Bioinformática aplicada al estudio genómico de las cepas endófitas *Kosakonia* sp. UYSO10 y *Rhizobium* sp. UYSO24

Martín Beracochea

Tesis de Maestría presentada al Programa de Posgrado en Bioinformática, PEDECIBA, como parte de los requisitos necesarios para la obtención del título de Magíster en Bioinformática.

Directores:

Dr. Federico Battistoni

Dr. Álvaro Martín

Codirector:

Dr. José Sotelo-Silveira

Montevideo – Uruguay

Diciembre de 2018

INTEGRANTES DEL TRIBUNAL DE DEFENSA DE TESIS

Dr. Andrés Iriarte

Dr. Héctor Romero

Dr. Gustavo Vázquez

Montevideo – Uruguay
Diciembre de 2018

A Cintia y Emi.

Agradecimientos

A mis tutores Federico, Álvaro y José. Fue un largo camino pero gracias a su apoyo constante aprendí mucho.

A todos los compañeros de laboratorio. Especialmente a Ceci con la cual compartimos no solamente el objeto de estudio sino el proceso de escritura.

A mi familia por todo el aguante. También a mi familia extendida.

A mis amigos.

A Cintia y Emi, ellas hicieron posible terminar este trabajo. A ellas les dedico esta tesis.

"I love deadlines. I like the whooshing sound they make as they fly by".

Douglas Adams

RESUMEN

Los estudios genómicos empleando plataformas de secuenciación del ADN de segunda generación es un área en constante desarrollo. Uno de los principales problemas que presentan dichas plataformas es la alta tasa de error. En particular, datos de secuenciación de genomas bacterianos pueden ser sensiblemente mejorados mediante el empleo de herramientas de corrección de errores sin incurrir en costos económicos adicionales. La genómica bacteriana aplicada al estudio de bacterias endófitas tiene el potencial de brindar información muy útil para entender la interacción con la planta huésped. Mediante esta aproximación se han caracterizado bacterias endófitas asociadas a cultivos de interés agronómico tales como la caña de azúcar, el arroz y el maíz con el fin de aportar a su aplicación biotecnológica. Estudios previos demostraron que las cepas *Kosakonia* sp. UYSO10 y *Rhizobium* sp. UYSO24 son promotores del crecimiento vegetal y endófitos verdaderos, de variedades de caña de azúcar cultivadas en Uruguay y son empleados como modelo de estudio en el laboratorio. El objetivo general de este trabajo de tesis fue desarrollar una herramienta informática para la corrección de errores de secuenciación, capaz de ser aplicada a un conjunto de datos genómicos de dos cepas bacterianas modelos, proveniente de la plataforma Ion Torrent PGM. Los resultados mostraron que la herramienta de corrección de errores de secuenciación de ADN desarrollada mejora algunos aspectos de los datos. Pero es superada por otras alternativas, las cuales fueron utilizadas para el desarrollo del resto del trabajo de tesis. Por otro lado, se realizó la corrección, ensamblado y anotación de los datos de secuenciación de los genomas de las cepas *Kosakonia* sp. UYSO10 y *Rhizobium* sp. UYSO24. Dichos genomas fueron caracterizados, poniendo énfasis en las características genómicas relacionadas a la interacción planta-endófito y la promoción del crecimiento vegetal.

Palabras claves:
bioinformática, ion-torrent, endófito, genómica.

Lista de figuras

Lista de tablas

Tabla de contenidos

Capítulo 1

Introducción

1.1. Secuenciación de ADN

Desde el descubrimiento de la estructura del ácido desoxirribonucleico (ADN), se han logrado numerosos avances en el entendimiento de la diversidad y complejidad de los genomas de los seres vivos. Posteriormente, con el desarrollo de las técnicas que permitieron la secuenciación del ADN, se logró dar un salto significativo en el conocimiento sobre su funcionamiento. Particularmente, la secuenciación de los genomas bacterianos aportó información valiosa sobre cómo las bacterias interaccionan entre sí, con sus huéspedes y con el medio ambiente. El primer genoma bacteriano secuenciando fue el de la cepa *Haemophilus influenzae* (50; 48) y al día de hoy el número ha aumentado exponencialmente debido al desarrollo de nuevas técnicas. El mencionado genoma fue secuenciado mediante la técnica *shotgun*, desarrollada como alternativa al proceso de secuenciación empleado para el genoma humano el cual fue considerado lento y laborioso (50). El método de *shotgun* se basa en la fragmentación del ADN total, el clonado de los fragmentos obtenidos en vectores adecuados y la posterior secuenciación de dichos fragmentos clonados. El desarrollo de este tipo de técnica es considerada la primera revolución en la secuenciación de ADN (95). El siguiente hito se alcanzó en el año 2004 con el desarrollo de la primera plataforma de secuenciación de alto rendimiento, o de segunda generación. Esta tecnología produce enormes cantidades de datos con un costo inferior a las metodologías de primera generación basadas en la secuenciación de Sanger (141). El término secuenciación de segunda generación comprende a las plataformas de lectura corta y gran rendimiento como: Roche 454 (103),

