UNIVERSIDAD
NACIONAL
DE COLOMBIA

Ensamblaje del genoma de *Leuconostoc mesenteroides* IBUN 91.2.98. por medio de herramientas bioinformáticas

César Augusto Núñez Campos

Universidad Nacional de Colombia

Facultad de ciencias

Maestría en Ciencias Químicas-Modalidad Profundización

Bogotá, Colombia

2022

Ensamblaje del genoma de *Leuconostoc mesenteroides* IBUN 91.2.98. por medio de herramientas bioinformáticas

César Augusto Núñez Campos

Trabajo final presentado como requisito para optar al título de:

Magister en Ciencias - Química

Director (a):

Sonia Amparo Ospina Sánchez

QF. MSc. PhD. En biotecnología

Instituto de Biotecnología

Grupo de investigación de Biopolímeros y Biofuncionales

Universidad Nacional de Colombia

Facultad de Ciencias

Maestría en Ciencias Químicas-Modalidad Profundización

Bogotá, Colombia

2022

Este trabajo lo dedico a mi madre por su gran labor.

Agradecimientos

Debo agradecer a las personas que me han acompañado durante mi formación académica. Especialmente a la Doctora Sonia Amparo Ospina Sánchez, por su apoyo, permanente apoyo e infinita paciencia durante el proceso de elaboración del presente proyecto de profundización.

Debo también agradecer al Instituto de Biotecnología de la Universidad Nacional y al Grupo de investigación de Biopolímeros y Biofuncionales, que me permitieron profundizar en temas que no manejaba, hasta que inicie este proceso de profundización en el tema de ensamblaje y uso de herramientas bioinformáticas.

Finalmente, quiero agradecer a mi familia que siempre me apoyo y estuvo atenta al desarrollo de mi postgrado.

Resumen

Titulo: Ensamblaje del genoma de *Leuconostoc mesenteroides* IBUN 91.2.98. por medio de herramientas bioinformáticas.

A pesar de la importancia del dextrano en varias aplicaciones industriales y la necesidad de masificar su producción, no se tiene muy documentado el mecanismo de expresión y regulación de las dextransacarasas en cepas productoras como el *Leuconostoc mesenteroides* cepa IBUN 91.2.98. El desarrollo de métodos de secuenciación de segunda generación y el uso de herramientas bioinformáticas permitirán secuenciar, ensamblar y evaluar genomas completos a un costo relativamente bajo guiados por un genoma de referencia de *L. mesenteroides subsp. mesenteroides* ATCC 8293, para ayudar al proceso de ensamblaje.

La secuenciación produjo un total de 1.47 Gb de datos crudos, que después del trimming y control de calidad generaron 1,40 Gb (7.78 X de profundidad) que se usaron para el ensamblaje. Se obtuvo un ensamblaje para el *Leuconostoc mesenteroides* cepa IBUN 91.2.98 de una longitud de 2,064 Mpb. La longitud del ensamblaje represento el 85% del tamaño estimado del genoma de referencia de *L. mesenteroides subsp. mesenteroides* ATCC 8293.

Palabras claves: reads, contigs, ensamblaje, tecnologías de secuenciación, traducción.

Abstract

Title: Assembly of the genome of *Leuconostoc mesenteroides* IBUN 91.2.98. through bioinformatics tools.

Despite the importance of dextran in various industrial applications and the need to massify your production, the mechanism of expression and regulation of dextransucrases in producer strains such as *Leuconostoc mesenteroides* strain IBUN 91.2.98 has not been well documented. The development of second-generation sequencing methods and the use of bioinformatic tools will make it possible to sequence, assemble, and evaluate complete of genomes at relatively low cost guided by a reference genome of *L. mesenteroides subsp. mesenteroides* ATCC 8293, to help the assembly process.

Sequencing produced a total of 1.47 Gb of raw data, which after trimming and quality control were generated 1.40 Gb (7.78 X deep) which was used for assembly. An assembly for *Leuconostoc mesenteroides* strain IBUN 91.2.98 with a length of 2,064 Mpb was obtained. The assembly length represented 85% of the estimated size of the reference genome of *L. mesenteroides subsp. mesenteroides* ATCC 8293.

Keywords: reads, contigs, assembly, sequencing technologies, translation.

Contenido

Lista de figuras

Lista de Tablas

Introducción

La información proporcionada por la secuencia de un genoma es fundamental y solo es posible utilizarla científicamente si es anotada con datos biológicos relevantes. Las anotaciones de un genoma permiten entender que significa la secuencia obtenida, como está estructurada y su funcionalidad [1]. Dado los avances tecnológicos actuales el desarrollo de la secuenciación masiva se ha vuelto un gran reto debido a la variedad de los genomas secuenciados y a la gran cantidad de datos manejados.

La parte fundamental de los métodos de anotación es la localización de genes en un genoma, así como la determinación de su estructura y de las proteínas que producen. Las herramientas de anotación intentan realizar esto a través de análisis computacionales y enfoques experimentales. A pesar de que existe una gran diversidad de procedimientos para anotar un genoma, todos comparten un conjunto de características esenciales. Al final, el mejor enfoque depende del tiempo y recursos disponibles [2].

Para iniciar el proceso es necesario tener un primer ensamble del genoma, el cual debe presentar alta calidad, cuya secuencia tenga al menos un 90% de la información del genoma, el cual se comparará inicialmente con bases de datos de genomas similares provenientes de tecnologías de secuenciación.

Inicialmente se necesita preparar el ensamble identificando y descartando las partes del genoma que no contienen genes, estas regiones suelen complicar el proceso y pueden generar anotaciones erróneas [1].

La primera fase consta de un proceso computacional donde se identifican los elementos del genoma con base en información de experimentos de expresión y en genes de otros genomas. Esta información es mapeada a la secuencia del genoma de referencia y utilizada de manera conjunta para predecir genes [3]. Una alternativa complementaria es la predicción *ab initio* mediante modelos matemáticos [4], sobre todo cuando no se cuenta con evidencia externa.

La fase de anotación permite combinar los elementos mapeados con información biológica relevante, para finalmente definir un conjunto óptimo de anotaciones. Esta última es una tarea difícil por lo que se necesitan combinaciones de distintos procedimientos computacionales para abordarla con precisión [3].

Para obtener óptimos resultados es necesario incluir un paso de validación de estas dos últimas fases de obtención de la información, mediante inspecciones manuales, comprobaciones experimentales y medidas de calidad.

El resultado obtenido es un conjunto diverso de datos biológicos localizados a lo largo del genoma de interés. Al principio se cuenta con una simple secuencia y al final se puede saber qué genes se encuentran en dicha secuencia y cuál es su función [1], mediante el uso de herramientas de búsqueda de secuencias homólogas como (Blast, HHMER, Diamond) se comparan las secuencias del genoma contra bases de datos como NCBI (National Center for Biotechnology Information) y COG (Cluster of Orthologous Groups). Una vez que se han alineado e identificado estas secuencias se les asigna una función si es posible. Estas herramientas son útiles para hacer comparaciones metabólicas e identificar genes de interés [3].

No filtrar las regiones en el genoma que no contienen genes, utilizar programas computacionales que no evalúen claramente el genoma de interés o usar bases de datos de genomas de referencia con anotaciones erróneas son las fuentes de error más comunes en el proceso.

Objetivos

Objetivo general

- Realizar el ensamblaje del genoma de la cepa *Leuconostoc mesenteroides IBUN 91.2.98*. mediante el uso de herramientas bioinformáticas.

Objetivos específicos

- Realizar el ensamblaje del genoma por metodologías de novo y por referencia.
- Validar el ensamblaje realizando una comparación detallada mediante el manejo de herramientas computacionales.
- Anotar los genes presentes en la secuencia del genoma de la cepa *Leuconostoc mesenteroides* IBUN 91.2.98. combinando métodos ab initio y métodos basados en homología.

1. Estado del arte

1.1 Dextrano y dextransacarasas

El nombre dextrano fue acuñado por primera vez por Scheibler en 1874 realizando un estudio de la caña de azucar y la remolacha, en el cual observo un espesamiento el cual es causado por la presencia de un carbohidrato de fórmula $C_6H_{10}O_6$ de rotación óptica positiva. Posteriormente Pasteur, demostró que dicho espesamiento se da por acción bacteriana la cual van Tieghem llamo a dicha bacteria *Leuconostoc mesenteroides* [5].

El dextrano actualmente se define como una homopolisacárido que contiene más del 50% de enlaces (α-1-6), y diferentes ramas vinculadas a (α-1-3), (α-1-2) o (α-1-4) dependiendo de la glicosiltransferasa [6] y de la cepa microbiana porque los dextranos pueden diferir en la longitud de las cadenas y el grado de ramificación [7]. Este compuesto tiene importantes aplicaciones industriales en la producción de productos químicos como sustitutos de plasma, usado como antitrombótico o para reducir la viscosidad de la sangre, también tiene aplicación en la industria alimentaria, en las fermentaciones lácteas utilizado como prebiótico común que incluyen fructooligosacáridos (FOS), galactooligosacáridos (GOS), lactulosa y lafinosa [8].

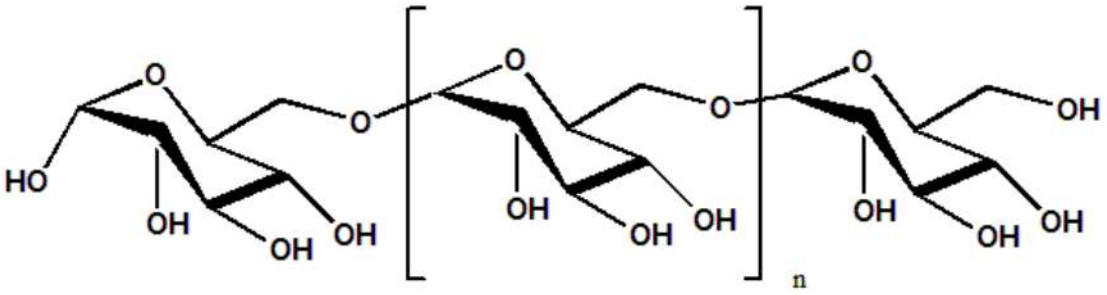


Figura 1-1. Estructura del dextrano, teniendo en cuenta su configuración lineal de enlaces (α-1-6) [9].

A pesar de la importancia del dextrano en varias aplicaciones industriales y la necesidad de masificar su producción, no se tiene muy documentado el mecanismo de expresión y regulación de las dextransacarasas en cepas productoras, usadas comercialmente, de dextrano como el *Leuconostoc mesenteroides* y el *L.Dextranicum (Lactobacteriaceae)* [10].

Las dextransacarasas o también conocidas como glucansacarasas o glucosiltranferasas son las enzimas encargadas de catalizar la transferencia de azúcar para así formar enlaces glicosídicos, al contener un sitio alostérico que induce una conformación favorable para la síntesis de dextrano a partir de sacarosa [11] . Son enzimas que catalizan la síntesis de glicoconjugados como glicolípidos, polisacáridos y glicoproteínas por transferencia de una molécula u oligosacárido a otra aceptora, para la iniciación o elongación de una cadena carbohidratos [12]

Muchos de los microorganísmos enunciados por el National Center for Biotecnology Information, a pesar de producir dextrano a partir de la conversión de dextrinas, no son de importancia industrial en la producción de dextrano. Normalmente se trata de bacterias epifitas que bien en los tallos, hojas y frutos de muchas plantas y juegan un papel importante en la fermentación de los frutos de las plantas donde habitan [13].

Tabla 1-1. Número de reportes de dextransacarasas publicados y microrganismos que las producen.

Microorganísmo	Número de reportes	Porcentaje (%)
Escherichia coli	752	36,75
Serratia marcescens	274	13,39
Proteus mirabilis	146	7,14
Klebsiella pneumoniae	146	7,14
Yersinia pestis	100	4,89
Morganella morganii	93	4,55
Limosilactobacillus reuteri	87	4,25
Serratia entomophila	64	3,13
Providencia rettgeri	52	2,54
Serratia quinivorans	43	2,10
Leuconostoc mesenteroides	40	1,96
Salmonella enterica	37	1,81
Klebsiella quasipneumoniae	36	1,76
Streptococcus salivarius	31	1,52
Leuconostoc citreum	29	1,42
Serratia fonticola	25	1,22
Providencia stuartii	24	1,17
Serratia liquefaciens	23	1,12
Citrobacter werkmanii	22	1,08
Limosilactobacillus fermentum	22	1,08

Tomado de: Dextransucrase DSrb - Protein - NCBI (nih.gov)

Las dextransacarasas son enzimas extracelulares hidrofílicas de gran tamaño 170 KDa, hacen parte de la familia de las amilasas, aunque el mecanismo de reacción no está totalmente comprendido es similar al mecanismos de las α-amilasas [10].

Las dextransacarasas catalizan la transferencia de unidades de D-glucopiranosil desde la sacarosa hasta las moléculas aceptoras, encontrando así dos productos los α-glucanos o oligosacáridos como la maltosa. Si se da la presencia de aceptores como maltosa o isomaltosa, las glucansacarasas catalizan la síntesis de oligosacáridos de bajo peso molecular en lugar de dextranos [14].

La síntesis de polisacáridos se da por un mecanismo de polimerización enzimática, donde se evidencia la actividad hidrolítica sobre la sacarosa que actúa como sustrato de la

reacción librando así una molécula de fructosa **(IV) (Figura 1-2**). Los restos de glucosil se transfieren luego al extremo reductor de una cadena de glucanosilo en crecimiento, la cual se une covalentemente al sitio activo de la Figura 1-2 **(I) (II) (III)** [11].

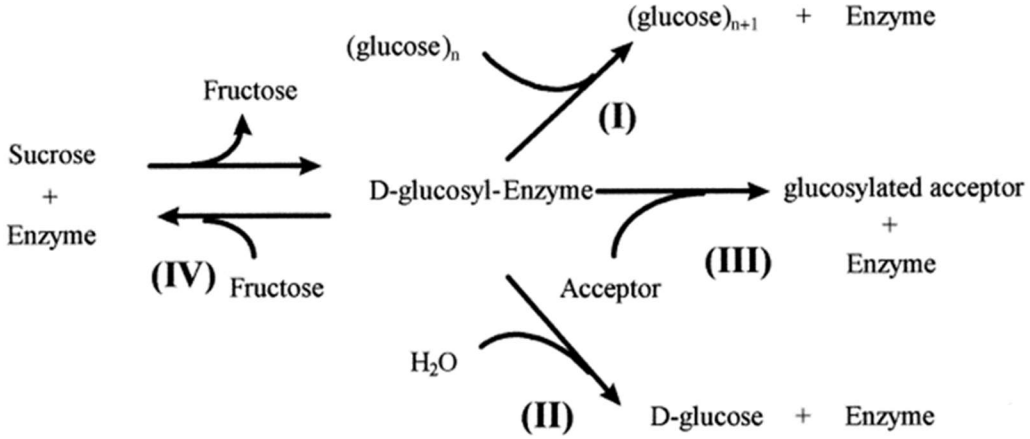


Figura 1-2. Reacciones catalizadas por glucansacarasas. (I) síntesis de glucano por transferencia sucesiva de unidades glucosil. (II) Hidrólisis de la sacarosa transfiriendo un grupo glucosil al agua como aceptor. (III) Síntesis de oligosacarido transfiriendo unidades glucosil a la molécula del polímero. (IV) Reacción reversa por introducción de unidad fructosil. Tomado de: Monchois y col [10].

Para que la enzima funcione satisfactoriamente se necesitan niveles bajos de calcio para una producción y actividad enzimática óptima, y la reacción tiene como requisito la transferencia de un ión hidrogenión al resto fructosilo desplazado de la sacarosa [11].

