PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

DESARROLLO DE UN *PIPELINE* BIOINFORMÁTICO QUE PERMITE EL ENSAMBLAJE Y LA ANOTACIÓN DEL GENOMA DE LA BACTERIA *RICKETTSIA ASEMBONENSIS*

Tesis para obtener el título profesional de Ingeniero Informático

AUTOR:

Ronie Paolo Arauco Alarcon

ASESOR:

Dr. Edwin Rafael Villanueva Talavera

Lima, febrero, 2022

Resumen

En las últimas décadas, el surgimiento y resurgimiento de las bacterias infecciosas se han convertido en amenazas de importancia para la salud pública. Este es el caso de la bacteria de la especie *Rickettsia asembonensis* -identificada en Asembo, Kenia- que, en los últimos años, ha sido detectada en pulgas (*Ctenocephalides felis* y *Ctenocephalides canis*), en regiones anteriormente no reportadas y en casos de síndromes febriles agudos inespecíficos. Este patógeno emergente -así como muchos otros- sigue siendo relativamente desconocido. Por lo que, se convierte en una necesidad sustancial no subestimarlo y expandir su estudio no solo epidemiológico, sino también relacionado a su biología molecular.

En la actualidad, el esfuerzo científico a fin de incrementar la eficiencia de la obtención de la biología molecular de las especies a nivel global ha generado la aparición de tecnologías de secuenciación de última generación. En ese sentido, la gran cantidad de datos genómicos deben ser manipulados con técnicas bioinformáticas. Estas últimas, han permitido un mejor entendimiento y uso de los datos que generan las tecnologías de secuenciación. Siendo que, recientemente, la aplicación de protocolos y *pipelines* ha generado resultados favorables. En consecuencia, la aplicación de técnicas bioinformáticas con la finalidad de obtener la información genómica de la bacteria *R. asembonensis* representa una oportunidad para contribuir al conocimiento científico de este microorganismo.

Por lo tanto, el presente trabajo tiene como objetivo principal el ensamblaje y la anotación del genoma de la bacteria *R. asembonensis* a través de un pipeline bioinformático, que hará uso de datos secuenciados de la pulga de la especie *C. felis* positivas para *R. asembonensis*, a partir de unas muestras recolectadas en un estudio llevado a cabo en la ciudad de Iquitos. El presente trabajo generará también un precedente y referente metodológico para otras especies de interés con la misma problemática.

Dedicatoria

A mis padres, Rosana y Oscar, que siempre me apoyaron y motivaron a crecer como persona y profesional.

A mis amigos que siempre me alentaron en todo momento.

A mi asesor, Edwin, por guiarme a lo largo del desarrollo de esta tesis.

Tabla de Contenido

Índice de Figuras

Índice de Tablas

Capítulo 1. Generalidades

1.1 Problemática

El surgimiento y resurgimiento de las bacterias infecciosas se han convertido en amenazas de importancia para la salud pública en las últimas décadas (Vouga & Greub, 2016). Esto se debe a que estas bacterias se esparcen rápidamente por todo el mundo y las infecciones que producen no presentan reducciones en los índices de casos nuevos y existentes (Rowley et al., 2019; Vouga & Greub, 2016; World Health Organization, 2015). Actualmente, la propagación de las bacterias infecciosas ha incrementado en regiones endémicas[1] y en otras anteriormente no reportadas, y una de las principales causas de su esparcimiento son los insectos transmisores de patógenos, también conocidos como vectores (Cantas & Suer, 2014; Stehman et al., 2014).

Los vectores tienen el potencial de propagar las infecciones bacterianas, ya que su densidad poblacional está incrementando (Stehman et al., 2014). El cambio climático está alterando su distribución y transmisión, permitiendo a los vectores esparcir las bacterias infecciosas en zonas geográficas de todo el mundo, causando un mayor número de casos emergentes (Fang, Blanton, & Walker, 2017; Rust, 2017; Vouga & Greub, 2016). Este es el caso de gran parte de las bacterias del género *Rickettsia*, las cuales tienen presencia a nivel global debido a que sus principales vectores son las garrapatas, pulgas, ácaros y piojos (Fang et al., 2017).

En las últimas tres décadas, las enfermedades rickettsiosas se han convertido en un tema de interés en la salud pública, debido a su histórica relación con los vectores, el aumento de los casos reportados y a la amenaza de que sigan creciendo a causa del calentamiento global (Rust, 2017; Troyo, Jose, & Rica, 2018). Si bien existen especies de *Rickettsia* que

[1] La situación en la que una población de organismos de la misma especie se encuentra en una zona geográfica específica debido a diversos factores (Morrone J.J., 1994)

producen infecciones que pueden ser diferenciadas e identificadas (Faccini-Martínez, García-Álvarez, Hidalgo, & Oteo, 2014), existe un grupo cuyas implicancias en enfermedades en humanos aún no son concluyentes (Fang et al., 2017; Maina et al., 2019). Por ejemplo, la bacteria de la especie *R. asembonensis*, que fue identificada en el año 2013, en Asembo, Kenia (Jiang et al., 2013; Maina et al., 2019).

En los últimos años, la bacteria *R. asembonensis* ha sido detectada en pulgas (*Ctenocephalides felis y Ctenocephalides canis*) (Maina et al., 2019), en casos de síndromes febriles agudos inespecíficos (Palacios-Salvatierra, Cáceres-Rey, Vásquez-Domínguez, Mosquera-Visaloth, & Anaya-Ramírez, 2018; Ramal, Díaz, & López, 2007) y en regiones anteriormente no reportadas (Palacios-Salvatierra et al., 2018), por lo que es un claro ejemplo de una especie emergente de *Rickettsia* (Palacios-Salvatierra et al., 2018; Troyo et al., 2018; Vouga & Greub, 2016). Se convierte en una necesidad sustancial no subestimarla y expandir sus estudios, ya que estos permitirían generar alternativas para mejorar su diagnóstico en los casos febriles inespecíficos (Odhiambo, Maina, Taylor, Jiang, & Richards, 2014; Palacios-Salvatierra et al., 2018; Troyo et al., 2018).

Por esta razón, es sumamente importante contribuir al conocimiento científico de la bacteria *R. asembonensis*. La información epidemiológica[2] que se tiene es insuficiente para determinar el verdadero riesgo que representa esta bacteria emergente en la salud pública (Troyo et al., 2018). Asimismo, en países como Perú, esta especie de bacteria sigue siendo relativamente desconocida. Esto genera escasez de alternativas de caracterización y diferenciación de las enfermedades que este microorganismo produce (Kocher et al., 2016; Odhiambo et al., 2014). Por lo tanto, sería una gran mejora si se obtuviese un detallado conocimiento de su biología molecular para conocer más sobre esta bacteria (Loyola et al., 2018; Maina et al., 2019; Odhiambo et al., 2014).

[2] Es el estudio de la distribución y factores determinantes relacionados a la salud y el control de enfermedades.

En la actualidad, el esfuerzo científico a nivel global por hacer más eficiente la obtención de la biología molecular de las especies (Low & Tammi, 2017) ha generado la aparición de tecnologías de secuenciación de última generación (Shendure & Ji, 2008), que han permitido la reducción de costo y tiempo en comparación de técnicas anteriores (Sheridan, 2014). En consecuencia, la gran cantidad de datos genómicos como resultado de la secuenciación tienen que ser manipuladas con técnicas bioinformáticas (Zhao, Wang, Wang, Jia, & Zhao, 2013). La información significativa de estas muestras es obtenida mediante algoritmos y mientras las tecnologías de secuenciación se hagan más robustas, el análisis de datos genómicos se volverá más desafiante (Zhao, Liu, & Qu, 2017).

La bioinformática ha permitido un mejor entendimiento y uso de la cantidad de datos que generan las tecnologías de secuenciación de última generación. Recientemente, la aplicación de protocolos y flujos de trabajo (*pipelines*), en el proceso de obtención de la información genómica de las especies, ha generado resultados favorables. Por ejemplo, ahora es posible ofrecer una mejor cobertura en datos de baja calidad, una mejor anotación de genes funcionales e incluso realizar un mejor análisis metagenómico[3] (Mengoni, Galardini, & Fondi, 2015). El hecho de aplicar técnicas bioinformáticas para obtener la información genómica de la bacteria *R. asembonensis* representa una oportunidad para contribuir al conocimiento científico de este microorganismo (Odhiambo et al., 2014; Troyo et al., 2018).

El desarrollo de un *pipeline* bioinformático también ayudaría a disminuir los errores presentes en el procesamiento de datos genómicos que contienen diversas especies (Mengoni et al., 2015; Visconti, Martin, & Falchi, 2018). Uno de los desafíos en el proceso de obtención de la secuencia genómica de un organismo de interés es lidiar con la cantidad y variedad de datos de otras especies detectadas por las tecnologías de secuenciación de última

[3] La metagenómica es el estudio a profundidad del material genético de una gran cantidad de microorganismos que se encuentran en una muestra (Mengoni et al., 2015).

generación (Merchant, Wood, & Salzberg, 2014; Treangen et al., 2013). Definir de manera correcta y organizada el análisis de calidad de las secuencias que debe seguir el *pipeline* bioinformático, ayudaría a evitar desajustes con otras secuencias de referencia (Visconti et al., 2018) y minimizar los errores durante el ensamblaje de la secuencia genómica del objetivo de interés (Merchant et al., 2014).

Por lo tanto, el presente proyecto tiene como objetivo principal el ensamblaje y la anotación de la secuencia genómica de la bacteria *R. asembonensis* a través de un *pipeline* bioinformático, que hará uso de datos secuenciados de la pulga de la especie *C. felis* positivas para *R. asembonensis*. Las muestras se recolectaron en un estudio llevado a cabo en el año 2013 en la ciudad de Iquitos, Perú (Kocher et al., 2016), que determinó la proporción de enfermedades en humanos producidas por las bacterias del género *Rickettsia* e identificó vectores y reservorios de este patógeno en la Amazonía peruana (Kocher et al., 2016).

Debido a que se trata de muestras secuenciadas de un vector, se encontrará material genético diverso (principalmente bacteria *R. asembonensis*, hospederos y microbiota intestinal) (Bitam, Dittmar, Parola, Whiting, & Raoult, 2010), por lo que la eliminación de los datos de organismos distintos a la bacteria se convierte en una etapa crucial en el proceso de obtención de la secuencia genómica de *R. asembonensis* (Merchant et al., 2014).

El presente trabajo, desarrollará un *pipeline* bioinformático con el fin de generar una secuencia genómica anotada y que pueda ser publicada en GenBank (base de datos que funciona como banco de genes) para investigaciones posteriores. El presente trabajo generará también un precedente y referente metodológico para otras especies de interés con la misma problemática.

1.2 Objetivos

1.2.1 Objetivo general

Desarrollar un *pipeline* bioinformático que permita el ensamblaje y la anotación de la secuencia genómica de la bacteria *R. asembonensis*

1.2.2 Objetivos específicos

O1: Implementar un flujo de preprocesamiento de los datos secuenciados de la bacteria *R. asembonensis*.

O2: Definir la secuencia genómica consenso de la bacteria *R. asembonensis*.

O3: Identificar los genes de la secuencia genómica de la bacteria *R. asembonensis* obtenida en O2.

1.2.3 Resultados esperados

R1: Flujo de trabajo para el preprocesamiento de los datos secuenciados (O1).

R2: Datos preprocesados de la bacteria *R. asembonensis* (O1).

R3: Flujo de trabajo para el ensamblaje de los datos preprocesados (O2).

R4: Secuencia genómica consenso de la bacteria *R. asembonensis* (O2).

R5: Flujo de trabajo para la anotación de genes (O3).

R6: Secuencia genómica anotada de la bacteria *R. asembonensis* (O3).

1.2.4 Mapeo de objetivos, resultados esperados, herramientas y métodos

Tabla 1

Mapeo de objetivos, resultados esperados, herramientas y métodos (elaboración propia)

Objetivo 1: Implementar un flujo de preprocesamiento de los datos secuenciados de la bacteria *R. asembonensis*.			
Resultado	Meta física	Medio de verificación	Herramientas o métodos
Flujo de trabajo para el preprocesamiento de los datos secuenciados	Documento	Resultados de las revisiones que ofrece la literatura.	• Análisis metagenómico
Datos preprocesados de la bacteria *R. asembonensis*	Conjunto de datos	Reporte de calidad de las secuencias.	• FastQC • Fastp • BBDuk • Bowtie2
Objetivo 2: Definir la secuencia genómica consenso de la bacteria *R. asembonensis*.			
Resultado	Meta física	Medio de verificación	Herramientas o métodos
Flujo de trabajo para el ensamblaje de los datos preprocesados	Documento	Resultados de las revisiones que ofrece la literatura.	• Análisis metagenómico
Secuencia genómica consenso de la bacteria *R. asembonensis*	Conjunto de datos	Reporte de métricas de desempeño de los ensambladores.	• Megahit • Ray Meta • Abyss • Quast • Samtools • BLAST
Objetivo 3: Identificar los genes de la secuencia genómica de la bacteria *R. asembonensis* obtenida en O2.			
Resultado	Meta física	Medio de verificación	Herramientas o métodos
Flujo de trabajo para la anotación de genes	Documento	Resultados de las revisiones que ofrece la literatura.	• Análisis metagenómico
Secuencia genómica anotada de la bacteria *R. asembonensis*	Conjunto de datos	Comparación con secuencia genómica de referencia de *R. asembonensis.*	• NCBI Prokaryotic Genome Annotation Pipeline

1.3 Herramientas y Métodos

En esta sección se detallan las herramientas y métodos que se utilizan en el presente proyecto.

1.3.1 Descripción de herramientas y métodos

Análisis metagenómico

Metagenómica es el conjunto de técnicas que tienen como fin estudiar a los organismos encontrados en una muestra. Con los últimos esfuerzos científicos, se han acoplado nuevas tecnologías que han ayudado a realizar un análisis correcto y efectivo (Garrido-Cardenas & Manzano-Agugliaro, 2017). El presente caso de estudio exige la manipulación de datos que contienen diversos organismos, ya que el comportamiento por naturaleza del vector consiste en desplazarse entre sus hospederos. Por lo tanto, resulta necesario un análisis metagenómico previo con el objetivo de lidiar este reto.

FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

Es una herramienta que brinda información sintetizada sobre la calidad de la secuencia. Es decir, facilita la detección de problemas en las secuencias a través de gráficos y datos estadísticos (Low & Tammi, 2017).

Fastp (https://github.com/OpenGene/fastp)

Ofrece funcionalidades que comúnmente son propias de otras herramientas de preprocesamiento de datos secuenciados (Chen, Zhou, Chen, & Gu, 2018). Se caracteriza por ser rápida en comparación con las demás y cubre la mayoría de las necesidades de los usuarios en la etapa de preprocesamiento.

BBDuk (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/)

Es una herramienta que pertenece al paquete de herramientas de BBTools. Ofrece parámetros flexibles y la variedad de funcionalidades la convierte en una herramienta versátil en la etapa de preprocesamiento.

Bowtie2 (https://github.com/BenLangmead/bowtie2)

Es una herramienta de alineación a secuencias de referencia caracterizada por ser rápida y eficiente en memoria. Los datos de entrada pueden ser secuencias muy cortas hasta muy largas, dando soporte a tecnologías de NGS de segunda y tercera generación.

Megahit (https://github.com/voutcn/megahit)

Megahit es una herramienta de ensamblaje para conjuntos de datos metagenómicos grandes y complejos. Fue desarrollada por el mismo grupo de investigadores de las herramientas SOAPdenovo y SOAPdenovo 2, por lo que es considerada su predecesora. Megahit utiliza un "grafo *de Brujin* sucinto", una estructura basada en este grafo, pero más liviana. Por esta razón, es capaz de lidiar con el consumo exigente de recursos computacionales que realiza la mayoría de ensambladores metagenómicos. El enfoque que utiliza es similar a IDBA-UD, otra herramienta de ensamblaje, ya que itera el ensamblaje en un rango de *k-mers* en orden creciente. A diferencia de esta herramienta, Megahit elimina los *k-mers* que no son "sólidos" y de esta forma reduce el consumo de memoria. Para que un *k-mer* sea sólido debe repetirse más de una cantidad de veces establecida (el valor predeterminado es 2). Para no perder datos relevantes, utiliza una estrategia llamada "compasión de *k-mers*", que consiste en volver a considerar los *k-mers* eliminados si estos están contenidos en *k-mers* sólidos y si son necesarios para conectar estos *k-mers* sólidos. Esto contrarresta el consumo de memoria y la pérdida de información.

Ray Meta (https://github.com/sebhtml/ray)

Ray Meta es una herramienta de ensamblaje capaz de ser utilizada de forma distribuida. Este ensamblador está basado en el grafo *de Brujin* y su enfoque

principal consiste en obtener patrones confiables del grafo a partir de las coberturas de *k-mers* mínimas y promedio. Estos patrones semillas necesitan ser alargados y utiliza un algoritmo *greedy* para poder extenderlos. Si ocurre el caso de que existen dos o más patrones subsecuentes al patrón semilla, se escoge el más confiable y el patrón alargado resultante pasa a ser el patrón semilla. Si no existe ningún patrón subsecuente suficientemente confiable, el alargamiento se detiene y ambos patrones son considerados *contigs* separados. De esta forma, Ray Meta se asegura de construir múltiples *contigs* confiables de los genomas más abundantes del conjunto de datos.

Abyss (https://github.com/bcgsc/abyss)

Abyss es una herramienta de ensamblaje que puede ser utilizado para genomas largos y secuencias unicelulares. Su algoritmo se ejecuta en dos etapas, la primera consiste en generar todos los *k-mers* de las secuencias de entrada y empezar a generar *contigs* verificando y removiendo los errores que se van generando en el ensamblaje. La segunda etapa, utiliza información adicional de las secuencias para poder resolver ambigüedades al momento de ensamblar los *contigs*.

Quast (http://bioinf.spbau.ru/metaquast)

Quast es una herramienta que evalúa y compara los ensambladores con métricas descriptivas. Una de las virtudes de Quast, es su capacidad de procesar datos metagenómicos. En los últimos años, los datos producidos por las tecnologías de última generación tienden a contener una gran variedad de fragmentos que pertenecen a diferentes organismos y esta herramienta puede lidiar con ello (Mikheenko, Saveliev, & Gurevich, 2016).

Samtools (https://github.com/samtools/samtools)

Samtools es una herramienta para leer, escribir, editar, indexar y visualizar archivos de tipo SAM/BAM/CRAM. Se caracteriza por ser eficiente en el uso de recursos computacionales y muy rápida.

BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi)

BLAST es una herramienta de búsqueda de regiones de similitud entre secuencias: una considerada *query* y la otra *subject*. Es posible realizar búsquedas genómicas, proteómicas y transcriptómicas, y posee un sistema web fácil de usar que está al alcance de todos los investigadores.

NCBI Prokaryotic Genome Annotation Pipeline (https://github.com/ncbi/pgap)

Las últimas tecnologías de secuenciación han exigido a los científicos a realizar estudios de secuencias largas y diversas. Con el objetivo de lidiar con el desafío de analizar estos datos eficientemente, NCBI y Georgia Tech desarrollaron un *pipeline* para la anotación de genes de las secuencias procariotas. Esta herramienta combina la predicción de genes sin información a través del algoritmo *ab initio* y el uso de métodos de homología. Asimismo, en las últimas versiones se han añadido mejoras con Modelos Ocultos de Markov y bases de datos de proteínas curadas (por ejemplo, TIGRFAMS).

Capítulo 2. Marco Conceptual

En este capítulo se abordan conceptos necesarios para el entendimiento de la problemática y desarrollo de la tesis. Además, debido a que el presente trabajo contempla distintas áreas, también se mencionan definiciones sobre temas de Biología y Bioinformática.

2.1 Biología

2.1.1 La bacteria *R. asembonensis*

Es una bacteria Gram negativa intracelular de la orden Rickettsiales (familia Rickettsiaceae) y es parte del grupo transicional de Rickettsias. Está relacionada estrechamente con la bacteria *R. felis*, pero mediante una tipificación de genes realizada en el año 2013 se demostró que es suficientemente distinta y se consideró una especie aparte. Los organismos similares a *R. felis* (RFLO: *Rickettsia felis-like organisms*) se encuentran en vectores como pulgas, garrapatas, ácaros y moscas tsé-tsé, los cuales son muy poco estudiados y el conocimiento de su biología es escaso (Maina et al., 2019). Hasta el año 2018 la patogenicidad de la bacteria *R. asembonensis* era desconocida, pero ese mismo año un artículo peruano confirmó su relación con síndromes febriles agudos inespecíficos en cuatro regiones del territorio de la selva peruana (Palacios-Salvatierra et al., 2018).

2.1.2 El ciclo biológico de la pulga (Hospedero de la *R. asembonensis*)

Las pulgas son insectos que succionan sangre y se caracterizan por ser pequeñas, no tener alas y saltar. Su ciclo de vida consiste en 4 etapas: huevo, larva, pupa y adultez (Rozendaal, 1997). Las larvas miden de 4 a 10 mm y son blancas. No tienen patas, pero tienen gran movilidad. Estas se alimentan de heces del hospedero, insectos más pequeños o sangre no digerida y desechada por las adultas. Luego del periodo larvario, la larva se cubre en un capullo y entra en el ciclo pupa. Su textura es pegajosa y es capaz de camuflarse porque se cubre de polvo, arena y otras partículas finas. Luego de 1-2 semanas, la pulga alcanza la adultez y sale del capullo a través de estímulos como la vibración producida por el

movimiento del hospedero. Si no hay estímulo, pueden estar dentro del capullo hasta por 1 año y es muy común que personas que se mudan a una casa desocupada puedan ocasionar que estas salgan simultáneamente y ataquen a personas o animales en grandes cantidades (Rozendaal, 1997). La pulga adulta tiene patas bien desarrolladas que están adaptadas para saltar y logran alcanzar una altura de hasta 30 cm. Su color varía de marrón claro a oscuro y miden de 1 a 4 mm. Las pulgas hembra y macho se alimentan de sangre y se aparean en lugares de descanso del hospedero (polvo, tierra, basura, grietas de pisos o paredes, alfombra, madriguera de animales o nidos de pájaros). Estos organismos evitan la luz, se desarrollan en humedad y son encontradas mayormente en cabello, pelaje de animales, ropa o camas. Si es posible, estas pueden succionar sangre repetidas veces al día y son capaces de sobrevivir sin alimento por varios meses (Rozendaal, 1997). Por lo general, pican a mamíferos, pero también son capaces de picar aves. Si bien sus picaduras causan irritación, incomodidad y pérdida de sangre, también es una vía de transmisión de patógenos. Por lo general, los humanos son picados por pulgas de la especie *Ctenocephalides felis*, también llamada pulga de gato. Suelen saltar desde el suelo, por lo que es común que ataquen la zona inferior del cuerpo (Rozendaal, 1997). Esta especie es muy común en perros y gatos, pero también infestan zarigüeyas, mapaches y ratas. Por esta razón, se pueden encontrar en gran cantidad en los hogares de todo el mundo (Bitam et al., 2010).

2.1.3 Genética

La genética es el estudio de los rasgos biológicos que se heredan, incluso aquellos que son influenciados en parte por el medio ambiente. Por ejemplo, el peso es una característica heredada. Sin embargo, también es influenciado por la cantidad y calidad de comida que se consume o por el estilo de vida que se lleva (Hartl & Ruvolo, 2011).

2.1.4 Ácido desoxirribonucleico

El ácido desoxirribonucleico (ADN) es el material genético que contiene las instrucciones para producir los diferentes procesos biológicos que sustentan la vida. Es fundamental para la existencia de todos los organismos, porque se encuentran todas las características heredadas de la mayoría de los seres vivos y porque su producto final son las proteínas (Hartl & Ruvolo, 2011). Químicamente, está compuesto por 4 moléculas: adenina (A), timina (T), citosina (C) y guanina (G). Estas también son llamadas bases nitrogenadas que al añadirle azúcar y fósforo forman nucleótidos. El modelo de ADN está representado por dos tiras llamadas hebras, que están direccionadas de forma opuesta y al entrecruzarse forman una doble hélice como se puede observar en la Figura 1. Sus hebras están compuestas por nucleótidos y están unidas por las leyes de Watson y Crick. Estas leyes son las siguientes: la Adenina se enlaza con la Timina y la Citosina con la Guanina. El enlazamiento de estas moléculas es llamado par base (Dale & Schantz, 2002).

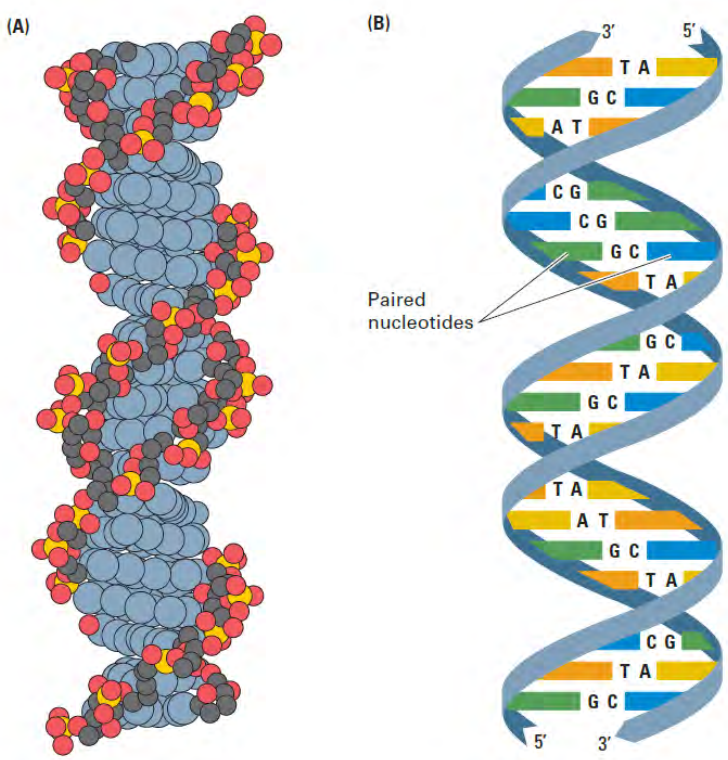

Figura 1. ADN. (A) Modelo donde los átomos son representados como esferas. (B) Modelo doble hélice de hebras direccionadas opuestamente y compuestas por nucleótidos que forman pares base.

Adaptado de Hartl & Ruvolo (2011) Genetics.

2.1.5 ARN

El ácido ribonucleico (ARN) contiene material genético al igual que el ADN, pero con dos diferencias químicas. Primero, el azúcar dentro del ARN es ribosa en lugar de desoxirribosa y, segundo, contiene uracilo (U) en vez de timina (T) (Hartl & Ruvolo, 2011). Estas diferencias tienen un efecto fuerte en el ARN, ya que no es tan estable como el ADN. El modelo de ARN en las células está representado por una sola hélice (Figura 2) y su función principal es la de sintetizar proteínas. Se le considera al ARN el primer producto del genoma y el proceso por el cual el ARN es producido se llama transcripción (Brown, 2006).

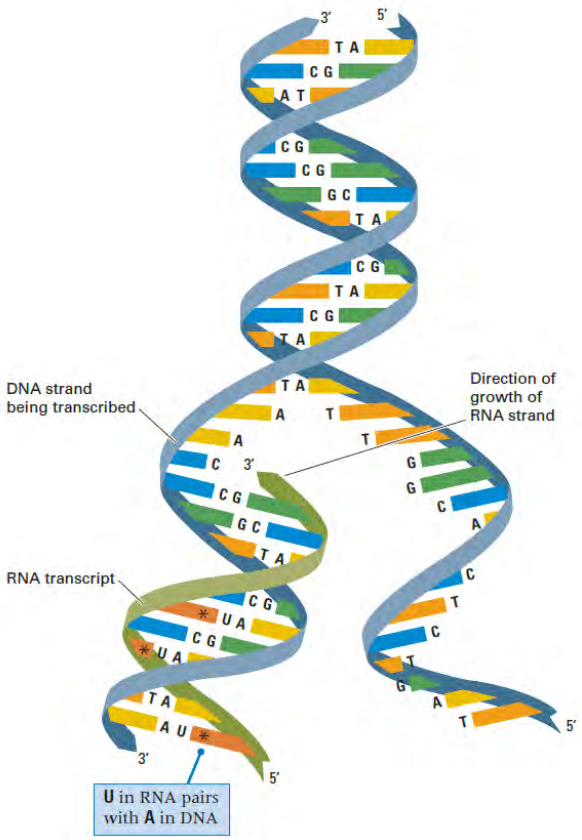

Figura 2. Se muestra el proceso de transcripción, es decir, la producción de la molécula de ARN que contiene uracilo en vez de timina y solo posee una hélice.