Illumina (19) y Ion Torrent (138). Desde su aparición las tecnologías de secuenciación de segunda generación han progresado rápidamente, incrementando su rendimiento en términos de velocidad y volumen de datos obtenidos, en un factor de 100 a 1000 veces (57). Existen tres diferencias principales entre estos dos tipos de tecnologías: 1- no es necesario clonar los fragmentos de ADN en vectores biológicos porque las muestras se preparan con kits libres de células; 2- es un proceso que se lleva a cabo en paralelo en el cual millones de moléculas son secuenciadas al mismo tiempo; y 3- las bases secuenciadas son detectadas directamente sin necesidad de realizar una electroforesis. El alto rendimiento y bajo costo han permitido que la secuenciación masiva forme parte de la infraestructura de laboratorios pequeños y no solamente de centros especializados. Gracias a este tipo de tecnologías se han podido secuenciar genomas complejos como el de *Sorghum bicolor* (117). Sin embargo, genomas más complejos, como el de *Amoeba dubia* con un tamaño de 670 gigabases, siguen siendo difíciles de ensamblar (3). A pesar de las ventajas antes mencionadas, este conjunto de técnicas presentan algunos problemas como la alta tasa de error en la secuenciación. En este sentido se han desarrollado múltiples estrategias con el fin de mitigarlo y poder aprovechar así todo su potencial.

1.1.1. Aplicaciones de la secuenciación de ADN

Las plataformas de secuenciación de segunda generación han transformado profundamente la biología ya que estudios que hasta hace pocos años eran impracticables, han sido posibles gracias al gran rendimiento de esta tecnología. En biología humana, por ejemplo, se han completado ambiciosos proyectos de genómica de poblaciones (1; 168), proporcionando valiosa información sobre enfermedades como el cáncer y la diabetes (122). Particularmente en el área de la microbiología, la genómica microbiana ha sido ampliamente potenciada por el desarrollo de las técnicas de secuenciación de segunda generación. Esto se ve reflejado por ejemplo en el banco de datos del *National Center for Biotechnology Information* (NCBI) de Estados Unidos, donde están depositados a la fecha 9061 genomas bacterianos secuenciados completamente y 118.062 genomas secuenciados a nivel de borrador. Asimismo, el gran desarrollo de la genómica ha potenciado y hecho posible el desarrollo de otras disciplinas complementarias como: la proteómica, la trancriptómica, la metabolómica, el modelado metabólico (159), la biovigilancia (86) y la epidemiología (38). El

estudio de un microorganismo, o comunidad microbiana, mediante el análisis genómico o metagenómico, en complementación con el de sus perfiles transcripcionales, es capaz de brindar información respecto a las vías metabólicas o genes expresados en determinado momento por el o los microorganismos en estudio. En ese sentido, la metagenómica ha sido empleada para el estudio de las comunidades presentes en distintos ambientes como el rumen vacuno (109), los suelos (31), el cuerpo humano (76), así como a las bacterias asociadas a las plantas (123).

1.2. Aplicación de la genómica al estudio de la interacción planta-bacterias

Las plantas interactúan con comunidades microbianas complejas las cuales juegan un rol determinante en la salud vegetal así como en su productividad (20). Entender los procesos que gobiernan esa interacción permite aportar conocimientos imprescindibles para desarrollar estrategias que lleven a combatir enfermedades y/o mejorar el rendimiento de los sistemas agrícolas (53). Las bacterias asociadas a las plantas ocupan distintos ambientes, tales como el suelo cercano a las raíces (rizósfera) (20), la superficie de las raíces (rizoplano) (66) o sus tejidos internos (77). Particularmente, el conjunto de bacterias capaces de colonizar el interior de los tejidos vegetales sin causar daño aparente a su huésped son llamados **endófitos** bacterianos (64). A su vez, dentro del conjunto de bacterias endófitas existe un sub-grupo que pueden mantener una relación mutualista con la planta, llamadas **bacterias promotoras del crecimiento vegetal** (BPCV) (65).