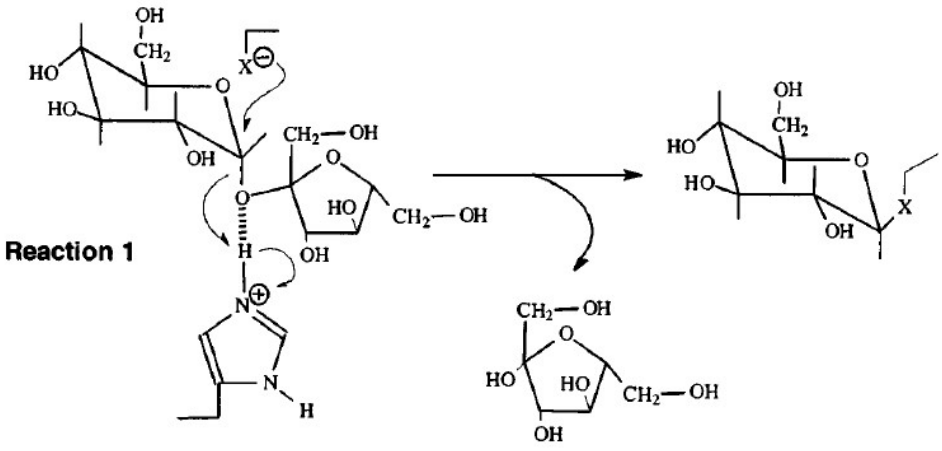


Figura 1-3. Mecanismo de formación de un enlace α (1,6) por dextransacarasa. *Reacción 1:* desplazamiento nucleofílico y protonación de la fracción fructosa para formar un intermedio glucosil-enzima [11].

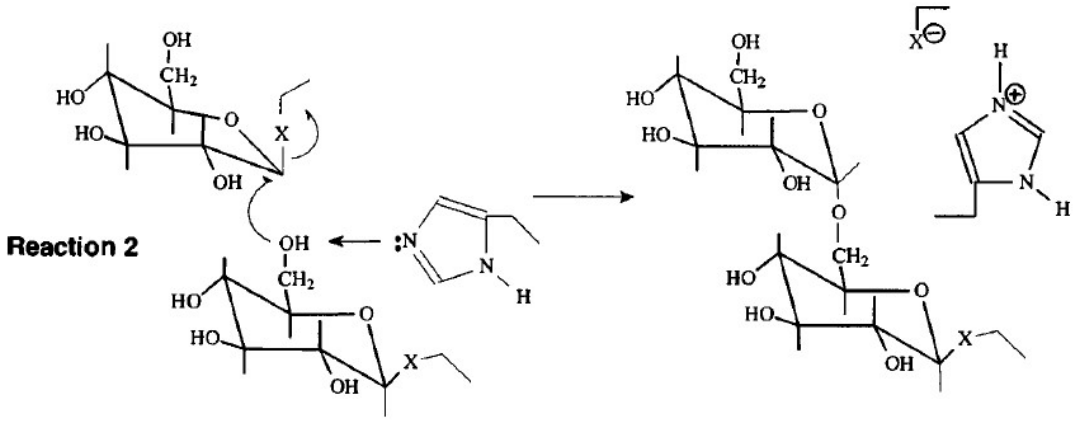


Figura 1-4. *Reacción 2:* formación de un α (1,6) glucosídico por ataque de un grupo hidroxilo C-6 sobre el C-1 del complejo glucosil-enzima; el ataque se facilita mediante la abstracción de un protón del grupo hidroxilo por el grupo imidazol [11].

1.2 *Leuconostoc mesenteroides*

El género *Leuconostoc* comprende actualmente 14 especies y ocho subespecies, todas bacterias grampositivas, no formadoras de esporas y productora del ácido láctico (BAL) [15]. *Leuconostoc mesenteroides* comprende bacterias del ácido láctico (LAB) grampositivas, catalasa-negativas, anaerobias facultativas, no formadoras de esporas y heterofermentativas esféricas y, en su mayoría, productoras de dextrano, con formas de coco [16]. Es un bacilo Gram positivo no esporulado, fenotípicamente esta especie presenta tolerancia al oxígeno, temperatura y pH de crecimiento, así como su capacidad de fermentación de carbohidratos [14].

Leuconostoc es una bacteria heterofermentativa con un contenido de G+C inferior al 50% utilizado en diversas manufacturas de alimentos fermentativos. Metaboliza glucosa para producir ácido láctico y dióxido de carbono por fermentación heteroláctica a través de la vía de la fosfocetolasa (PKP) lo que resulta en la formación de compuestos de aroma y sabor, lo cual beneficia la textura de los productos y se producen diversos compuestos como ácido láctico, ácido acético [17] y compuestos aromáticos como diacetilo, acetoína, acetato y etanol, que son beneficiosos para mejorar el sabor del producto [18].

La producción comercial de dextrano es principalmente realizado por bacterias facultativas como la L. *mesenteroides* en presencia de sacarosa, una fuente orgánica de nitrógeno oligoelementos y fosfatos [11], cuando se realiza la separación y extracción del dextrano, se utilizan solventes polares como el etanol y el metanol, ya que el polisacárido no presenta solubilidad en estos solventes o a través de microfiltración [7].

El genoma del *Leuconostoc mesenteroides* tiene un tamaño promedio de 1,90138 millones de pares de bases, 1762 genes protéicos, con un contenido promedio de G+C del 37,8, los elementos genéticos involucrados en la producción de compuestos de aroma y sabor incluyen el gen de la acetoína reductasa y genes implicados en la biosíntesis de aminoácidos ramificados como el gen de la acetolactato sintasa [17] . Estos compuestos contribuyen al sabor de los productos, mientras que la biosíntesis de bacteriocinas tiene potencial de conservación al inhibir el crecimiento de bacterias patógenas como Listeria spp., Escherichia coli, estafilococo aureus, o Salmonella entérica, beneficiando la calidad e inocuidad del producto final [19].

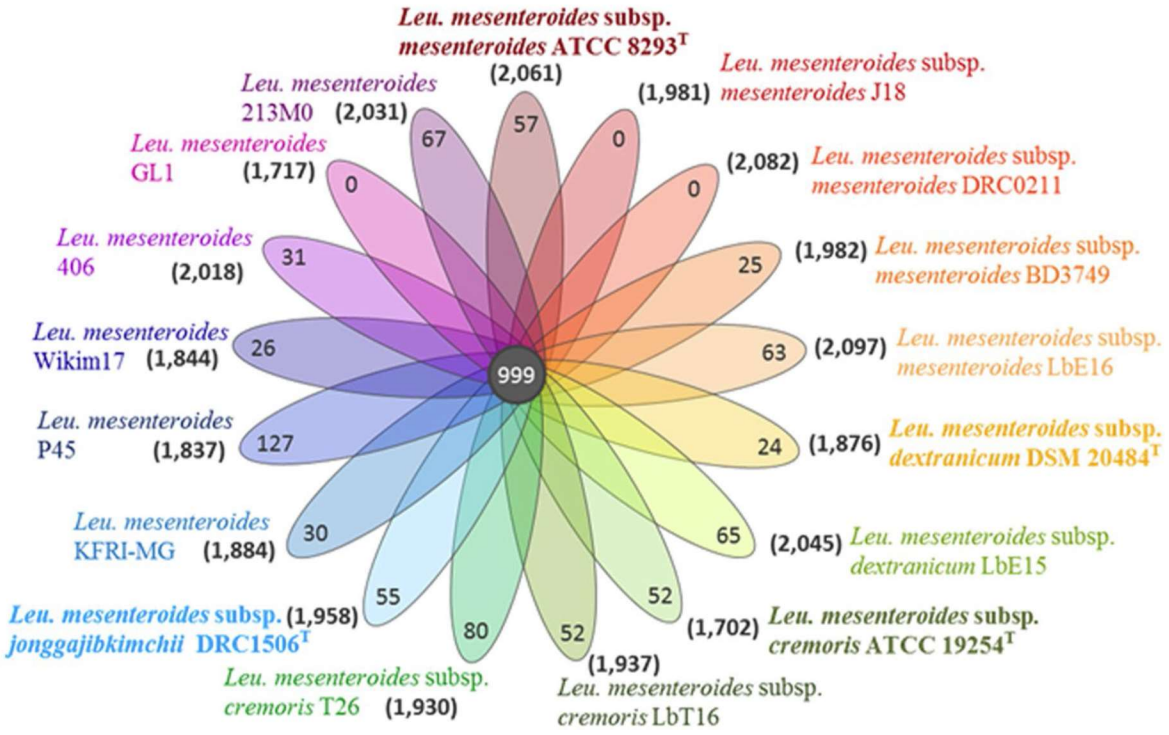


Figura 1-5. Diagrama de parcelas de flores del estudio del genoma de 17 cepas de *Leuconostoc mesenteroides*

1.3 Fundamentos de las tecnologías de secuenciamiento de ADN

1.3.1 Secuenciación de genomas

La secuenciación de genomas completos es un método poderoso para la rápida identificación de genes en un organismo, y sirve como herramienta básica para posteriores análisis funcionales de los nuevos genes descubiertos.

El método de secuenciación automática de Sanger dominó la industria de secuenciación por casi 20 años, llevando a innumerables logros en este campo, como fue la secuenciación del primer genoma bacteriano *Haemophilus influenzae* y la primera

secuencia completa del genoma humano [20]. A pesar que se han implementado muchas mejoras en el proceso de secuenciación, las limitaciones de la tecnología de Sanger trajo consigo la necesidad de desarrollar nuevas y mejores alternativas para la secuenciación de un gran número de genomas en corto tiempo, denominando estas nuevas alternativas como tecnologías de secuenciación de nueva generación [21].

Se ha utilizado un método luminiscente descubierto recientemente para medir la síntesis de pirofosfato, este proceso consiste en un proceso de dos enzimas en el que se usa ATP sulfurilasa para convertir el pirofosfato en ATP, que luego se usa como sustrato para la luciferasa, produciendo así luz en proporción a la cantidad de pirofosfato [22], esta técnica conocida como pirosecuenciación, fue la primera NGS en salir al mercado entre los años 2004 y 2005. A continuación surgieron Illumina en 2006, basada en secuenciación por síntesis, SOLiD en 2007, basada en secuenciación por ligación, y Ion Torrent en el año 2010 basada en detección de pH, las cuales necesitan de la amplificación del ADN previamente a su secuenciación [21]. Además, se han desarrollado tecnologías que no necesitan del paso inicial de amplificación, sino que secuencian directamente una sola molécula de ADN, entre las que se encuentran Helicos, salida al mercado en 2008 y SMRT Pacific Biosciences (PacBio) en 2010 [21].

1.3.2 Características generales de las tecnologías de secuenciación de nueva generación (NGS)

La mayor ventaja ofrecida por las NGS es la capacidad para producir un inmenso volumen de datos de forma económica, logrando llegar a millones o billones de lecturas en solamente una corrida del equipo para un único genoma, en comparación con la secuenciación automática de Sanger, que puede llegar solo hasta cientos de reads [23], pero con una longitud de hasta 1000 pb aproximadamente. Por tanto, con las tecnologías NGS se incrementa considerablemente la cobertura del genoma, que no es más que la cantidad promedio de veces que un nucleótido es representado en un conjunto de secuencias crudas al azar [21]. Sin embargo, los datos de NGS generalmente son secuencias más cortas (a excepción de los producidos por PacBio), que representan un reto desde el punto de vista computacional para su ensamblaje, debido a la longitud y la enorme cantidad de secuencias.

Dentro de las tecnologías de secuenciación de nueva generación, se encuentra Illumina, una de las plataformas de secuenciación de ADN más populares y ampliamente adoptadas en la investigación genómica. El proceso Illumina SBS (Sequencing by Synthesis) es una técnica de secuenciación que implica la carga de la biblioteca de ADN preparada en un sustrato sólido o celda de flujo [24].

La celda de flujo está recubierta con pequeños oligómeros complementarios a las secuencias adaptadoras utilizadas en la preparación de la biblioteca. Estos oligómeros sirven como cebadores para la amplificación del ADN en la celda de flujo [25]. A continuación, se lleva a cabo el proceso de síntesis, donde se incorporan bases de nucleótidos fluorescentes de forma secuencial en la cadena de ADN, utilizando la secuencia de ADN de la biblioteca como plantilla [26].

Después de cada incorporación de base, se realiza una lectura de la señal fluorescente para determinar la base añadida. Este proceso se repite varias veces, generando millones de lecturas de secuencias cortas de ADN en paralelo [27]. Estas lecturas son posteriormente procesadas y ensambladas para reconstruir la secuencia completa del ADN original de la biblioteca.

El proceso Illumina SBS es conocido por su alta precisión y capacidad de generación de grandes cantidades de datos de secuenciación en paralelo, lo que lo hace adecuado para una amplia gama de aplicaciones en la investigación genómica, como estudios de genómica comparativa, identificación de variantes genéticas, metagenómica, transcriptómica, epigenómica, entre otros [28].

La química SBS (Sequencing by Synthesis) de Illumina comparte algunos conceptos con la secuenciación de Sanger, que es otra técnica de secuenciación de ADN. Ambos métodos implican la síntesis de la cadena de ADN y la detección de las bases incorporadas utilizando la fluorescencia [29].

1.3.3 Aplicaciones de las tecnologías NGS en el estudio de genomas

Lo más importante de un proceso de secuenciación es entender las dinámicas y conceptos biológicos que están detrás del campo de estudio, antes de enfrentarse a un proyecto de secuenciación, es importante tener claras las respuestas a las siguientes preguntas: ¿Cuál es el microorganismo en cuestión? ¿Existe un genoma que se pueda utilizar como referencia? ¿Qué características tiene el genoma? A partir de todo lo anterior se decide la tecnología más adecuada para la secuenciación y, por consiguiente, el método a emplear para el análisis de los datos, incluido el ensamblaje del genoma.

Las aplicaciones de las tecnologías NGS se extienden a muchos campos de la biología, como la genómica, la transcriptómica, la epigenética, la genómica de poblaciones, la metagenómica, entre otros. Inicialmente, la aplicación más obvia fue la secuenciación de genomas completos, incluyendo resecuenciación o secuenciación de novo. Proyectos de resecuenciación requieren un genoma de referencia en el que se alinean los reads para detectar variaciones. Por el contrario, los proyectos de novo no tienen genoma de referencia disponible, y los reads se utilizarán únicamente para la reconstrucción de todo el genoma [30]. En la actualidad, las NGS se emplean, además, para la secuenciación de ARN con el objetivo de cuantificar la expresión génica.