Adaptado de Hartl & Ruvolo (2011) Genetics.

2.1.6 Proteína

El producto final del complejo y diverso ADN es la proteína. Es una macromolécula muy importante debido a que ejecuta casi todas las actividades bioquímicas de la célula. De hecho, la célula misma está hecha de proteínas. Esto incluye su estructura que otorga rigidez y movilidad, los poros en la membrana que controla el tráfico de moléculas dentro y fuera de esta, y receptores que regulan sus actividades. Las proteínas son importantes porque intervienen en las actividades metabólicas y otorgan energía a la célula (Hartl & Ruvolo, 2011). Químicamente, la proteína, al igual que el ADN, es un polímero, pero, a diferencia de este, sus monómeros no son nucleótidos, sino aminoácidos. Estos polímeros son llamados también polipéptidos y su longitud no supera las 2000 unidades. El proceso por el cual las

proteínas son producidas se llama traducción y se puede apreciar en la Figura 3 (Brown, 2006).

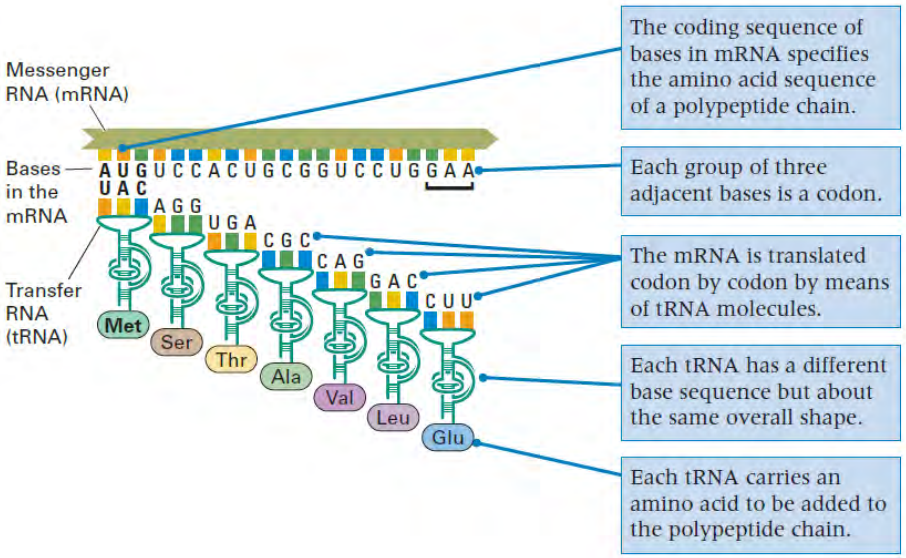

Figura 3. Se muestra la molécula de ARN siendo traducida a bloques llamados aminoácidos. El producto final es el conjunto de aminoácidos que forman una proteína.

Adaptado de Hartl & Ruvolo (2011) Genetics.

2.1.7 Gen

Desde los inicios de la Genética, se concibió al gen (Figura 4) como la unidad de herencia de una característica observable (Dale & Schantz, 2002). Sin embargo, desde 1960 se comparte una definición más específica. Un gen es una secuencia lineal (no necesariamente contigua) de ADN que codifica la cadena de moléculas que conforman una proteína, también llamados aminoácidos, y que en un plazo determinado produce efectos en las características de herencia de un organismo. (Portin & Wilkins, 2017). En pocas palabras, es una secuencia de nucleótidos en el ADN que determina las características bioquímicas de las células y organismos. Si se representa a la secuencia de nucleótidos del ADN como una cadena de caracteres en una hoja de papel, un gen está hecho de palabras distintas que pueden formar oraciones o párrafos que dan sentido al conjunto de caracteres. (Hartl & Ruvolo, 2011).

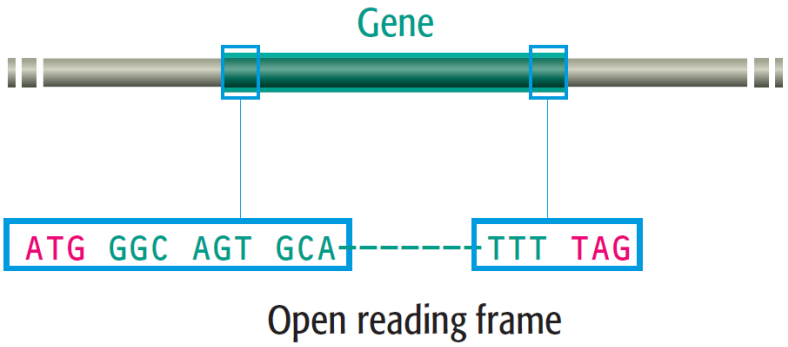

Figura 4. Representación de una secuencia de nucleótidos, también llamado gen, limitado por dos grupos llamados codones de inicio y codones de terminación.

Adaptado de Brown (2006) Genomes 3.

2.1.8 Pseudogen

Los pseudogenes son genes que ya no se encuentran activos debido a una mutación o una adjunción anormal en la expresión de un gen. Esto implica que es la copia no funcional de un gen y es considerado una reliquia, aparte de ser una muestra de que los genomas se encuentran en constante cambio (Brown, 2006).

2.1.9 Genoma/Secuencia genómica

Se considera genoma a toda la información biológica que tiene un organismo. La mayoría de los genomas, incluyendo el ser humano y toda forma de vida celular, está compuesto por la totalidad de ADN (Brown, 2006) que se encuentra en una célula, ya sea en los cromosomas o en organelos llamados mitocondrias (Figura 5). Por ejemplo, si se habla del genoma humano, este se refiere a la totalidad de ADN presente en una célula reproductiva normal. En los últimos años, se ha perfeccionado la tecnología para obtener la secuencia completa de ADN de una especie de una forma rápida y eficiente. Por tanto, actualmente es posible saber el genoma de muchos organismos vivientes (Hartl & Ruvolo, 2011).

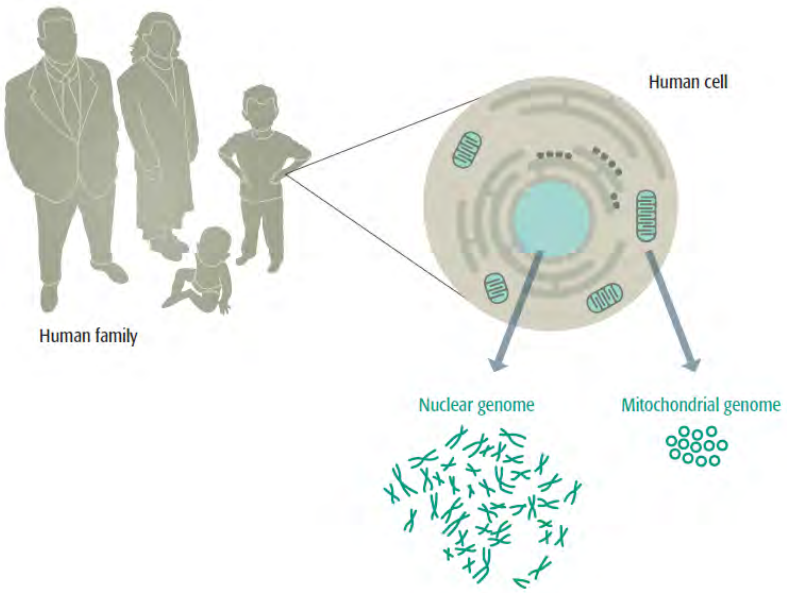

Figura 5. Representación de una secuencia de nucleótidos, también llamado gen, limitado por dos grupos llamados codones de inicio y codones de terminación.

Adaptado de Brown (2006) Genomes 3.

2.1.10 Genómica

Se encarga de determinar y analizar la secuencia completa de ADN de un organismo, es decir, de un genoma (Pevsner, 2015). Por lo tanto, es el estudio de la totalidad de los genes de un organismo para comprender su organización, función, interacción y evolución molecular (Hartl & Ruvolo, 2011).

2.1.11 Secuenciación de última generación

Debido a que los científicos siempre han estado en la búsqueda de descifrar la estructura del ADN, surgió la necesidad de desarrollar herramientas y métodos para poder leer estas secuencias. En 1977, Sanger y sus colegas desarrollaron un método para secuenciar el ADN. Si bien se fue mejorando con el paso de los años, los problemas principales eran su bajo rendimiento y alto costo. Luego de que se terminara de secuenciar el genoma humano en el 2001, los científicos siguieron buscando alternativas para mejorar la eficiencia de este método. La secuenciación de última generación (o NGS, por sus siglas en inglés), nació como

solución a las dos principales desventajas de sus predecesoras: rendimiento y costo. Es una tecnología que permitió la secuenciación de grandes cantidades de ADN en forma paralela y, actualmente, sigue en constante evolución, como se muestra en la Figura 6 (Low & Tammi, 2017).

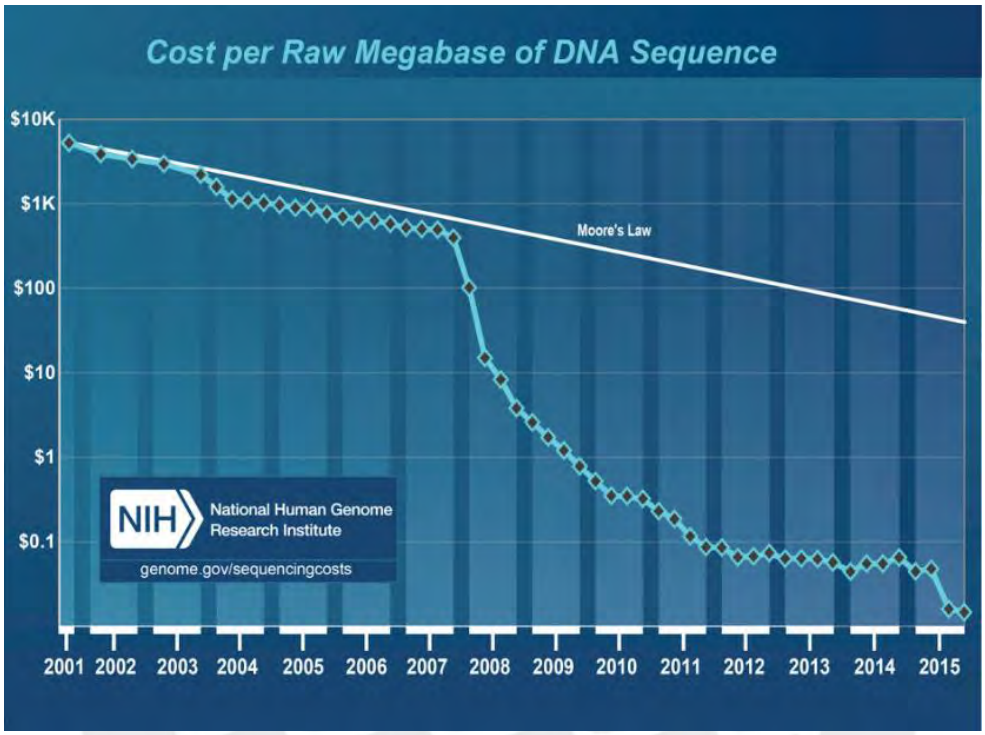

Figura 6. Gráfica que muestra la reducción del costo de la secuenciación de ADN a una calidad específica.

Adaptado de Low & Tammi (2017) Bioinformatics: A Practical Handbook of Next Generation Sequencing and Its Applications.

2.2 Bioinformática

2.2.1 Preprocesamiento

En los últimos años, el avance de la secuenciación de alto rendimiento ha permitido que la secuencia de genomas disminuya en costo y aumente en calidad. Sin embargo, para poder obtener información de las variables finales, es necesario realizar previamente la identificación de elementos considerados innecesarios y sin utilidad, ya que, al fin y al cabo, la secuencia inicial cruda es obtenida de una máquina que puede ser afectada por agentes externos. El preprocesamiento consiste en la remoción de adaptadores, duplicados y contaminación. Por esta razón, es importante el uso correcto de técnicas específicas (Wright, Gola, & Ziegler, 2017).

2.2.2 Ensamblaje

Si bien las tecnologías de secuenciación de alto rendimiento han permitido generar secuencias de una forma eficiente, estas solo son capaces de leer secuencias de aproximadamente 5000 nucleótidos. En contraste, hasta los genomas de los virus más pequeños están compuestos de varios miles de nucleótidos y los genomas más complejos, como el de los mamíferos, están compuestos por millones. Por lo tanto, se necesitan métodos computacionales que permitan la unión de estas pequeñas secuencias o fragmentos. El ensamblaje consiste en el desarrollo de un algoritmo que permita la unión de estos fragmentos de modo que produzca una secuencia consenso bien estructurada en el tiempo razonable (Masoudi-Nejad, Narimani, & Hosseinkhan, 2013).

2.2.3 Anotación

Luego del preprocesamiento y ensamblaje, se obtiene una secuencia ADN consenso. La anotación consiste en la inspección de la secuencia para identificar y localizar genes, ya que estos no se encuentran esparcidos aleatoriamente, sino están puestos de tal forma que codifican características importantes del organismo (Brown, 2006). Este procedimiento se ha

llevado a cabo varios años de forma experimental. Sin embargo, en la nueva generación de la biología molecular, este se convierte en un problema computacional en el que se deben construir métodos de identificación de los límites de los genes (Frishman & Valencia, 2008).

2.2.4 El problema del ensamblaje del genoma

Para poder entender el problema del ensamblaje del genoma, se propone una analogía llamada "El problema del periódico". Se tienen varias copias del periódico New York Times junto a una pila de dinamita a punto de estallar. Luego de la explosión se forman pequeños rezagos de papel y el problema es el siguiente: ¿qué decía el periódico? La única información que se tiene es la superposición de las pequeñas piezas de papel que se formaron tras la explosión como se muestra en la Figura 7 (Phillip & Pavel, 2015).

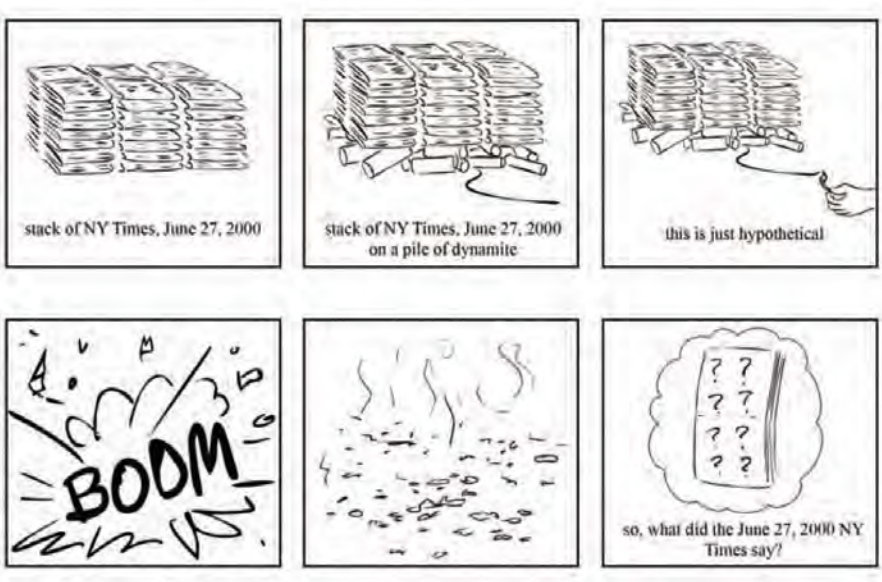

Figura 7. Analogía del problema del ensamblaje: El problema del periódico.
Tomado de Phillip & Pavel (2015) Bioinformatics Algorithms: An Active Learning Approach.

Análogamente, para obtener el genoma de una especie se utilizan varias copias de secuencias de ADN (Figura 8) que son particionadas en lecturas pequeñas llamadas *reads*. Las últimas tecnologías de secuenciación establecen una longitud determinada a los *reads* y es posible llamarlos *k-mers* para describir su longitud en base a k nucleótidos. Al igual que el

ejemplo mencionado, la única información que se tiene es la superposición de los nucleótidos de los *reads*, ya que incluso, no es posible saber la posición de cada una de estas lecturas (Phillip & Pavel, 2015).

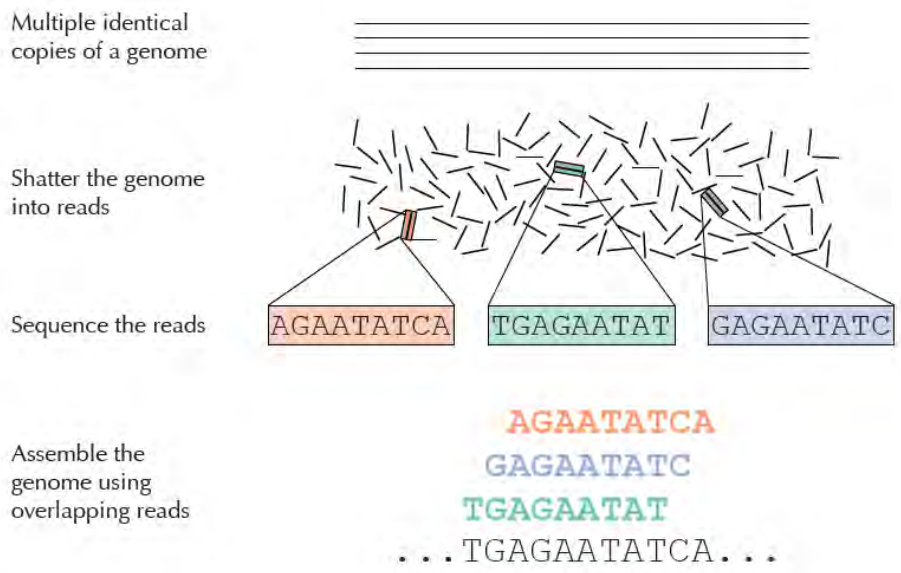

Figura 8. Problema del ensamblaje del genoma: *reads* que están dispersos y que se busca ensamblar para formar el genoma.

Tomado de Phillip & Pavel (2015) Bioinformatics Algorithms: An Active Learning Approach.

2.2.5 ¿Cómo se ensambla una secuencia?

El ensamblaje de secuencias de nucleótidos se relaciona con el problema de la reconstrucción de una cadena de caracteres. Es decir, unir subgrupos de k caracteres (*k-mers*) que pertenecen a una cadena larga con el fin de obtener su forma original (Phillip & Pavel, 2015). En la Figura 9 se puede apreciar una secuencia y sus sub grupos de 3 caracteres.

$$\text{COMPOSITION}_3(\text{TATGGGGTGC}) = \{\text{ATG}, \text{GGG}, \text{GGG}, \text{GGT}, \text{GTG}, \text{TAT}, \text{TGC}, \text{TGG}\}.$$

Figura 9. Cadena de caracteres y sus *k-mers*, donde k es igual a 3.

Tomado de Phillip & Pavel (2015) Bioinformatics Algorithms: An Active Learning Approach.

Para resolver este problema es posible colocar cada sub cadena en los nodos de un grafo y recorrer cada nodo a lo más una vez (Camino Hamiltoniano). Sin embargo, actualmente esta solución no tiene un algoritmo eficiente. Por otro lado, en 1946, Nicolaas de Brujin resolvió este problema con una cadena de números binarios y la hizo universal. Propuso colocar las sub cadenas en las aristas del grafo y el prefijo y sufijo en 2 nodos a los extremos como se puede observar en la Figura 10 (Phillip & Pavel, 2015).

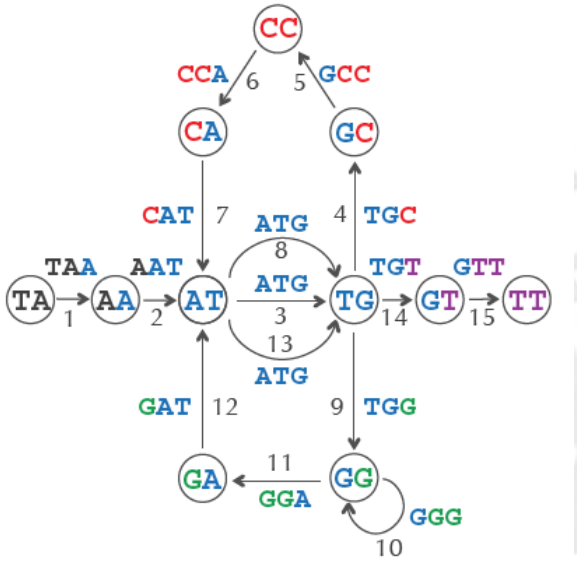

Figura 10. Solución del ejemplo utilizando el grafo de De Brujin.

Tomado de Phillip & Pavel (2015) Bioinformatics Algorithms: An Active Learning Approach.

Los nodos que están repetidos se unen para formar uno solo y de esta manera se logra simplificar la cantidad de nodos. Para obtener la cadena final que vendría a ser el genoma de la especie de interés, se realiza una trayectoria en el grafo tal que cada arista se recorra a lo más una sola vez (Camino de Euler). De esta forma, se pueden ensamblar los *reads* y obtener un genoma consenso (Phillip & Pavel, 2015).

Capítulo 3. Estado del Arte

En este capítulo se detallarán las últimas investigaciones y métodos relacionados al preprocesamiento, ensamblaje y anotación de datos secuenciados, con el fin de mostrar las últimas tendencias en el campo de esta investigación.

3.1 Estrategia de búsqueda

3.1.1 Preguntas de revisión

Se busca responder las siguientes preguntas en la revisión de los artículos obtenidos:

- Pregunta 1: ¿Cuáles son las últimas tendencias respecto al preprocesamiento, ensamblaje y anotación de datos secuenciados?
- Pregunta 2: ¿Con qué finalidad se llevaron a cabo las investigaciones sobre la bacteria de la especie *R. asembonensis* o similares?
- Pregunta 3: ¿Es posible encontrar el genoma de la bacteria de la especie *R. asembonensis* en el banco público de genomas? ¿Cómo se obtuvieron los resultados?
- Pregunta 4: ¿De qué forma se validaron los resultados obtenidos?

3.1.2 Palabras clave

Las palabras que serán incluidas en la cadena de búsqueda guardan relación directa a las 3 etapas principales del presente trabajo (preprocesamiento, ensamblaje y anotación). Es decir, si en la primera fase se realiza control de calidad, filtrado y cortado, se deben incluir estas palabras. De la misma forma, para las siguientes etapas. Todos los pasos constituyen el *pipeline* bioinformático propuesto, por lo que también se añadirá el término y sus derivados como flujo de trabajo y marco de trabajo (*framework*). Asimismo, debido a que se utilizan temas de Biología Molecular e Informática, cada consulta deberá relacionarse al área de Bioinformática:

- *Bioinformatics*

- *Quality control, filtering, trimming*
- *Assembler, metagenome, metagenomics*
- *Annotation*
- *Pipeline, workflow, framework*

3.1.3 Cadenas de búsqueda

Utilizando las palabras claves mencionadas, se obtuvieron las siguientes cadenas de búsqueda desde el año 2016. La primera cadena de búsqueda está enfocada en artículos sobre el preprocesamiento de datos secuenciados completos, ya que este tema está directamente relacionado con la primera etapa del objetivo del presente proyecto. La segunda cadena de búsqueda tiene como objetivo obtener resultados acerca de las herramientas de ensamblaje metagenómicas, ya que el presente caso de uso exige un análisis variado de material genético en las secuencias. Por último, la tercera cadena se relaciona al proceso total que cubre todos los pasos: el *pipeline* bioinformático. Asimismo, se le añade la última fase del proyecto, la anotación.

- *TITLE-ABS-KEY ("quality control" AND "bioinformatics" AND ("filtering" OR "trimming"))*
- *TITLE-ABS-KEY ("assembler" AND "bioinformatics" AND ("metagenome" OR "metagenomics"))*
- *TITLE-ABS-KEY (("pipeline" OR "workflow" OR "framework") AND "bioinformatics" AND ("metagenome" OR "metagenomics") AND ("assembly" OR "annotation"))*

3.1.4 Selección de bases de datos

Los artículos se obtuvieron de las siguientes bases de datos relacionadas a Ciencia y de los recursos que posee la Pontificia Universidad Católica del Perú:

- Scopus[4]
- PubMed[5]
- Web of Science[6]
- Springer[7]

3.1.5 Estrategia de extracción

Los artículos del estado del arte fueron seleccionados en base a los siguientes criterios de selección y exclusión:

Criterios de selección

- Artículos recomendados por expertos.
- Artículos relacionados a la pulga u organismos similares.
- Artículos relacionados a la biología molecular de la *Rickettsia asembonensis* o microorganismos similares.
- Artículos relacionados a la biología molecular de la pulga u organismos similares.
- Artículos relacionados al preprocesamiento, ensamblaje o anotación de datos genómicos de la *Rickettsia asembonensis* o microorganismos similares.
- Artículos relacionados al preprocesamiento, ensamblaje o anotación de datos genómicos de la pulga u organismos similares.
- Artículos que contemplen temas de Ingeniería, Bioinformática, Ciencia de la Computación y Biología Molecular.
- Artículos relacionados al análisis metagenómico

Criterios de exclusión

- Artículos de más de 6 años de antigüedad.

[4] https://www.scopus.com/
[5] https://www.ncbi.nlm.nih.gov/pubmed/
[6] https://www.webofknowledge.com/
[7] https://link.springer.com/

- Artículos que no estén escritos en inglés o español.
- Artículos relacionados a bacterias que no sean transmitidas por vectores.
- Artículos enfocados a la molécula ARN.

3.1.6 Selección de artículos

En la siguiente tabla se muestran los resultados obtenidos en las bases de datos de acuerdo a los criterios de selección y exclusión definidos:

Tabla 2

Cantidad de resultados obtenidos en la revisión sistemática (elaboración propia)

Orden	Base de datos	Resultados de las cadenas de búsqueda		
		Cadena 1	Cadena 2	Cadena 3
1	Scopus	23	17	20
2	PubMed	27	25	23
3	Web of Science	12	21	22
4	Springer	190	0	0

Finalmente, los artículos a revisar pertenecen a la Cadena 3, ya que es la que abarca la totalidad de temas relacionados al presente proyecto.

3.2 Revisión y discusión

En esta sección se detalla la revisión y discusión de los artículos seleccionados y que están relacionados al preprocesamiento, ensamblaje y anotación de datos genómico de vectores.

A new coronavirus associated with human respiratory disease in China (Wu et al., 2020)

El surgimiento de enfermedades infecciosas como el Síndrome Respiratorio Agudo Severo (SARS) o el Zika siguen siendo algunas de las principales amenazas de la salud pública mundial. Las causas, el lugar y el momento en el que aparecen estas enfermedades siguen siendo una incertidumbre. El artículo describe las características biomoleculares del virus emergente SARS-CoV-2, así también como su criticidad, ya que, hasta esa fecha (25 de

enero de 2020) se habían reportado, por lo menos, 1975 casos desde el primer paciente hospitalizado a causa de este patógeno. Para obtener el genoma del virus se secuenciaron las muestras (fluido de lavado broncoalveolar) de un paciente que trabajaba en el mercado de comida marina de Wuhan, lugar que se cree que se inició la actual pandemia. En primer lugar, el procesamiento de datos consistió en la remoción de adaptadores y un corte por calidad usando la herramienta Trimmomatic. Los *reads* restantes fueron ensamblados por *de novo* con Megahit (v.1.1.3) y Trinity (v.2.5.1) con los parámetros predeterminados. Megahit generó menos *contigs* que Trinity, lo cual es lo ideal. En total se obtuvieron 384,096 y 1,329,960 *contigs* respectivamente. Estos fueron comparados con nucleótidos no redundantes y las bases de datos de proteínas usando BLASTn y Diamond BLASTx. Los parámetros *e* que utilizaron fueron $1x10^{-10}$ y $1x10^{-5}$, respectivamente. A diferencia de la bacteria del presente trabajo, el virus es de RNA, por lo que los investigadores estimaron la abundancia de transcripción usando la herramienta RSEM incorporada en Trinity. Realizaron una remoción del host (en este caso, el ser humano) con la herramienta Bowtie2 y los *reads* remanentes fueron sometidos al proceso de estimación de abundancia. Luego, los *contigs* más grande de Megahit y Trinity mostraron similitudes con una cepa de la familia de coronavirus (Bat SARS-like coronavirus, SL-CoVZC45). Este *contig* formaba casi la mayor parte del virus, así que fue utilizado para el diseño de un *primer* que iba a dar soporte a otro proceso PCR para la confirmación del genoma final. Este nuevo resultado de secuenciación se utilizó para darle una mayor cobertura al genoma y se realizó con las herramientas Bowtie2 y Samtools.