1.3. Interacción entre los endófitos y las plantas

La interacción entre los endófitos y las plantas constituye un proceso dinámico que depende de un ambiente adecuado y de factores genéticos. La principal vía de colonización de los tejidos vegetales se da por bacterias del suelo, dividiéndose el proceso de interacción en tres etapas: 1- acercamiento y adhesión a la superficie radicular, 2- colonización del rizoplano y 3- infección

y colonización de los tejidos radiculares. El primer paso de la interacción involucra el reconocimiento inicial de señales moleculares, seguido del movimiento de la bacteria en dirección a la planta hospedera, su adhesión a la superficie vegetal, colonización del rizoplano y posterior penetración y multiplicación en el interior de los tejidos, pudiendo alcanzar una dispersión sistémica a través de los tejidos vasculares (101) (Figura 1.1).

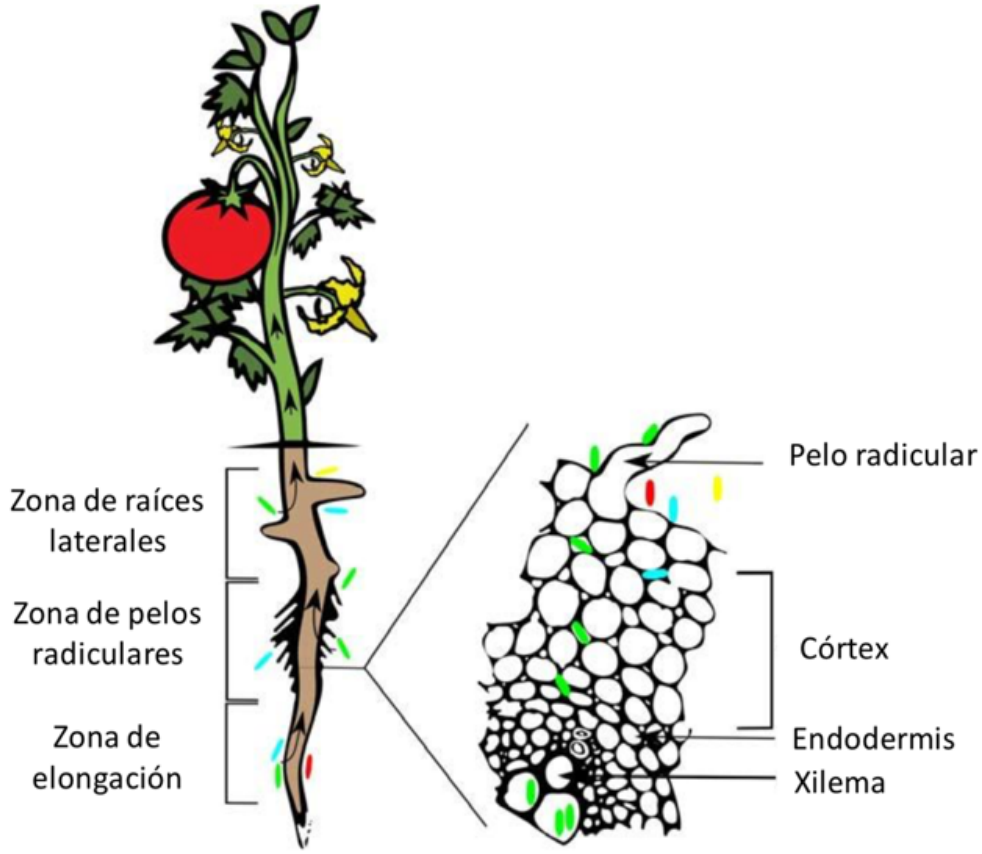

Figura 1.1: Principales rutas de colonización de la planta por bacterias endófitas. Los puntos de ingreso de las bacterias a las raíces de las plantas son múltiples. Una vez que ingresan los endófitos pueden permanecer en la zona de entrada (celeste) o migrar y colonizar el espacio intercelular del córtex y los vasos del xilema (verde). En rojo y amarillo se representan las bacterias rizósfericas que son incapaces de colonizar los tejidos internos de la planta. Tomado con modificaciones de Malfanova y colaboradores (2013).

1.3.1. Acercamiento y adhesión a la superficie radicular

Las bacterias pueden aproximarse a las raíces de forma pasiva, si la bacteria se encuentra en la semilla o en la yema de propagación vegetativa, o activa

a través del movimiento por atracción quimiotáctica (137). El desplazamiento hacia las raíces se da generalmente por medio de flagelos. Las moléculas exudadas por las raíces como los ácidos orgánicos, carbohidratos y aminoácidos; son los quimioatrayentes más comunes (65). Después que la bacteria se acerca lo suficiente a la raíz, estas comienzan a interaccionar con la pared celular vegetal a través de uniones electrostáticas débiles, las cuales son rápidas, inespecíficas y reversibles (170).