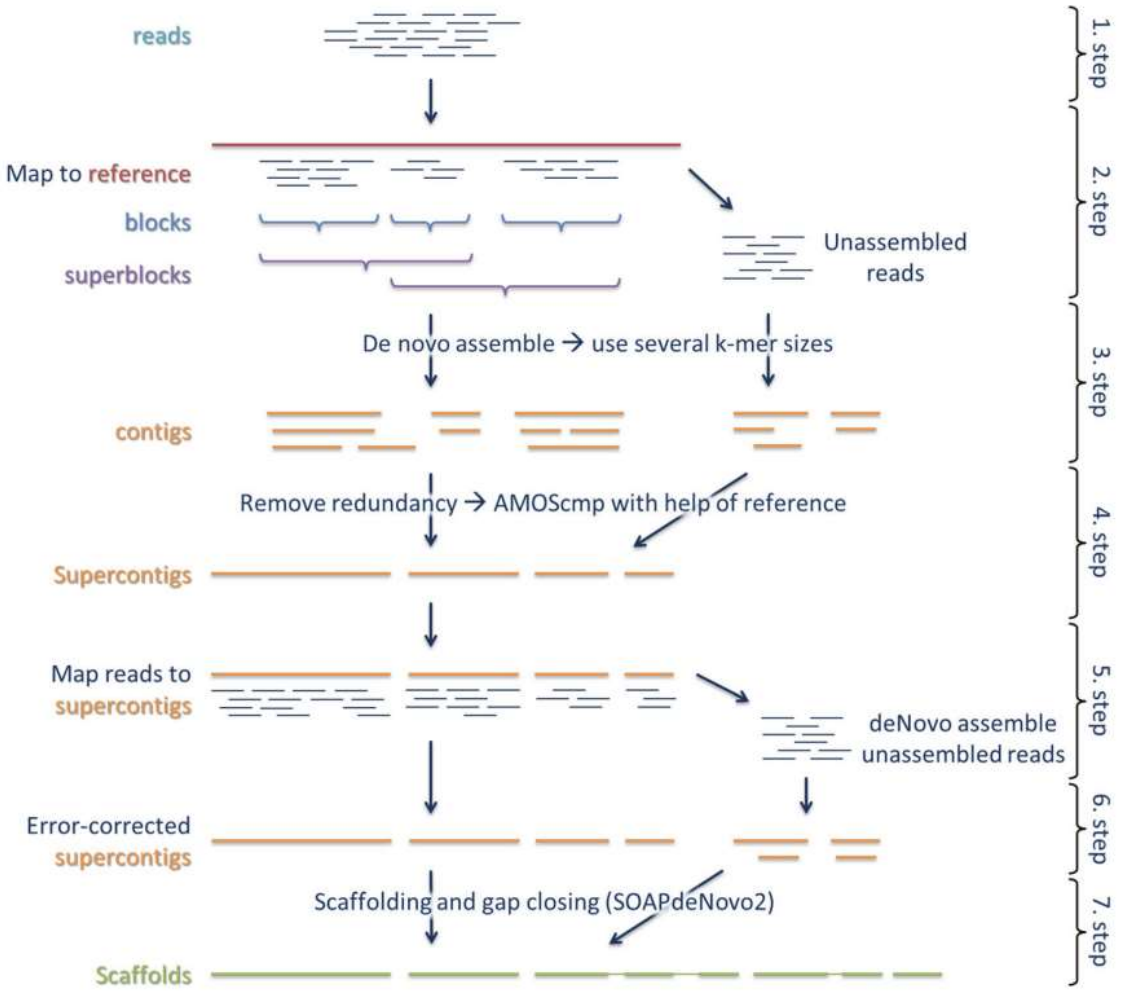


Figura 1-6. Diagrama para el ensamblaje de novo y referencia [31].

El proceso lógico propuesto por Lischer para el ensamblaje de novo y por referencia es primero recortar las lecturas sin procesar para mejorar la calidad (1 paso), asignación de una referencia (2 paso), las lecturas asignadas de referencia se agrupan en bloques con cobertura de lectura continua y estos bloques se combinan en superbloques hasta que se alcanza una longitud total de al menos 12 kb. Los superbloques se superponen en al menos un bloque. Cada superbloque y todas las lecturas no asignadas se ensamblan de novo por separado (3 paso). Los contigs resultantes se fusionan en supercontigs no redundantes (4 paso). En el quinto paso, las lecturas se asignan de nuevo a los supercontigs y las lecturas no asignadas se ensamblan de novo para obtener supercontigs adicionales. Todos los supercontigs se corrigen con lecturas asignadas hacia atrás (6 paso) y luego se usan para andamios y cierre de espacios (7 paso) [31].

1.4 Ensamblaje de genomas

Al proceso de descifrar la secuencia genómica a partir de pequeños fragmentos de ADN con la información biológica adicional disponible, se le denomina ensamblaje de genomas. Las estrategias para el ensamblaje de genomas se pueden dividir en dos categorías: ensamblaje por comparación, en el que se 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 a priori de la organización del mismo [32]. Sin embargo, en esta última estrategia algunas informaciones previas son útiles, como la talla esperada del genoma, el contenido de GC y el contenido de regiones repetitivas, ya que ayudan a elegir la mejor estrategia a seguir [4]. Estos datos pueden ser inferidos a partir de secuencias de organismos relacionados. El ensamblaje de novo con datos de NGS normalmente se da cuando el objeto de estudio son genomas bacterianos que son generalmente pequeños [30].

1.4.1 Control de calidad y corrección de los datos

El control de calidad de los datos crudos sirve como un chequeo rápido para identificar y excluir datos con serios problemas de calidad, lo cual permite ahorrar tiempo en los pasos posteriores. Las herramientas empleadas evalúan la probabilidad de que la base asignada sea la correcta, errores sistemáticos en la técnica de secuenciación, distribución del contenido de GC, secuencias repetitivas, contaminación de ADN exógeno, contenido de bases añadidas e indicación de la presencia de adaptadores entre otros [33].

En los casos de secuencias cortas, producidas por tecnologías NGS de segunda generación (Illumina, IonTorrent, SOLiD), la tendencia es a filtrar los reads que tengan poca calidad, o cortarlos a partir de la posición en la cual la calidad comienza a decaer.

En plataformas de tercera generación, como PacBio RS, se construyen dos tipos de bibliotecas: CLR (Continuous Long Reads, por sus siglas en inglés, reads largos continuos) y CCS (Circular Consensus Sequences, por sus siglas en inglés, secuencias consenso circulares). Durante la corrección de los errores de la biblioteca CLR, se usan reads cortos de alta calidad, como los CCS o los producidos por Illumina. Se estima que el porcentaje de error de los datos CLR crudos es de aproximadamente del 15% [34]; por ello se recomienda corregirlos, ya sea con datos de segunda generación o con secuencias CCS.

1.4.2 Metodologías empleadas en el ensamblaje

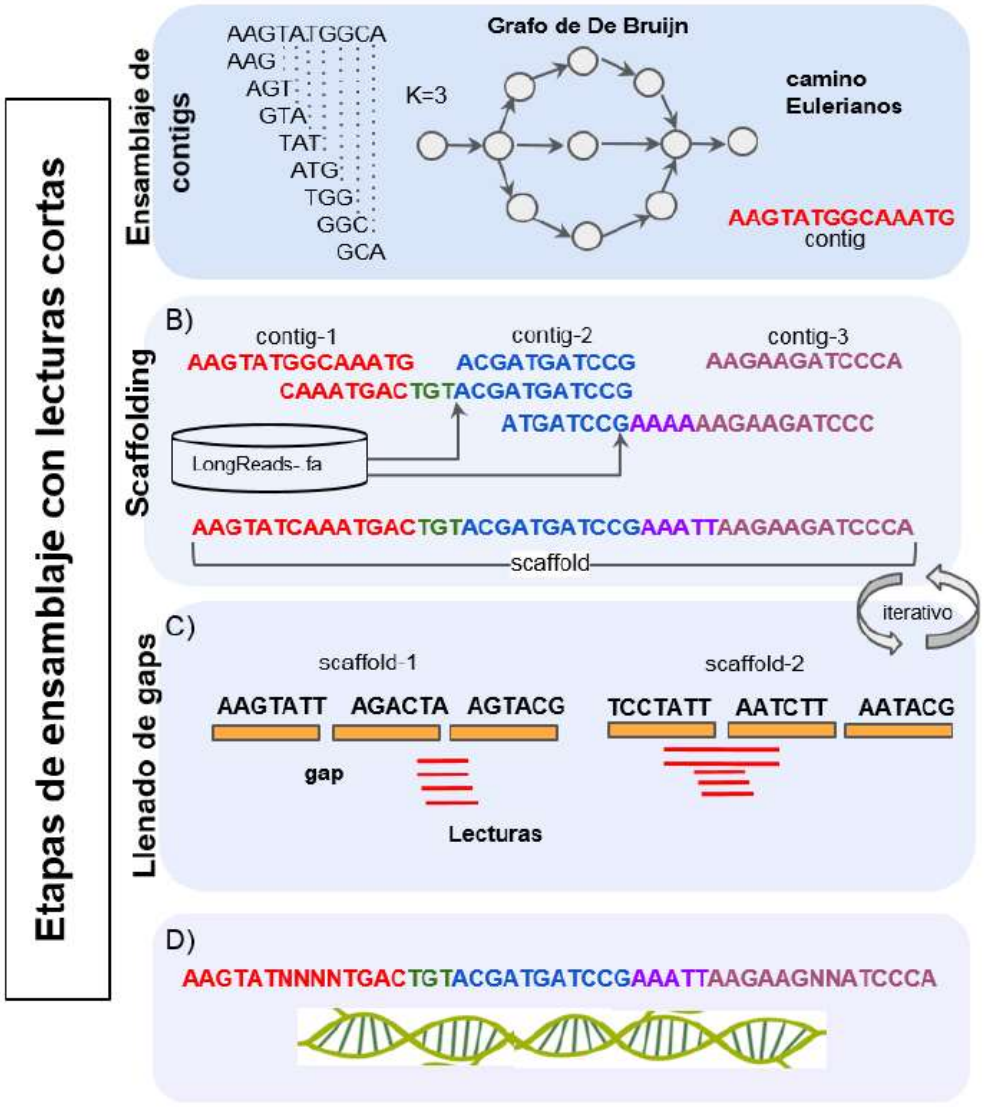


Figura 1-7. Aproximación metodológica para el ensamblaje genómico a partir de lecturas cortas. Se distinguen tres etapas principales: A) Ensamblaje de contigs, donde las lecturas se dividen de acuerdo a un tamaño de subsecuencia k, conocido como k-mers. A partir de estos k-mers se construye el grafo de Bruijn, el recorrido del camino Euleriano del grafo genera el ensamblaje de los contigs. B) Sca olding, permite la unión de contigs. C) Llenado de gaps, usa la información de las lecturas pareadas para resolver regiones con secuencias 'N' [35].

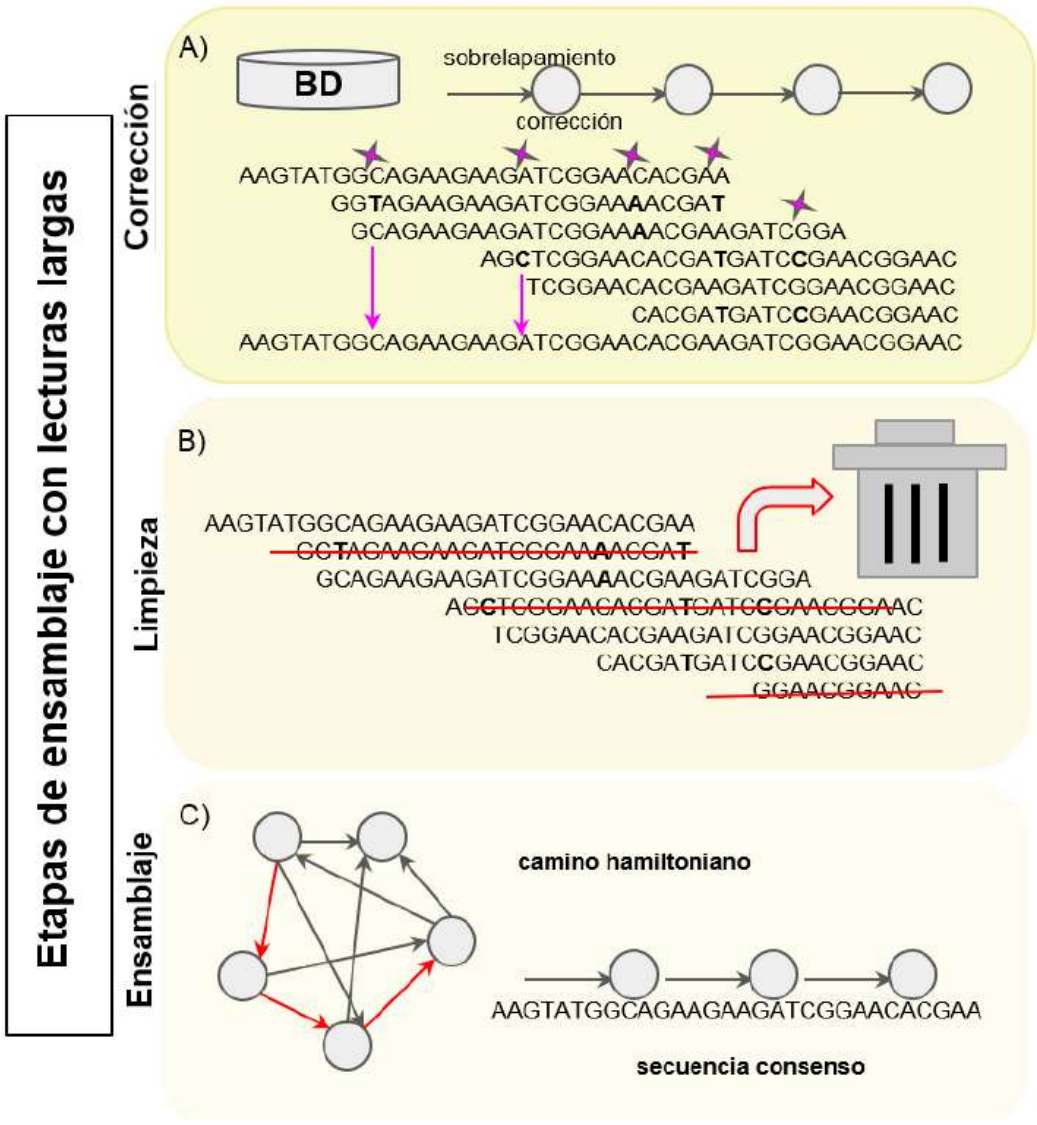


Figura 1-8. Aproximación metodológica para el ensamblaje genómico a partir de lecturas largas. Se distinguen tres etapas principales: A) Corrección, las lecturas se sobrelapan buscando la corrección de errores. B) Limpieza, las lecturas sin soporte identicadas en la fase anterior se eliminan. C)Ensamblaje, se realiza empleando el grafo Overlap-Layout-Consensus, buscando el camino Hamiltoniano que reconstruye la secuencia consenso [35].

Todos los métodos de ensamblaje se basan en la simple suposición de que fragmentos de ADN altamente similares se originan de la misma posición dentro del genoma. De esta manera, la similitud entre secuencias de ADN se usa para conectar fragmentos individuales en secuencias contiguas más largas, denominadas contigs o secuencias consenso obtenidas a partir del ensamblaje de los reads [36].

Las tecnologías NGS han reformado la biología en la actualidad, incluyendo el ensamblaje de genomas. En comparación con el método tradicional de Sanger, el rendimiento de los datos obtenidos por estas nuevas tecnologías es significativamente mayor a bajo costo

[22]. Sin embargo, representan un nuevo reto desde el punto de vista computacional para el ensamblaje de novo debido a la corta longitud de los fragmentos secuenciados.

Los reads con alta repetición son segmentos de ADN que aparecen más de una vez a lo largo del genoma. Cuando un read proviene de una región repetitiva, y es más corto que esta, no se sabe con certeza de cuál copia de la repetición se obtuvo. Es por ello que durante el ensamblaje, se pueden crear falsas uniones en el genoma en las regiones de repeticiones [33]. Además, debido a las regiones repetitivas, muchas veces resulta difícil ensamblar todos los fragmentos para que se logre, en un solo evento, reconstruir la secuencia del genoma completo, incluso en genomas microbianos pequeños.

La tercera generación permite tener reads más largos, manteniendo el rendimiento (cantidad de ADN que puede ser procesado por unidad de tiempo) y permite resolver muchos de los problemas de reads repetitivos [37].

Las estrategias empleadas por los programas ensambladores de secuencias pueden agruparse en tres paradigmas principales: Greedy, Overlap-Layout-Consensus y gráficos de Bruijn [38].

Siendo Greedy el algoritmo más sencillo e intuitivo, ya que siempre conecta los reads que mejor se solapan, de manera iterativa, mientras no contradigan el ensamblaje ya construido. Sin embargo, esta metodología no es ampliamente empleada (Sanger), ya que es inherentemente un proceso de ensamblaje local, no emplea información global, y no resuelve de manera eficiente largas regiones repetitivas en los genomas.

OLC (por sus siglas en inglés, Overlap-Layout-Consensus) es un método que primero identifica todos los pares de reads que se solapan lo suficientemente bien y organiza esta información en un gráfico en el cual hay un nodo por cada uno de ellos y un conector (edge) por cada solapamiento entre los mismos. Esta estructura del gráfico permite el desarrollo de complejos algoritmos de ensamblaje que tienen en cuenta la relación global entre los reads. Para finalmente reconstruir el genoma mediante la búsqueda de un único camino que atraviese todos los nodos solo una vez. Este paradigma dominó el mundo del ensamblaje hasta la emergencia de las tecnologías NGS [38].