YAMP: a containerized workflow enabling reproducibility in metagenomics research (Visconti et al., 2018)

YAMP (Yet Another Metagenomics Pipeline) es un *pipeline* bioinformático para el análisis metagenómico de *single-end* y *paired-end reads*. Ataca la problemática que tienen

los demás *pipelines*: la reproducibilidad y repetividad. Los *pipelines* o flujos de trabajo desarrollados en años anteriores no son capaces de lidiar con este tipo problema. Asimismo, dependen de servicios de terceros como EBI y Galaxy para poder ejecutars sus *pipelines*. Esto agudiza la preocupación sobre la privacidad de datos de los investigadores, ya que tienen que compartir sus datos para poder adquirir estos servicios.

Por otro lado, con el creciente avance tecnológico de las plataformas de última generación que permiten el análisis de datos metagenómicos, surge la necesidad de una herramienta que permite realizar este tipo de estudio. La complejidad de este tipo de conjunto de datos es que es abundante y presenta diversidad de material genético, por lo que es necesario el uso de herramientas especializadas en esta problemática. La solución que plantea YAMP es el desarrollo de un *pipeline* bioinformático para el análisis metagenómico, utilizando contenedores. Las herramientas para el procesamiento de los datos que utiliza YAMP son FastQC, BBMap, MetaPhlAn2, HUMAnN2 y QIIME. Además, las herramientas que utiliza para su funcionamiento son Docker y Nextflow. Este *pipeline* enfoca su esfuerzo en el preprocesamiento de los datos, ya que argumenta que es uno de los puntos débiles de los demás *pipelines*.

MOCAT2: a metagenomic assembly, annotation and profiling framework (Kultima et al., 2016)

MOCAT2 es un *pipeline* bioinformático para análisis metagenómico. Está enfocado en las etapas de ensamblaje y anotación, por lo que utiliza un conjunto de bases de datos públicas debidamente priorizadas por su contenido largo y amplio. El *pipeline* pasa por las siguientes fases: preprocesamiento de los *reads* iniciales, en donde realiza un filtro por calidad, luego, el ensamblaje de los *reads*, obteniendo *contigs* de gran tamaño y, por último, la anotación y obtención de genes predichos. Para la etapa de anotación, utiliza DIAMOND y lo combina con el uso de la anotación convencional BLAST.

Practical evaluation of 11 de novo assemblers in metagenome assembly (Forouzan, Shariati, Mousavi Maleki, Karkhane, & Yakhchali, 2018)

Se explica la problemática y dificultad del cultivo de microorganismos en los laboratorios. Menos del 1% de los organismos procariotas pueden ser cultivados en estos ambientes, por lo que los experimentos que necesitan realizar este proceso tienen que recurrir a otras alternativas. Una de estas alternativas es el estudio de estos microorganismos bajo el enfoque metagenómico, que es el análisis de la totalidad de estos organismos en las muestras.

La introducción de la secuenciación de última generación ha permitido que estos experimentos reduzcan sus costos, ya que es posible secuenciar las muestras y analizarlas computacionalmente. Para lograr esto, se debe utilizar herramientas de ensamblaje especiales, ya que los ensambladores convencionales presentan limitaciones con los datos genómicos complejos (algunas secuencias son muy repetitivas y cortas). Por esta razón, se compararon 11 ensambladores: ABySS, Eden, IDBA-UD, MaSuRCA, Ray Meta, SGA, SOAPdenovo2, MetaSPAdes, Velvet, MetaVelvet y Megahit.

Para poder realizar la comparación se utilizó tres conjuntos de datos metagenómicos, uno real y dos simulados. Los datos fueron secuenciados con tecnología de Illumina (HiSeq 100), ya que son más baratos y prevalentes en el campo. Luego de probar todos los ensambladores, según el artículo, MaSuRCA obtuvo el mayor puntaje en la métrica AQI (Assembly Quality Index) que propusieron, seguido de MetaSPAdes, ABySS y IDBA-UD. Sin embargo, en cuestión de contigüidad y longitud, el que mejor desempeño tuvo, fue la herramienta de ensamblaje Megahit. Asimismo, el consumo de memoria que demostró fue sobresaliente.

Por lo tanto, se recomienda Megahit en el contexto donde los datos metagenómicos son de complejidad baja y cuando la capacidad computacional es relativamente baja, ya que su estrategia y algoritmo permite conservar la menor cantidad de memoria RAM.

Preprocessing and quality control for whole-genome sequences from the Illumina HiSeq X platform (Wright et al., 2017)

Se presenta el flujo de trabajo de preprocesamiento y control de calidad luego de la secuenciación de un genoma completo utilizando la plataforma Illumina HiSeq X. Se detallan las principales desventajas de la tecnología de secuenciación de alta calidad: lecturas pequeñas, dispersas y contaminadas. Por esta razón, los autores recomiendan un flujo de trabajo antes de empezar cualquier análisis con el fin de lograr una secuencia final de calidad.

En primer lugar, los datos secuenciados crudos pueden estar contaminados por ADN de biomaterial externo del propio laboratorio o secuencias de otras especies. Los autores sugieren tener conocimiento de la tecnología usada, ya que existen reportes de genomas publicados que contienen librerías propias de la plataforma de secuenciación que no fueron desechadas por posible desconocimiento de los investigadores. Los autores recomiendan el uso del programa Picard[8]. En segundo lugar, una vez obtenidos los datos secuenciados sin contaminación, se procede con la alineación de las secuencias utilizando un genoma referencial. La alineación también incluye coincidencias aproximadas y repetición de secuencias en el genoma referencial por lo que la complejidad computacional se incrementa. Los autores recomiendan el uso de la herramienta BWA-MEM[9].

Por último, en esta etapa también es posible identificar lecturas duplicadas. Las duplicaciones pueden propagar errores y en genomas referenciales donde existen secuencias repetidas son difíciles de reconocer. En este caso, también se utiliza el programa Picard. Una

[8] https://broadinstitute.github.io/picard/
[9] https://github.com/lh3/bwa

vez reconocidas las lecturas duplicadas, se tienen que eliminar. Para esto se utiliza el programa GATK[10] y es necesario ejecutar una realineación, ya que las inserciones o eliminaciones pueden producir alineaciones incorrectas. Además, desde el inicio hasta la finalización de cada etapa, se tiene que evaluar la calidad de los datos. Los autores recomiendan el uso de la herramienta Fast QC[11] que permite la visualización de la calidad de los datos con múltiples métricas y con módulos de estadística básica.

A high-quality de novo genome assembly from a single mosquito using PacBio sequencing (Kingan et al., 2019)

Este artículo presenta el uso del ensamblaje de alta calidad FALCON-Unzip para obtener el genoma del mosquito *Anopheles coluzzii*. Las características resaltantes de esta secuencia consenso fueron: alta contigüidad y completitud (más del 98% de genes conservados y con longitud total). La complejidad para obtener una secuencia genómica de alta calidad de este vector y de otros organismos similares radica en su alta variabilidad genética, el hecho que es una especie diploide (presenta pares de cromosomas, al igual que los humanos) y el poco material de ADN que contiene. Sin embargo, la presente investigación propone un enfoque que permite lidiar con estas complicaciones.

Los autores resaltan que es importante utilizar una sola tecnología de secuenciación y un individuo a secuenciar. Esto resulta en un nivel mayor de contigüidad, integridad y precisión, porque se obtienen lecturas más largas (se sugiere mayor a 20 kb en promedio) y simplifica el proceso de ensamblaje e interpretación. Además, el ensamblaje FALCON-Unzip[12] proporciona una base excelente para generar información genómica de los cromosomas. Puede ser mejorado con el uso de herramientas que expliquen las regiones de

[10] https://software.broadinstitute.org/gatk/blog?id=7712
[11] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[12] https://github.com/PacificBiosciences/FALCON/

variabilidad genética. Los resúmenes de las estadísticas obtenidas se pueden observar en la Tabla 3 (Kingan et al., 2019).

Tabla 3

Estadísticas del ensamblaje de datos crudos y curados de PacBio, comparado con el ensamblaje de secuencias obtenidas con Sanger que se encuentra publicado

		Datos crudos de PacBio	Datos curados de PacBio	Ensamblaje Sanger
Ensamblaje de *contig* primario	Tamaño (Mb)	266	251	224
	N° de *contigs*	372	206	27 063
	Contig N50 (Mb)	3.52	3.47	0.025
Haplotigs alternativos	Tamaño (Mb)	78.5	89.2	Sin resolver
	N° de *contigs*	665	830	N/A
	Contig N50 (Mb)	0.22	0.199	N/A

El resultado de este artículo fue el genoma de alta calidad del mosquito *Anopheles coluzzii* y se puede encontrar como secuencia referencial en NCBI. La variación genética fue resuelta en un tercio del genoma, proveyendo información adicional que ensambladores anteriores no habían producido.

MELC genomics: A framework for de novo genome assembly (Costa, 2017)

El autor presenta un marco de trabajo llamado Genómica MELC[13], un programa que abarca las etapas de preprocesamiento y ensamblaje de datos secuenciados, de una forma amigable y a través de una interfaz web. Al igual que otros marcos de trabajo y servicios web como Genome Analysis Toolkit, BioExtract y Galaxy, el autor busca integrar en un *pipeline* fácil de usar las siguientes capacidades: verificación de calidad, recorte de adaptadores, eliminación de nucleótidos de baja calidad, ensamblaje *de novo*, verificación de calidad del ensamblaje y la comparación con otros enfoques de ensamblaje.

En el módulo de calidad, el *pipeline* usa el software FastQC en su versión 0.11.5. Esta etapa es importante porque constantemente se realiza una verificación de la calidad de las

[13] https://github.com/evaldocosta/melc

secuencias. En el módulo de tratamiento de datos, se realiza el recorte de adaptadores y la eliminación de nucleótidos de baja calidad. Para esto, se utiliza el programa Trimmomatic en su versión 0.36. Luego, se estima la mejor longitud de una subsecuencia (k-mer) de las secuencias leídas (*reads*). El software que se utilizó fue KmerGenie en su versión 1.7016 y su uso es importante para realizar un ensamblaje de alta calidad.

Por último, en el módulo de ensamblaje, el *pipeline* ofrece dos alternativas dependiendo de la naturaleza de los datos secuenciados. Las opciones de ensamblajes son las siguientes: SOAPdenovo assembler y SPAdes assembler. El resultado ensamblado depende del genoma de la especie. Por esta razón, se tiene que verificar la calidad. El autor hace uso del programa Quast en su versión 4.1, ya que ofrece métricas para verificar la calidad del genoma ensamblado.

Genome scaffolding and annotation for the pathogen vector Ixodes ricinus by ultra-long single molecule sequencing (Cramaro, Hunewald, Bell-Sakyi, & Muller, 2017)

Los autores obtuvieron el genoma de la especie *Ixodes ricinus*, el tipo de garrapata más común en Europa y que transmite una gran cantidad de patógenos. Los datos que se utilizaron fueron secuencias obtenidas con la tecnología PacBio e Illumina. La primera fue generada por los autores y la segunda fue obtenida del genoma de referencia publicado. Cabe resaltar que las lecturas que genera la tecnología de PacBio son más largas que las de Illumina y, por esta razón, las lecturas que se generaron contenían algunos *contigs* de esta última. Ambas secuencias generadas se complementaron para eliminar los espacios vacíos o brechas (*gaps*). Los resultados se pueden observar en la siguiente tabla:

Tabla 4

Comparación de los andamios creados y los contigs originales de referencia

	Contigs	Andamios (*Scaffolds*)
Número de secuencias	235 953	204 904
N50 (bp)	1643	3067
Secuencia más larga (bp)	32 538	38 109
Longitud total abarcada	392 924 918	515 788 051

Luego de eliminar los espacios vacíos (*gaps*) en la secuencia y realizar el ensamblaje con ALLORA (tecnología de PacBio), se elaboró la anotación del genoma. Se utilizó TBLASTX para relacionar la secuencia obtenida con la secuencia de referencia de una especie muy cercana (*Ixodes scapularis*) publicada en NCBI. Como resultado, el artículo realizó la primera secuenciación híbrida PacBio-Illumina agrupando *contigs* con andamios (scaffolding). Esto permitió que el genoma completo se extendiera y que la anotación identifique proteínas que describen funciones importantes de la garrapata.

3.3 Conclusiones

- Una de las últimas tendencias para preprocesar datos secuenciados de alta calidad es el uso de lecturas largas. Una de las tecnologías que permite esto es PacBio y probablemente en los próximos años esta plataforma tenga más impacto en la comunidad científica. Por otro lado, el uso del ensamblaje FALCON-Unzip será de gran ayuda, ya que toma ventaja de las longitudes largas de las lecturas. De esta manera, el genoma consenso se hace más contiguo. Por último, la anotación del genoma se realiza utilizando paquetes de BLAST y es posible utilizar especies cercanas de referencia para obtener una mayor cantidad de coincidencias.

- Dejar de lado un patógeno con alta posibilidad de manifestarse puede desencadenar problemas posteriores en la salud pública como, por ejemplo, el caso del surgimiento del virus SARS-CoV-2. Luego de los brotes de SARS en las últimas 2 décadas, se dejaron estudios inconclusos sobre métodos y fármacos para enfrentar esta enfermedad, debido a que ya se había controlado la propagación. Comparando esta situación con el presente caso de estudio, los autores de los artículos relacionados a la bacteria *R. asembonensis* coinciden en la escasez de investigación sobre este microorganismo. Asimismo, otro problema principal que abordan es la relación de incertidumbre entre la bacteria y la salud pública. El hecho de que este patógeno pueda ser una amenaza para el ser humano (por ser una bacteria emergente) es razón suficiente para investigar esta bacteria. Por lo tanto, es muy importante aportar al conocimiento científico para evitar consecuencias graves en la salud de los seres humanos.

- Los artículos relacionados a la información genómica sobre la bacteria de la especie *R. asembonensis* es escasa. Existe un genoma disponible en NCBI, sin embargo, está presente el riesgo de que los datos contengan material genético de otras especies. A partir de estos estudios, es posible reconocer los problemas que los autores tuvieron que lidiar y tomar esas experiencias como oportunidades de aprendizaje para poder desarrollar un flujo de trabajo enfocado en la problemática que presenta este proyecto.

- Las validaciones que los autores utilizaron fueron métricas de contigüidad, completitud e integridad. Para lograr genomas de alta calidad es recomendable hacer uso de los softwares empleados por la comunidad científica, ya que existe evidencia de su buen funcionamiento. Sin embargo, es posible generar métodos y flujos de trabajos diferentes, basados en ensayo y error y documentando evidencia de su funcionamiento.

Capítulo 4. El flujo de trabajo para el preprocesamiento de los datos secuenciados

4.1 Introducción

La secuenciación de última generación es una poderosa herramienta capaz de generar datos biológicos de diversas especies; sin embargo, también es propensa a errores. Obtener datos de calidad depende del procesamiento inicial de los datos, siendo la definición del flujo de trabajo del preprocesamiento la etapa crucial para obtener una secuencia genómica de calidad.

En el presente capítulo, se muestra el desarrollo del resultado esperado 1. Para este fin, se ha propuesto un flujo de trabajo que garantiza la mayor cantidad de datos preprocesados de calidad. Los resultados para cada paso del flujo se obtuvieron mediante la identificación y uso de herramientas bioinformáticas. Asimismo, se seleccionaron criterios de comparación y a partir de estos, se puntuaron las herramientas. Esto permitió realizar una comparación cuantitativa en base a un puntaje y peso de los criterios, con el fin de seleccionar la mejor herramienta para cada paso del flujo de trabajo.

4.2 Descripción del resultado

El flujo de trabajo para el preprocesamiento está compuesto por pasos que incluyen la remoción de adaptadores, el filtrado por calidad, el corte por calidad, el recorte por ruido (sesgos y fluctuaciones), el filtro por longitud y el análisis de calidad. Se recomienda aplicar constantemente el último paso para tener un entendimiento visual de cada etapa. Los resultados serán obtenidos a través de herramientas bioinformáticas, que serán elegidas en base a los criterios más destacados de la literatura.

4.3 Desarrollo del resultado

A continuación, se explican los pasos seleccionados del flujo de trabajo para el preprocesamiento:

4.3.1 Pasos del flujo de trabajo

- **Análisis de calidad**

  En el inicio y en cada etapa del preprocesamiento se recomienda realizar la revisión del reporte de calidad de las muestras secuenciadas. Es necesario tener conocimiento del estado actual de las bases (calidad y contenido), secuencias (longitud y redundancia) y presencia de adaptadores, ya que permite proyectar acciones que se podrían realizar posteriormente. De este modo, es posible tomar decisiones en base a un estado controlado, teniendo la información estadística necesaria para comprender los datos secuenciados que se tienen.

  Cada vez que se realiza una acción de preprocesamiento, se debe evaluar la calidad de las secuencias, con el fin de saber si una decisión fue correcta o si se debe ajustar algún parámetro que mejore el resultado de la acción ejecutada. Por lo tanto, la revisión del reporte de calidad de las secuencias debe ser una actividad que se repita constantemente durante todo el flujo de trabajo de preprocesamiento.

- **Remoción de adaptadores**

  La siguiente acción que se debe ejecutar es la remoción de adaptadores. La presencia de adaptadores en los datos de las muestras secuenciadas ocurre cuando la preparación de las librerías no es lo suficientemente robusta y llegan a ser secuenciadas por la plataforma. Los protocolos establecidos para evitar la permanencia de adaptadores en los datos secuenciados solo reducen el riesgo de su presencia, por lo que no aseguran su ausencia en la totalidad de los casos.

Por lo tanto, es importante remover los adaptadores en los datos secuenciados, a pesar de que las herramientas bioinformáticas no las detecten. La permanencia de los adaptadores produce problemas en el ensamblaje y en el uso de recursos computacionales, pues son secuencias redundantes que no aportan información valiosa del organismo de interés.

- **Filtro por calidad**

El siguiente paso del *pipeline* bioinformático es la ejecución del filtro por calidad. Como se mencionó, es posible que las plataformas de secuenciación de última generación cometan errores al momento de secuenciar las muestras de ADN, ocasionando que las bases detectadas no necesariamente representen a las bases reales (puntaje *Phred*[14]). Incluso, también es posible que las plataformas no las detecten y aquellas que no fueron determinadas tomen el valor de "N". Estas secuencias que contienen las bases de baja calidad no son confiables y pueden sesgar los datos secuenciados al momento de analizarlos contra secuencias genómicas de referencia, detectando especies o genes falsos.

La presencia de material genético diverso en el presente caso de estudio exige detectar con precisión las especies contaminantes para poder aislarlas de la especie de interés. Debido a que el objetivo principal de la presente investigación es una secuencia genómica para el uso de investigadores, se convierte en una necesidad sustancial asegurar la mejor calidad posible, teniendo en cuenta también que exagerar el valor del umbral de calidad puede remover datos valiosos. Por lo tanto, se recomienda asignar un valor moderado en casos de estudios metagenómicos.

[14] Ver diccionario de términos.

- **Corte por calidad**

En el proceso de secuenciación es común que las secuencias tengan una caída de calidad en las bases finales. Este error es propio de las plataformas de secuenciación, por lo que es importante la aplicación de este paso al terminar el filtro por calidad. La caída de calidad en las bases no permite un buen ensamblaje de las secuencias, ya que los algoritmos buscan la coincidencia entre la parte final de una secuencia y la parte inicial de otra. Por lo tanto, es importante mantener una calidad uniforme en la totalidad de las bases, para realizar un ensamblaje apropiado que conecte secuencias que son confiables en la orientación 5' a 3' o viceversa.

- **Recorte por ruido (opcional)**

Un detalle que se puede notar en el análisis de calidad de los datos genómicos, es la presencia de los sesgos y fluctuaciones del contenido de bases al inicio y final de las secuencias. Por un lado, los sesgos de las 12 primeras bases de las secuencias ocurren debido a los procesos enzimáticos aplicados a las muestras de ADN en el laboratorio. El comportamiento "no tan aleatorio" de los dos *primers* (de 6 bases de longitud) que se utilizan para la secuenciación de las muestras inducen a un sesgo en las primeras 12 bases de las secuencias. Por esta razón, es posible notar un contenido no uniforme de bases al momento de realizar el análisis de calidad. A pesar de esto, el sesgo de las primeras bases no impacta de manera significativa y cortarlas no representa una mejora en la etapa de ensamblaje.

Por otro lado, las fluctuaciones que se pueden observar en el contenido de las bases al final de las secuencias sí son de importancia. Las razones de este ruido pueden ser diversas, pero comúnmente ocurren debido a los adaptadores. La

presencia de las fluctuaciones al final de las secuencias puede traer problemas en la etapa de ensamblaje, por lo que es importante removerlas. En este caso, previamente se ejecutó un paso de remoción de adaptadores, por lo cual, si las variaciones en el contenido de bases se siguen dando, se pueden deber a otros motivos. Por lo tanto, el presente paso se considera opcional, ya que es posible que existan diferentes casos con respecto al contenido inusual de las bases y el recorte de estas depende de su impacto en las siguientes etapas.

- **Filtro por longitud**

  El último paso del preprocesamiento es el filtro por longitud. Las secuencias que tienen una longitud bastante corta dificultan a los algoritmos de ensamblaje, ya que mientras más pequeña es la secuencia, más ambiguo se vuelve el ensamblado. El tamaño corto de las secuencias genera una mayor cantidad de ocurrencias con otras, dificultando la etapa de ensamblaje y la hace más propensa a errores. Por esta razón, desechar las secuencias pequeñas es importante y se propone como último paso, ya que previamente las secuencias han sido sometidas a diferentes filtros y cortes, provocando que algunas reduzcan sustancialmente su tamaño.

Los pasos definidos se pueden visualizar en la Figura 11, donde se muestra el *pipeline* o flujo de trabajo para el preprocesamiento:

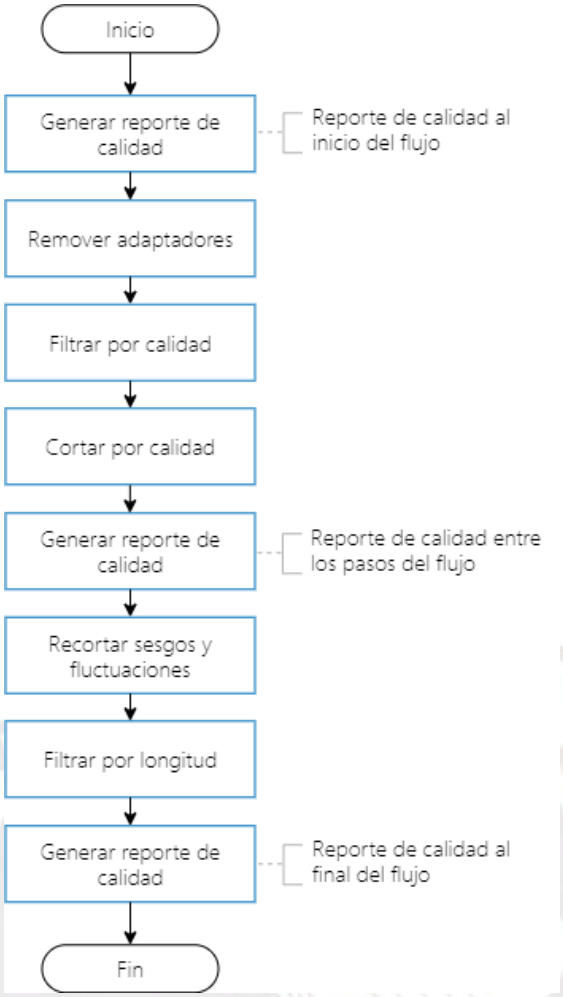

Figura 11. Diagrama de flujo del p*ipeline* bioinformático definido para el preprocesamiento.

4.3.2 Elección de herramientas del flujo de trabajo

Una vez definidos los pasos del flujo de trabajo para el preprocesamiento, se procedió a identificar las herramientas del estado del arte avaladas por la comunidad científica. La siguiente tabla muestra la selección de los 5 criterios con sus respectivos pesos y se establecieron en base a las características resaltadas por los artículos revisados.

Tabla 5

Criterios y pesos establecidos para la comparación de herramientas de preprocesamiento

Parámetros	Descripción	Peso
Uso por la comunidad	La presencia de reseñas de la herramienta en artículos y en foros de la comunidad.	0.4
Última versión	El último lanzamiento o última versión de la herramienta.	0.2
Rapidez	El tiempo de ejecución de la herramienta.	0.2
Flexibilidad	La capacidad de generar un resultado relevante a partir de la manipulación de los parámetros de entrada.	0.1
Documentación	La presencia de un manual completo y detallado.	0.1

El puntaje que se le asignará a cada herramienta será del 0 (refiriéndose al puntaje más bajo) al 5 (puntaje más alto). De esta forma, el puntaje final de cada herramienta se hallará con la siguiente fórmula:

$$p = \sum_{i=1}^{n} w_i s_i \,, \qquad s_i \in \{0, 1, 2, 3, 4, 5\}$$

$$Donde:$$

$$n: cantidad\ de\ criterios\ de\ comparación\ definidos$$

$$w_i: peso\ del\ criterio\ de\ comparación\ i$$

$$s_i: puntaje\ del\ criterio\ i\ asignado\ a\ una\ herramienta$$

$$p: puntaje\ total\ asignado\ a\ una\ herramienta$$

- **Análisis de calidad**

Las herramientas de análisis de calidad que se identificaron fueron FastQC, Fastp y FQC Dashboard, de las cuales, se seleccionó la primera. FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) es una de las herramientas más utilizadas por la comunidad científica y es mencionada en gran parte de los artículos relacionados al análisis de calidad. Se seleccionó esta herramienta, ya que es altamente utilizada por la comunidad bioinformática desde el año 2010 y desde esa fecha se mantiene actualizada (la última versión fue en el 2020). El enfoque principal de FastQC no es aprobar ni desaprobar la calidad de las secuencias. Los resultados que se obtienen se consideran experimentos y por cada uno de ellos se toma una muestra aleatoria de la totalidad de los datos. Esto implica, que los resultados obtenidos por FastQC, se debe tomar como "lo que se espera" de la totalidad de las secuencias. Se considera una herramienta rápida, con buena documentación, pero no tan flexible en la generación de reportes, ya que siempre ha mostrado la misma información estadística en los últimos años y esta no depende de los parámetros ingresados. Entre los principales gráficos que muestra se tiene la información estadística básica, la calidad promedio de cada base según posición, el contenido de las bases, el porcentaje GC, la presencia de adaptadores, entre otros.

La siguiente herramienta es Fastp (Chen et al., 2018) que, a pesar de ser una herramienta relativamente nueva, ha recibido la aceptación de la comunidad científica. Es una herramienta de alta velocidad de respuesta y contiene múltiples funcionalidades de preprocesamiento, por lo que su función principal no es el control de calidad. Por esta razón, ofrece menos información que los reportes generados por FastQC y carece de flexibilidad en este campo. Por último, la

herramienta FQC Dashboard, también es relativamente nueva, ha sido aceptada por la comunidad y sigue recibiendo actualizaciones periódicas. Esta herramienta consume la información generada por FastQC y genera un reporte con más detalle y flexibilidad. A pesar de esto, está elaborada en el lenguaje de programación Python, por lo que el hecho de obtener información de FastQC y añadir funcionalidades extra, la hace una herramienta menos veloz que las primeras ya mencionadas. La siguiente tabla explica el análisis comparativo de las herramientas de análisis de calidad:

Tabla 6

Herramientas de análisis de calidad con sus puntuaciones, utilizando los criterios de comparación establecidos. La herramienta FastQC tiene el puntaje máximo

Herramientas / Criterios	Uso por la comunidad	Última versión	Rapidez	Flexibilidad	Documentación	Puntaje
FastQC	5	5	4	4	5	4.7
Fastp	3	5	5	3	5	4.0
FQC Dashboard	3	5	3	5	5	3.8

- **Remoción de adaptadores**

El primer paso de preprocesamiento es la remoción de adaptadores y se identificaron las herramientas BBDuk, Cutadapt y Trimmomatic, de las cuales, se seleccionó BBDuk. BBDuk (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/) es una herramienta que tiene una gran presencia en foros de bioinformática y es recomendado para la remoción de adaptadores. Fue desarrollada por Joint Genome Institute y recibe mantenimiento y actualizaciones cada cierto periodo. Se considera una herramienta rápida y flexible, ya que su tiempo de ejecución es bajo y permite el ingreso varios parámetros al igual que las herramientas Cutadapt y Trimmomatic. A diferencia de estas herramientas, los

parámetros de BBDuk tienen funcionalidades diferentes y tienen un mayor impacto en el resultado final.

