJ., Yu, WB., Yang, JB. *et al.* GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol* **21**, 241 (2020).
8) Lukhele, S. T., & Motadi, L. R. (2016). Cannabidiol rather than Cannabis sativa extracts inhibit cell growth and induce apoptosis in cervical cancer cells. *BMC complementary and alternative medicine*, *16*(1), 335.
9) BThors. (2022, diciembre 28). Using SPAdes for genome assembly: A step-by-step guide. Medium.
10) Simiyu, D. C., Jang, J. H., & Lee, O. R. (2022). Understanding *Cannabis sativa* L.: Current Status of Propagation, Use, Legalization, and Haploid-Inducer-Mediated Genetic Engineering. *Plants (Basel, Switzerland)*, *11*(9), 1236.
11) Ishida, K., Matsuda, H., Murakami, M., & Yamaguchi, K. (1997). Kawaguchipeptin B, an antibacterial cyclic undecapeptide from the cyanobacterium Microcystis aeruginosa. *Journal of natural products*, *60*(7), 724–726.

12) Abed, R.M.M., Dobretsov, S. and Sudesh, K. (2009), Applications of cyanobacteria in biotechnology. Journal of Applied Microbiology, 106: 1-12.
13) Raven, JA, Allen, JF Genómica y evolución del cloroplasto: ¿qué hicieron las cianobacterias por las plantas?. *Genoma Biol* 4 , 209 (2003).
14) Raven, J.A., Allen, J.F. Genomics and chloroplast evolution: what did cyanobacteria do for plants?. *Genome Biol* 4, 209 (2003).
15) Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic acids research*, *45*(4), e18.
16) in, JJ., Yu, WB., Yang, JB. *et al.* GetOrganelle: un conjunto de herramientas rápido y versátil para el ensamblaje preciso de novo de genomas de orgánulos. *Genoma Biol* 21 , 241 (2020).
17) AUTHOR=Wang Yu, Xu Jing, Hu Bin, Dong Chunxing, Sun Jin, Li Zixian, Ye Kangzhuo, Deng Fang, Wang Lulu, Aslam Mohammad, Lv Wenliang, Qin Yuan, Cheng Yan, Assembly, annotation, and comparative analysis of Ipomoea chloroplast genomes provide insights into the parasitic characteristics of Cuscuta species. Frontiers in Plant Science 13
18) Anabalón, L., Solano, J., Encina-Montoya, F., Bustos, M., Figueroa, A., & Gangitano, D. (2022). Cannabis Seeds Authentication by Chloroplast and Nuclear DNA Analysis Coupled with High-Resolution Melting Method for Quality Control Purposes. *Cannabis and cannabinoid research*, *7*(4), 548–556.
19) Liu, D., Cui, Y., Li, S., Bai, G., Li, Q., Zhao, Z., Liang, D., Wang, C., Wang, J., Shi, X., Chen, C., Feng, G., & Liu, Z. (2019). A New Chloroplast DNA Extraction Protocol Significantly Improves the Chloroplast Genome Sequence Quality of Foxtail Millet (Setaria italica (L.) P. Beauv.). Scientific reports, 9(1), 16227.
20) Ángeles López, Guadalupe Esther, Brindis, Fernando, Cristians Niizawa, Sol, & Ventura Martínez, Rosa. (2014). Cannabis sativa L., una planta singular. Revista mexicana de ciencias farmacéuticas, 45(4), 1-6. Recuperado en 22 de noviembre de 2023.
21) Gloss D. (2015). An Overview of Products and Bias in Research. Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics, 12(4), 731–734.