Finalmente, los ensambladores basados en gráficos De Bruijn modelan la relación entre subcadenas exactas de longitud k dentro de los reads. De manera similar al método OLC, los nodos en el gráfico representan k-mers, y los conectores indican que k-mers adyacentes se solapan por k-1 letras, por lo que la longitud del k-mer correlaciona con la longitud del solapamiento que el ensamblador es capaz de detectar. En esta metodología no se modelan directamente los reads, sino que están implícitamente representados por los conectores en el gráfico de Bruijn. La mayoría de los ensambladores usan la información global de los reads para refinar la estructura del gráfico, resolver repeticiones y eliminar patrones no consistentes. Además, incorporan métodos de corrección de errores para mejorar la calidad del ensamblaje.

Muchos de los programas ensambladores incluyen además, el proceso de scaffolding, mediante el cual intentan conectar los contigs obtenidos empleando la información que brindan las bibliotecas, cuando varios reads en el extremo de un contigs «apuntan» todos hacia otro contigs. A pesar de no conocer la secuencia entre ellos, se pueden conectar, dejando una distancia aproximada, determinada por la longitud del inserto. Por tanto, se puede inferir cuándo dos contigs son adyacentes, si cada reads PE se ubica en cada contigs a la distancia y orientación esperadas [38].

1.4.3 Métricas de evaluación de la calidad del ensamblaje

Aunque utópico, el objetivo final del ensamblaje de es obtener un número de fragmentos igual al número total de cromosomas (menor número de contigs posibles), en el caso de las bacterias un cromosoma y plásmidos que se obtienen como fragmentos separados.

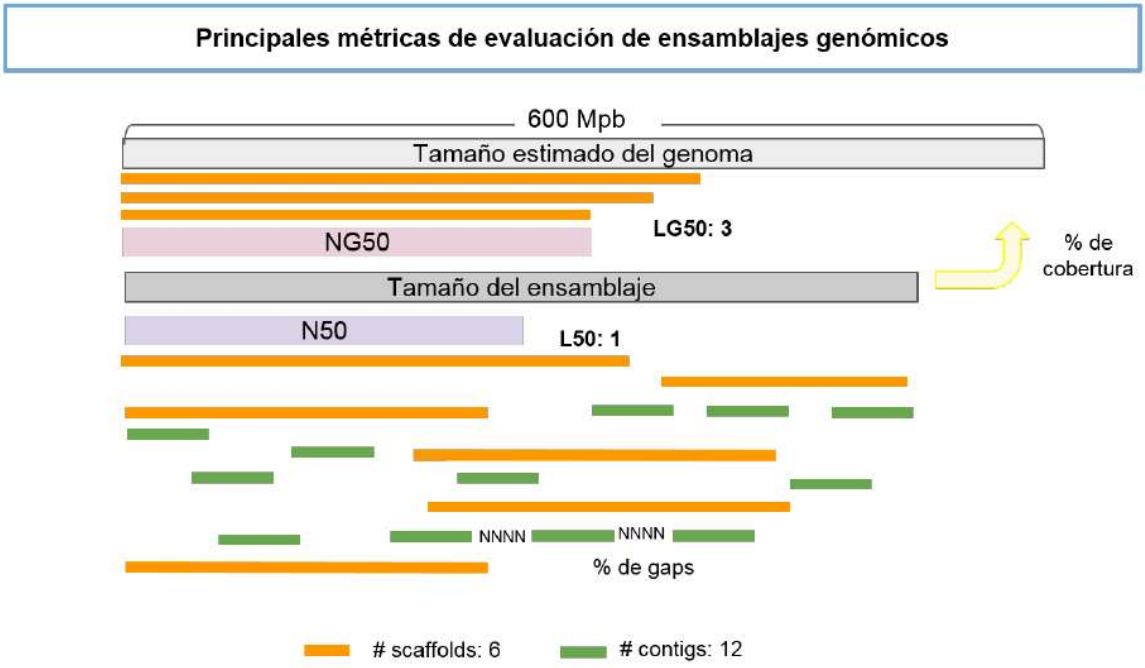


Figura 1-9. Métricas de evaluación de ensamblajes genómicos [35].

Existen algunos indicadores métricos que permiten evaluar la calidad del ensamblaje cuantitativamente. Se calcula generalmente la talla mínima, máxima y media de los contigs, así como la talla total del ensamblaje, la cual debe coincidir con la talla esperada del genoma. Pero el principal valor estadístico es el valor N50, el cual corresponde con el menor de los mayores contigs que cubren la mitad del genoma. Aunque el valor N50 constituye un indicador acerca de la contigsüidad del genoma (acierto del programa ensamblador en unir las secuencias contigsuas en una única secuencia más larga), no es una señal de precisión y calidad del genoma ensamblado. Tampoco brinda una correcta estimación de si la talla del ensamblaje difiere o no de la talla esperada [38].

Para una mejor valoración de la calidad, desde el punto de vista cualitativo, se puede realizar el alineamiento de los reads con los contigs obtenidos, proceso denominado remapeo [38]. La visualización de estos alineamientos permite analizar la consistencia del genoma y si los contigs son confiables, permitiendo así identificar potenciales regiones mal ensambladas. Otra estrategia es comparar la secuencia obtenida con otras secuencias genómicas, ya sea una secuencia de referencia, o genomas de organismos relacionados [39].

1.4.4 Evaluación de la calidad del ensamblaje

Aunque utópico, el objetivo final del ensamblaje de es obtener un número de fragmentos igual al número total de cromosomas (menor número de contigs posibles), en el caso de las bacterias un cromosoma y plásmidos que se obtienen como fragmentos separados.

Existen algunos indicadores métricos que permiten evaluar la calidad del ensamblaje cuantitativamente. Se calcula generalmente la talla mínima, máxima y media de los contigs, así como la talla total del ensamblaje, la cual debe coincidir con la talla esperada del genoma. Pero el principal valor estadístico es el valor N50, el cual corresponde con el menor de los mayores contigs que cubren la mitad del genoma. Aunque el valor N50 constituye un indicador acerca de la contigüidad del genoma (acierto del programa ensamblador en unir las secuencias contiguas en una única secuencia más larga), no es una señal de precisión y calidad del genoma ensamblado. Tampoco brinda una correcta estimación de si la talla del ensamblaje difiere o no de la talla esperada [38].

Para una mejor valoración de la calidad, desde el punto de vista cualitativo, se puede realizar el alineamiento de los reads con los contigs obtenidos, proceso denominado remapeo [38]. La visualización de estos alineamientos permite analizar la consistencia del genoma y si los contigs son confiables, permitiendo así identificar potenciales regiones mal ensambladas. Otra estrategia es comparar la secuencia obtenida con otras secuencias genómicas, ya sea una secuencia de referencia, o genomas de organismos relacionados [39].

1.4.5 Anotación del genoma

Después de un exitoso ensamblaje del genoma, el próximo reto consiste en interpretar la información que contiene. Para ello es necesaria la identificación de las principales características del genoma, proceso conocido como anotación. La anotación de genomas comprende dos etapas fundamentales: la anotación estructural (predicción de regiones codificantes) y la anotación funcional (asignación de información biológica a los genes previamente predichos).

Los métodos para la anotación estructural en un genoma se dividen en dos categorías: método ab initio o de novo, y método por comparación [7]. El método ab initio utiliza

algoritmos estadísticos o de reconocimiento de patrones para determinar si la secuencia de interés es codificante o no, mediante la detección de patrones o motivos específicos en la secuencia. Por otro lado, el método por comparación identifica zonas de alta similitud en organismos relacionados o en bases de datos de proteínas para reconocer las regiones codificantes. Sin embargo, este método es menos exitoso en la identificación de nuevos genes y en nuevos organismos, ya que las bases de datos están sesgadas hacia los genes altamente expresados en los organismos más estudiados.

Para la anotación funcional, también se emplean distintos métodos. La función de un gen se puede inferir mediante la búsqueda de secuencias homólogas en bases de datos, empleando algoritmos de alineamiento local, como BLAST. Esta asignación se realiza tomando como base la premisa de que genes con secuencias compartidas, también comparten su función. Otra metodología empleada es la búsqueda de motivos y dominios funcionales. Aunque un dominio funcional no permite asignar un nombre directamente al gen, sí puede dar una idea de la familia génica a la que pertenece el gen, o indicar el grupo de procesos en los que pueda estar involucrado.

2. Ensamblaje genómico de *Leuconostoc mesenteroides* IBUN 91.2.98

Previamente se realizó el cultivo de la cepa *L. mesenteroides* IBUN 91.2.98, se obtuvo el ADN cromosomal por un proceso estandarizado con una selección conocida de la cantidad de muestra de cultivo y la fase de crecimiento para la extracción de ADN.

2.1 Conjunto de datos de secuenciación

El proceso de secuenciación genero 3,3 millones de lecturas con longitudes que van desde las 35 a 301 bases, contienen 1,46 giga bases (Gb), a través del uso de la tecnología de secuenciamiento Illumina. Se obtuvo una cobertura promedio de 7,78 X teniendo en cuenta la expresión de Lander Waterman [40], y teniendo un tamaño de 2,1 Mb del genoma del *Leuconostoc mesenteroides subsp mesenteroides ATCC 8293* tenemos un cubrimiento de 97.72 X,

Libreria	Plataforma	Longitud promedio de lectura (pb)	Número de secuencias (M)	Número de reads	Bases totales (Mb)	Datos crudos (Gb)
1	Illumina 1.9	168	3,3	95	2,05	1,47

pb: pares de bases, M: millones, Gb: gigabase

Tabla 2-1. Datos de secuenciación genómico obtenidos por la tecnología Illumina.

2.2 Evaluación de la calidad de las lecturas

Al realizar la evaluación de las lecturas se encontró que para la librería secuenciada de la hebra 5´- 3´que la calidad con la que han sido secuenciados cada una de las más de 300 bases que componen nuestras secuencias, comienzan a presentar un resultado de baja calidad para las últimas 10 bases con un q-score inferior a un nivel de calidad PHRED de 28, por lo que es necesario el tratamiento de esta secuencia para optimizarla (Figura 4-1).

Razón por la cual, se realizó a través del uso del Shell de Linux un trimming en la modalidad single end sobre las librerías, ya que al ser no ser secuencias cortas permiten una superposición mínima sustancial entre las secuencias obtenidas. Lo importante de la realización del trimming en la secuencia es perder la mínima cantidad de información secuenciada, evidenciando la mayor calidad posible en las librerías evaluadas, por esta razón se decidió utilizar la siguiente programación en el proceso:

TrimmomaticSE -phred33 G5-3_S18_L001_R1_001.fastq R1_clean42035.fastq LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:35

Tabla 2-2. Fórmula utilizada en el proceso de trimming para la librería obtenida por secuenciación, con un LEADING:3 (Remueve bases de baja calidad del extremo 5´ por debajo de un phred score 30); TRAILING:3 (Remueve bases de baja calidad del extremo 3´ por debajo de un phred score 30); SLIDINGWINDOW 4:20 realiza un corte de ventana deslizante cada 4 bases si la ventana cae por debajo del umbral (score 20-99%); MINLEN:50 elimina las lecturas que caen por debajo de la longitud mínima.

Los resultados del trimming permitieron delimitar un nivel PHRED de calidad de 20, ya que con este obtenemos una precisión en el nombramiento de la base del 99%, con pérdida de información secuenciada no superior a 8.19%.

Niveles de calidad PHRED	**15**	**20**	**25**	**30**
Precisión nombramiento de la base	**97%**	**99%**	**99,5%**	**99,9%**
Pérdida de información Librería 1	1.20	2.72	5.69	11.8
Pérdida de información Librería 2	4.53	8.19	13.2	21.9

Tabla 2-3. Evaluación de los datos de trimming de la librería teniendo en cuenta diferentes niveles de calidad PHRED en comparación a la pérdida de la información secuenciada.

Con los resultados del trimming se evidencia en la gráfica de calidad de la secuencia por base una calidad >20 (Figura 4-8), con una calidad por secuencia superior a 36 para als secuencias superiores a 250.000 pb (Figura 4-10), teniendo la mayor cantidad de frecuencias en dichas proporciones de pb (Figura 4-13). En la Figura 4-11 de contenido de bases por secuencia se observan las condiciones ideales de este tipo de gráficas, con líneas casi paralelas entre sí, lo que indica que la correspondencia en la proporción de las bases nitrogenadas se está presentando. Aunque en las primeras quince bases se presenta una distorsión que presuntivamente se puede estar dando por el uso de primers en el proceso de secuenciación por Illumina (Tabla 2-4), la inclusión de dichos primers no hace que se presenten secuencias sobrerrepresentadas o adaptadores a las librería.

Nombre secuencia	**Secuencia Primer**	**Número de nucleótidos**
>DIR_LM_IBUN_91.2.98	ATGCCATTTACAGAAAAAGTAA	20
>REV_LM_IBUN_91.2.98	TGTGTCAGCATAAGCTTGTA	22

Tabla 2-4. Secuencias de primer utilizadas para el proceso de amplificación mediante el método de secuenciación por Illumina.

2.3 Estrategia del ensamblaje de novo del genoma de la cepa *Leuconostoc mesenteroides* IBUN 91.2.98

Los contigs obtenidos y evaluados se ensamblan de novo a través de las plataformas Velvet y Megahit, estas dos plataformas aplican para el ensamblaje de genomas y metagenomas.

Velvet es un paquete de algoritmos diseñado para tratar las alineaciones de secuenciación de lectura corta, esto se logra a través de la manipulación de gráficos de Bruijn para el ensamblaje de secuencias genómicas a través de la eliminación de errores y la simplificación de regiones repetidas por gráficas de Bruijn, se utilizan las librerías limpias con el comando descrito en la **Tabla 2-5** para la ejecución del ensamblaje.

velveth vel_31 31 -fastq -shortPaired R1_clean42035.fastq R2_clean42035.fastq
velvetg vel_31 -unused_reads yes

Tabla 2-5. Comando para llevar a cabo el ensamblaje por Velvet (http://www.ebi.ac.uk/~zerbino/velvet/Manual.pdf)

El ensamblador MEGAHIT es un ensamblador NGS rápido y eficiente, esta optimizado para metagenomas, pero también funciona bien en el ensamblaje de un genoma pequeño como el del *Leuconostoc mesenteroides* IBUN 91.2.98, trabaja con un tamaño de kmer máximo de 255.

megahit --mem-flag 0 -1 R1_clean42035.fastq -2 R2_clean42035.fastq --k-min 105 --k-max 255 --k-step 20 -o outdir

Tabla 2-6. Comando para llevar a cabo el ensamblaje por Velvet (https://github.com/voutcn/megahit)

2.4 Estrategia del ensamblaje por referencia del genoma de la cepa *Leuconostoc mesenteroides* IBUN 91.2.98

Los contigs obtenidos y evaluados se ensamblaron por referencia usando como molde el genoma de L. mesenteroides subsp. mesenteroides ATCC 8293. Lo primero que se realizó fue la indexación del genoma de referencia de L. mesenteroides subsp. mesenteroides ATCC 8293 a través de la herramienta Bowtie2, la cual está orientada a alinear lecturas de secuenciación relativamente cortas con genomas largos y esta optimizado para las longitudes de lectura y los modos de error producidos por los secuenciadores típicos de Illumina.