La siguiente herramienta, Cutadapt (https://cutadapt.readthedocs.io/en/stable/), también es usada y recomendada por la comunidad bioinformática. Recibe mantenimiento y actualizaciones cada cierto periodo y se considera una herramienta igual de rápida que BBDuk. Si bien, es una herramienta flexible, los parámetros que posee también las tienen otras herramientas, por lo que no se diferencia de las demás como lo hace BBDuk. Estas dos herramientas están documentadas detalladamente; sin embargo, Trimmomatic (Bolger, Lohse, & Usadel, 2014) no alcanza el nivel de detalle que las otras, a pesar de tener un manual correctamente hecho.

Trimmomatic es una herramienta con mucha popularidad en la comunidad científica. En los últimos años, no se ha liberado una versión nueva, por lo que no es constantemente actualizada como BBDuk y Cutadapt. Utiliza una técnica llamada *sliding window* para poder remover los adaptadores; sin embargo, no es lo suficientemente rápida como las dos primeras herramientas. Asimismo, también es considerada una herramienta flexible, pero el nivel de detalle de sus parámetros y el nivel de impacto en el resultado es menor o igual que BBDuk y Cutadapt. La siguiente tabla explica el análisis comparativo de las herramientas para la remoción de adaptadores:

Tabla 7

Herramientas para la remoción de adaptadores con sus puntuaciones, utilizando los criterios de comparación establecidos. La herramienta BBDuk tiene el puntaje máximo

Herramientas / Criterios	Uso por la comunidad	Última versión	Rapidez	Flexibilidad	Documentación	Puntaje
BBDuk	5	5	5	5	5	5.0
Cutadapt	5	5	5	4	5	4.9
Trimmomatic	5	4	3	3	4	4.1

- **Filtro por calidad**

El siguiente paso de preprocesamiento es el filtro por calidad y se identificaron las herramientas Fastp, FastX-Toolkit, EA-Utils y Prinseq, de las cuales, se seleccionó Fastp (Chen et al., 2018). Esta herramienta es relativamente nueva y ha sido aceptada por la comunidad científica. Como se mencionó en el análisis de calidad, es actualizada constantemente, es rápida y ofrece una flexibilidad por encima de las demás. Los parámetros que permite ingresar son variados e impacta positivamente en el resultado final. Por ejemplo, a comparación de las demás herramientas, Fastp permite realizar el filtro por calidad utilizando un porcentaje de las bases de las secuencias, mientras que las otras toman el total de estas.

Por otro lado, las herramientas FastX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) y EA-Utils (https://expressionanalysis.github.io/ea-utils/), son usadas con menor frecuencia para el filtro por calidad. Las versiones de ambas no han sido actualizadas hace mucho tiempo y no llegan a ser tan rápidas como las demás. Se considera a FastX-Toolkit relativamente flexible al permitir escoger el porcentaje de las bases al igual que Fastp; sin embargo, carece de otros parámetros adicionales y de rapidez de ejecución. Asimismo, EA-Utils ofrece modificar pocos parámetros y solo

permite controlar la calidad de las secuencias a través del promedio total de las bases.

Por último, Prinseq (Schmieder & Edwards, 2011) es una herramienta que busca ser interactiva y fácil de usar. Si bien, fue aceptada por la comunidad bioinformática, no está adaptada a las exigencias de los datos de la última generación de tecnologías de secuenciación. Recientemente, el uso de datos para estudios metagenómicos requiere de herramientas versátiles, rápidas y escalables, capacidades que Prinseq no posee actualmente. Una alternativa es su predecesora, Prinseq++ (Sadural & Edwards, 2019), que es una versión de Prinseq-lite 16 veces más rápida; sin embargo, es una herramienta recién introducida que debe ser aprobada por la comunidad científica. La siguiente tabla explica el análisis comparativo de las herramientas para el filtro por calidad:

Tabla 8

Herramientas para el filtro por calidad con sus puntuaciones, utilizando los criterios de comparación establecidos. La herramienta Fastp tiene el puntaje máximo

Herramientas / Criterios	Uso por la comunidad	Última versión	Rapidez	Flexibilidad	Documentación	Puntaje
Fastp	3	5	5	5	5	4.2
FastX-Toolkit	3	1	3	4	2	2.6
EA-Utils	3	2	3	4	2	2.8
Prinseq	5	4	2	4	4	4.0

- **Corte por calidad, recorte por ruido y filtro por longitud**

Para los últimos pasos del preprocesamiento se identificaron las herramientas Fastp, BBDuk, FastX-Toolkit, Cutadapt, EA-Utils, Trimmomatic y Prinseq, de los cuales, se escogió BBDuk. Como se mencionó, la herramienta BBDuk es versátil, rápida y se diferencia de las otras, debido a que aplica algoritmos basados en *k-mers* en sus funcionalidades (por ejemplo, utiliza el algoritmo *Phred* para el corte por calidad). Se eligió esta herramienta, debido a

que alcanzó el máximo puntaje respecto a las otras en los criterios de comparación de este paso del flujo. Asimismo, se consideró mantener la menor cantidad de herramientas y no utilizar varias, ya que la dependencia de un amplio repertorio de estas podría llevar a problemas asociados a la descontinuación o falta de apoyo del grupo desarrollador. La siguiente tabla explica el análisis comparativo de las herramientas para el corte por calidad:

Tabla 9

Herramientas para el corte por calidad, recorte por ruido y filtro por longitud con sus puntuaciones, utilizando los criterios de comparación establecidos. La herramienta BBDuk tiene el puntaje máximo

Herramientas / Criterios	Uso por la comunidad	Última versión	Rapidez	Flexibilidad	Documentación	Puntaje
Fastp	3	5	5	5	5	4.2
BBDuk	5	5	5	5	5	5.0
FastX-Toolkit	3	1	3	4	2	2.6
Cutadapt	5	5	5	4	5	4.9
EA-Utils	3	2	3	4	2	2.8
Trimmomatic	5	4	3	3	4	4.1
Prinseq	5	4	2	4	4	4.0

Capítulo 5. El preprocesamiento de los datos secuenciados a través del flujo de trabajo propuesto

5.1 Introducción

En el presente capítulo se muestra el resultado número 2. Para lograr este resultado, se ejecutaron los pasos del flujo de trabajo para el preprocesamiento, definidos en el capítulo 4 del presente trabajo, utilizando las herramientas seleccionadas bajo los criterios más destacados por la literatura. Los parámetros ingresados en las herramientas también fueron determinados realizando una revisión de los manuales y de los artículos de preprocesamiento. Asimismo, se ajustaron algunos valores de los parámetros para poder obtener mejores resultados en el presente caso de estudio. Por lo tanto, este capítulo desarrolla el paso de los datos secuenciados a través del flujo de trabajo y el uso de las herramientas de preprocesamiento con sus respectivos parámetros.

5.2 Descripción del resultado

Los datos secuenciados preprocesados son las muestras de la pulga *C. felis* positiva para *R. asembonensis* que fueron procesados por las herramientas bioinformáticas. Este conjunto de datos destaca por mantener la mayor cantidad de secuencias de calidad y tener una longitud deseable para la etapa de ensamblaje.

5.3 Datos de entrada

En el año 2013 se realizó un estudio (Kocher et al., 2016) en Iquitos, Perú, con el objetivo de determinar la presencia de enfermedades rickettsiosas en esta zona geográfica del país. Para lograr esto, una de las actividades que se llevaron a cabo fue la recolección de muestras de ectoparásitos. Entre estos, se recolectaron pulgas de la especie *C. felis* positivas para *R. asembonensis*. Cinco pulgas seleccionadas fueron secuenciadas con las tecnologías de secuenciación Ion PGM y Ion Proton (Thermo Fisher Scientific), mediante la técnica *whole*

genome shotgun sequencing, que consiste en la secuenciación de la totalidad de ADN presente en la muestra. Esto implica que todo el contenido de las pulgas fue secuenciado y es posible encontrar datos genómicos correspondientes a diferentes especies hospederas de la pulga o de su propio microbiota intestinal.

Como resultado de la secuenciación se generaron 18 archivos que contienen *single-end reads* (128 gigabytes en total). El tipo de archivo de origen es bam, pero se realizó la conversión a fastq para poder preprocesar los datos con las herramientas de preprocesamiento seleccionadas. Por lo tanto, en el inicio del flujo de trabajo se tuvo un *pool* de 18 archivos sin preprocesar (*raw data*) y todos fueron sometidos al *pipeline* bioinformático.

5.4 Desarrollo del resultado

A continuación, se muestra la ejecución de los pasos del flujo de trabajo definido para el preprocesamiento, con el objetivo de obtener la mayor cantidad de secuencias preprocesadas de calidad de la bacteria *R. asembonensis*.

5.4.1 Análisis de calidad inicial

Se realizó el análisis de calidad inicial de cada uno de los archivos para poder visualizar la información estadística básica de las secuencias, como se puede observar en la Figura 12. El comando principal que se utilizó para generar el reporte de calidad fue el siguiente:

$$\text{fastqc} \ -\text{o outputdir} \ -\text{j /usr/bin/java} \ -\text{f bam} \ -\text{t 20}$$
$$-\text{a adapters input1.bam input2.bam input3.bam}$$

Una vez obtenidos los reportes de calidad, se detectó la presencia de adaptadores, por lo que debían ser eliminados en el siguiente paso del *pipeline* (se puede visualizar en el Anexo C). Una de las características que los archivos tenían en común era que la calidad de las bases

iniciales se mantenía por encima de Q 20[15], mientras que la calidad de las bases finales caía por debajo de Q 15. Otra característica común, era que el contenido de bases al inicio y al final de las secuencias no era uniformes, por lo que el paso opcional de recorte por ruido tenía que ejecutarse necesariamente. Las acciones para mejorar el estado de los datos serán ejecutadas en los siguientes pasos. La siguiente figura muestra el análisis de calidad inicial de las bases:

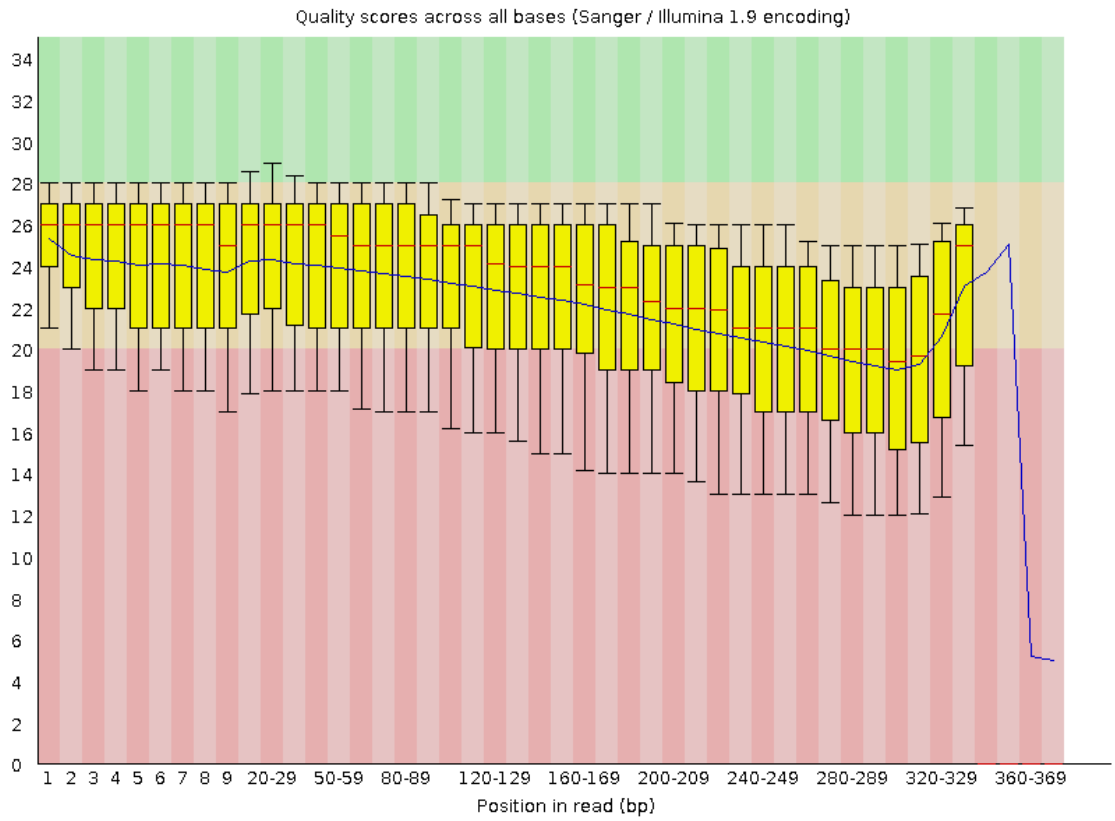

Figura 12. Calidad de las bases y sus posiciones en las secuencias. La imagen representa un diagrama de las posiciones de las bases vs. su calidad

5.4.2 Remoción de adaptadores

La presencia de adaptadores es uno de los problemas más comunes de contaminación en los datos genómicos. Por esta razón, es importante removerlos a pesar de que las

[15] *Phred score* o *Q score*. Ver Diccionario de términos.

herramientas de análisis de calidad no las detecten. El comando principal que se utilizó para remover los adaptadores con la herramienta BBDuk fue el siguiente:

$$\text{bbduk.sh in} = \text{input.fq out} = \text{output.fq ref} = \text{adapters.fa ktrim} = \text{r k} = 10 \text{ mink} = 5 \text{ hdist} = 1 \text{ threads} = 20$$

Los adaptadores que se removieron se pueden observar en la Tabla 10.

Tabla 10

Los adaptadores removidos y sus respectivos identificadores

Id	Secuencia
IonXpress_002	TAAGGAGAAC
IonXpress_035	TAAGCCATTGTC
IonXpress_038	TGGAGGACGGAC
IonXpress_041	TTCCACTTCGC
IonXpress_096	TTAAGCGGTC

El parámetro “ktrim” permite definir el lado de corte de los adaptadores. En este caso, solo se requiere recortar las secuencias de la parte derecha (3’), por lo que no es necesario recortar los *primers* en el lado izquierdo (5’).

Finalmente, se volvió a generar un reporte de calidad con FastQC para poder observar si se habían eliminado los adaptadores y si el contenido de las bases al final de las secuencias había mejorado. Al ver el nuevo reporte (Anexo C), se notó la ausencia de los adaptadores y una leve mejora en el contenido de las bases finales; sin embargo, las fluctuaciones todavía eran exageradas y evidentes. Esto implicó la necesidad de recorte de las bases finales de las secuencias, con el fin de poder eliminar las fluctuaciones generadas por posibles adaptadores no detectados.

5.4.3 Filtro por calidad

El siguiente es el filtro por calidad. Como se pudo observar en el análisis inicial, las secuencias tenían una caída de calidad en una parte de sus bases, por lo que era necesario filtrarlas. Para este fin, se utilizó la herramienta Fastp y se filtraron todas las secuencias con más del 20% de las bases con calidad menor a Q 15 (Low & Tammi, 2017; Santiago, 2015; Simon Andrews, 2018) con el siguiente comando principal:

$$\text{fastp } -\text{i input.fq } -\text{o output.fq } -\text{V } -\text{L } -\text{q 15 } -\text{u 20 } -\text{w 20}$$

Las secuencias que se mantuvieron son aquellas que tienen por lo menos el 80% de sus bases con una calidad mayor a Q 15. Se realiza la generación del reporte de análisis de calidad para poder apreciar el nuevo estado de las secuencias. Se puede observar la presencia de secuencias de mayor calidad, pero las bases todavía siguen presentando una caída de calidad al final de la secuencia.

5.4.4 Corte por calidad

Debido a la presencia de bases con baja calidad al final de las secuencias (3') se procedió a realizar el corte por calidad. Este paso consiste en hacer una revisión base por base desde el extremo final de la secuencia (3') y hacer una medición de su calidad a través de la probabilidad de error *Phred*. El algoritmo que usa BBDuk consiste en mantener una calidad acumulada conforme revisa las bases. Si la calidad acumulada cae por debajo de la establecida (Q 15), la secuencia es cortada en esa base. Para lograr esto se utilizó el siguiente comando principal de la herramienta BBDuk:

$$\text{bbduk.sh in} = \text{input.fq out} = \text{output.fq qtrim} = \text{r trimq} = 15 \text{ threads} = 20$$

Sealizó la generación del reporte de análisis de calidad para observar el estado de las secuencias y bases. Se pudo observar que ambas tenían una calidad mayor a Q 15. El filtro y el corte por calidad permitió mantener los datos secuenciados de calidad a partir de los datos

iniciales (*raw data*) que se tenían. El siguiente problema era las fluctuaciones presentes en el contenido de las bases al final de la secuencia, cuyo origen probablemente sea la presencia de adaptadores no detectados. Esto iba a afectar la etapa de ensamblaje y se decidió realizar el recorte como siguiente paso.

5.4.5 Recorte por ruido y filtro por longitud

Como último paso se realizó el recorte por ruido y el filtro por longitud. El recorte por ruido consiste en volver a hacer un corte en la secuencia, pero esta vez a partir de una posición entre las bases. El ruido estaba presente en diferentes posiciones de las bases en los archivos que se tenían, por lo que se realizó un recorte especial a cada uno, a partir del inicio de las fluctuaciones observadas en el análisis de calidad. Debido a que el recorte hace que las secuencias reduzcan su tamaño, algunas quedaron muy cortas para la etapa de ensamblaje. Entonces, era necesario un filtro por longitud, en donde se desecharon las secuencias que tenían una longitud menor a 30. Para lograr estas acciones se utilizó la herramienta BBDuk y el siguiente comando principal:

$$\text{bbduk.sh in} = \text{input.fq out} = \text{output.fq ftr} = 200 \text{ minlen} = 30 \text{ threads} = 20$$

Los dos primeros parámetros hacen referencia a los archivos de entrada y salida. El siguiente parámetro, “ftr”, define la posición de la base inicial para realizar el recorte hacia la derecha (dirección 5’ a 3’). Debido a que las secuencias quedaron muy cortas para el ensamblaje, se realizó un filtro por longitud con el parámetro “minlen”. Todas aquellas secuencias con una longitud menor a 30, se desecharon. Por último, se estableció que el comando utilice 20 hilos para ejecutarse.

5.4.6 Análisis de calidad final

Por último, se generó el reporte de análisis de calidad final, con el fin de observar alguna anomalía adicional luego de haber realizado la etapa de preprocesamiento. En este

caso, se logró lidiar con todos los problemas de limpieza de datos genómicos y se dejaron listos los datos para poder ensamblarlos en la siguiente fase del presente proyecto. En la Figura 13 se puede observar la calidad y contenido final de las secuencias preprocesadas:

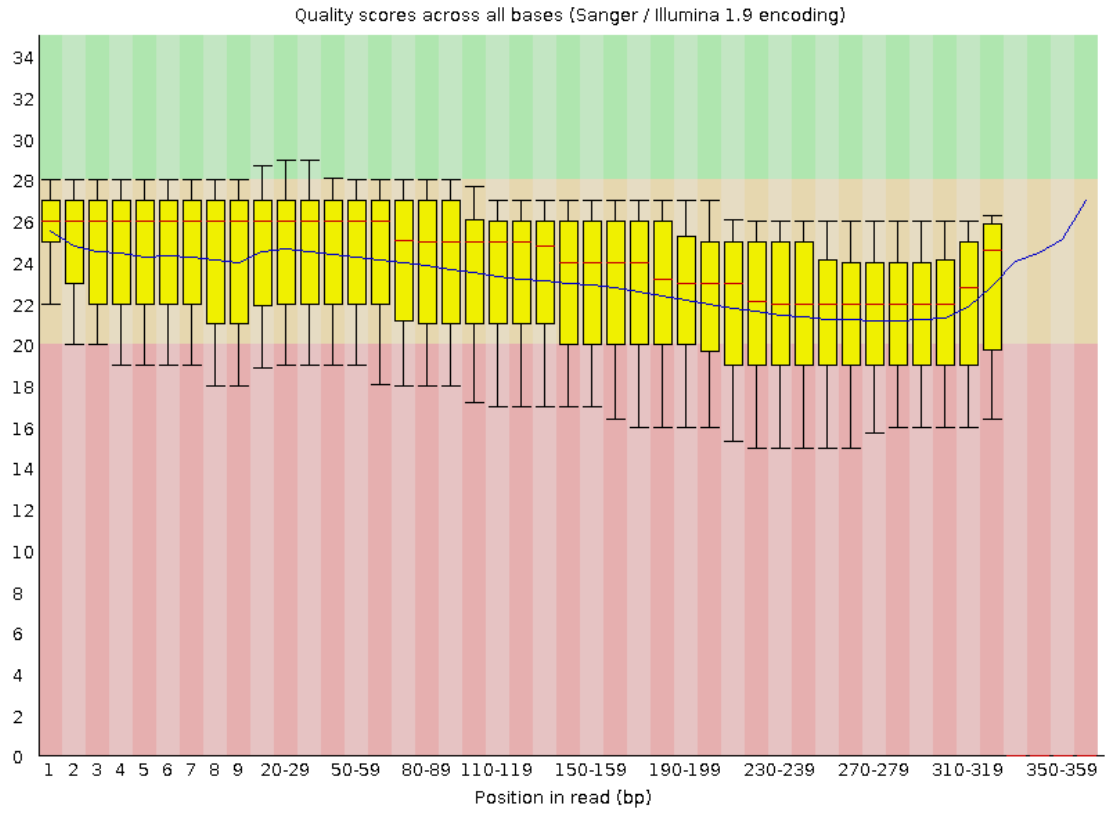

Figura 13. Calidad de las bases y sus posiciones en las secuencias.

Capítulo 6. *Host removal*: Manejo de la contaminación presente en las secuencias

6.1 Introducción

En un conjunto de datos obtenidos por *whole genome shotgun sequencing* (WGS), el material genético presente en los datos secuenciados es diverso. Los datos de estas especies pueden provocar alteraciones en las siguientes etapas y generar ensamblajes contaminados del organismo de interés. Por ejemplo, las secuencias pertenecientes al hospedero (la pulga *C. felis*) es un ejemplo de contaminación en el conjunto de datos, ya que el objetivo es la bacteria *R. asembonensis*.

En el presente capítulo, se detalla el manejo de la contaminación del conjunto de datos y la importancia de incluir este paso en el *pipeline* bioinformático. Las acciones tomadas con respecto a las secuencias pertenecientes a la pulga *C. felis* son relevantes, ya que, si bien el conjunto de datos completo fue obtenido a partir de la secuenciación de esta especie, no se requiere su presencia en la secuencia genómica consenso. Por lo tanto, el hecho de ser mayoría en el número de secuencias representa un riesgo en el resultado final. El proceso de eliminar el hospedero del conjunto de datos se denomina *host removal*.

6.2 Descripción del resultado

El conjunto de datos obtenido en la etapa de preprocesamiento no contendrá los datos del material genético de la pulga *C. felis*. Asimismo, este paso, denominado *host removal*, será incluido en el *pipeline* bioinformático con el objetivo de que sea un referente para casos de estudio similares.

6.3 Desarrollo del resultado

El conjunto de datos contiene una gran diversidad de especies en los que predomina la pulga *C. felis* y las bacterias *Ricketssia Felis Like Organisms* (RFLO). Para poder eliminar

los datos de la pulga *C. felis* se deben identificar todos los *reads* del conjunto de datos que se relacionan con esta especie de pulga. De esta forma, es posible realizar una filtración en los datos y mantener los *reads* no relacionados a la pulga.

Para identificar los *reads* pertenecientes a la especie *C. felis* se debe realizar una alineación con una secuencia genómica de referencia. En este caso, se utilizó el ensamblaje que se encuentra en la base de datos pública de NCBI. Esta secuencia genómica fue obtenida por West Virginia University utilizando la tecnología de última generación PacBio, en el año 2018.

La incertidumbre de la presencia de contaminación de esta referencia se convierte en el primer desafío de este proceso, ya que, si no se evalúa, es posible que afecte al conjunto de datos que se tiene. Por ejemplo, si la secuencia genómica de referencia está contaminada, habrían *reads* que se alinearían a esta y se filtrarían más de lo debido. En cambio, si la referencia es estable, los *reads* alineados serán los correctos y se filtraría la cantidad precisa.

Para determinar la presencia de contaminación en la secuencia genómica de referencia se propone un método de detección directa e indirecta.

6.3.1 Detección directa de contaminación

La detección directa de contaminación consiste, en primer lugar, en realizar una alineación entre el conjunto de datos y la secuencia genómica de referencia de la pulga *C. felis*. Los *reads* que se alineen a la referencia volverán a ser alineados, pero esta vez a la secuencia genómica de referencia de la bacteria *R. asembonensis*. Por último, se debe tomar una decisión: si la cantidad de *reads* alineados a la bacteria *R. asembonensis* es considerable (dependiendo de la cantidad de *reads* total) se concluye la presencia de contaminación, y si la cantidad es mínima, se asume que no hay contaminación. En la Figura 14 se puede observar el flujo de trabajo del método directo.

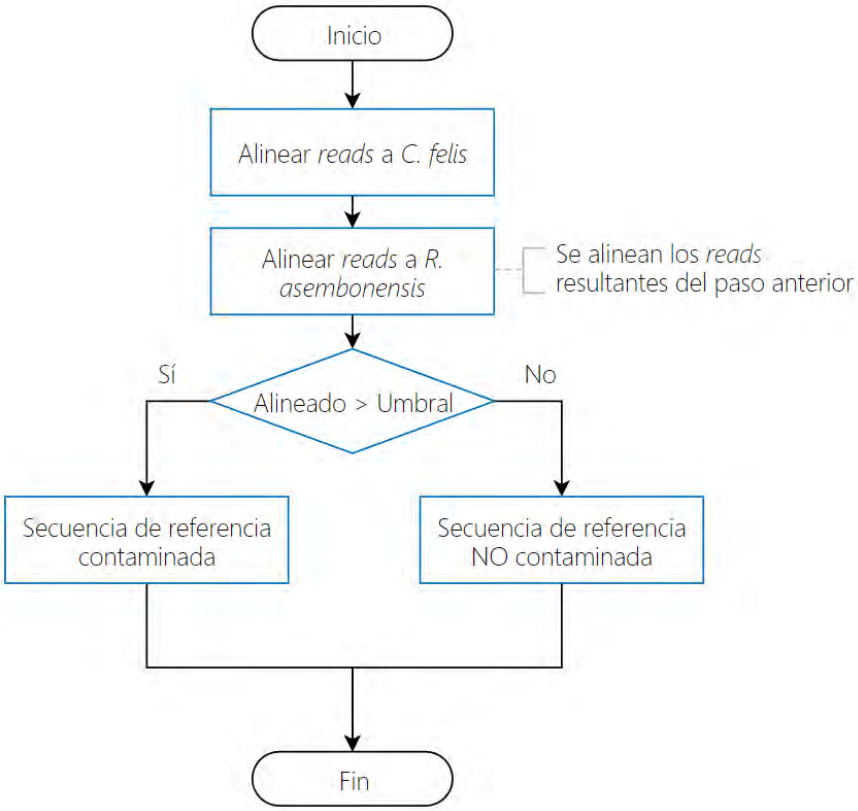

Figura 14. Diagrama de flujo del método directo de detección de contaminación.

Para realizar la alineación se utilizó la herramienta Bowtie 2, la cual destaca por ser precisa y rápida. Antes de ejecutar la herramienta, se descargaron las secuencias genómicas de referencia:

- *C. felis*, con código de acceso GCA_003426905.1 (https://www.ncbi.nlm.nih.gov/assembly/GCA_003426905.1)
- *R. asembonensis*, con código de acceso GCA_000828125.2 (https://www.ncbi.nlm.nih.gov/assembly/GCA_000828125.2)

El primer paso que se realizó fue alinear los *reads* del conjunto de datos al hospedero (pulga *C. felis*). De esta forma, se podían extraer los *reads* que mapearon y no mapearon a la especie *C. felis*. El único caso de interés es el primero, ya que se desea saber la cantidad de

reads de la pulga *C. felis* que mapean a la bacteria *R. asembonensis*, con el fin de determinar si la secuencia genómica de referencia está contaminada.