22) Yodphaka, S., Boonpragob, K., Lumbsch, H. T., & Kraichak, E. (2018). Evaluation of six regions for their potential as DNA barcodes in epiphyllous liverworts from Thailand. Applications in plant sciences, 6(8), e01174.
23) Schofs, L., Sparo, M. D., & Sánchez Bruni, S. F. (2021). The antimicrobial effect behind Cannabis sativa. Pharmacology research & perspectives, 9(2), e00761.
24) Efectos sociales y para la salud del consumo de cannabis sin fines médicos. Washington, D.C.: Organización Panamericana de la Salud; 2018. Licencia: CC BY-NC-SA 3.0 IGO.
25) Karas, J. A., Wong, L. J. M., Paulin, O. K. A., Mazeh, A. C., Hussein, M. H., Li, J., & Velkov, T. (2020). The Antimicrobial Activity of Cannabinoids. Antibiotics (Basel, Switzerland), 9(7), 406.
26) Schofs, L., Sparo, M.D. and Sánchez Bruni, S.F. (2021), The antimicrobial effect behind Cannabis sativa. Pharmacol Res Perspect, 9: e00761
27) Schofs, L., Sparo, M. D., & Sánchez Bruni, S. F. (2021). The antimicrobial effect behind Cannabis sativa. Pharmacology research & perspectives, 9(2), e00761.
28) BARRERA, L & SAHAGÚN, S. (2020). E De cannabis sin fines médicos, C. (s/f). Efectos sociales y para la salud
29) Dong, W., Liu, H., Xu, C. et al. A chloroplast genomic strategy for designing taxon specific DNA mini-barcodes: a case study on ginsengs. BMC Genet 15, 138 (2014).
30) López-Muñoz, F., González, E., Serrano, M.D., Antequera, R., & Alamo, C.. (2011). Una visión histórica de las drogas de abuso desde la perspectiva criminológica (Parte I). Cuadernos de Medicina Forense, 17(1), 21-33.
31) El sistema endocannabinoide. (s/f). Fundacion-canna.es.
32) Daniell, H., Lin, C.-S., Yu, M., & Chang, W.-J. (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biology, 17(1).
33) Kang, Y., Deng, Z., Zang, R. et al. DNA barcoding analysis and phylogenetic relationships of tree species in tropical cloud forests. Sci Rep 7, 12564 (2017).
34) Rubio, Santiago; Pacheco-Orozco, Rafael Adrián; Milena Gómez, Ana; Perdomo, Sandra & García-Robles, Reggie (2020). Secuenciación de nueva generación (NGS) de ADN: presente y futuro en la práctica clínica. Universitas Medica, 61(2).

35) Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 1–8.
36) Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 4, 357–359 (2012).
37) Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. Genomics, proteomics & bioinformatics, 13(5), 278–289
38) Aguilar-Bultet, Lisandra, & Falquet, Laurent. (2015). Secuenciación y ensamblaje de novo de genomas bacterianos: una alternativa para el estudio de nuevos patógenos. Revista de Salud Animal, 37(2), 125-132. Recuperado en 03 de diciembre de 2023
39) Albero, A. (2021, enero 20). *NCBI – National Center for Biotechnology Information*. Enfocatss.
40) Anabalón, L., Solano, J., Encina-Montoya, F., Bustos, M., Figueroa, A., & Gangitano, D. (2022). Cannabis Seeds Authentication by Chloroplast and Nuclear DNA Analysis Coupled with High-Resolution Melting Method for Quality Control Purposes. Cannabis and cannabinoid research, 7(4), 548–556.
41) Freudenthal, JA, Pfaff, S., Terhoeven, N. *et al.* Una comparación sistemática de herramientas de ensamblaje del genoma del cloroplasto. *Genoma Biol* **21** , 254 (2020).
42) Stephan Greiner, Pascal Lehwark, Ralph Bock, OrganellarGenomeDRAW (OGDRAW) versión 1.3.1: kit de herramientas ampliado para la visualización gráfica de genomas organellares, Nucleic Acids Research , volumen 47, número W1, 2 de julio de 2019, páginas W59–W64
43) Ryan R. Wick, Mark B. Schultz, Justin Zobel, Kathryn E. Holt, Bandage: visualización interactiva de ensamblajes de genoma *de novo , Bioinformática* , volumen 31, número 20, octubre de 2015, páginas 3350–3352
44) Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq - versatile and accurate annotation of organelle genomes. *Nucleic acids research*, *45*(W1), W6–W11.
45) Hernández, M., Quijada, N. M., Rodríguez-Lázaro, D., & Eiros, J. M. (2020). Aplicación de la secuenciación masiva y la bioinformática al diagnóstico microbiológico clínico. *Revista Argentina de microbiologia*, *52*(2), 150–161.
46) Crossley, B. M., Bai, J., Glaser, A., Maes, R., Porter, E., Killian, M. L., Clement, T., & Toohey-Kurth, K. (2020). Guidelines for Sanger sequencing and molecular

assay monitoring. *Journal of veterinary diagnostic investigation : official publication of the American Association of Veterinary Laboratory Diagnosticians, Inc*, *32*(6), 767–775.

47) D. (2003). Chloroplast research in the genomic age. Trends Genet. 19: 47-56.

48) Jiao, Y., & Guo, H. (2014). Prehistory of the angiosperms. En A. H. Paterson (Ed.), *Genomes of Herbaceous Land Plants* (Vol. 69, pp. 223–245). Elsevier.

49) *Basic Local Alignment Search Tool*. (s/f). Nih.gov. Recuperado el 4 de diciembre de 2023

50) . Illumina.com. Technology spotlight: Illumina Sequencing [Sede Web]. Illumina. 2010

51) *Sequence alignment*. (s/f). Vectorbuilder.com. Recuperado el 4 de diciembre de 2023