Para poder llevar a cabo el ensamblaje es necesario realizar la indexación del genoma de referencia, mediante la transformación por Burrows Wheeler (Burrows wheeler transform o BWT) para la compresión del archivo, la generación del LF mapping y el aumento de la eficiencia en el ensamblado (Tabla 2-7**Tabla 2-2**).

bowtie2-build -f ATCC8293.fa refindex

Tabla 2-7. Comando utilizado en el proceso indexación por “sufijos” del genoma de referencia mediante el uso de bowtie el sufijo –f se incluye porque se va a indexar un archivo fasta, al no contener valores de calidad los obvia en la indexación. (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#the-bowtie2-build-indexer)

Bowtie2 toma como índice el genoma de referencia de L. mesenteroides subsp. mesenteroides ATCC 8293 y las dos librerías de lectura de secuenciación y genera un conjunto de alineaciones en formato sequence aligment map (SAM), este formato consiste

en el mapeo de los reads y exporta un archivo de dos secciones, sección del encabezado y sección de los alineamientos posibles.

bowtie2 -x refindex -1 R1_clean42035.fastq -2 R2_clean42035.fastq -S align42035.sam

Tabla 2-8. Comando utilizado para la generación del archivo SAM con los alineamientos.(http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml) (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#sam-output)

```
samtools view -bS align42035.sam > align42035.unsorted.bam
```

Tabla 2-9. Comando para convertir el archivo SAM a BAM mediante samtools (http://www.htslib.org/)

En el archivo BAM obtenido en el paso anterior los alineamientos se encuentran organizados en el orden en el que estos aparecen en el archivo FASTQ original. Con el fin de mejorar cualquier otro análisis que se realice sobre el archivo BAM obtenido, éste tiene que ser reorganizado, de tal manera que los alineamientos sean organizados de acuerdo al orden del genoma, no por el orden de aparición de los alineamientos. El siguiente comando realiza dicho proceso, y es aquí donde realmente arranca el ensamblaje del genoma (Tabla 2-10).

samtools sort align42035.unsorted.bam -o align42035.sorted.bam

Tabla 2-10. Comando para convertir el archivo SAM a BAM mediante samtools

El archivo BAM permite también obtener el listado de variantes (Tabla 2-11) que se encuentran entre nuestro genoma y el genoma de referencia. El formato estándar para obtener estas variantes es llamado VCF (variant call format). Este último archivo permite evidenciar los contigss que no se ensamblaron sobre la referencia, al tratarse de una secuenciación de ADN puro de *L. Mesenteroides* sepa IBUN 91.2.98, dichos contigsns refieren a las diferencias genómicas respecto *del L. mesenteroides subsp. mesenteroides ATCC 8293.*

samtools mpileup -uf ATCC8293.fa align.sorted.bam | bcftools view -Ov - > variants.vcf

Tabla 2-11. Comando para convertir el archivo SAM a BAM mediante samtools (https://samtools.github.io/hts-specs/VCFv4.2.pdf)

Mediante el comando (samtools mpileup -uf reference.fa align.sorted.bam | bcftools call -c | vcfutils.pl vcf2fq > consenso42035.fq) finalmente exporto un archivo llamado consenso en formato fastq, donde en un solo contigs posiciona los reads de secuenciación de acuerdo a la secuencia del genoma de referencia. Una vez finalizado el proceso de ensamblaje se obtienen los archivos de la Tabla 2-12.

Peso (bytes)	Peso (Mb)	Nombre archivo generado
1427202871	1427.2	R1_clean42035.fastq
1350478284	1350.4	R2_clean42035.fastqcccc
2096724	2.09	ATCC8293.fa
3357806592	3357.8	align42035.sam
511136611	511.1	align42035.sorted.bam
16808	0.016	align42035.sorted.bam.bai
895015829	895.0	align42035.unsorted.bam
7580865	7.58	variants.vcf
2064418	2.06	consenso42035.fa
4128825	4.12	consenso42035.fq

Tabla 2-12. Archivos obtenidos en el proceso de ensamblaje mediante el uso de bowtie, archivos SAM y BAM.

Comparando el tamaño del archivo consenso42035.fa y del archivo ATCC8293.fa se encuentra una cobertura en el ensamblaje de 98,45% al comparar los tañamos en los archivos ensamblados.

2.5 Evaluación de la calidad del genoma de la cepa *Leuconostoc mesenteroides* IBUN 91.2.98

Para confirmar la cobertura en el conjunto de contigs finales del ensamblaje realizado por cada una de las metodologías y plataformas utilizadas, las tasas de alineamiento fueron del 127% para el ensamblaje por velvet, 99.0% para el ensamblaje por megahit y de 98.45% para el ensamblaje por referencia.

Peso (bytes)	Peso (Mb)	Nombre archivo generado
2096724	2.09	ATCC8293.fa
2663220	2.66	denovo_velvet.fa
2075822	2.07	denovo_megahit.fa
2064418	2.06	referencia_bowtie.fa

Tabla 2-13. Archivos de ensamblaje obtenidos en el proceso de ensamblaje por metodologías de novo y por referencia.

El objetivo final del ensamblado es obtener un número de fragmentos igual al número total de cromosomas, en el caso de las bacterias un cromosoma, y también puede haber plásmidos que se obtienen como fragmentos separados. Por lo tanto, se buscará obtener el menor número de contigs posible. Como se mencionó anteriormente existen algunas métricas interesantes para evaluar la calidad del ensamblado N50, L50, entre otras a partir de genomas de organismos con genomas similares [38].

Para esto se utilizará el programa QUAST (Quality Assessment Tool for Genome Assemblies: http://bioinf.spbau.ru/quast: http://quast.sourceforge.net/quast). QUAST evalúa los ensamblados funcionando con y sin genoma de referencia, mediante el comando de la Tabla 2-14.

```
quast.py consensos.fa -r ATCC8293.fa
```

Tabla 2-14. Comando para evaluar a través de quast por el lenguaje de programación Python

La Figura 4-18 muestra el número de bases en los primeros x contigs, ya que x varía de cero al número de contigs totales [41], pero no realiza una comparación tan acertada entre los ensamblajes por realizados por metodologías de-novo y referencia, ya que por ensamblador Velvet se tiene un número de contigs (842) mucho mayor a los obtenidos por megahit (40) o el ensamblado por referencia (1). El tamaño de dichos contigs es mayor en los ensamblados por megahit, presentado un N50 y N75 mayor al de los otros dos modelamientos (Figura 4-19), y a pesar que la fracción del genoma en esta plataforma es de 85.412% (Tabla 2-15) y que el tamaño de los contigs es mayor solamente alinearon correctamente algunos contigs largos, teniendo así que en dicha cobertura en su mayoría los contigs estén mal ensamblados o con porciones de contigs no alineados (Figura 4-20).

Worst Median Best

Genome statistics	referencia_bowtie	denovo_velvet	denovo_megahit
Genome fraction (%)	85.336	25.68	85.412
Duplication ratio	1.044	1.014	1.009
Largest alignment	72 290	3307	437 336
Total aligned length	72 290	525 373	1 754 100
NG50	2 030 562	-	253 858
NG75	2 030 562	-	166 379
NA50	-	669	87 308
NA75	-	535	37 147
NGA50	-	-	87 308
NGA75	-	-	36 960
LG50	1	-	2
LG75	1	-	4
LA50	-	339	6
LA75	-	594	14
LGA50	-	-	6
LGA75	-	-	15
Misassemblies			
# misassemblies	0	1	48
# relocations	0	1	46
# translocations	0	0	0
# inversions	0	0	2
# misassembled contigs	0	1	23
Misassembled contigs length	0	788	1 834 038
# local misassemblies	0	2	12
# scaffold gap ext. mis.	0	0	0
# scaffold gap loc. mis.	236	0	0
# unaligned mis. contigs	0	1	0
Unaligned			
# fully unaligned contigs	0	106	2
Fully unaligned length	0	75 425	3393
# partially unaligned contigs	0	1	11
Partially unaligned length	0	578	234 895

Tabla 2-15. Tabla de resultados de QUAST.

Mismatches			
# mismatches	333	3286	11 114
# indels	0	37	260
Indels length	0	111	1232
# mismatches per 100 kbp	460.64	627.75	638.36
# indels per 100 kbp	0	7.07	14.93
# indels (<= 5 bp)	0	33	221
# indels (> 5 bp)	0	4	39
# N's	1 960 680	0	0
# N's per 100 kbp	96 558	0	0
Statistics without reference			
# contigs	1	842	40
# contigs (>= 0 bp)	1	10 490	237
# contigs (>= 1000 bp)	1	78	16
# contigs (>= 5000 bp)	1	0	12
# contigs (>= 10000 bp)	1	0	9
# contigs (>= 25000 bp)	1	0	7
# contigs (>= 50000 bp)	1	0	7
Largest contig	2 030 562	3308	923 490
Total length	2 030 562	606 813	1 994 271
Total length (>= 0 bp)	2 030 562	2 234 685	2 066 613
Total length (>= 1000 bp)	2 030 562	102 924	1 979 695
Total length (>= 5000 bp)	2 030 562	0	1 972 073
Total length (>= 10000 bp)	2 030 562	0	1 954 130
Total length (>= 25000 bp)	2 030 562	0	1 922 158
Total length (>= 50000 bp)	2 030 562	0	1 922 158
N50	2 030 562	707	253 858
N75	2 030 562	583	166 379
L50	1	323	2
L75	1	559	4
GC (%)	38.81	37.73	37.53
Similarity statistics			
# similar correct contigs	0	0	0
# similar misassembled blocks	0	0	0

Tabla 2-16. Tabla de resultados de QUAST.

La Figura 4-21 evidencia la distribución del contenido GC en los contigs obtenidos en cada una de las plataformas de ensamblaje. El eje X muestra el porcentaje de GC y Y muestra el número de ventanas de 100 pb no superpuestas cuyo contenido de GC es el valor X [41]. Esta distribución normalmente es gaussiana [42], es posible que el número de ventanas sea mayor en el ensamblaje por megahit, por la presencia de contaminantes con un contenido de GC diferente, lo cual genera una superposición de múltiples contenidos GC.

Los datos generales presentados en la **Tabla 2-15** muestran una mayor cobertura del genoma para el ensamble realizado por referencia, evidenciando así la similaridad genómica entre el *Leuconostoc mesenteroides IBUN 91.298* con el genoma de referencia del *L. mesenteroides subsp. mesenteroides* ATCC 8293, en comparación a los ensamblajes de novo realizados por las plataformas velvet y megahit.

2.6 Anotación del genoma de la cepa *Leuconostoc mesenteroides* IBUN 91.2.98

La anotación de genomas es el proceso de identificar y describir las características funcionales de los genes en un genoma. Esto implica la predicción de genes, la identificación de regiones codificantes y no codificantes, la asignación de funciones a los genes y la anotación de otras características genómicas, como regiones reguladoras, elementos repetitivos, sitios de inicio y terminación de la replicación, entre otros.

Dentro de las múltiples herramientas que ofrece BV-BRC (Bacterial and Viral Bioinformatics Resource Center), el cual es un recurso en línea que proporciona una amplia variedad de herramientas bioinformáticas para el análisis y la anotación de genomas bacterianos y virales. Tenemos la herramienta PATRIC (Pathosystems Resource Integration Center), la cual es una plataforma de análisis y anotación de genomas bacterianos y virales que proporciona una amplia gama de herramientas de análisis, visualización y comparación de genomas.

A través del uso de dicha herramienta se realizó el proceso de anotación del genoma ensamblado por referencia, comparándolo con el genoma de *Leuconostoc subsp mesenteroides* ATCC 8293. Obteniendo una completeness de 99.7%, esto último refiere a una medida que evalúa el grado en que el genoma ha sido anotado o descrito de manera completa en términos de sus características genéticas y funcionales, dentro de los 95 contigs (Figura 4-22) .

En PATRIC, la completeness se calcula utilizando la herramienta CheckM, que es una herramienta ampliamente utilizada para estimar la calidad y la completitud de los genomas bacterianos y archaeales. CheckM compara los genes predichos en un genoma con un conjunto de marcadores universales de genes conservados, y luego calcula un índice de completitud basado en la presencia o ausencia de estos marcadores, con un nivel de contaminación de 1.4%, el cual es bajo teniendo en cuenta los datos reportados por BV-BRC.

En la Figura 4-23 se evidencia que la longitud del genoma obtenido fue de 1896561 pares de bases con un contenido de GC de 37.77% lo cual es característico para un genoma de *Leuconostoc mesenteroides*. Obteniendo a su vez un coarse consistency de 98.7%, lo cual refiere una medida que evalúa la consistencia general de la anotación de un genoma bacteriano o viral en términos de la presencia o ausencia de genes y funciones esperadas en función de la taxonomía del organismo evaluado, lo anterior determina la poca presencia de errores en la anotación o en la identificación taxonómica del organismo.

Microorganismo	% GC	Tamaño de genoma (pb)	Secuencias codificantes	Referencia
Leuconostoc mesenteroides LT-38	37,56	2022184	2005	Kato & Oikawa, 2017
L. mesenteroides subsp. cremoris cepa T26	38,4	1833933	1687	Pedersen *et al.* 2014
L. mesenteroides ATCC 8293 con	37,5	2038396	1598	Kim et al., 2018
L. mesenteroides subsp. mesenteroides J18	37,77	1896561	1942	Jung et al., 2012

Tabla 2-17. Estudios previos realizados a microorganismos también del grupo de *Leuconostoc mesenteroides.*

En la Figura 4-24 se pueden observar que el genoma ensamblado por referencia tiene 1819 CDS o regiones codificantes homologas en la plataforma BV-BRC y 1803 CDS en la plataforma GenBank, lo anterior muestra que las regiones codificantes obtenidas son significativas teniendo en cuenta los estudios realizados a otras cepas de *Leuconostoc* (Tabla 2-17).

3. Conclusiones y recomendaciones

En la presente revisión documental y uso de herramientas bioinformáticas se generó el primer ensamblaje del genoma de alta calidad para el *Leuconostoc mesenteroides* cepa IBUN 91.2.98. Adicionalmente, el presente ensamblaje da inicio a la posibilidad de realizar estudios a nivel comparativo con otras cepas de *Leuconostoc*, o en la interpretación misma del genoma obtenido a través de procesos de anotación estructural y funcional.

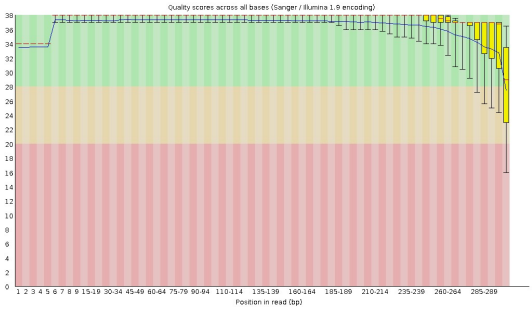
El proceso de ensamblaje conllevo una serie de retos computacionales, al intenta reconstruir la secuencia original del genoma de *Leuconostoc mesenteroides* cepa IBUN 91.2.98 a partir de lecturas cortas producidas por la tecnología de secuenciación Illumina, lo cual se puede mejorar en futuros proyectos generando secuenciaciones con tecnologías con Pac-Bio para obtener contigs largos y así poder generar futuros modelos de genomas más robustos.

Adicional a la generación del genoma, es necesario realizar su caracterización estructural y funcional. Sin embargo, este proceso conlleva un análisis cuidadoso que permita obtener un alto nivel de anotación, para ello se requieren múltiples rondas de re-anotación, prestando especial cuidado con los procesos de identificación de elementos repetitivos y prediciones de novo.