El proceso empezó cuando se construyó la base de datos del hospedero (pulga *C. felis*). Esto se realiza con el siguiente comando de Bowtie 2:

$$\text{bowtie2} - \text{build} \; - -\text{threads } 20 \; - \text{f GCA\_003426905.1\_ASM342690v1\_genomic.fna cfelis\_flea}$$

Este comando describe la construcción de la base de datos del host utilizando la secuencia genómica de referencia. Una vez construida, se procedió a alinear el conjunto de *reads* total a la secuencia genómica de referencia:

$$\text{bowtie2} \; - -\text{threads } 20 \; - \text{x cfelis\_flea} \; - \text{U \$files} \; - \text{S cfelis\_alignment.sam}$$

El comando está realizando la alineación de los *reads* del conjunto de datos (archivo fastq) y la visualización de los datos se puede observar en el Anexo. El resultado es un archivo de tipo sam, que contiene una estructura especial para el entendimiento de cuáles *reads* se han mapeado y cuáles no. El siguiente paso es extraer los datos que alinearon a la especie *C. felis*, y para lograr esto, se ejecutó un comando de extracción de la herramienta Samtools. Esta herramienta permite realizar operaciones de modificación a los *reads* (corte, filtro, extracción, ordenación, entre otros), y a comparación de las demás como Bedtools o Bamtools, es bastante rápida:

$$\text{samtools view} \; - \text{b} \; - \text{F 4} \; - -\text{threads } 20 \text{ cfelis\_alignment.sam}$$
$$> \text{ extract\_cfelis\_alignment.bam}$$

El comando de Samtools explica la extracción de los *reads* mapeados, donde, el parámetro F es un comando específico que indica la extracción de *reads* alineados y b indica que el output será de tipo bam. Una vez extraídas las secuencias, se alinearon a la secuencia

genómica de la bacteria *R. asembonensis*. Para este fin, se creó la base de datos de esta especie de bacteria de la siguiente forma:

$$\text{bowtie2} - \text{build} \; - -\text{threads } 20$$
$$- \text{f GCA\_000828125.2\_ASM82812v2\_genomic. fna rasembonensis\_bacteria}$$

También se convirtió el archivo que contiene los *reads* mapeados a tipo fastq, ya que el comando próximo utilizará el archivo en este formato.

$$\text{samtools fastq} \; - -\text{threads } 20 \text{ extract\_cfelis\_alignment. bam}$$
$$> \text{ extract\_cfelis\_alignment. fq}$$

Por último, estos *reads* se alinean a la bacteria *R. asembonensis* con el comando de alineación descrito anteriormente:

$$\text{bowtie2} \; - -\text{threads } 20 \; - \text{x rasembonensis\_bacteria} \; - \text{U extract\_cfelis\_alignment. fq}$$
$$- \text{S cfelis\_rasembonensis\_alignment. sam}$$

A continuación, en la Figura 15 se muestran las estadísticas resultantes de la contaminación en la secuencia genómica de referencia de la pulga *C. felis*:

$$119699350 \text{ reads; of these:}$$
$$119699350 \; (100.00\%) \text{ were unpaired; of these:}$$
$$119699258 \; (100.00\%) \text{ aligned 0 times}$$
$$92 \; (0.00\%) \text{ aligned exactly 1 time}$$
$$0 \; (0.00\%) \text{ aligned} > 1 \text{ times}$$
$$0.00\% \text{ overall alignment rate}$$

Figura 15. Resultado de alineación de *reads* de pulga *C. felis* a *R. asembonensis*.

Esto significa que solo se encontraron 92 de 119699350 secuencias que alinearon a la bacteria. Por lo tanto, se comprobó que la secuencia genómica de referencia de la pulga *C. felis* no contiene contaminación, ya que la cantidad de *reads* presentes de la bacteria *R. asembonensis* es sumamente pequeña. Esto implica que la referencia puede ser utilizada para ejecutar el *host removal*.

6.3.2 Detección indirecta de contaminación (opcional)

La detección indirecta de contaminación utiliza la totalidad de los *reads* del conjunto de datos y los *reads* restantes luego de realizar el *host removal*. En cada caso, se hace una alineación con la secuencia genómica de referencia de la bacteria *R. asembonensis* y la cantidad de *reads* alineados debería coincidir aproximadamente. Si no coinciden, significa que la secuencia genómica de referencia de la pulga *C. felis* está contaminada, ya que se concluye que el *host removal*, está quitando reads de la bacteria *R. asembonensis* que no debería. En la Figura 16 se puede observar el flujo de trabajo del método indirecto.

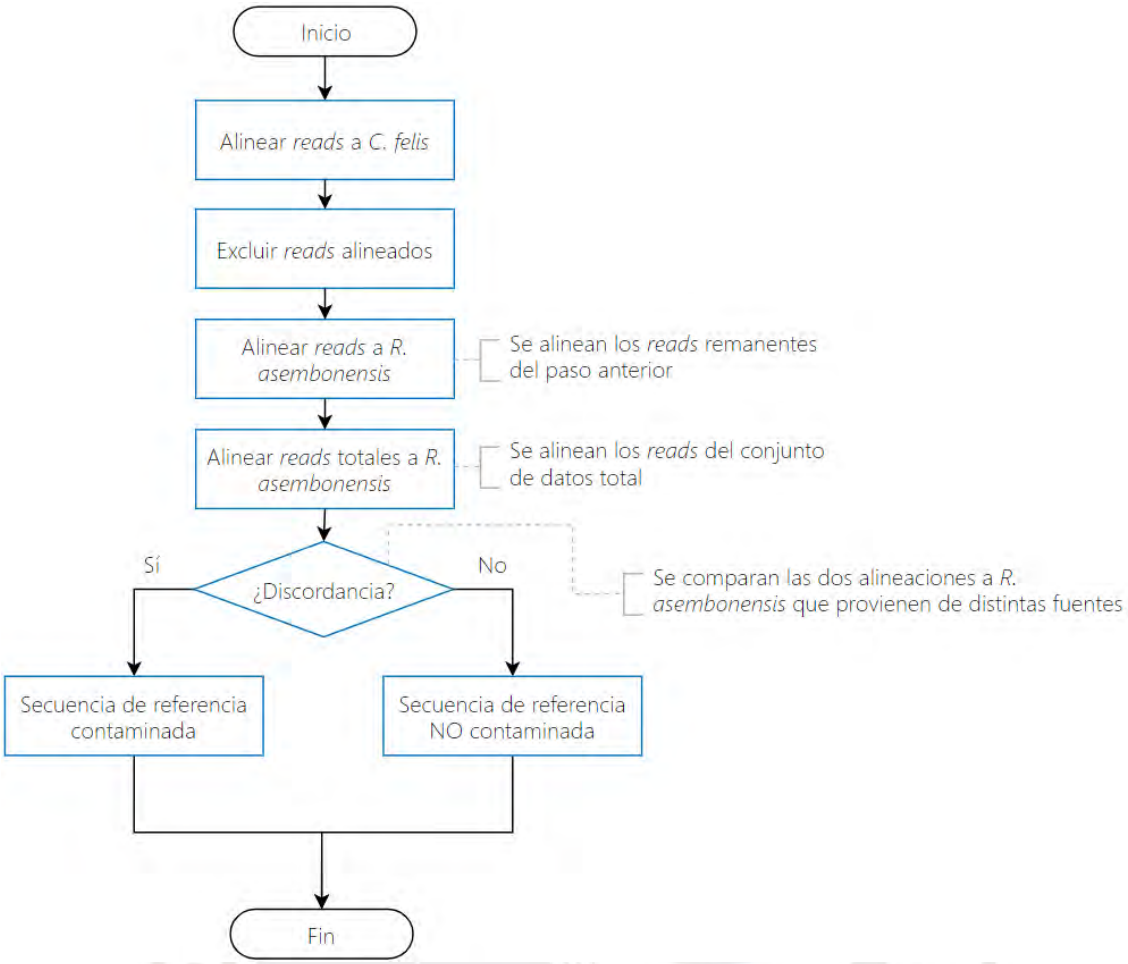

Figura 16. Diagrama de flujo del método indirecto de detección de contaminación.

Entonces, el primer paso que se debe realizar en este método es la remoción del hospedero utilizando la alineación que se hizo a la pulga *C. felis* en el anterior método.

$$\text{samtools view } -\text{b } -\text{f 4 } -\text{F 256 } --\text{threads 20 cfelis\_alignment.sam}$$

$$> \text{ extract\_not\_cfelis\_alignment.bam}$$

El parámetro f 4 indica que solo se extraen del archivo los *reads* que no se alinearon a la secuencia genómica de referencia, mientras que el parámetro F 256, aquellos que no se extraen porque sí lograron alinear.

Una vez obtenidos los archivos que contienen los *reads* sin *host removal* y con *host removal*, respectivamente, se procedió a convertir los archivos a tipo fastq para alinear cada uno a la bacteria *R. asembonensis* con los siguientes comandos:

$$\text{bowtie2} \; -\!-\text{threads } 20 \; -\text{x rasembonensis\_bacteria} \; -\text{U \$files}$$
$$-\text{S rasembonensis\_alignment.sam}$$

$$\text{bowtie2} \; -\!-\text{threads } 20 \; -\text{x rasembonensis\_bacteria}$$
$$-\text{U extract\_not\_cfelis\_alignment.fq}$$
$$-\text{S not\_cfelis\_rasembonensis\_alignment.sam}$$

La visualización de la alineación se puede apreciar en el Anexo y los resultados obtenidos se parecían en la Figura 17 y 18:

191245954 reads; of these:

191245954 (100.00%) were unpaired; of these:

190605393 (99.67%) aligned 0 times

561915 (0.29%) aligned exactly 1 time

78646 (0.04%) aligned $>$ 1 times

0.33% overall alignment rate

Figura 17. Resultados de la alineación de *reads* sin *host removal* a la secuencia genómica de referencia *R. asembonensis*.

71546604 reads; of these:

71546604 (100.00%) were unpaired; of these:

70906135 (99.10%) aligned 0 times

561823 (0.79%) aligned exactly 1 time

78646 (0.11%) aligned $>$ 1 times

0.90% overall alignment rate

Figura 18. Resultados de la alineación de *reads* con *host removal* a la secuencia genómica de referencia *R. asembonensis*.

El resultado indicó que el total de *reads* alineados coincide aproximadamente (hay una diferencia de 92 secuencias) y, por lo tanto, se concluye por el método indirecto que la secuencia genómica de referencia de la pulga *C. felis* no presenta contaminación. Esto implica que, si se intenta extraer los *reads* que alinean a la referencia, se asegura que esta cantidad es la necesaria y no contiene *reads* que pueden ser útiles para los pasos posteriores.

6.3.3 *Host removal* como parte de la etapa de preprocesamiento

Para casos de estudio similares al presente trabajo, es importante realizar la remoción del hospedero y considerarla como paso crucial en el preprocesamiento (la visualización del *Host Removal* se puede observar en el Anexo D). La cantidad de datos relacionados al hospedero es notoriamente superior a las demás especies y puede resultar un problema en pasos posteriores. Si se realiza el ensamblaje de las secuencias sin haber realizado el *host removal*, se corre el riesgo de que el algoritmo de ensamblaje se enfoque en los datos que tienen mayor presencia en el conjunto de datos. Por lo tanto, es sustancial eliminar los datos

que hacen referencia a la especie *C. felis* y es un paso importante dentro del flujo de preprocesamiento.

En la Figura 19 se muestra el flujo del preprocesamiento incluyendo la remoción del hospedero en el último paso.

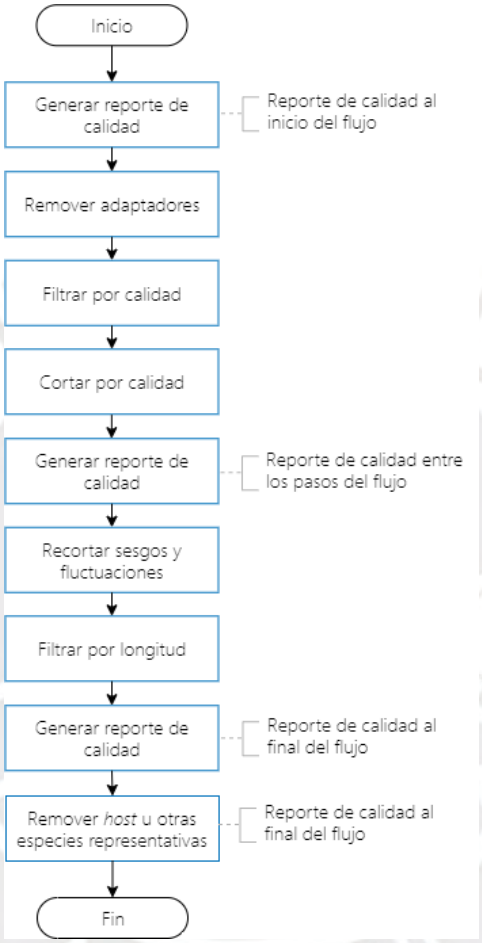

Figura 19. Diagrama de flujo del p*ipeline* para la fase de preprocesamiento.

Capítulo 7. El ensamblaje de los datos secuenciados preprocesados a través de un flujo de trabajo

7.1 Introducción

Construir un resultado sin saber cómo luce es el desafío principal de la fase de ensamblaje. Los factores como la elección de la herramienta o las características de los datos genómicos (tamaño, redundancia y baja calidad de las secuencias), afectan el resultado de esta fase. Por lo tanto, no solo el ensamblaje mismo es importante, sino también los pasos previos.

En el presente capítulo, se describe el desarrollo del objetivo específico 2. Para este fin, se describirán los pasos necesarios para la ejecución del ensamblaje *de novo* de la secuencia genómica de la bacteria *R. asembonensis*, así como las herramientas que se identificaron para llevar a cabo el ensamblaje. A comparación de la fase de preprocesamiento, el ensamblaje depende del nivel de éxito de la ejecución de las herramientas. Por lo tanto, la selección de la mejor, se realizará en base a criterios destacados por la literatura y a las métricas obtenidas luego de ensamblar los datos preprocesados con cada herramienta.

7.2 Descripción del resultado

El flujo de trabajo definido para el ensamblaje debe generar una secuencia genómica ensamblada consenso.

7.3 Desarrollo del resultado

El ensamblaje, a diferencia del preprocesamiento, necesita la identificación de las herramientas antes de definir los pasos específicos. Asimismo, los datos secuenciados que se tienen pertenecen a un organismo que contiene diversas especies en su interior, por lo que también es necesario el uso de herramientas metagenómicas. Por lo tanto, se realizaron diversas pruebas con diferentes herramientas y se escogieron aquellas que produjeron

ensamblajes con métricas sobresalientes. Esto aseguró cubrir la mayor parte de la secuencia genómica consenso.

7.3.1 Identificación de las herramientas de ensamblaje

Las herramientas que se utilizaron para el ensamblaje de datos fueron descritas en la sección Herramientas y Métodos del capítulo 1. Estos ensambladores fueron destacados en las reseñas de la literatura revisada en el capítulo 3 y son los siguientes:

- Megahit (metagenómica)
- Ray Meta (metagenómica)
- Abyss

7.3.2 Flujo de trabajo

El producto final del preprocesamiento fueron 18 archivos de tipo fastq preprocesados. A continuación, se definen los pasos del flujo de trabajo para el ensamblaje y su ejecución con los archivos obtenidos del preprocesamiento.

- **Unión de los archivos (opcional)**

  Algunas herramientas de ensamblaje solo aceptan un solo archivo de tipo fastq o fasta como datos de entradas. Si se tiene un *pool* de archivos, será necesario unirlos. Para lograr esto, se utiliza el comando *cat*, disponible en las terminales de las distribuciones de Linux:

$$\text{cat} \ *.\text{fq} \ > \ \text{input.fq}$$

- **Generación de una muestra del conjunto de datos (opcional)**

  Si el conjunto de datos es grande, es viable obtener una muestra del total. El tamaño depende de la cantidad de *reads* que se tenga, ya que si es relativamente grande es más factible obtener una muestra dado un número específico (por

ejemplo, 100 000 *reads* de 200 millones de *reads*). En cambio, si la cantidad de *reads* no es relativamente grande es factible obtener la muestra dado un porcentaje (por ejemplo, 10% de 1 millón de *reads*) (Géron, 2019).

$$\text{reformat.sh in} = \text{extract\_rasembonensis\_alignment.fq out} = \text{sample\_rasembonensis.fq samplereadstarget} = 10000000$$

Se puede utilizar la herramienta reformat del kit de BBTools. Se recomienda tener la mayor cantidad de bases para un correcto ensamblaje (Illumina recomienda una longitud de 150bp por *read* para tecnologías NGS de segunda generación, por lo que se puede considerar como referencia un conjunto de datos de 10 000 000 *reads* con más de 1 500 000 000 bp).

- **Determinar el conjunto de datos de entrada**

  Se utilizaron 2 conjuntos de datos. El primero fue el resultado del *Host Removal* alineado a la bacteria *R. asembonensis*. La característica de este conjunto de datos es que los *reads* pertenecen a una sola especie, ideal para todas las herramientas, en especial para Abyss que no tiene soporte metagenómico. El segundo, es el resultado del *Host Removal* sin alinear a la bacteria *R. asembonensis*. Este conjunto de datos contiene esta especie además de otras, por lo que es un caso de uso para las herramientas metagenómicas (Megahit y Ray Meta).

- **Ensamblaje**

  Conjunto de datos 1: *Host Removal* alineado a la bacteria *R. asembonensis*

  La primera herramienta fue Abyss:

$$\text{abyss} - \text{pe name} = \text{abyss\_rasembonensis\_XX se}$$
$$= \text{extract\_rasembonensis\_alignment.fq k} = \text{XX}$$

No es posible asignarle un rango de *k-mers*, por lo que se tiene que ejecutar varias veces por *k-mer* si es que se quieren realizar varias pruebas. El parámetro "name" indica el prefijo de los archivos de salida y el parámetro "se" especifica que el archivo de entrada contiene *single-end reads*.

La siguiente herramienta fue Megahit:

$$\text{megahit} \; - -\text{min} - \text{count 3} \; - -\text{k} - \text{min 27} \; - -\text{k} - \text{max 127} \; - -\text{k} - \text{step 2}$$
$$- \text{t 20} \; - \text{r extract\_rasembonensis\_alignment.fq} \; - \text{o megahit}$$

Megahit te permite utilizar un k-mer mínimo, un máximo y la cantidad de unidades para recorrer el rango (*steps*). El parámetro "-t" se ingresa para especificar la cantidad de hilos que utilizará el ensamblador. Luego, como las muestras fueron secuenciadas como secuencias *single-end*, se ingresó el parámetro "-r" para detallar esta característica. Por último, el parámetro "-o" indica la carpeta de salida de la herramienta Megahit.

La última herramienta fue Ray Meta:

$$\text{Ray mpiexec} \; - \text{n 80 Ray} \; - \text{k XX} \; - \text{s extract\_rasembonensis\_alignment.fq}$$
$$- \text{o ray\_XX}$$

El parámetro "-s" significa que el archivo de entrada es de tipo *single-end* y, el parámetro "-o" especifica el directorio de salida, donde se colocarán los resultados de la ejecución de la herramienta. Ray no puede iterar en un rango de *k-mers* de forma nativa (como Megahit) por lo que se ejecutó varias veces con diferentes longitudes de *k-mer*.

Conjunto de datos 2: *Host Removal* sin alinear a la bacteria *R. asembonensis*

Los experimentos utilizaron un rango de k-mers desde la longitud 27 a 127. Debido a que el conjunto de datos es más grande que el primero, se utilizó una muestra de 10 millones de *reads* para elegir los mejores parámetros de cada herramienta.

El ensamblaje utilizando un conjunto de datos de muestra como entrada es útil para poder realizar pruebas de ensayo y error y determinar los parámetros más óptimos para cada herramienta dado nuestro caso de uso. Una vez definidos estos parámetros, se procedió a ejecutar el ensamblaje con el conjunto de datos total.

Se utilizó la herramienta Megahit:

$$\text{megahit} \; -\!-\text{k} - \text{min } 27 \; -\!-\text{k} - \text{max } 127 \; -\!-\text{k} - \text{step } 2 \; - \text{t } 18 \; - \text{r extract\_not\_cfelis\_alignment.fq} \; - \text{o megahit\_meta}$$

También se utilizó la herramienta Ray:

$$\text{Ray mpiexec} \; - \text{n } 80 \; \text{Ray} \; - \text{k } 27 \; - \text{s extract\_not\_cfelis\_alignment.fq} \; - \text{o ray\_meta}$$

Finalmente, se logró obtener 6 ensamblajes que serán comparados en los siguientes pasos con respecto a una secuencia genómica de referencia.

- **Determinar el mejor ensamblaje**

Para determinar el mejor ensamblaje, se escogieron 5 criterios con sus respectivos pesos, que permitirán realizar una comparación cuantitativa entre las herramientas del estado del arte. Se pueden observar en las siguientes tablas:

Tabla 11

Primera parte de criterios y pesos establecidos para la comparación de herramientas de ensamblaje

Parámetros	Descripción	Peso
Uso por la comunidad	La presencia de reseñas de la herramienta en artículos y en foros de la comunidad.	0.2
Última versión	El último lanzamiento o última versión de la herramienta.	0.2
Consumo de recursos	El uso eficiente de los recursos computacionales	0.1

Por otro lado, también se definieron tres criterios destacados de comparación: la métrica N50, L50 y la fracción del genoma, cuyos valores son obtenidos de los resultados del ensamblaje:

Tabla 12

Segunda parte de criterios y pesos establecidos para la comparación de herramientas de ensamblaje

Parámetros	Descripción	Peso
Métrica NG50	Longitud del *contig* que usándolo o escogiendo una longitud mayor, representa el 50% del genoma de referencia.	0.2
Fracción del genoma	Porcentaje del número de bases alineadas al genoma de referencia.	0.2
Métrica LG50	El menor número de *contigs* cuya longitud es mayor al 50% del tamaño del genoma de referencia.	0.1

Análogo al capítulo 5, se realizó una comparación en base al puntaje obtenido con la fórmula. El puntaje que se le asignó a cada herramienta fue del 0 (refiriéndose al puntaje más bajo) al 5 (puntaje más alto). En la siguiente tabla se muestra la evaluación de las herramientas:

Tabla 13

Herramientas para el ensamblaje con sus puntuaciones, utilizando los criterios de comparación establecidos. La herramienta Megahit tiene el puntaje máximo

	Uso por la comunidad	Última versión	Consumo de recursos	NG50	LG50	Fracción del genoma	Puntaje
Megahit	5	5	5	5	5	5	5
Ray Meta	5	4	2	4	4	4	4
Abyss	4	4	2	2	2	4	3.2

Por un lado, según la literatura, todos los ensambladores identificados han sido aceptados por lo comunidad científicas y tienen una cantidad significativa de citas. La última modificación que tuvo Megahit fue en octubre de 2019 y Ray Meta, en julio de 2017, por lo que son herramientas que tienen un soporte relativamente constante.

Las herramientas de ensamblaje manejan algoritmos diferentes y utilizan los recursos computacionales de forma distinta. Entre estos, el único ensamblador que realiza una estrategia para disminuir el consumo de recursos es Megahit, utilizando "compasión de *k-mers*". Asimismo, utiliza hilos que ayudaron a disminuir el tiempo de ejecución del ensamblaje. Por otro lado, Ray tuvo un buen desempeño en el ensamblaje no metagenómico (conjunto de datos 1). Sin embargo, al ingresarle el conjunto de datos 2, que era mucho más grande, utilizaba una gran cantidad de memoria y el programa se detenía inesperadamente.

Finalmente, para la justificación de los valores asignados a las métricas NG50, LG50 y la fracción del genoma se utilizó la herramienta MetaQUAST, que permitió realizar la comparación de los resultados de las herramientas. Se pueden

observar los valores en la Figura 20 y Megahit es superior en casi todas las métricas (la visualización completa de las métricas se puede observar en el Anexo E):

Worst Median Best ☑ Show heatmap

Genome statistics	Abyss	Megahit (No Meta)	Ray (No Meta)	Megahit (Meta)
Genome fraction (%)	80.452	96.465	86.283	96.412
GCA_000828125.2_ASM82812v2_genomic	78.399	94.046	83.835	94.022
Duplication ratio	1.649	1.31	1.012	1.501
GCA_000828125.2_ASM82812v2_genomic	1.59	1.214	1.011	1.329
# genomic features	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part
GCA_000828125.2_ASM82812v2_genomic	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part
Largest alignment	1786	22 910	8103	12 823
GCA_000828125.2_ASM82812v2_genomic	1786	22 910	8103	12 823
Total aligned length	1 710 845	1 557 610	1 164 849	1 648 921
GCA_000828125.2_ASM82812v2_genomic	1 710 120	1 556 180	1 164 230	1 647 538
NA50	230	2899	1675	-
GCA_000828125.2_ASM82812v2_genomic	231	2911	1681	2140
NA75	205	1256	948	-
GCA_000828125.2_ASM82812v2_genomic	205	1282	954	548
LA50	2302	161	215	-
GCA_000828125.2_ASM82812v2_genomic	2287	160	214	221
LA75	4361	366	444	-
GCA_000828125.2_ASM82812v2_genomic	4335	363	442	616
NG50	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	274	3855	1435	3528
NG75	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	207	2169	623	2148
NGA50	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	272	3297	1431	2961
NGA75	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	206	1830	603	1634
LG50	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	1595	112	280	120
LG75	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	3073	231	639	247
LGA50	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	1605	129	282	140
LGA75	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	3093	266	646	294

Figura 20. Reporte de MetaQUAST. Los colores azules más intensos significan valores favorables, mientras que los rojos, no favorables.

De esta manera, se justifican los valores asignados en la puntuación de cada herramienta. El ensamblaje hecho con Megahit utilizando el primer conjunto de datos es el elegido debido a que presenta una alta contigüidad, (métricas NG50 y LG50) y también porque posee la alineación más larga a la secuencia genómica de referencia.

- **Explorar resultados**

En el último paso del flujo de trabajo del ensamblaje se busca que los resultados obtenidos sean coherentes, comparando el resultado de las secuencias construidas y las especies que se esperan. En este caso de estudio, las secuencias ensambladas tuvieron un gran parecido a los genes de la secuencia de *R. asembonensis* de Kenia; es decir, lo que se esperaba encontrar en los datos que se tenían. Este paso es necesario, ya que permite tener una visión descriptiva del resultado (sin llegar a la fase de anotación), en contraste a lo que se observaría si se omitiera: secuencias ensambladas que carecen de sentido.

La herramienta que se utilizó para este paso fue BLAST, un buscador capaz de encontrar similitudes entre secuencias.

A continuación, se muestra una tabla que compara el porcentaje de identidad entre las secuencias consenso formadas y la secuencia de referencia:

Tabla 14

Comparación de porcentajes de identidad entre las secuencias consenso y la secuencia de referencia de R. asembonensis

Contig	Longitud (bases)	*Contig* de Referencia	Porcentaje de Identidad	Confiabilidad de la identidad (1 - e-value)
k127_624	22935	JWSW01000020.1	99.26%	100%
k127_1318	15634	JWSW01000087.1	99.15%	100%
k127_984	12828	JWSW01000087.1	99.45%	100%
k127_269	11409	JWSW01000001.1	99.57%	100%

Finalmente, se muestra el resultado del flujo de trabajo del ensamblaje en la Figura 21.

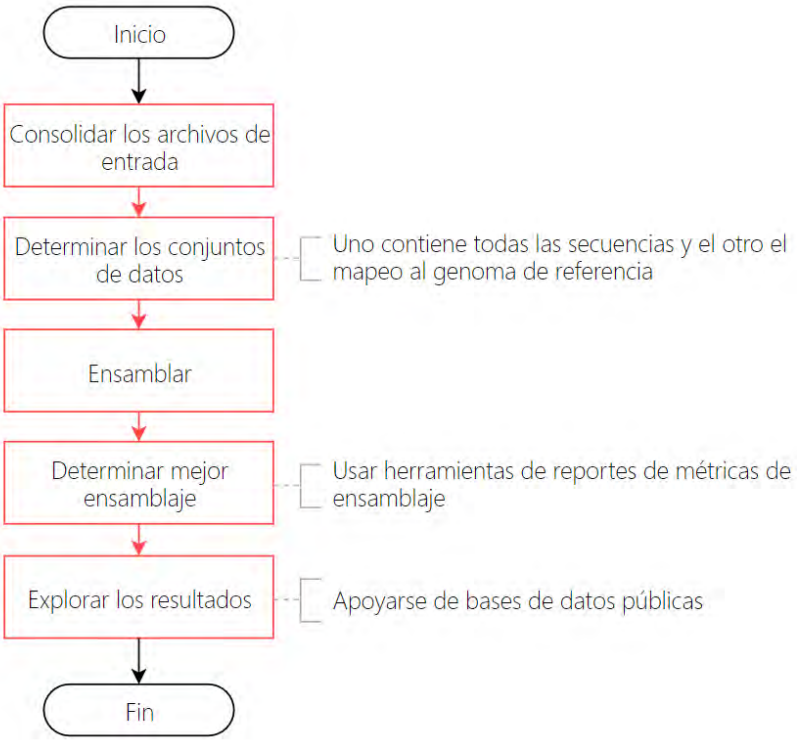

Figura 21. Diagrama de flujo del p*ipeline* bioinformático definido para el ensamblaje.

Capítulo 8. La anotación de la secuencia ensamblada a través de un flujo de trabajo

8.1 Introducción

En el presente capítulo, se detallan los resultados 5 y 6. Para este fin, se describe el desarrollo del *pipeline* bioinformático de la anotación y su ejecución luego de haber obtenido la secuencia genómica ensamblada consenso. La fase de ensamblaje da como resultado un conjunto de secuencias que carece de sentido y que no puede ser estudiada a profundidad. Por esta razón, se realizan pasos que dan significado a la secuencia genómica, identificando y etiquetando los genes que codifican las proteínas. De esta manera, se completa la última etapa del *pipeline* bioinformático.

8.2 Descripción del resultado

La secuencia genómica ensamblada con los genes identificados y etiquetados, con el fin de ser utilizado como referencia para investigaciones posteriores.

8.3 Desarrollo del resultado

El éxito de la anotación de la secuencia genómica formada depende de los pasos previos al punto actual. Es decir, el preprocesamiento y ensamblaje de las secuencias se deben ejecutar siguiendo protocolos pre establecidos, con valores de parámetros adecuados al tipo de análisis y eliminando la contaminación identificada. Si se ha realizado un estudio y seguimiento apropiado, el resultado de la anotación no arrastrará errores.

Debido a que se está proponiendo un *pipeline* o flujo de trabajo bioinformático, todos los pasos propuestos aseguran calidad y cantidad de datos secuenciados sustentados en la literatura. Los desafíos de preprocesamiento y ensamblaje han sido concretados y aquellos restantes dependen de la automatización de esta última fase. Los resultados de las

herramientas de anotación son archivos necesarios para enviarlos a las bases de datos públicas.

Desde 1982, las secuencias de ADN se han subido en bases de datos públicas y las 3 principales a nivel global son GenBank, ENA y DDBJ (todas comparten los datos a través de International Nucleotide Sequence Database Collaboration). Cada una, necesita de archivos específicos para poder enviar las anotaciones de las secuencias genómicas ensambladas. Por lo tanto, necesitan de herramientas propias de estas organizaciones para poder obtener los archivos requeridos por cada base de datos.

NCBI desarrolló la herramienta NCBI Prokaryotic Genome Annotation Pipeline para los usuarios que hacen uso y envíos de datos genómicos a su base de datos (Tatusova et al., 2016). Asimismo, DDBJ cuenta con la herramienta DFAST (Tanizawa, Fujisawa, & Nakamura, 2018), un *pipeline* flexible y rápido para la anotación y publicación de genomas de organismos procariotas. EMBL-EBI recomienda utilizar conversores que permiten convertir los archivos generados por herramientas construidas por terceros a archivos admisibles por la base de datos.

El presente *pipeline* bioinformático propone el uso de la herramienta desarrollada por NCBI: NCBI Prokaryotic Genome Annotation Pipeline (PGAP) (Tatusova et al., 2016). Esta, fue desarrollada en el año 2001 y ha sido mejorada en los años siguientes hasta el presente. Una de las ventajas principales de la anotación es que evita el envío apresurado de ensamblajes y los investigadores pueden darles significado a sus resultados en una etapa temprana.

Debido a que la herramienta es bastante práctica, los pasos que se deben realizar son mínimos y el tiempo invertido en esta fase se concentra más en la etapa de análisis. Se propone el siguiente flujo de trabajo que se puede apreciar en la Figura 22:

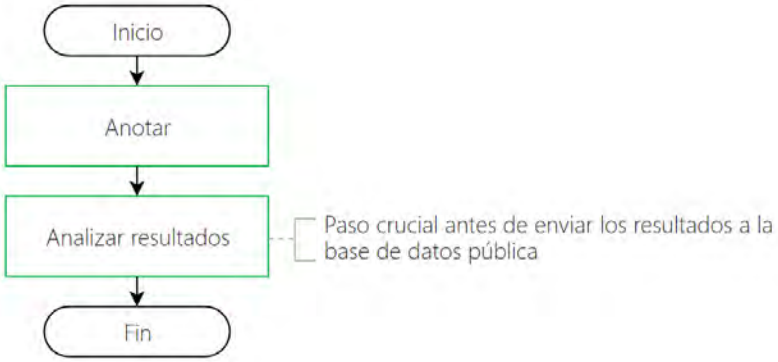

Figura 22. Diagrama de flujo del p*ipeline* para la anotación de genes.

8.3.1 Ejecución del flujo de trabajo

PGAP necesita de metadatos para el reporte de anotación. Para esto, utiliza archivos yaml, un lenguaje de serialización comúnmente utilizado para la organización de datos. En primer lugar, las bases de datos que utiliza la herramienta pueden ser descargados de la siguiente forma:

$$\text{curl} - \text{OL https://github.com/ncbi/pgap/raw/prod/scripts/pgap.py}$$

$$./\text{pgap.py} \; - -\text{update}$$

Luego se crean dos archivos yaml, el primero que indica dónde se encuentran los datos de entrada que la herramienta utilizará y el segundo que contendrá la descripción de los investigadores involucrados en la anotación (los ejemplos de los archivos yaml utilizados se pueden encontrar en el Anexo F). Una vez configurados estos archivos, se ejecuta el siguiente comando para empezar con la anotación:

$$./\text{pgap.py} \; - \text{r} \; - \text{o output input.yaml}$$

El resultado de la anotación de la secuencia genómica consenso y de referencia se muestran en las siguientes figuras (la visualización de una muestra de los genes encontrados se puede observar en el Anexo G):

```
DEFINITION  Rickettsia asembonensis chromosome, complete genome.
ACCESSION
VERSION
KEYWORDS    .
SOURCE      Rickettsia asembonensis
  ORGANISM  Rickettsia asembonensis
            Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsiales;
            Rickettsiaceae; Rickettsieae; Rickettsia; spotted fever group.
REFERENCE   1  (bases 1 to 22935)
  AUTHORS   Arauco,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-MAR-2021) Ingenieria Informatica, Pontificia
            Universidad Catolica del Peru, Av. Universitaria 1801, Lima 15088,
            Peru
COMMENT     The annotation was added by the assembly submitters using the NCBI
            Prokaryotic Genome Annotation Pipeline (PGAP). Information about
            stand-alone PGAP can be found here: https://github.com/ncbi/pgap/

            ##Genome-Annotation-Data-START##
            Annotation Provider               :: Pontificia Universidad
                                                 Catolica del Peru
            Annotation Date                   :: 03/06/2021 21:31:04
            Annotation Pipeline               :: NCBI Prokaryotic Genome
                                                 Annotation Pipeline (PGAP)
            Annotation Method                 :: Best-placed reference protein
                                                 set; GeneMarkS-2+
            Annotation Software revision      :: 2021-01-11.build5132
            Features Annotated                :: Gene; CDS; rRNA; tRNA; ncRNA;
                                                 repeat_region
            Genes (total)                     :: 1,802
            CDSs (total)                      :: 1,762
            Genes (coding)                    :: 665
            CDSs (with protein)               :: 665
            Genes (RNA)                       :: 40
            rRNAs                             :: 1, 1, 1 (5S, 16S, 23S)
            complete rRNAs                    :: 1, 1, 1 (5S, 16S, 23S)
            tRNAs                             :: 33
            ncRNAs                            :: 4
            Pseudo Genes (total)              :: 1,097
            CDSs (without protein)            :: 1,097
            Pseudo Genes (ambiguous residues) :: 0 of 1,097
            Pseudo Genes (frameshifted)       :: 945 of 1,097
            Pseudo Genes (incomplete)         :: 418 of 1,097
            Pseudo Genes (internal stop)      :: 185 of 1,097
            Pseudo Genes (multiple problems)  :: 399 of 1,097
            ##Genome-Annotation-Data-END##
```

Figura 23. Anotación de la secuencia genómica consenso de la bacteria *R. asembonensis*.

```
DEFINITION  Rickettsia asembonensis chromosome, complete genome.
ACCESSION
VERSION
KEYWORDS    .
SOURCE      Rickettsia asembonensis
  ORGANISM  Rickettsia asembonensis
            Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsiales;
            Rickettsiaceae; Rickettsieae; Rickettsia; spotted fever group.
REFERENCE   1  (bases 1 to 21692)
  AUTHORS   Arauco,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (25-MAR-2021) Ingenieria Informatica, Pontificia
            Universidad Catolica del Peru, Av. Universitaria 1801, Lima 15088,
            Peru
COMMENT     The annotation was added by the assembly submitters using the NCBI
            Prokaryotic Genome Annotation Pipeline (PGAP). Information about
            stand-alone PGAP can be found here: https://github.com/ncbi/pgap/

            ##Genome-Annotation-Data-START##
            Annotation Provider               :: Pontificia Universidad
                                                 Catolica del Peru
            Annotation Date                   :: 03/25/2021 03:30:07
            Annotation Pipeline               :: NCBI Prokaryotic Genome
                                                 Annotation Pipeline (PGAP)
            Annotation Method                 :: Best-placed reference protein
                                                 set; GeneMarkS-2+
            Annotation Software revision      :: 2021-01-11.build5132
            Features Annotated                :: Gene; CDS; rRNA; tRNA; ncRNA;
                                                 repeat_region
            Genes (total)                     :: 1,678
            CDSs (total)                      :: 1,637
            Genes (coding)                    :: 1,365
            CDSs (with protein)               :: 1,365
            Genes (RNA)                       :: 41
            rRNAs                             :: 1, 1, 1 (5S, 16S, 23S)
            complete rRNAs                    :: 1, 1, 1 (5S, 16S, 23S)
            tRNAs                             :: 33
            ncRNAs                            :: 5
            Pseudo Genes (total)              :: 272
            CDSs (without protein)            :: 272
            Pseudo Genes (ambiguous residues) :: 0 of 272
            Pseudo Genes (frameshifted)       :: 147 of 272
            Pseudo Genes (incomplete)         :: 116 of 272
            Pseudo Genes (internal stop)      :: 62 of 272
            Pseudo Genes (multiple problems)  :: 50 of 272
            ##Genome-Annotation-Data-END##
```

Figura 24. Anotación de la secuencia genómica de referencia de la bacteria *R. asembonensis*.

La herramienta de anotación identifica genes y pseudogenes en los *contigs* del conjunto de datos de entrada y estos pueden repetirse a lo largo del genoma. Por esta razón, se desarrolló un código en Python para lograr un mejor análisis e interpretabilidad.

Tabla 15

Genes y pseudogenes de la R. asembonensis de Perú y Kenia

	Perú	Kenia
Genes	160	775
Pseudogenes	725	107
Total	885	882

La Tabla 15. muestra la comparativa de la anotación de la especie de interés y su referencia. El genoma del presente caso de estudio presenta 160 genes y 725 pseudogenes, mientras que el genoma de referencia anotado, 775 y 107. La precisión de la anotación también puede estar sujeta a los protocolos de secuenciación utilizados, y esta puede ser la explicación de la sobrevaluación de la cantidad de pseudogenes en el genoma peruano (Berrios & Ely, 2018). Por esta razón, se generó un código en Python que analizó los reportes de anotación generados por la herramienta PGAP (extensión .gbk) y al consultar la presencia de los pseudogenes de Perú en el genoma de Kenia, se confirmó que gran parte eran funcionales. Esto permitió reclasificar los productos de proteína. En la Figura 25 se puede observar el resultado de la reclasificación.

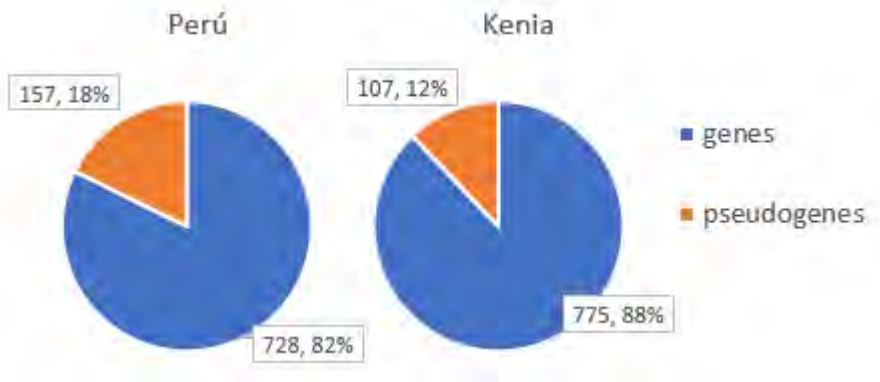

Figura 25. Reclasificación de los pseudogenes de la *R. asemobenensis* peruana. En cada recuadro, se muestra la cantidad y el porcentaje que representa.

Al comparar ambos resultados, se observó que la *R. asembonensis* de Perú tiene 728 genes funcionales, de los cuales, 12 no están presentes en el genoma de referencia. Del mismo modo, de los 157 pseudogenes, 72 tampoco se encontraron en la especie de Kenia. La Figura 26, muestra los nuevos hallazgos del genoma peruano.

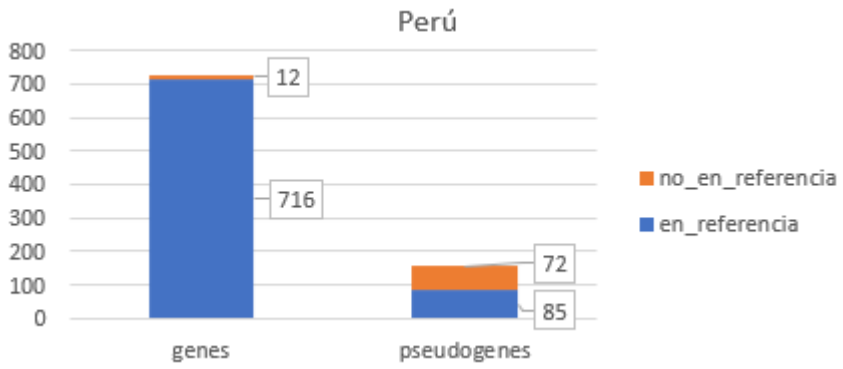

Figura 26. Presencia de genes y pseudogenes de la R. asembopnensis peruana en la referencia de Kenia.

La fase de la anotación concluye con el análisis y entendimiento de los resultados finales, así como la propuesta de genes y pseudogenes producto de los hallazgos de la anotación del genoma de interés. La Tabla 16 y Tabla 17 describen la participación de algunos genes y pseudogenes encontrados y que se proponen en el presente estudio (la lista completa de genes y pseudogenes propuestos se encuentra en el Anexo H).

Tabla 16

Ejemplos de genes identificados en la R. asembonensis peruana

Proteína	Participación
DprA	Transformación de ADN
YidC/Oxa1	Inserción de proteínas en membrana
MazF	Adaptación en ambientes hostiles
peptidase	División de péptidos

Tabla 17

Ejemplos de pseudogenes identificados en la R. asembonensis peruana

Proteína	Participación
YbgF	Mantención de la membrana
PHP	Estabilización de la polimerasa
zinc ABC transporter	Adquisición de zinc
FtsK	División de la bacteria

La fase de la anotación concluye con el análisis previamente descrito: se propone el hallazgo de genes y pseudogenes del genoma de interés anotado como resultado de la ejecución completa del *pipeline*, que se puede encontrar en el Anexo I. La exploración biológica de los datos e información generados serán realizados por estudios posteriores de los científicos y profesionales de la salud.

Capítulo 9. Conclusiones y trabajos futuros

9.1 Conclusiones

La solución propuesta en el presente trabajo es el desarrollo de un *pipeline* bioinformático que permita la anotación del genoma de la bacteria *R. asembonensis*. La primera fase, de preprocesamiento, fue diseñada para lidiar con la diversidad de especies presentes en el conjunto de datos. Los valores de los parámetros especificados no fueron exigentes, con el fin de mantener la mayor cantidad de datos posible, sin dejar de lado la calidad.

Asimismo, se aplicaron métodos de detección de contaminación que actuaron como diferenciador con respecto a otros *pipelines* descritos en el estado del arte. Esta parte es crucial cuando se está analizando datos metagenómicos y sobre todo cuando las secuencias han sido obtenidas del hospedero. Reconocer la presencia de contaminación en el conjunto de datos permite al investigador tomar una pausa y replantear el manejo de esta. También asegura que no se esparzan datos contaminados en las bases de datos públicas que suelen ser usadas por muchos investigadores. Las herramientas utilizadas en el preprocesamiento son diversas, pero se recomiendan: FastQC, Fastp, BBDuk y Bowtie2.

En la fase de ensamblaje, se utilizaron herramientas avaladas por la literatura. Estas también tienen como característica principal ensamblar secuencias con diversidad de material genético. Las herramientas planteadas también fueron sometidas a una comparación cuantitativa utilizando el conjunto de datos total y una muestra para poder medir el desempeño de cada una rápidamente. En este caso, la mejor herramienta que se identificó fue Megahit.

Por último, el presente estudio culmina con el *pipeline* completado (se puede visualizar en el Anexo I) y el análisis de la anotación del genoma de la *R. asembonensis* peruana, en

donde se proponen 12 genes y 72 pseudogenes que no se habían identificado anteriormente en la referencia de Kenia. Este nuevo hallazgo resulta fundamental en la aspiración de contribuir con información genómica para científicos y profesionales de la salud, así como el *pipeline* genera un precedente para casos de estudio similares.

9.2 Trabajos futuros

Con las últimas mejoras en las tecnologías de secuenciación, la tercera generación de estas puede ofrecer resultados más contiguos y más seguros, debido a las lecturas superlargas que se logran secuenciar (por ejemplo, PacBio). En una siguiente iteración del trabajo, la inversión en una tecnología de tercera generación sería necesaria para obtener nuevos resultados que complementen el valor generado del desarrollo del presente *pipeline*.

Los siguientes pasos no solo apuntarían a una mejora de generación en las tecnologías, sino también a los softwares utilizados para la obtención del genoma. El uso de biocontenedores darían soporte a los equipos de los profesionales de la salud y científicos en general, al momento de establecer un ambiente de trabajo homogéneo y evitar la discordancia de resultados a causa del uso de diferentes infraestructuras o versiones de las herramientas.

Asimismo, para un uso más eficiente y más práctico de la solución propuesta se apuntaría a automatizar el proceso completo considerando distintos *frameworks* y herramientas que se vienen trabajando en el campo de *pipelines* bioinformáticos. Algunos ejemplos son Cromwell, un sistema de gestión de flujos de trabajo que soporta WDL (Workflow Description Language) y CWL (Common Workflow Language), y Nextflow, un framework de flujos de trabajo que cohesiona *pipelines* bioinformáticos con su propio lenguaje DSL (Domain Specific Language).

Así como se podría dar el pase a la automatización, también se podría implementar un sistema con inspiración en MGnify o hacer uso de esta misma. MGnify es una plataforma

usada como una base de datos de microbiomas y en donde también se pueden automatizar *pipelines*. Este recurso libre y desarrollado por European Molecular Biology Laboratory (EMBL) puede complementar el desarrollo del *pipeline* y la exploración de casos de estudios similares al presente trabajo, ya que su basto repositorio de datos podría potenciar la identificación de nuevos hallazgos.

Finalmente, un *pipeline* bioinformático automático y desplegado en un contenedor, así como también uno situado en una plataforma con inspiración en MGnify o en esta misma, serían los trabajos a considerar en el futuro. También es importante tomar en cuenta el uso de tecnologías de la tercera generación de secuenciación y que la exploración de los resultados generados puede ser extrapolada a casos de estudios similares al presente trabajo.

Bibliografía

Berrios, L., & Ely, B. (2018). Achieving Accurate Sequence and Annotation Data for Caulobacter vibrioides CB13. *Current Microbiology*, *75*(12), 1642–1648. https://doi.org/10.1007/s00284-018-1572-3

Bitam, I., Dittmar, K., Parola, P., Whiting, M. F., & Raoult, D. (2010). Fleas and flea-borne diseases. *International Journal of Infectious Diseases*, *14*(8), e667–e676. https://doi.org/10.1016/j.ijid.2009.11.011

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

Brown, T. (2006). *Genomes 3* (3rd ed.). Garland Science.

Cantas, L., & Suer, K. (2014). Review: The important bacterial zoonoses in "One Health" concept. *Frontiers in Public Health*, *2*(OCT), 1–8. https://doi.org/10.3389/fpubh.2014.00144

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Costa, E. B. (2017). MELC Genomics: A Framework for De Novo Genome Assembly. *Journal of Computational Biology*, *25*(2), 194–199. https://doi.org/10.1089/cmb.2017.0102

Cramaro, W. J., Hunewald, O. E., Bell-Sakyi, L., & Muller, C. P. (2017). Genome scaffolding and annotation for the pathogen vector Ixodes ricinus by ultra-long single molecule sequencing. *Parasites and Vectors*, *10*(1), 1–9. https://doi.org/10.1186/s13071-

017-2008-9

Dale, J., & Schantz, M. (2002). *From genes to genomes : concepts and applications of DNA technology*.

Djebali, S., Wuches, V., Foissac, S., Hitte, C., Corre, E., & Derrien, T. (2016). *Bioinformatics Pipeline for Transcriptome Sequencing Analysis*. https://doi.org/10.1007/978-1-4939-4035-6_14

Faccini-Martínez, Á. A., García-Álvarez, L., Hidalgo, M., & Oteo, J. A. (2014). Syndromic classification of rickettsioses: An approach for clinical practice. *International Journal of Infectious Diseases*, *28*, 126–139. https://doi.org/10.1016/j.ijid.2014.05.025

Fang, R., Blanton, L. S., & Walker, D. H. (2017). Rickettsiae as Emerging Infectious Agents. *Clinics in Laboratory Medicine*, *37*(2), 383–400. https://doi.org/10.1016/j.cll.2017.01.009

Forouzan, E., Shariati, P., Mousavi Maleki, M. S., Karkhane, A. A., & Yakhchali, B. (2018). Practical evaluation of 11 de novo assemblers in metagenome assembly. *Journal of Microbiological Methods*, *151*(February), 99–105. https://doi.org/10.1016/j.mimet.2018.06.007

Frishman, D., & Valencia, A. (2008). *Modern Genome Annotation: The BioSapiens Network*. Springer-Verlag Wien.

Garrido-Cardenas, J. A., & Manzano-Agugliaro, F. (2017). The metagenomics worldwide research. *Current Genetics*, *63*(5), 819–829. https://doi.org/10.1007/s00294-017-0693-8

Géron, A. (2019). Hands-on Machine Learning. In *Journal of Chemical Information and Modeling* (Vol. 53).

Graña, O., López-Fernández, H., Fdez-Riverola, F., González Pisano, D., & Glez-Peña, D.

(2018). Bicycle: A bioinformatics pipeline to analyze bisulfite sequencing data. *Bioinformatics*, *34*(8), 1414–1415. https://doi.org/10.1093/bioinformatics/btx778

Hartl, D. L., & Ruvolo, M. (2011). *Genetics*. Jones & Bartlett Publishers.

Henry, J. G., & Heinke, G. W. (1999). *Ingeniería Ambiental*.

Ishihara, S., Kotomura, N., Yamamoto, N., & Ochiai, H. (2017). Ligation-mediated PCR with a back-to-back adapter reduces amplification bias resulting from variations in GC content. *Analytical Biochemistry*, *531*, 37–44. https://doi.org/10.1016/j.ab.2017.05.011

Jiang, J., Maina, A. N., Knobel, D. L., Cleaveland, S., Laudisoit, A., Wamburu, K., … Richards, A. L. (2013). Molecular detection of Rickettsia felis and Candidatus Rickettsia Asemboensis in Fleas from Human Habitats, Asembo, Kenya. *Vector-Borne and Zoonotic Diseases*, *13*(8), 550–558. https://doi.org/10.1089/vbz.2012.1123

Kingan, S. B., Heaton, H., Cudini, J., Lambert, C. C., Baybayan, P., Galvin, B. D., … Lawniczak, M. K. N. (2019). A high-quality de novo genome assembly from a single mosquito using pacbio sequencing. *Genes*, *10*(1). https://doi.org/10.3390/genes10010062

Kocher, C., Morrison, A. C., Leguia, M., Loyola, S., Castillo, R. M., Galvez, H. A., … Richards, A. L. (2016). Rickettsial Disease in the Peruvian Amazon Basin. *PLoS Neglected Tropical Diseases*, *10*(7), 1–13. https://doi.org/10.1371/journal.pntd.0004843

Kultima, J. R., Coelho, L. P., Forslund, K., Huerta-Cepas, J., Li, S. S., Driessen, M., … Bork, P. (2016). MOCAT2: A metagenomic assembly, annotation and profiling framework. *Bioinformatics*, *32*(16), 2520–2523. https://doi.org/10.1093/bioinformatics/btw183

Low, L., & Tammi, M. (2017). *Bioinformatics: A Practical Handbook of Next Generation Sequencing and Its Applications*. Retrieved from

https://www.worldscientific.com/worldscibooks/10.1142/10159#t=aboutBook

Loyola, S., Flores-mendoza, C., Torre, A., Kocher, C., Melendrez, M., Luce-fedrow, A., … Leguia, M. (2018). *Rickettsia asembonensis Characterization by Multilocus Sequence Typing of Complete Genes, Peru*. *24*(931). https://doi.org/https://doi.org/10.3201/eid2405.170323

Maina, A. N., Jiang, J., Luce-Fedrow, A., St. John, H. K., Farris, C. M., & Richards, A. L. (2019). Worldwide presence and features of flea-borne Rickettsia asembonensis. *Frontiers in Veterinary Science*, *5*(JAN). https://doi.org/10.3389/fvets.2018.00334

Masoudi-Nejad, A., Narimani, Z., & Hosseinkhan, N. (2013). *Next generation sequencing and sequence assembly: methodologies and algorithms*. https://doi.org/10.1007/978-1-4614-7726-6

Mengoni, A., Galardini, M., & Fondi, M. (2015). *Bacterial Pangenomics: Methods and Protocols*. Retrieved from http://www.springer.com/series/7651

Merchant, S., Wood, D. E., & Salzberg, S. L. (2014). Unexpected cross-species contamination in genome sequencing projects. *PeerJ*, *2*, e675. https://doi.org/10.7717/peerj.675

Mikheenko, A., Saveliev, V., & Gurevich, A. (2016). MetaQUAST: Evaluation of metagenome assemblies. *Bioinformatics*, *32*(7), 1088–1090. https://doi.org/10.1093/bioinformatics/btv697

Morrone J.J. (1994). On the identification of areas of endemism. *Systematic Biology*, *43*(3), 438–441.

Odhiambo, A. M., Maina, A. N., Taylor, M. L., Jiang, J., & Richards, A. L. (2014). Development and validation of a quantitative real-time polymerase chain reaction assay

specific for the detection of Rickettsia felis and not Rickettsia felis-like organisms. *Vector-Borne and Zoonotic Diseases*, *14*(7), 476–481. https://doi.org/10.1089/vbz.2013.1518

Palacios-Salvatierra, R., Cáceres-Rey, O., Vásquez-Domínguez, A., Mosquera-Visaloth, P., & Anaya-Ramírez, E. (2018). Especies rickettsiales en casos humanos con síndrome febril agudo inespecífico en Perú. *Revista Peruana de Medicina Experimental y Salud Pública*, *35*(4), 630. https://doi.org/10.17843/rpmesp.2018.354.3646

Parola, P., Socolovschi, C., Jeanjean, L., Bitam, I., Fournier, P. E., Sotto, A., … Raoult, D. (2008). Warmer weather linked to tick attack and emergence of severe Rickettsioses. *PLoS Neglected Tropical Diseases*, *2*(11), 1–8. https://doi.org/10.1371/journal.pntd.0000338

Pevsner, J. (2015). Bioinformatics and Functional Genomics. In *Briefings in Functional Genomics and Proteomics* (Vol. 3). https://doi.org/10.1093/bfgp/3.2.187

Phillip, C., & Pavel, P. (2015). *Bioinformatics Algorithms: An Active Learning Approach* (A. L. Publ., Ed.).