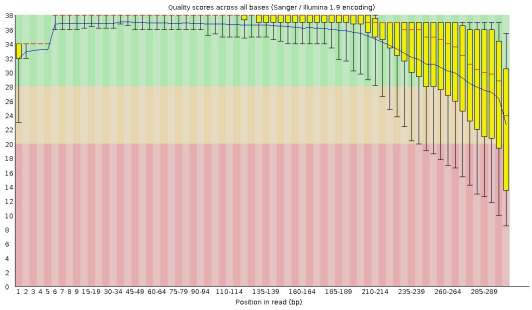
4. Anexos

Anexo A. Control de calidad de la librería genómica datos crudos

Figura 4-1. Calidad de la secuencia por base

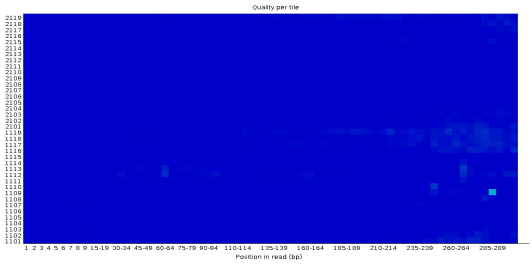


G5-3_S18_L001_R1_001.fastqc

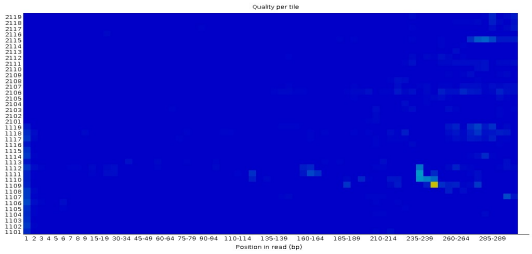


G5-3_S18_L001_R2_001.fastqc

Figura 4-2. Calidad de la secuencia por tile

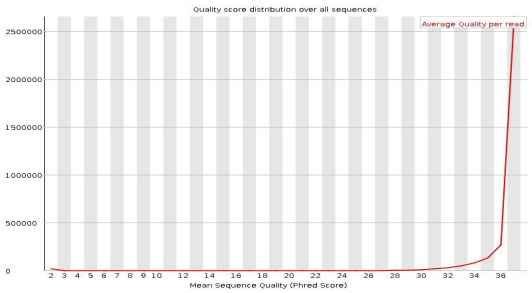


G5-3_S18_L001_R1_001.fastqc

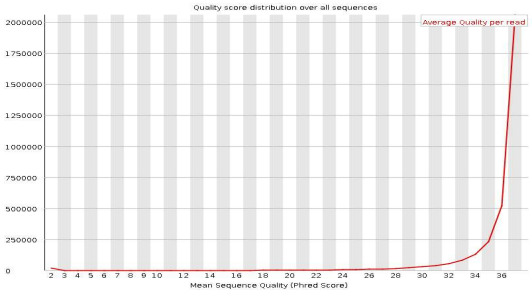


G5-3_S18_L001_R2_001.fastqc

Figura 4-3. Puntuación de calidad por secuencia

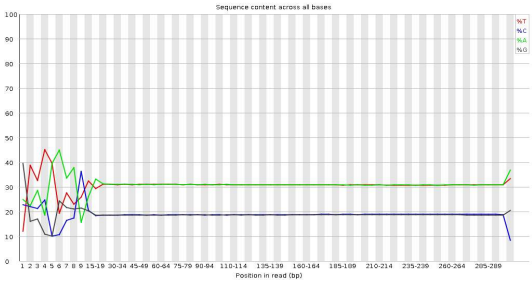


G5-3_S18_L001_R1_001.fastqc

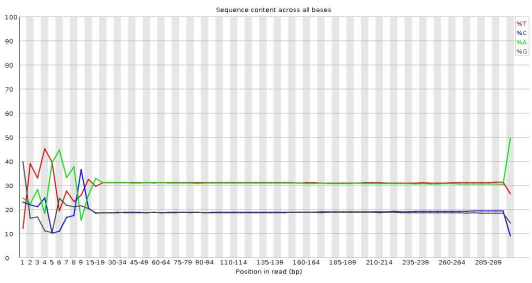


G5-3_S18_L001_R2_001.fastqc

Figura 4-4. Contenido de bases por secuencia

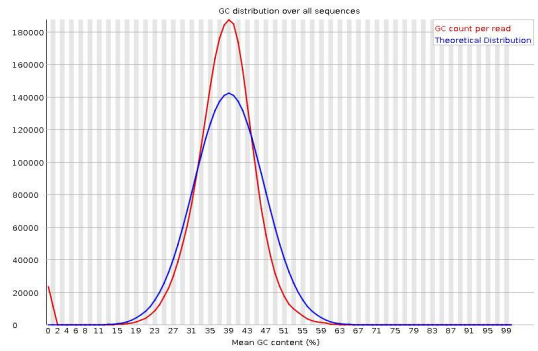


G5-3_S18_L001_R1_001.fastqc

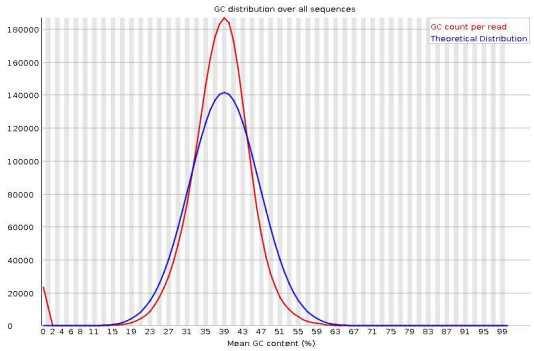


G5-3_S18_L001_R2_001.fastqc

Figura 4-5. Contenido de GC por secuencia

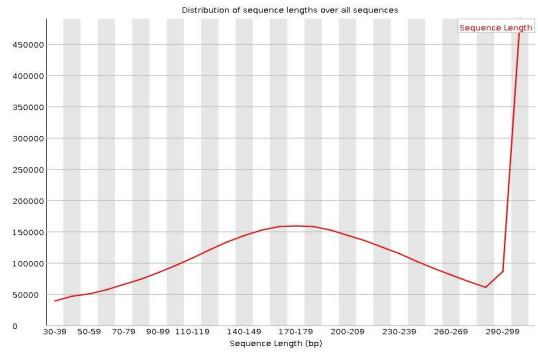


G5-3_S18_L001_R1_001.fastqc

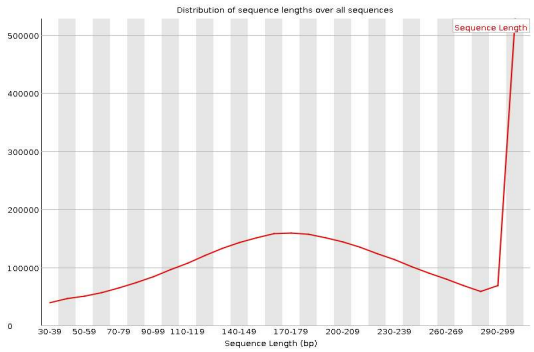


G5-3_S18_L001_R2_001.fastqc

Figura 4-6. Distribución longitud de la secuencia

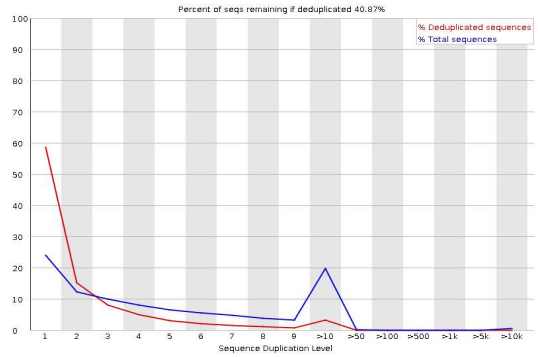


G5-3_S18_L001_R1_001.fastqc

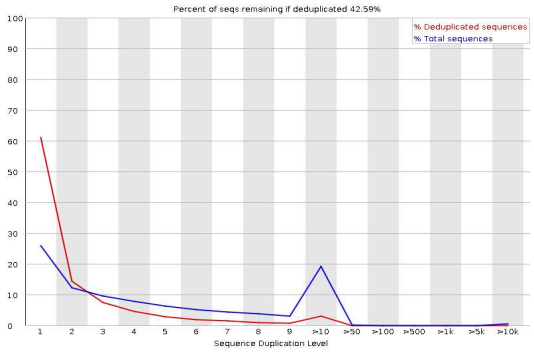


G5-3_S18_L001_R2_001.fastqc

Figura 4-7. Niveles de duplicación de secuencia



G5-3_S18_L001_R1_001.fastqc

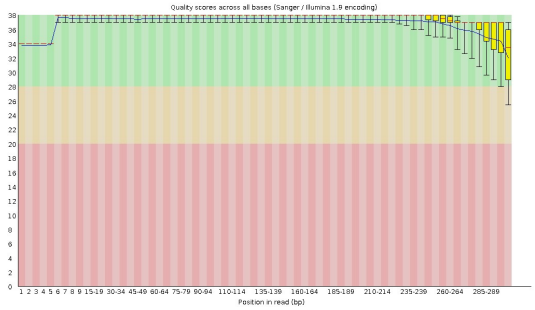


G5-3_S18_L001_R2_001.fastqc

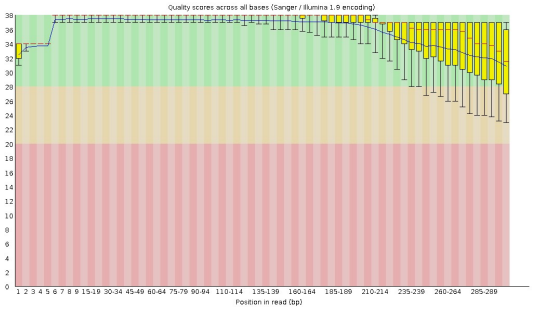
Tabla 4-1. Control de calidad de la librería genómica datos crudos.

Anexo B. Control de calidad de las librerías genómica una vez realizado el trimming con un nivel PHRED de 20 (99%)

Figura 4-8. Calidad de la secuencia por base

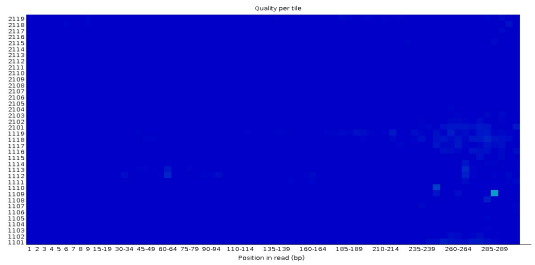


R1_clean42035_fastqc

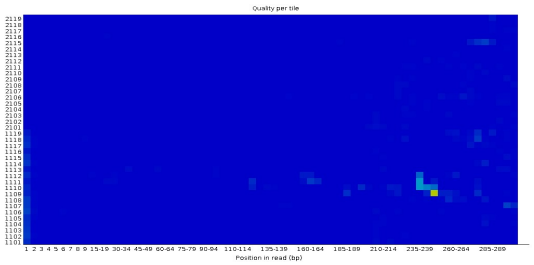


R2_clean42035_fastqc

Figura 4-9. Calidad de la secuencia por tile

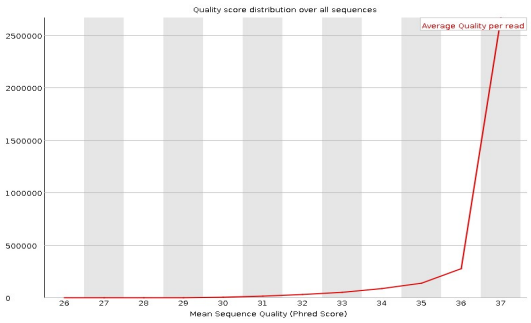


R1_clean42035_fastqc

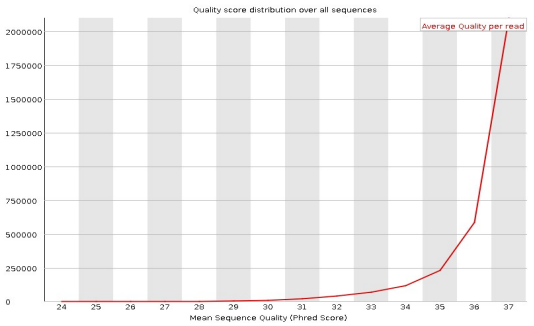


R2_clean42035_fastqc

Figura 4-10. Puntuación de calidad por secuencia

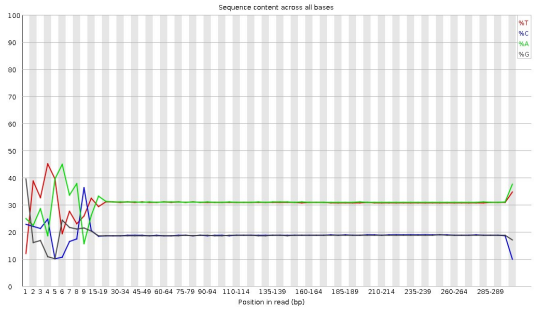


R1_clean42035_fastqc

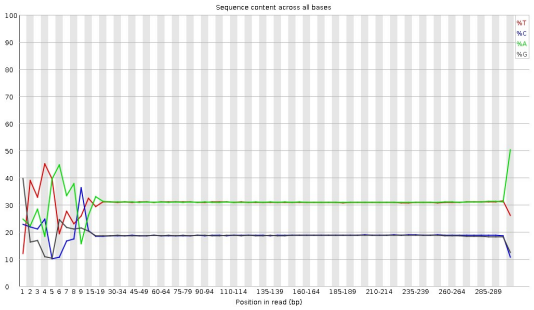


R2_clean42035_fastqc

Figura 4-11. Contenido de bases por secuencia

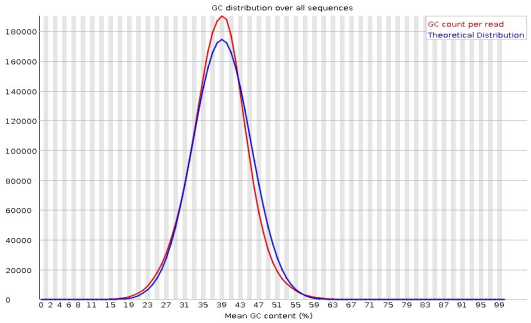


R1_clean42035_fastqc

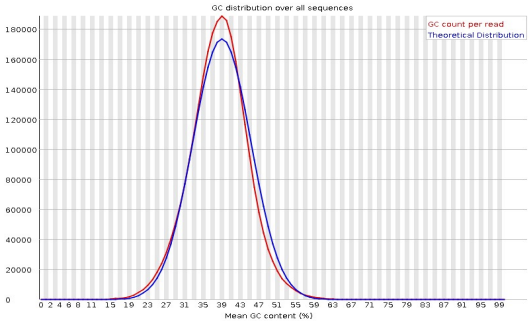


R2_clean42035_fastqc

Figura 4-12. Contenido de GC por secuencia

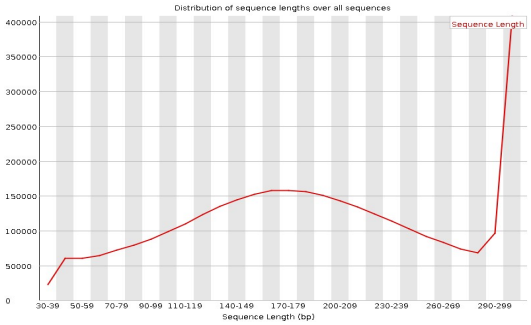


R1_clean42035_fastqc

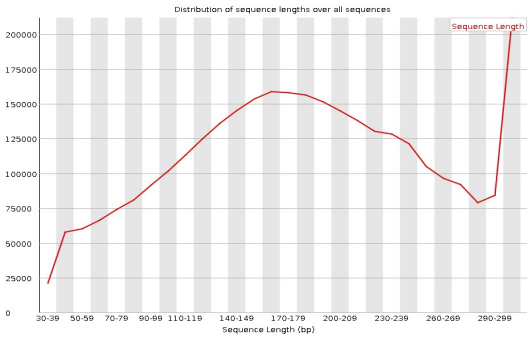


R2_clean42035_fastqc

Figura 4-13. Distribución longitud de la secuencia

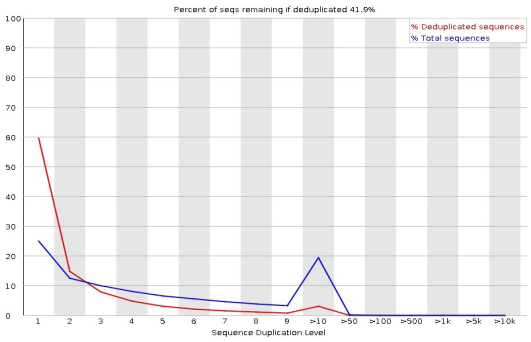


R1_clean42035_fastqc

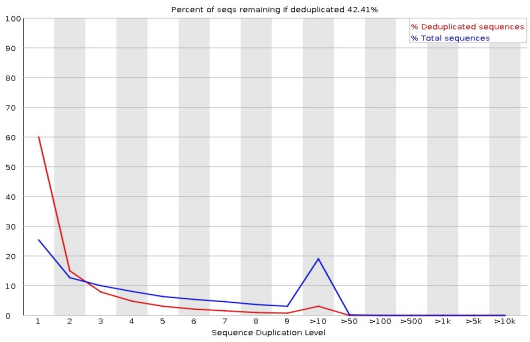


R2_clean42035_fastqc

Figura 4-14. Niveles de duplicación de secuencia



R1_clean42035_fastqc



R2_clean42035_fastqc

Tabla 4-2. Control de calidad de las librerías genómica una vez realizado el trimming con un nivel PHRED de 20 (99%).