Portin, P., & Wilkins, A. (2017). The Evolving Definition of the Term "Gene." *Genetics*, *205*(4), 1353–1364. https://doi.org/10.1534/genetics.116.196956

Ramal, C., Díaz, E., & López, J. (2007). *Rickettsiosis, enfermedad emergente en Loreto. Evidencia serológica de 20 casos*. *24*(1), 99–100.

Rowley, J., Vander Hoorn, S., Korenromp, E., Low, N., Unemo, M., Abu-Raddad, L. J., … Taylor, M. M. (2019). Chlamydia, gonorrhoea, trichomoniasis and syphilis: global prevalence and incidence estimates, 2016. *Bulletin of the World Health Organization*, *97*(8), 548-562P. https://doi.org/10.2471/blt.18.228486

Rozendaal, J. A. (1997). Vector control: methods for use by individuals and communities. In *World Health Organization*. Retrieved from http://www.who.int/water_sanitation_health/resources/vector237to261.pdf

Rust, M. K. (2017). The biology and ecology of cat fleas and advancements in their pest management: A review. *Insects*, *8*(4). https://doi.org/10.3390/insects8040118

Sadural, J., & Edwards, R. (2019). PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets. *PeerJ Preprints*, 43–45. https://doi.org/10.7287/peerj.preprints.27553

Santiago, I. De. (2015). *Quality assessment of NGS data*. (July), 1–7. https://doi.org/10.1177/1465712002005001170

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863–864. https://doi.org/10.1093/bioinformatics/btr026

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. https://doi.org/10.1038/nbt1486

Sheridan, C. (2014). Illumina claims $1,000 genome win. *Nature Biotechnology*, *32*(2), 115–115. https://doi.org/10.1038/nbt0214-115a

Simon Andrews. (2018). *FastQC v0.11.8 Manual*. Retrieved from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

Stehman, D. S., Smith, M. C., Sani, R. A., Gray, G. D., Baker, R. L., Schantz, P. M., … GoI, M. (2014). A global brief on vector-borne diseases. *Symposium of Goat Health*, *11*(September), 1–9. https://doi.org/WHO/DCO/WHD/2014.1

Tanizawa, Y., Fujisawa, T., & Nakamura, Y. (2018). DFAST: A flexible prokaryotic genome annotation pipeline for faster genome publication. *Bioinformatics*, *34*(6), 1037–1039.

https://doi.org/10.1093/bioinformatics/btx713

Tatusova, T., Dicuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., … Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, *44*(14), 6614–6624. https://doi.org/10.1093/nar/gkw569

Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaya, I., Ondov, B., … Pop, M. (2013). MetAMOS: A modular and open source metagenomic assembly and analysis pipeline. *Genome Biology*, *14*(1), R2. https://doi.org/10.1186/gb-2013-14-1-r2

Troyo, A., Jose, S., & Rica, C. (2018). *A review of the genus Rickettsia in Central America*. 103–112.

Visconti, A., Martin, T. C., & Falchi, M. (2018). YAMP: A containerized workflow enabling reproducibility in metagenomics research. *GigaScience*, *7*(7), 1–9. https://doi.org/10.1093/gigascience/giy072

Vouga, M., & Greub, G. (2016). Emerging bacterial pathogens: The past and beyond. *Clinical Microbiology and Infection*, *22*(1), 12–21. https://doi.org/10.1016/j.cmi.2015.10.010

World Health Organization. (2015). *Managing possible serious bacterial infection in young infants when referral is not feasible*.

Wright, M. N., Gola, D., & Ziegler, A. (2017). *Preprocessing and Quality Control for Whole-Genome Sequences from the Illumina HiSeq X Platform*. https://doi.org/10.1007/978-1-61779-555-8

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., … Zhang, Y. Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, *579*(7798), 265–269. https://doi.org/10.1038/s41586-020-2008-3

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., & Madden, T. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. *Ncbi*, 13:134. https://doi.org/10.1186/1471-2105-13-134

Zhao, M., Liu, D., & Qu, H. (2017). Systematic review of next-generation sequencing simulators: computational tools, features and perspectives. *Briefings in Functional Genomics*, *16*(3), 121–128. https://doi.org/10.1093/bfgp/elw012

Zhao, M., Wang, Q., Wang, Q., Jia, P., & Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. *BMC Bioinformatics*, *14*(SUPPL11). https://doi.org/10.1186/1471-2105-14-S11-S1

Diccionario de términos

Pipeline bioinformático

Un *pipeline* bioinformático es un flujo de trabajo que describe una secuencia de pasos ordenados y organizados para lograr un objetivo concerniente a biología haciendo uso de herramientas informáticas (Costa, 2017; Djebali et al., 2016; Graña, López-Fernández, Fdez-Riverola, González Pisano, & Glez-Peña, 2018; Wright et al., 2017).

Patógeno

Los patógenos son agentes (entre ellos bacterias y virus) que causan enfermedades en un hospedero vivo (Henry & Heinke, 1999).

Rickettsiosis

Rickettsiosis son todas aquellas enfermedades infecciosas producidas por bacterias del género *Rickettsia* (Parola et al., 2008).

Bases

Son las moléculas adenina, timina, citosina y guanina. También son considerados bloques base que conforman el ADN (Brown, 2006).

Nucleótidos

Son bases unidas a otras moléculas de azúcar y fosfato (Brown, 2006).

Datos genómicos

Hace referencia a las secuencias que conforman la totalidad de moléculas de ADN de un organismo (Brown, 2006).

Read

Es una cadena de caracteres compuestas por nucleótidos (adenina, timina, citosina y guanina) como resultado de la secuenciación (Low & Tammi, 2017).

Contig

Es la abreviación de secuencia contigua y es resultado del ensamblaje (unión) de *reads* (Low & Tammi, 2017).

Adaptador

Es una secuencia de nucleótidos artificial y conocida que es añadida al final de la secuencia de la muestra de ADN fraccionada (Ishihara, Kotomura, Yamamoto, & Ochiai, 2017; Low & Tammi, 2017).

Primer

Es una secuencia de nucleótidos artificial y conocida, pero, a diferencia del adaptador, es añadida al inicio la lectura de la secuencia de ADN fraccionada (Low & Tammi, 2017; Ye et al., 2012).

Datos crudos

Son las secuencias (*reads*) obtenidas a través de la secuenciación de última generación (Low & Tammi, 2017).

Información genómica

Es el conocimiento obtenido luego de procesar los datos genómicos de una especie de interés (Brown, 2006).

***Phred score* (*Q score*)**

Es una medida de calidad basada en una escala logarítmica que mide la probabilidad de error de lectura de una base. Si bien, las plataformas de última generación son tecnologías potentes, presentan un margen de error que ocasiona falsos positivos en la secuenciación (Low & Tammi, 2017).

Anexos

Anexo A: Plan de proyecto

- **Justificación**

El panorama de la salud pública frente a las enfermedades transmitidas por vectores es alarmante. Cada año más de un billón de personas son infectadas y cerca de un millón mueren a causa de estas enfermedades. El resurgimiento de estas infecciones es reconocido por la Organización Mundial de la Salud - la agencia de salud más grande del mundo - y representa un problema de gran importancia para la salud pública en todos los países, sobre todo aquellos con zonas endémicas dentro de su geografía. Asimismo, el calentamiento global afecta la distribución y transmisión de los vectores, causando un incremento en su densidad poblacional. Como consecuencia, estos organismos amplían su horizonte de hospederos y propagan bacterias en zonas geográficas de todo el mundo, que anteriormente no eran afectadas. Este es el caso de las bacterias del género *Rickettsia*, que tienen presencia a nivel global, debido a que sus vectores son las pulgas, garrapatas, piojos y ácaros. Si bien existen especies de *Rickettsia* que producen infecciones que pueden ser diferenciadas e identificadas, existe un grupo cuyas implicancias en enfermedades en humanos aún no son concluyentes. Entre estas, la bacteria *Rickettsia asembonensis* ha sido detectada en pulgas, en casos febriles inespecíficos y en regiones que anteriormente no habían sido reportadas. Por tal motivo, es un claro ejemplo de una especie de *Rickettsia* emergente y subestimada, ya que se ha determinado su presencia en varias partes del mundo y, a pesar de tener presencia en casos febriles inespecíficos, la información concerniente a su diagnóstico y diferenciación sigue siendo escasa.

Por esta razón, es sumamente importante realizar esfuerzos para saber más de esta especie de bacteria. Dentro de la salud pública es realmente necesario definir su patogenicidad y comprobar si es un peligro para los seres humanos y animales. El presente proyecto ayudaría a contribuir al conocimiento científico con información genómica de importancia enfocada en la bacteria de la especie *R. asembonensis*. Este trabajo de fin de carrera pretende cubrir la escasez de información genómica de la bacteria *R. asembonensis* y beneficiaría a profesionales de la salud y científicos en general a desarrollar diagnósticos diferenciados en casos febriles inespecíficos relacionados a esta especie de microorganismo. Asimismo, el *pipeline* o flujo de trabajo a desarrollar, generará un precedente y referente metodológico para otras especies de interés con la misma problemática.

- **Viabilidad**
  - **Viabilidad técnica**

    El proyecto se considera viable técnicamente debido a que se recibe el apoyo del laboratorio de Genómica y el laboratorio de Inteligencia Artificial, que se encuentran dentro de la Pontificia Universidad Católica del Perú (PUCP). Es decir, se tiene acceso a los datos secuenciados de muestras de pulgas de la especie *C. felis* positivas para *R. asembonensis* que se recolectaron en un estudio llevado a cabo en Iquitos, en Perú. El conjunto de datos (que se encuentra en la escala de los cientos de gigabytes) está almacenado en discos duros y se cuenta con el equipamiento para procesarlo. Asimismo, se recibe apoyo de investigadores del grupo de Inteligencia Artificial de la PUCP que han realizado trabajos relacionados a bioinformática dentro y fuera del país.

Por otro lado, existe evidencia de que los conocimientos bioinformáticos están en pleno crecimiento y desarrollo, ya que solo una década atrás se obtuvo la totalidad del genoma humano. Por esta razón, es viable contar con herramientas y métodos desarrollados en los últimos años (como se mostró en el estado del arte) que contribuyen a la obtención del genoma de las especies. Además, otros conocimientos necesarios serán cubiertos con el apoyo de integrantes del laboratorio de genómica, de libros especializados (Pevsner, 2015; Phillip & Pavel, 2015) o cursos (*Bioinformatics Specialization* por la Universidad de San Diego y *Bioinformatics: Introductions and Methods* por la Universidad de Pekín) de gran reconocimiento.

- **Viabilidad temporal**

El proyecto consta de un plan de 5 meses. Las actividades del presente trabajo se centran en la investigación y aplicación de técnicas bioinformáticas, cuya curva de aprendizaje es compensada con el conocimiento previo de los expertos del grupo de Inteligencia Artificial de la PUCP y del laboratorio de Genómica. El desarrollo del *pipeline* consiste en la elaboración de los 3 pasos principales (procesamiento, ensamblaje y anotación), los cuales tienen una duración de 3 semanas cada uno en promedio. Asimismo, se está tomando en cuenta que los recursos computacionales estarán procesando un conjunto de datos en el rango de cientos de gigabytes. La implicancia del tiempo de ejecución del análisis de datos está incluida en el tiempo establecido.

- **Viabilidad económica**

El presente proyecto es viable económicamente, ya que se hará uso de herramientas y programas de código libre. Por ejemplo, los procesos se podrán

ejecutar en las distribuciones de Linux, las herramientas y programas a utilizar son de código libre e incluso los datos de referencia se encuentran en bases de datos públicas como GenBank. En cuanto a hardware, se cuenta con el apoyo del laboratorio de Genómica y del grupo de Inteligencia Artificial de la PUCP que tienen a disposición equipos con alto nivel de procesamiento.

- **Alcance, limitaciones y riesgos**
  - **Alcance**

El presente proyecto, el cual pertenece a la rama de Ciencia de la Computación, propone ensamblar y anotar el genoma de la bacteria *R. asembonensis* a través del desarrollo de un *pipeline* bioinformático. Se utilizarán datos secuenciados que fueron obtenidos en un estudio que se llevó a cabo en el año 2013 en la Amazonía peruana, en donde se recolectaron 284 ectoparásitos y de los cuales 190 eran pulgas de la especie *C. felis* positivas para *R. asembonensis*. En dicho estudio, se eligió la selva peruana porque es considerada zona endémica y la bacteria *R. asembonensis*, por ser un microorganismo emergente con potencial impacto negativo en la salud pública.

Todas las tareas por realizar se llevarán a cabo en el laboratorio de Genómica y en el laboratorio de Inteligencia Artificial, situados en el campus de la PUCP. Las secuencias se procesarán en cada etapa del *pipeline*: preprocesamiento, ensamblaje y anotación. Se utilizarán herramientas del estado del arte que permitirán asegurar la calidad de la secuencia y se comprobarán los resultados con métricas estadísticas. Una vez obtenidos los datos preprocesados, se realizará un análisis de herramientas de ensamblaje para, finalmente, anotar el genoma resultante. De esta manera, se logrará identificar la mejor opción respecto a los

demás ensambladores e identificar y proponer genes de la *R. asembonensis* peruana.

Por lo tanto, el presente proyecto busca obtener una secuencia genómica consenso anotada de la bacteria *R. asembonensis* a partir de un análisis comparativo de diferentes herramientas ensamblaje y anotación, garantizando un genoma sin contaminación y contiguo, con el objetivo de contribuir a la comunidad científica en general. Se partirá de los datos crudos de la pulga de la especie *C. felis* positivas para *R. asembonensis* para generar un precedente y referente metodológico de otras especies de interés con la misma problemática.

- **Limitaciones**

Se identificaron las siguientes limitaciones en el planteamiento del proyecto:

- Las tecnologías de última generación han reducido el costo y el tiempo de la secuenciación, sin embargo, generan *reads* muy pequeños y con una calidad que va decreciendo mientras ocurre el proceso de secuenciación.
- Las secuencias de nucleótidos de un vector como la pulga contienen material genético de diversos organismos. Es decir, debido a la naturaleza misma de los ectoparásitos, en el ADN es posible encontrar secuencias de microbiota intestinal o sangre de los hospederos que habitaron.
- La cantidad de datos que se procesarán se encuentran en el rango de los cientos de gigabytes. Por lo tanto, se tiene como limitación operativa la capacidad de los equipos que dispone la PUCP.
- Los conocimientos biológicos relacionados al presente proyecto son limitados, debido a la escasez de expertos en las áreas de bioinformática y biología molecular.

- **Riesgos**

Tabla A1

Riesgos identificados en el proyecto

Descripción	Síntomas	Proba-bilidad	Impac-to	Mitigación	Contingencia
Escasez de datos secuenciados para la generación del genoma de la bacteria *R. asembonensis*	Gran porcentaje de secuencias de ADN de baja calidad	Baja	Alto	Evitar la filtración y eliminación de los datos de baja calidad para retenerlos y emplearlos en la generación del genoma	Obtener datos secuenciados de la bacteria *R. asembonensis* desde las bases de datos públicas
Elevado tiempo de procesamiento de los datos en los equipos de cómputo de la universidad	Tiempos de ejecución real sobrepasa el tiempo de ejecución teórico	Media	Alto	Utilizar particiones o una muestra representativa del conjunto de datos	Utilizar los equipos que dispone el grupo de inteligencia artificial de la PUCP
Retraso en el desarrollo del proyecto	No finalizar una actividad en la semana establecida	Baja	Medio	Tomar en cuenta el tiempo que se invierte en actividades que afectan el cronograma	Reajustar el cronograma de actividades con el fin de evitar otro retraso en las semanas siguientes

- **Estructura de descomposición del trabajo (EDT)**

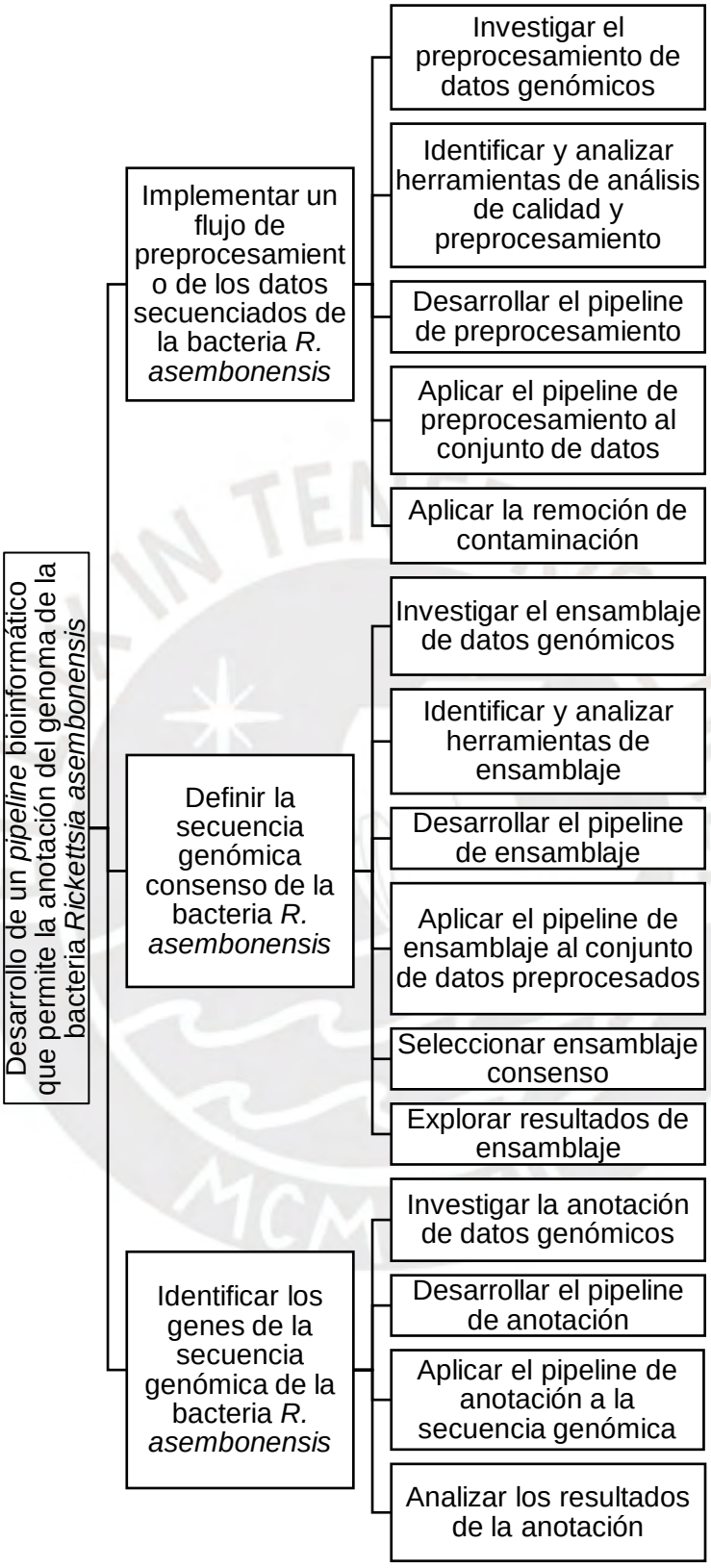

Figura A1.Diagrama de estructura de descomposición del trabajo (EDT)

- **Lista de tareas**

  - **Implementar un flujo de preprocesamiento de los datos secuenciados de la bacteria *R. asembonensis*.**

    - Investigar el preprocesamiento de datos genómicos
    - Identificar y analizar herramientas de análisis de calidad y preprocesamiento
    - Desarrollar el *pipeline* de preprocesamiento
    - Aplicar el *pipeline* de preprocesamiento al conjunto de datos
    - Aplicar la remoción de contaminación

  - **Definir la secuencia genómica consenso de la bacteria *R. asembonensis***

    - Investigar el ensamblaje de datos genómicos
    - Identificar y analizar herramientas de análisis de calidad y ensamblaje
    - Desarrollar el *pipeline* de ensamblaje
    - Aplicar el *pipeline* de ensamblaje al conjunto de datos preprocesados
    - Seleccionar ensamblaje consenso
    - Explorar resultados de ensamblaje

  - **Identificar los genes de la secuencia genómica de la bacteria *R. asembonensis***

    - Investigar la anotación de datos genómicos
    - Desarrollar el *pipeline* de anotación
    - Aplicar el *pipeline* de anotación a la secuencia genómica
    - Analizar los resultados de la anotación

- **Lista de recursos**
    - **Personas involucradas**
        - Autor del proyecto
        - Asesor
        - Integrantes del laboratorio de genómica de la Pontificia Universidad Católica del Perú
    - **Equipamiento**
        - Computadoras y herramientas instaladas
    - **Herramientas**
        - NCBI
        - Web of Science
        - Scopus
        - FastQC
        - Fastp
        - BBDuk
        - Bowtie2
        - Megahit
        - Ray Meta
        - Abyss
        - Quast
        - Samtools
        - BLAST
        - NCBI Prokaryotic Genome Annotation Pipeline

- **Costeo del proyecto**

Tabla A2

Costeo del proyecto

Ítem	Descripción	Unidad	Cantidad	Valor Unitario (S/.)	Monto Total (S/.)	Monto Acumulado (S/.)
0	Costo total del proyecto					4,540
1	Estudiantes					
1.1	Autor del proyecto	Horas	60	10	600	600
2	Otros participantes					
2.1	Integrante 1 del laboratorio de genómica	Horas	6	30	180	540
2.2	Integrante 2 del laboratorio de genómica	Horas	6	30	180	
2.3	Asesor del proyecto	Horas	6	30	180	
3	Materiales e insumos					
3.1	Discos de almacenamiento externos	Unidad	1	400	400	400
4	Bienes y equipos					
4.1	Computadora	Equipo	1	3,000	3,000	3,000

Anexo B: Cronograma de actividades

Tabla B1

Cronograma de actividades

Actividades	Semanas																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Implementar un flujo de preprocesamiento de los datos secuenciados de la bacteria *R. asembonensis*.																				
Investigar el preprocesamiento de datos genómicos	X	X																		
Identificar y analizar herramientas de análisis de calidad y preprocesamiento			X	X																
Desarrollar el *pipeline* de preprocesamiento					X															
Aplicar el *pipeline* de preprocesamiento al conjunto de datos						X														
Aplicar la remoción de contaminación							X													
Definir la secuencia genómica consenso de la bacteria *R. asembonensis*																				
Investigar el ensamblaje de datos genómicos								X	X											
Identificar y analizar herramientas de ensamblaje										X	X									
Desarrollar el *pipeline* de ensamblaje												X								
Aplicar el *pipeline* de ensamblaje al conjunto de datos preprocesados													X							
Seleccionar ensamblaje consenso														X						
Explorar resultados de ensamblaje															X					
Identificar los genes de la secuencia genómica de la bacteria *R. asembonensis*																				
Investigar la anotación de datos genómicos																X	X			
Desarrollar el *pipeline* de anotación																		X		
Aplicar el *pipeline* de anotación a la secuencia genómica																			X	
Analizar los resultados de la anotación																				X