Anexo C. Visualización del alineamiento mediante tview

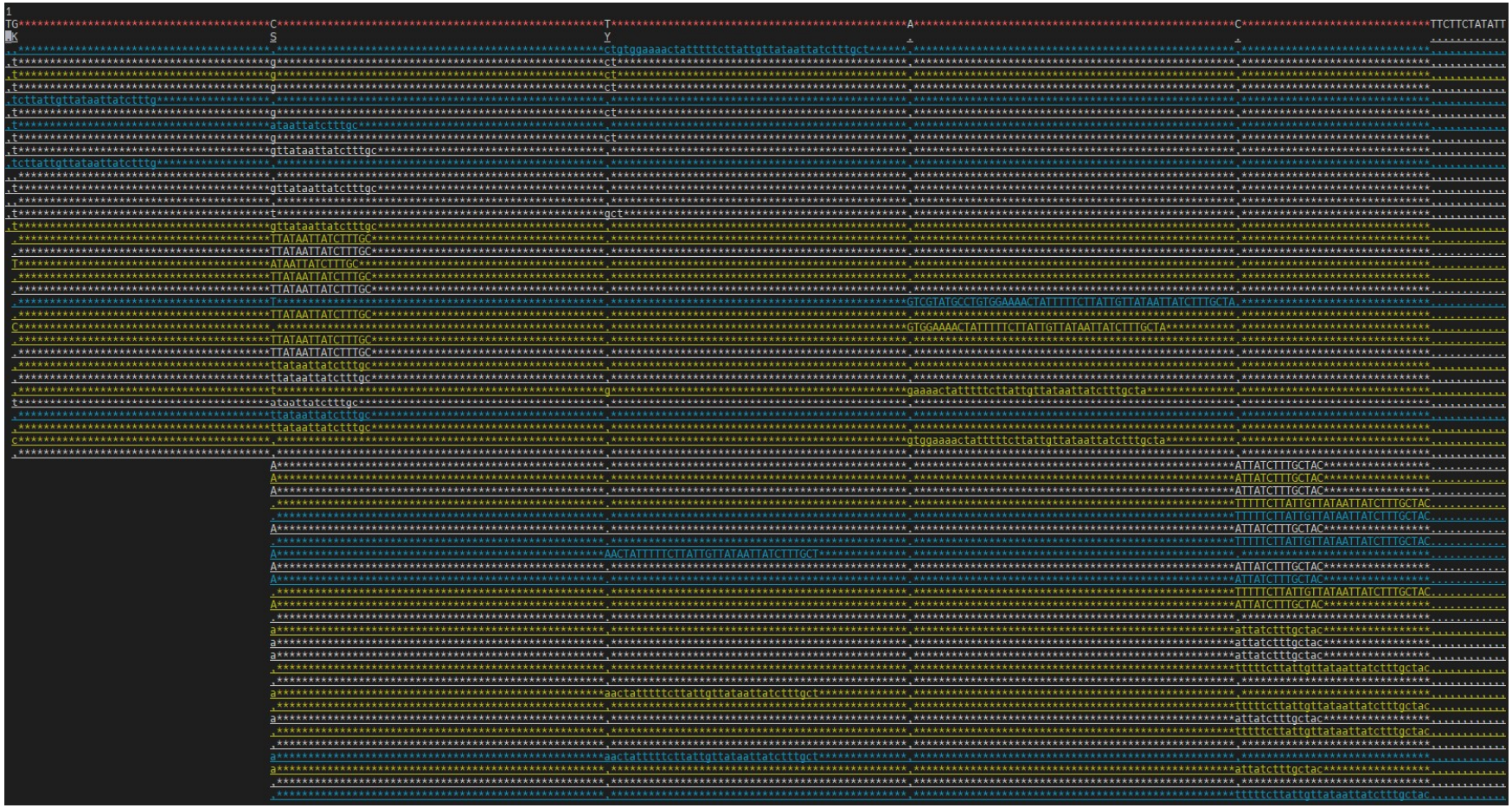


Figura 4-15. Visualización de alineamiento mediante tview

Comando: samtools tview align.sorted.bam reference.fa (http://www.htslib.org/doc/samtools-tview.html) (http://samtools.sourceforge.net/tview.shtml)

Anexo D. Visualización del alineamiento mediante ASCIIGenome

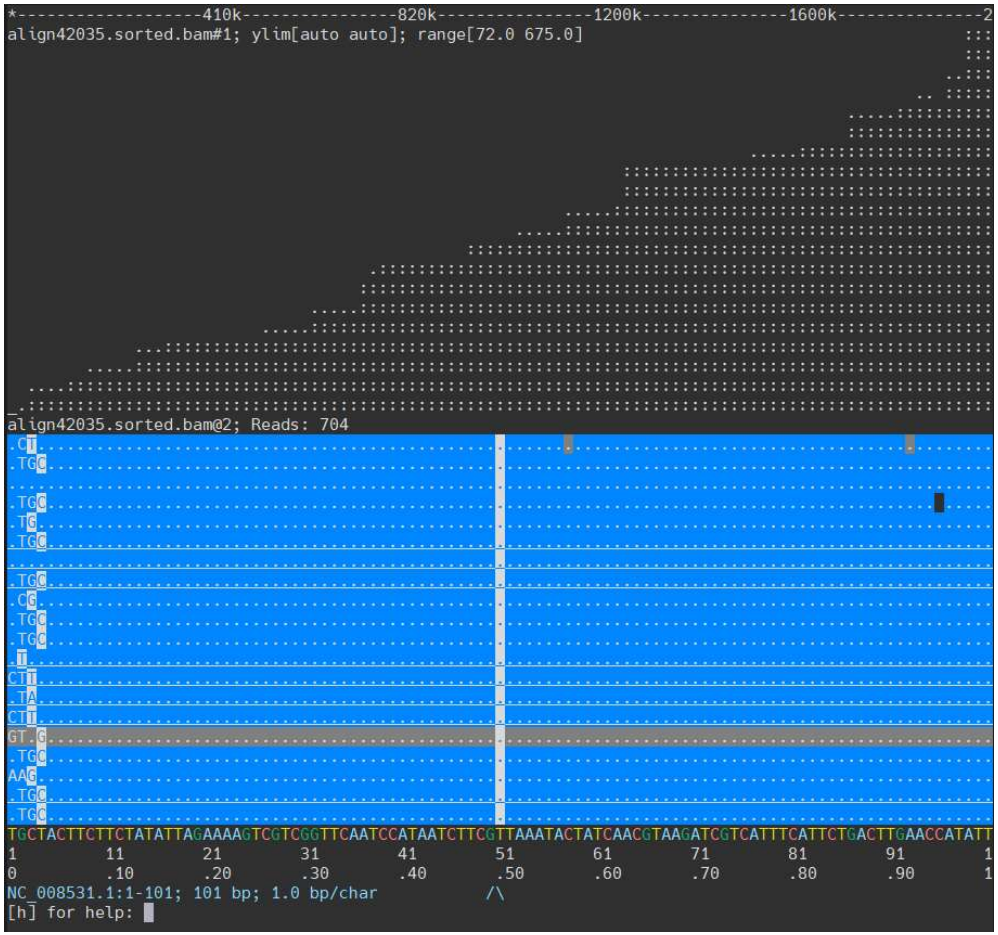


Figura 4-16. Visualización del alineamiento mediante ASCIIGenome

Comando: ASCIIGenome -fa reference.fa align.sorted.bam (https://asciigenome.readthedocs.io/en/latest/index.html)

Anexo E. Visualización del alineamiento mediante Tablet

Figura 4-17. Visualización del alineamiento mediante Tablet

(https://ics.hutton.ac.uk/tablet/)

Anexo F. Gráficos de resultados de QUAST (Quality Assessment Tool for Genome Assemblies)

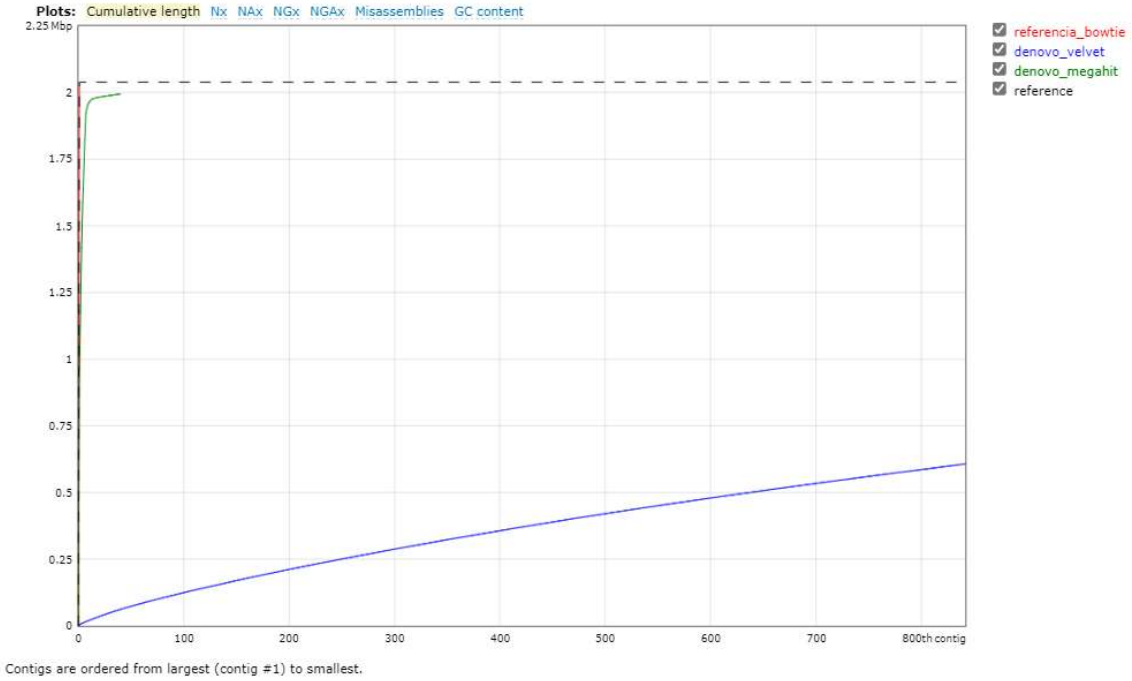


Figura 4-18. Longitud acumulada – Ordenamiento de contigs de mayor a menor.

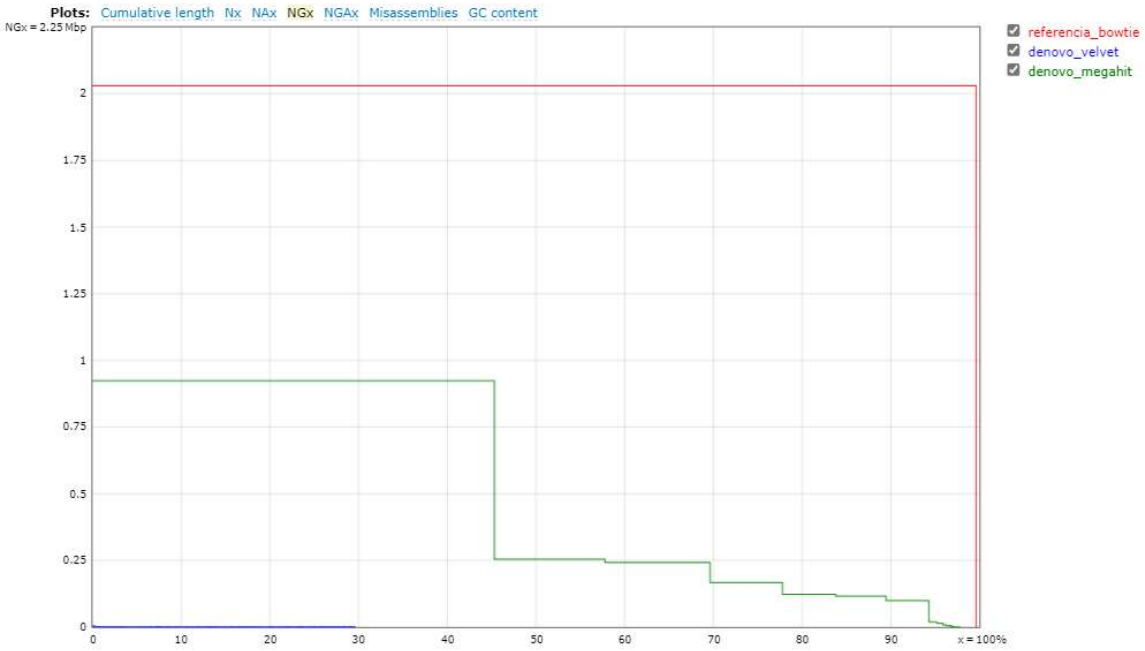


Figura 4-19. Tendencia de la métrica NGx

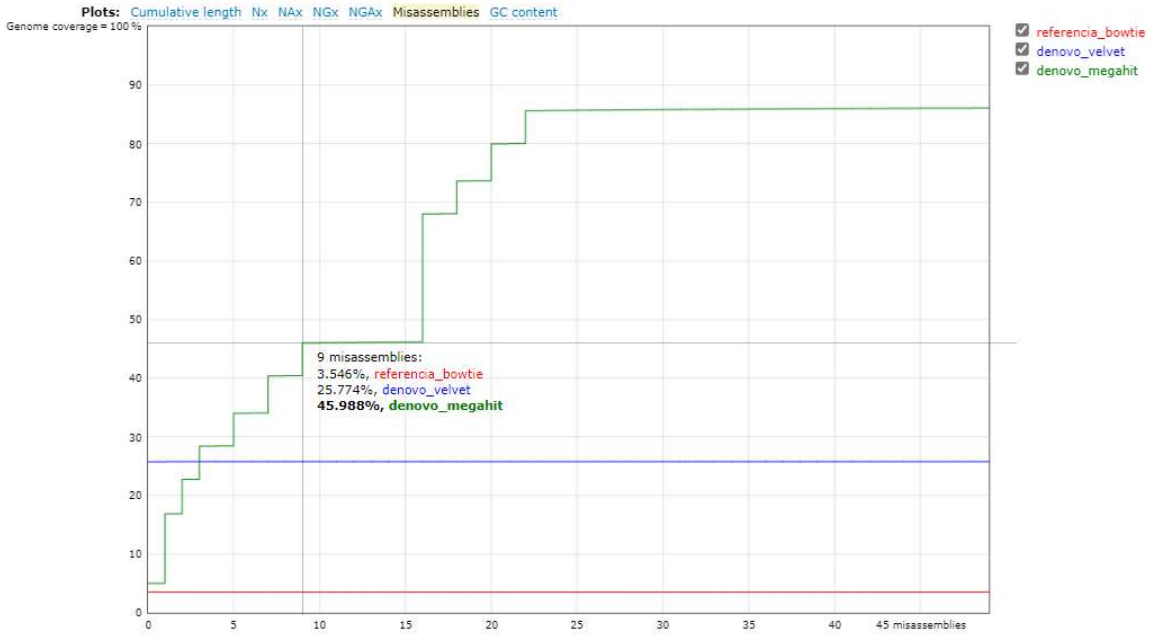


Figura 4-20. Montajes erróneos donde Y es el número total de bases alineadas dividido por la longitud de referencia, en los contigs teniendo el número total de ensamblajes incorrectos como máximo X.

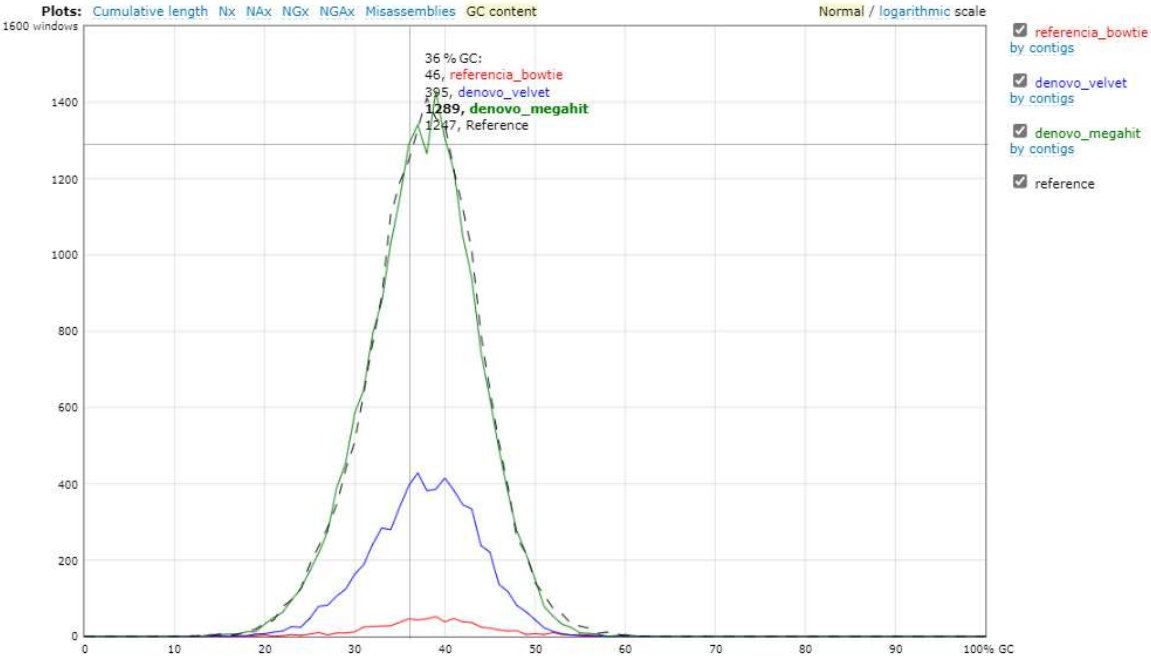


Figura 4-21. Contenido GC - Los contigs se dividen en ventanas de 100 pb no superpuestas. La gráfica muestra el número de ventanas para cada porcentaje de GC.

Anexo G. Resultados de anotación obtenido por la herramienta PATRIC (Pathosystems Resource Integration Center)

Genome ID	1245.593
Genome Name	Leuconostoc mesenteroides 91.2.98 Ensamblado
Reference Genomes	1107880.3
Coarse consistency (%)	99.9
Fine consistency (%)	99.3
Completeness (%)	99.7
Contamination (%)	1.4
Evaluation Group	R200 (Leuconostoc gasicomitatum LMG 18811)
Contig count	95
DNA size (bp)	2015455
Contigs N50 (bp)	254217

Figura 4-22. Resultados generales del proceso de anotación mediante PATRIC.

Genome Statistics	
Chromosomes	1
Contigs	1
Genome Length	1896561
GC Content	37.77

Figura 4-23. Estadísticas obtenidas del proceso de traducción del genoma ensamblado por referencia.

Genomic Features

	BV-BRC	GenBank / RefSeq
CDS	1819	1803
tRNA	69	71
pseudogene	29	0
rRNA	12	12
misc_feature	0	11

Figura 4-24. Tabla de resultados de funciones genómicas obtenidas con el genoma del *Leuconostoc mesenteroides 91.2.98*.

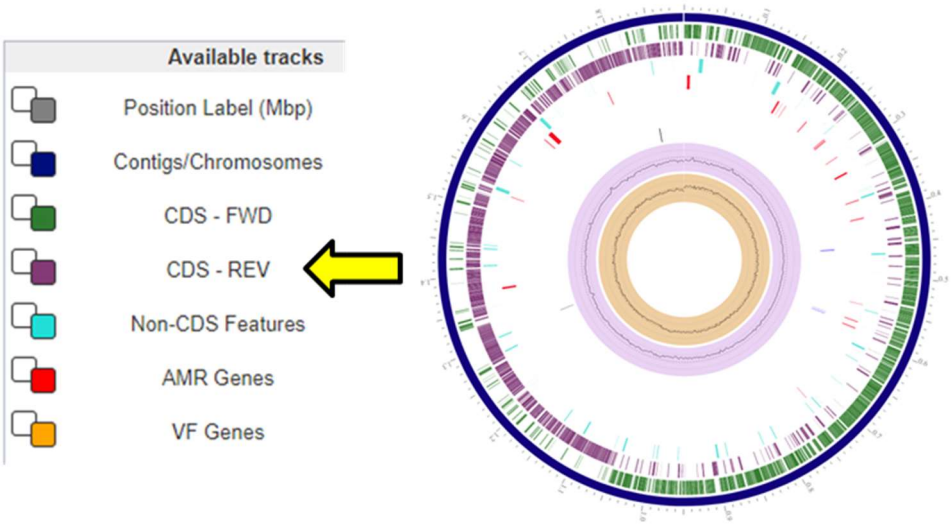


Figura 4-25. Grafica de subregiones del genoma traducido a través de la herramienta Patric.

5. Bibliografía

[1] *M. N. Ruiz, "Bioinformática: Conceptos y alcances en las fronteras de la ciencia," p. 88, 2004.*

[2] *J. A. Valverde, "Anotación de genoma," Conogasi, Conoc. para la vida, 2016.*

[3] *L. Brenes-Guillén, "Ensamblaje de genomas y anotación," vol. 22, no. 3, p. 2013, 2013.*

[4] *S. González de la Fuente, "Ensamblaje de novo y anotación génica del genoma de Leishmania major mediante secuenciación masiva," Uoc Univ. Oberta Catalunya, 2018, [Online]. Available: http://hdl.handle.net/10609/81889*

[5] *M. Naessens, A. Cerdobbel, and W. Soetaert, "Leuconostoc dextransucrasa y dextrano : producción , propiedades y aplicaciones," vol. 860, pp. 845–860, 2005.*

[6] *H. Neubauer, A. Bauché, and B. Mollet, "Molecular characterization and expression analysis of the dextransucrase DsrD of Leuconostoc mesenteroides Lcc4 in homologous and heterologous Lactococcus lactis cultures," Microbiology, vol. 149, no. 4, pp. 973–982, 2003, doi: 10.1099/mic.0.26029-0.*

[7] *E. Díaz-Montes, J. Yáñez-Fernández, and R. Castro-Muñoz, "Microfiltration-mediated extraction of dextran produced by Leuconostoc mesenteroides SF3," Food Bioprod. Process., vol. 119, pp. 317–328, 2020, doi: 10.1016/j.fbp.2019.11.017.*

[8] *G. S. Park, S. J. Hong, B. K. Jung, C. Lee, C. K. Park, and J. H. Shin, "The complete genome sequence of a lactic acid bacterium Leuconostoc mesenteroides ssp. dextranicum strain DSM 20484T," J. Biotechnol., vol. 219, pp. 3–4, 2016, doi: 10.1016/j.jbiotec.2015.12.009.*

[9] *F. Chen, G. Huang, and H. Huang, "Preparation and application of dextran and its derivatives as carriers," Int. J. Biol. Macromol., vol. 145, pp. 827–834, 2020, doi: 10.1016/j.ijbiomac.2019.11.151.*

[10] *V. Monchois, R. Willemot, and P. Monsan, "Glucansucrasas : mecanismo de acción y estructura ^ función relaciones," vol. 23, 1999.*

[11] *M. Naessens, A. Cerdobbel, W. Soetaert, and E. J. Vandamme, "Leuconostoc dextransucrase and dextran: Production, properties and applications," J. Chem. Technol. Biotechnol., vol. 80, no. 8, pp. 845–860, 2005, doi: 10.1002/jctb.1322.*

[12] *F. G. G. Yhon., "Estudio de la enzima dextransacarasa (DS) producida por Leuconostoc mesenteroides cepa IBUN 91.2.98." Bogotá, p. 46, 2014.*

[13] *"Universidad de San Carlos de Guatemala Facultad de Ingeniería Escuela de Ingeniería Química EMMETT ECHEVERRÍA VALENZUELA ASESORADO POR M . Sc . ZENÓN MUCH SANTOS," 2006.*

[14] *L. Alejandra and G. Galindo, "Caracterización molecular y funcional del gen codificante para la dextransacarasa de," pp. 1–36, 2018.*

[15] G. S. Park, S. J. Hong, B. K. Jung, C. Lee, C. K. Park, and J. H. Shin, "The complete genome sequence of a lactic acid bacterium Leuconostoc mesenteroides ssp. dextranicum strain DSM 20484T," J. Biotechnol., vol. 219, pp. 3–4, 2016, doi: 10.1016/j.jbiotec.2015.12.009.

[16] B. H. Chun, K. H. Kim, H. H. Jeon, S. H. Lee, and C. O. Jeon, "Pan-genomic and transcriptomic analyses of Leuconostoc mesenteroides provide insights into its genomic and metabolic features and roles in kimchi fermentation," Sci. Rep., vol. 7, no. 1, pp. 1–16, 2017, doi: 10.1038/s41598-017-12016-z.

[17] W. Ruppitsch et al., "Genetic diversity of leuconostoc mesenteroides isolates from traditional montenegrin brine cheese," Microorganisms, vol. 9, no. 8, pp. 1–16, 2021, doi: 10.3390/microorganisms9081612.

[18] P. Zhang, P. Zhang, J. Wu, D. Tao, and R. Wu, "Effects of Leuconostoc mesenteroides on physicochemical and microbial succession characterization of soybean paste, Da-jiang," Lwt, vol. 115, 2019, doi: 10.1016/j.lwt.2019.04.029.

[19] L. H. Deegan, P. D. Cotter, C. Hill, and P. Ross, "Bacteriocins: Biological tools for bio-preservation and shelf-life extension," Int. Dairy J., vol. 16, no. 9, pp. 1058–1071, 2006, doi: 10.1016/j.idairyj.2005.10.026.

[20] N. A. Vega Castro and E. A. Reyes Montaño, "Introducción al análisis estructural de proteínas y glicoproteínas," Introd. al análisis estructural proteínas y glicoproteínas, 2020, doi: 10.36385/fcbog-3-0.

[21] L. Liu et al., "Comparison of next-generation sequencing systems," J. Biomed. Biotechnol., vol. 2012, 2012, doi: 10.1155/2012/251364.

[22] J. M. Heather and B. Chain, "The sequence of sequencers: The history of sequencing DNA," Genomics, vol. 107, no. 1, pp. 1–8, 2016, doi: 10.1016/j.ygeno.2015.11.003.

[23] T. J. Treangen and S. L. Salzberg, "Repetitive DNA and next-generation sequencing: Computational challenges and solutions," Nat. Rev. Genet., vol. 13, no. 1, pp. 36–46, 2012, doi: 10.1038/nrg3117.

[24] N. B. Larson, A. L. Oberg, A. A. Adjei, and L. Wang, "A Clinician's Guide to Bioinformatics for Next-Generation Sequencing," J. Thorac. Oncol., vol. 18, no. 2, pp. 143–157, 2023, doi: 10.1016/j.jtho.2022.11.006.

[25] L. J. Fennell et al., "Comparative analysis of Illumina Mouse Methylation BeadChip and reduced-representation bisulfite sequencing for routine DNA methylation analysis," Cell Reports Methods, vol. 2, no. 11, p. 100323, 2022, doi: 10.1016/j.crmeth.2022.100323.

[26] Y. Guo et al., "Metagenomic next-generation sequencing to identify pathogens and cancer in lung biopsy tissue," EBioMedicine, vol. 73, p. 103639, 2021, doi: 10.1016/j.ebiom.2021.103639.

[27] M. Yermagambetova, S. Abugalieva, Y. Turuspekov, and S. Almerekova, "Illumina

sequencing data of the complete chloroplast genome of rare species Juniperus seravschanica (Cupressaceae) from Kazakhstan," Data Br., vol. 46, p. 108866, 2023, doi: 10.1016/j.dib.2022.108866.

[28] T. Soni, R. Pandit, D. Blake, C. Joshi, and M. Joshi, "Comparative analysis of two next-generation sequencing platforms for analysis of antimicrobial resistance genes," J. Glob. Antimicrob. Resist., vol. 31, pp. 167–174, 2022, doi: 10.1016/j.jgar.2022.08.017.

[29] Z. Liang et al., "Combined Illumina and Pacbio sequencing technology on transcriptome analysis reveals several key regulations during the early development of American shad (Alosa sapidissima)," Aquac. Reports, vol. 25, no. July, p. 101264, 2022, doi: 10.1016/j.aqrep.2022.101264.

[30] L. Aguilar-Bultet and L. Falquet, "Secuenciación y ensamblaje de novo de genomas bacterianos: una alternativa para el estudio de nuevos patógenos," Rev. Salud Anim, vol. 37, no. 2, pp. 125–132, 2015.

[31] 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, pp. 1–12, 2017, doi: 10.1186/s12859-017-1911-6.

[32] B. Wajid and E. Serpedin, "Review of General Algorithmic Features for Genome Assemblers for Next Generation Sequencers," Genomics, Proteomics Bioinforma., vol. 10, no. 2, pp. 58–73, 2012, doi: 10.1016/j.gpb.2012.05.006.

[33] S. Andrews, "Index of projects fastqc help 3 - Analysis Modules," 2010.

[34] M. J. Chaisson and G. Tesler, "Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory," BMC Bioinformatics, vol. 13, no. 1, 2012, doi: 10.1186/1471-2105-13-238.

[35] L. T. G. Navarrete, "Garcia Navarrete (2018) Estrategia computacional," Universidad Nacional de Colombia. 2018.

[36] K. J. McKernan et al., "Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding," Genome Res., vol. 19, no. 9, pp. 1527–1541, 2009, doi: 10.1101/gr.091868.109.

[37] M. Hernández, N. M. Quijada, D. Rodríguez-Lázaro, and J. M. Eiros, "Bioinformatics of next generation sequencing in clinical microbiology diagnosis," Rev. Argent. Microbiol., vol. 52, no. 2, pp. 150–161, 2020, doi: 10.1016/j.ram.2019.06.003.

[38] N. Nagarajan and M. Pop, "Sequence assembly demystified," Nat. Rev. Genet., vol. 14, no. 3, pp. 157–167, 2013, doi: 10.1038/nrg3367.

[39] S. Meader, L. D. W. Hillier, D. Locke, C. P. Ponting, and G. Lunter, "Genome assembly quality: Assessment and improvement using the neutral indel model," Genome Res., vol. 20, no. 5, pp. 675–684, 2010, doi: 10.1101/gr.096966.109.

[40] E. Port, F. Sun, D. Martin, and M. S. Waterman, "Genomic mapping by end-characterized random clones: a mathematical analysis," Genomics, vol. 26, no. 1,

pp. 84–100, 1995, doi: 10.1016/0888-7543(95)80086-2.

[41] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, "QUAST: Quality assessment tool for genome assemblies," Bioinformatics, vol. 29, no. 8, pp. 1072–1075, 2013, doi: 10.1093/bioinformatics/btt086.

[42] J. Bohlin et al., "Analysis of intra-genomic GC content homogeneity within prokaryotes," BMC Genojmics, vol. 11, no. 1, 2010, doi: 10.1186/1471-2164-11-464.