Anexo C: Análisis de calidad de los datos secuenciados

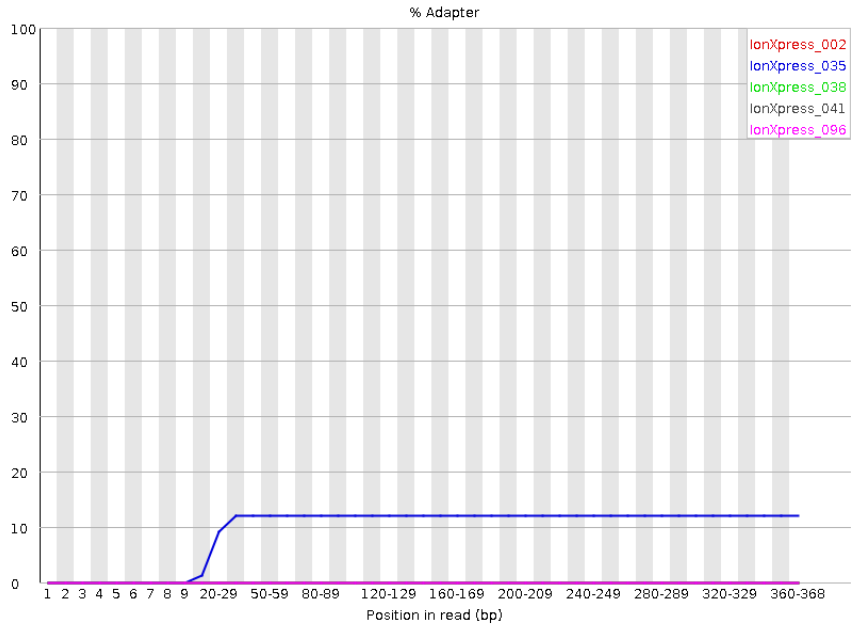

Figura C1. Presencia de adaptadores

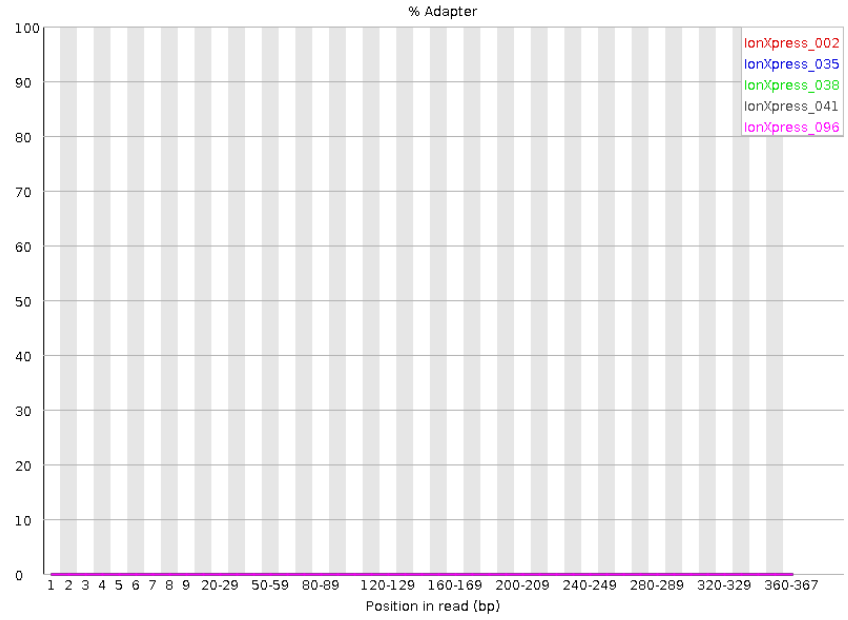

Figura C2. Remoción de adaptadores

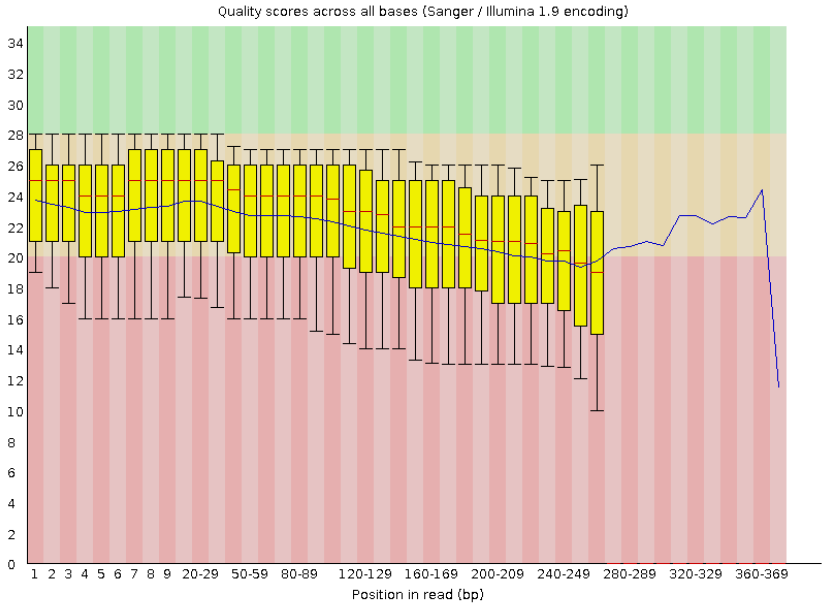

Figura C3. Filtro por calidad

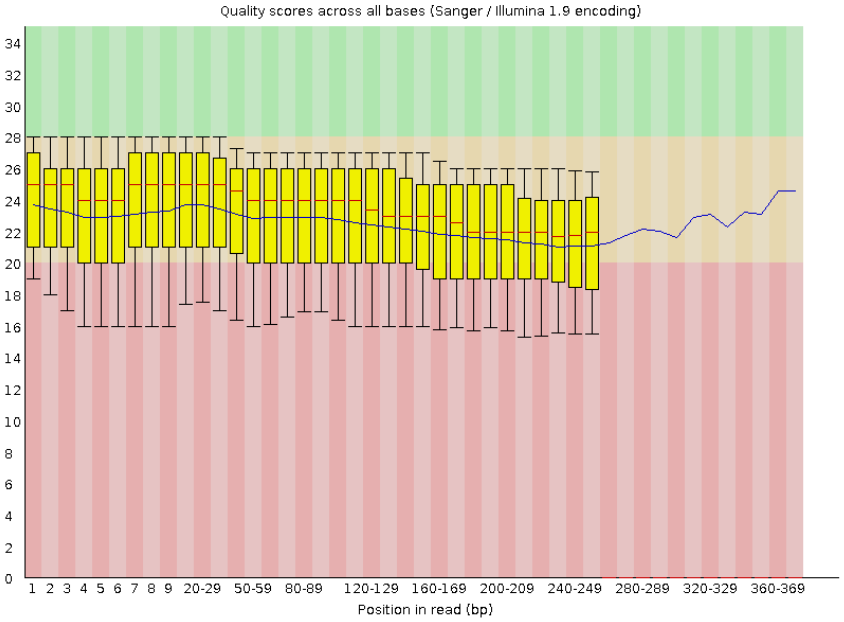

Figura C4. Corte por calidad

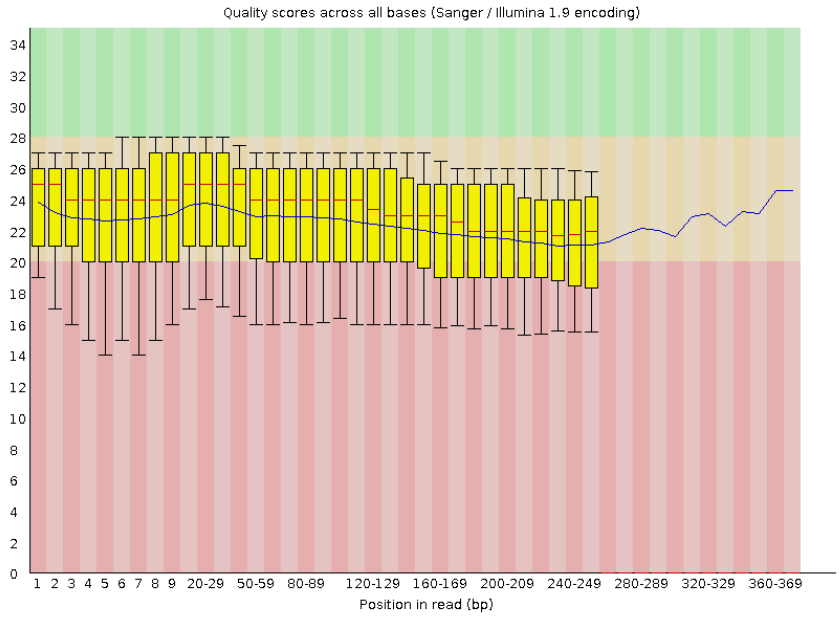

Figura C5. Filtro por longitud

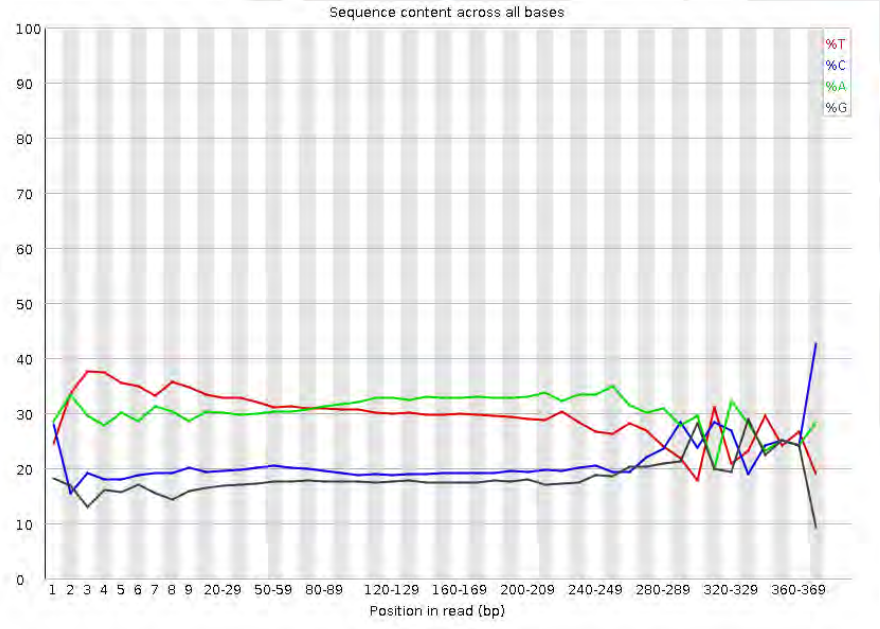

Figura C6. Recorte por ruido

Anexo D: Alineación de *reads* con *Host Removal* y sin *Host Removal* con la bacteria *R. asembonensis*

Figura D1. Vista general de los *reads* alineados. Ambos coinciden porque no existe contaminación

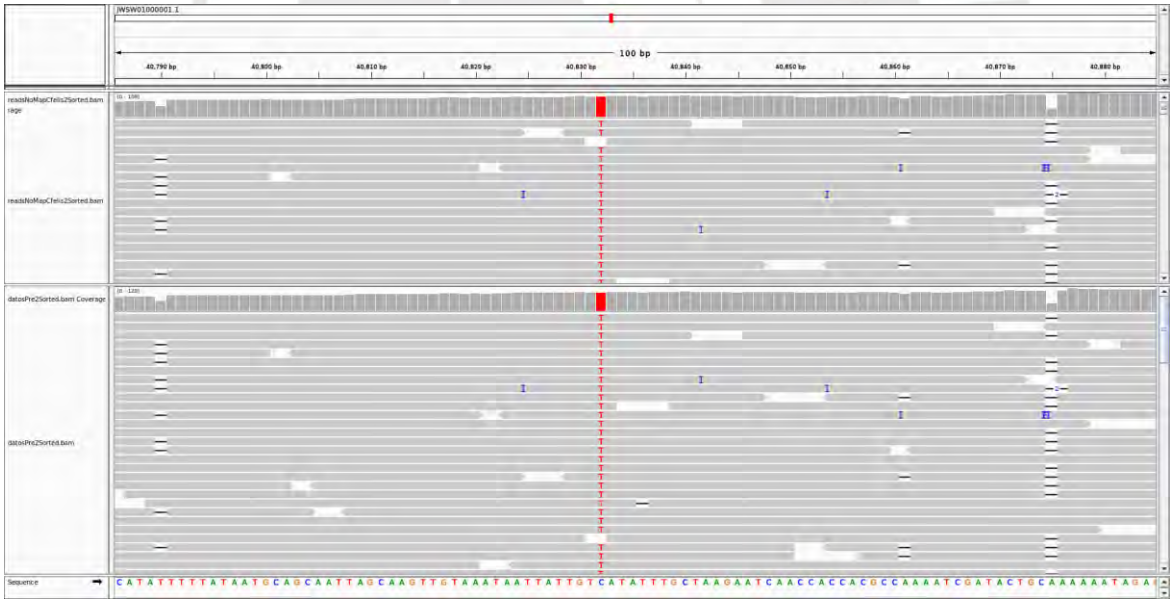

Figura D2. Acercamiento a una alineación que corrige la secuencia genómica de referencia. Todos los *reads* alineados especifican timina en esa posición, mientras que la referencia presenta una citosina.

Anexo E: Ensamblaje de los datos preprocesados

Worst Median Best ☑ Show heatmap

Genome statistics	Abyss	Megahit (No Meta)	Ray (No Meta)	Megahit (Meta)
Genome fraction (%)	80.452	96.465	86.283	96.412
GCA_000828125.2_ASM82812v2_genomic	78.399	94.046	83.835	94.022
Duplication ratio	1.649	1.31	1.012	1.501
GCA_000828125.2_ASM82812v2_genomic	1.59	1.214	1.011	1.329
# genomic features	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part
GCA_000828125.2_ASM82812v2_genomic	0 + 0 part	0 + 0 part	0 + 0 part	0 + 0 part
Largest alignment	1786	22 910	8103	12 823
GCA_000828125.2_ASM82812v2_genomic	1786	22 910	8103	12 823
Total aligned length	1 710 845	1 557 610	1 164 849	1 648 921
GCA_000828125.2_ASM82812v2_genomic	1 710 120	1 556 180	1 164 230	1 647 538
NA50	230	2899	1675	-
GCA_000828125.2_ASM82812v2_genomic	231	2911	1681	2140
NA75	205	1256	948	-
GCA_000828125.2_ASM82812v2_genomic	205	1282	954	548
LA50	2302	161	215	-
GCA_000828125.2_ASM82812v2_genomic	2287	160	214	221
LA75	4361	366	444	-
GCA_000828125.2_ASM82812v2_genomic	4335	363	442	616
NG50	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	274	3855	1435	3528
NG75	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	207	2169	623	2148
NGA50	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	272	3297	1431	2961
NGA75	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	206	1830	603	1634
LG50	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	1595	112	280	120
LG75	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	3073	231	639	247
LGA50	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	1605	129	282	140
LGA75	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	3093	266	646	294

Figura E1. Primera parte de las métricas de desempeño de los ensambladores

Misassemblies				
# misassemblies	5	151	7	236
GCA_000828125.2_ASM82812v2_genomic	2	145	6	233
# relocations	0	3	0	1
GCA_000828125.2_ASM82812v2_genomic	0	3	0	3
# translocations	4	144	7	227
GCA_000828125.2_ASM82812v2_genomic	1	138	6	222
# inversions	1	4	0	8
GCA_000828125.2_ASM82812v2_genomic	1	4	0	8
# interspecies translocations	0	0	0	0
GCA_000828125.2_ASM82812v2_genomic	0	0	0	0
# misassembled contigs	4	135	7	209
GCA_000828125.2_ASM82812v2_genomic	2	128	6	207
Misassembled contigs length	1245	350 160	14 397	426 062
GCA_000828125.2_ASM82812v2_genomic	400	343 707	13 667	425 462
# possibly misassembled contigs	0	1	0	56
GCA_000828125.2_ASM82812v2_genomic	0	1	0	56
# possible misassemblies	0	1	0	59
GCA_000828125.2_ASM82812v2_genomic	0	1	0	59
# local misassemblies	2	18	8	32
GCA_000828125.2_ASM82812v2_genomic	3	15	8	28
# scaffold gap ext. mis.	0	0	0	0
GCA_000828125.2_ASM82812v2_genomic	0	0	0	0
# scaffold gap loc. mis.	0	0	0	0
GCA_000828125.2_ASM82812v2_genomic	0	0	0	0
# unaligned mis. contigs	1	3	1	4
GCA_000828125.2_ASM82812v2_genomic	1	3	0	5
Unaligned				
# fully unaligned contigs	25	11	7	1 256 065
Fully unaligned length	7474	5833	3241	674 915 689
# partially unaligned contigs	0	1	1	58
GCA_000828125.2_ASM82812v2_genomic	0	0	2	58
Partially unaligned length	0	1177	1043	64 083
GCA_000828125.2_ASM82812v2_genomic	0	0	1788	64 110
Mismatches				
# mismatches	3259	5344	1436	7627
GCA_000828125.2_ASM82812v2_genomic	2388	3216	1368	3766
# indels	9328	13 383	4017	17 225
GCA_000828125.2_ASM82812v2_genomic	8613	10 011	3872	11 879
Indels length	9995	15 584	4821	20 166
GCA_000828125.2_ASM82812v2_genomic	9247	11 724	4658	13 895
# mismatches per 100 kbp	294.24	402.39	120.89	574.61
GCA_000828125.2_ASM82812v2_genomic	221.24	248.38	118.52	290.94
# indels per 100 kbp	842.17	1007.7	338.16	1297.71
GCA_000828125.2_ASM82812v2_genomic	797.98	773.19	335.47	917.69
# indels (<= 5 bp)	9288	13 307	3976	17 124
GCA_000828125.2_ASM82812v2_genomic	8574	9947	3833	11 812
# indels (> 5 bp)	40	76	41	101
GCA_000828125.2_ASM82812v2_genomic	39	64	39	67
# N's	0	0	0	0
GCA_000828125.2_ASM82812v2_genomic not_aligned	0	0	0	0
# N's per 100 kbp	0	0	0	0
GCA_000828125.2_ASM82812v2_genomic not_aligned	0	0	0	0

Figura E2. Segunda parte de las métricas de desempeño de los ensambladores

Statistics without reference				
# contigs	6445	1077	983	1 257 516
GCA_000828125.2_ASM82812v2_genomic	6420	1066	976	1451
not_aligned	25	11	7	1 256 065
# contigs (>= 0 bp)	18 023	1077	1078	1 257 516
# contigs (>= 1000 bp)	24	405	419	62 738
GCA_000828125.2_ASM82812v2_genomic	24	404	418	440
not_aligned	0	1	1	62 298
# contigs (>= 5000 bp)	0	65	10	98
GCA_000828125.2_ASM82812v2_genomic	0	65	10	65
not_aligned	0	0	0	33
# contigs (>= 10000 bp)	0	8	0	4
GCA_000828125.2_ASM82812v2_genomic	0	8	0	4
not_aligned	0	0	0	0
# contigs (>= 25000 bp)	0	0	0	0
GCA_000828125.2_ASM82812v2_genomic	0	0	0	0
not_aligned	0	0	0	0
# contigs (>= 50000 bp)	0	0	0	0
GCA_000828125.2_ASM82812v2_genomic	0	0	0	0
not_aligned	0	0	0	0
Largest contig	1816	22 935	8103	13 016
GCA_000828125.2_ASM82812v2_genomic	1816	22 935	8103	13 016
not_aligned	528	1305	1240	7690
Total length	1 724 677	1 579 025	1 171 871	676 701 462
GCA_000828125.2_ASM82812v2_genomic	1 717 203	1 573 192	1 168 630	1 785 773
not_aligned	7474	5833	3241	674 915 689
Total length (>= 0 bp)	3 379 899	1 579 025	1 185 858	676 701 462
Total length (>= 1000 bp)	29 771	1 292 785	858 408	82 097 414
GCA_000828125.2_ASM82812v2_genomic	29 771	1 291 480	857 168	1 330 371
not_aligned	0	1305	1240	80 767 043
Total length (>= 5000 bp)	0	483 647	61 344	645 553
Total length (>= 10000 bp)	0	103 986	0	47 808
GCA_000828125.2_ASM82812v2_genomic	0	103 986	0	47 808
not_aligned	0	0	0	0
Total length (>= 25000 bp)	0	0	0	0
Total length (>= 50000 bp)	0	0	0	0
GCA_000828125.2_ASM82812v2_genomic	0	0	0	0
not_aligned	0	0	0	0
N50	233	3321	1696	557
GCA_000828125.2_ASM82812v2_genomic	233	3326	1696	2641
not_aligned	339	501	434	556
N75	205	1451	965	419
GCA_000828125.2_ASM82812v2_genomic	205	1463	968	964
not_aligned	224	395	344	418
L50	2280	141	214	424 472
GCA_000828125.2_ASM82812v2_genomic	2272	140	213	188
not_aligned	10	4	2	424 903
L75	4330	317	440	775 871
GCA_000828125.2_ASM82812v2_genomic	4315	314	438	450
not_aligned	17	7	5	775 647
GC (%)	...	...	...	...
GCA_000828125.2_ASM82812v2_genomic	34.36	32.78	32.38	32.67
not_aligned	38.04	37.1	34.68	45.3

Figura E3. Tercera parte de las métricas de desempeño de los ensambladores

Anexo F: Archivos de entrada para la anotación con PGAP

```
fasta:
  class: File
  location: megahit_no_meta_dedupe_copia.fa
submol:
  class: File
  location: submol.yaml
```

Figura F1. Archivo de entrada (input.yaml) que hace referencia al archivo fasta a anotar y a sus metadatos

```
topology: circular
organism:
    genus_species: 'Rickettsia asembonensis'
contact_info:
    last_name: 'Arauco'
    first_name: 'Ronie'
    email: 'ronie.arauco@pucp.edu.pe'
    organization: 'Pontificia Universidad Catolica del Peru'
    department: 'Ingenieria Informatica'
    street: 'Av. Universitaria 1801'
    city: 'Lima'
    postal_code: '15088'
    country: 'Peru'

authors:
    -     author:
            first_name: 'Ronie'
            last_name: 'Arauco'
```

Figura F 2Archivo de entrada (submol.yaml) que contiene los metadatos del resultado

Anexo G: Visualización de los genes identificados del genoma de *R. asembonensis* en Artemis.

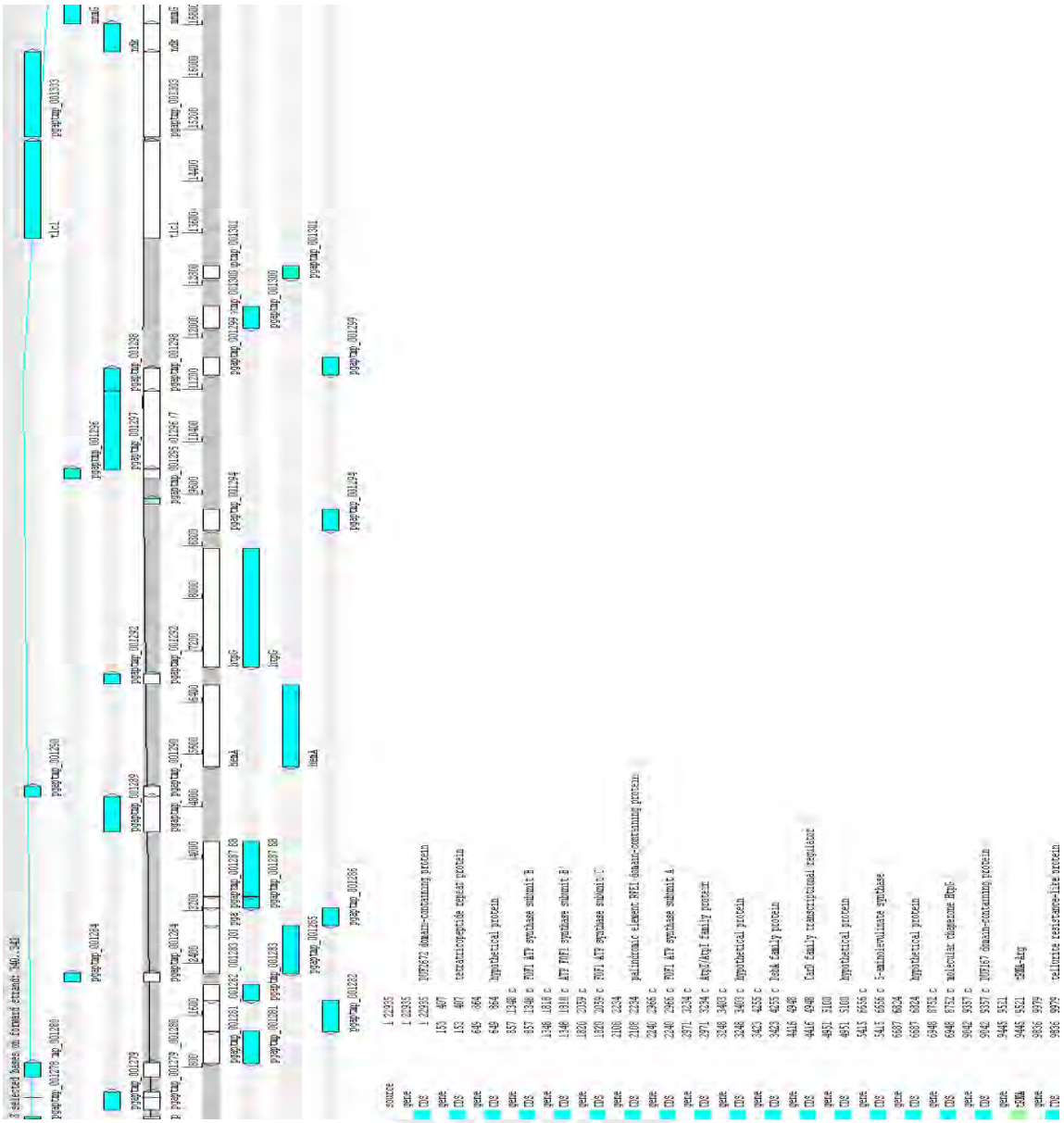

Figura G1. Visualización de genes identificados del genoma de *R. asembonensis*

Anexo H: Potenciales genes y pseudogenes de la *R. asembonensis* peruana.

Genes

- DNA processing protein DprA
- YidC/Oxa1 family membrane protein insertase
- UDP-2,3-diacylglucosamine diphosphatase LpxI
- MazF family transcriptional regulator
- DNA-processing protein DprA
- FAD-dependent monooxygenase
- DUF4258 domain-containing protein
- FAD-binding oxidoreductase
- peptidase
- tRNA-OTHER
- SDR family NAD(P)-dependent oxidoreductase
- enoyl-CoA hydratase/isomerase family protein

Pseudo

- DUF1189 domain-containing protein
- tol-pal system YbgF family protein
- threonine/serine dehydratase
- 50S ribosome-binding GTPase
- DNA replication and repair protein RecF
- 3-demethylubiquinone-9 3-O-methyltransferase
- 3-hydroxyacyl-CoA dehydrogenase family protein
- PHP domain-containing protein
- cytochrome b/b6 domain-containing protein

- cation:proton antiporter subunit C
- tRNA(Ile)-lysidine synthetase
- response regulator
- winged helix-turn-helix domain-containing protein
- UDP-phosphate alpha-N-acetylglucosaminyltransferase
- DUF4385 family protein
- nucleoside triphosphate hydrolase
- inorganic pyrophosphatase
- ATP-grasp domain-containing protein
- aminotransferase class I/II-fold pyridoxal phosphate-dependent enzyme
- DNA-binding response regulator
- aldehyde dehydrogenase family protein
- citrate (Si)-synthase
- 2-polyprenylphenol hydroxylase
- ATPase
- 16S rRNA (cytosine(1402)-N(4))-methyltransferase
- acyl-CoA desaturase
- SEC-C domain-containing protein
- zinc ABC transporter substrate-binding protein
- class I tRNA ligase family protein
- cell division protein FtsK
- DNA topoisomerase IV
- cytochrome C biogenesis protein CcmF
- GTP-binding protein
- PCRF domain-containing protein

- ATP-dependent protease
- TGS domain-containing protein
- CADD family putative folate metabolism protein
- 2-oxo acid dehydrogenase subunit E2
- lipopolysaccharide 1,2-glucosyltransferase
- YhcG family protein
- prolyl endopeptidase
- ribonucleotide-diphosphate reductase subunit alpha
- histidine kinase
- ribosomal protein L16
- AMP-binding protein
- RsmD family RNA methyltransferase
- gamma-glutamyl-gamma-aminobutyrate hydrolase
- methylmalonyl-CoA carboxyltransferase
- MerR family transcriptional regulator
- CBS domain-containing protein
- polysaccharide biosynthesis protein
- DUF1009 domain-containing protein
- poly(A) polymerase
- radical SAM family heme chaperone HemW
- DUF2460 domain-containing protein
- aconitate hydratase
- protein U
- DUF1311 domain-containing protein
- transglycosylase SLT domain-containing protein

- dihydropteroate synthase
- C4-dicarboxylate ABC transporter substrate-binding protein
- DUF21 domain-containing protein
- bifunctional 3-demethylubiquinol 3-O-methyltransferase/2-polyprenyl-6-hydroxyphenol methylase
- peptidase M15
- HAMP domain-containing protein
- 50S ribosomal protein L25
- TlyA family rRNA (cytidine-2'-O)-methyltransferase
- biotin synthase
- glutamine ABC transporter ATP-binding protein GlnQ
- Asp-tRNA(Asn)/Glu-tRNA(Gln) amidotransferase GatCAB subunit C
- mannose-1-phosphate guanyltransferase
- MobA/MobL family protein

Anexo I: *Pipeline* desarrollado en el presente trabajo.

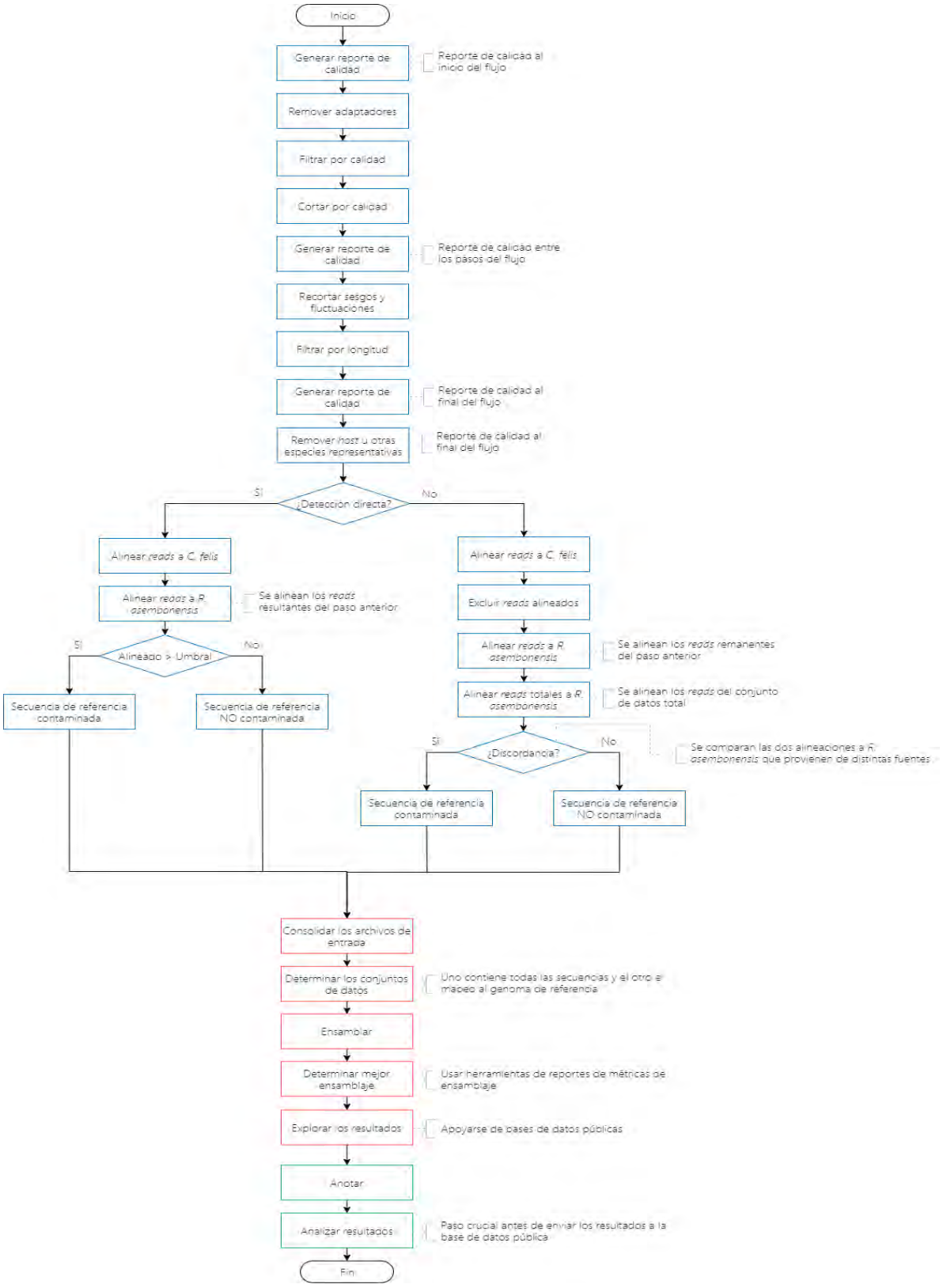

Figura I1. *Pipeline* desarrollado en el presente trabajo