1.3.2. Colonización del rizoplano

Una vez que la bacteria entra en contacto con la pared celular vegetal, comienza el proceso de adhesión de las células a los tejidos radiculares. Este proceso involucra distintas estructuras bacterianas como fimbrias, *pillus*, lipopolisacáridos (LPS), exopolisacáridos (EPS) y proteínas de membrana externa como porinas y la proteína mayor de membrana externa (105; 26; 42; 29). Se ha visto que las bacterias se establecen en forma de biopelículas o agregados en los sitios de adhesión. Las biopelículas son microambientes óptimos para la transferencia horizontal de genes, así como para el intercambio de señales de *quorum sensing*, donde las capas, compuestas en parte por EPS, cumplen la función de barrera de difusión de dichas moléculas.

1.3.3. Infección y colonización de los tejidos radiculares

Luego de que las bacterias colonizaron los tejidos de la raíz es posible que estas colonicen los tejidos internos. El ingreso a los tejidos de la planta generalmente se da por mecanismos pasivos, donde la infección se realiza a través de las aperturas naturales como el área de emergencia de las raíces laterales y de la cofia (101; 68), así como a través de heridas causadas por factores bióticos o abióticos. Por otro lado, la infección activa involucra la participación de vectores (insectos) o la participación de enzimas degradadoras de la pared celular (37). Las enzimas antes mencionadas están también implicadas en el desencadenamiento de las vías de defensa en la planta. La inducción de esta respuesta resulta en una disminución de la propagación dentro de la planta, por ejemplo, de patógenos. En el caso de las bacterias endófitas, estas deben ser capaces de mitigar la respuesta inmune de la planta para poder lograr una infección efectiva. Una vez dentro de la planta, es posible que las bacterias permanezcan en un tejido especifico o logren colonizar la planta sistemáticamente, trans-

portándose por los tejidos vasculares o el apoplasto (137). Aquellas bacterias endófitas que poseen los mecanismos para lograr infectar los tejidos internos de la planta y logran permanecen este ambiente son denominados endófitos competentes (65).

1.3.4. Promoción del crecimiento vegetal

La promoción del crecimiento vegetal (PCV) por los endófitos bacterianos se puede dar por diferentes mecanismos directos o indirectos. Los **mecanismos directos** incluyen el aumento de la capacidad de solubilizar nutrientes poco biodisponibles (P, K, Fe), de fijar biológicamente el nitrógeno atmosférico (FBN), así como de producir y/o modular fitohormonas (ej: auxinas, giberelinas y citoquininas). Por otro lado, los **mecanismos indirectos** incluyen el control biológico de fitopatógenos o la estimulación de la resistencia sistémica inducida en plantas (129; 36).

1.3.4.1. Producción de fitohormonas

Las fitohormonas regulan procesos básicos de la planta como la quiescencia y germinación de las semillas, la formación de raíces, floración, la ramificación, macollaje y la maduración de los frutos, e influyen en la resistencia de las plantas a factores ambientales (155). Las bacterias pueden producir algunas de las fitohormas estimulando así el crecimiento vegetal, esta característica es uno de los mecanismos que da mayor aporte a la capacidad PCV por bacterias endófitas (32).

Producción de auxinas. Se ha estimado que cerca del 80 % de las bacterias rizosféricas son capaces de sintetizar ácido indol acético (AIA) (118). Describiéndose 6 vías para la producción de AIA: la vía indol-3-piruvato (IPyA), indol-3- acetamida (IAM), triptamina (TAM), triptófano oxidasa de cadena lateral (TSO), indol-3-acetonitrilo (IAN), todas dependientes de triptófano y una vía independiente de este aminoácido (44).

Modulación de los niveles de etileno. El etileno es una hormona que es sintetizada en situaciones de estrés biótico, infección por patógenos, o condiciones de estrés abiótico como la sequía. La modulación de la concentración de etileno posee como consecuencia la elongación del sistema radicular y la disminución de la respuesta de defensa en algunas plantas. Esta modulación es realizada por la enzima 1-amino-ciclopropano-1-carboxilato (ACC) desaminasa,

la cual hidroliza el ACC que es el precursor inmediato del etileno, produciendo amonio y α-cetobutirato (52).

1.3.4.2. Fijación biológica de nitrógeno

El nitrógeno es un componente fundamental de una gran proporción de las moléculas biológicas. A pesar de ser uno de los elementos más comunes de la atmósfera, este se encuentra en estado gaseoso (N_2) y no es biodisponible para las plantas. La FBN es el proceso por el cual el N_2 es reducido a amonio (NH_4+) y es realizado por un grupo de procariotas denominados diazótrofos. La enzima encargada de llevar a cabo la FBN es la nitrogenasa. Esta enzima esta conformada por dos componentes, la dinitrogenasa (Fe proteína) y la dinitrogenasa reductasa (FeMo proteína), codificados por los genes *nifHDK* (110). Si bien todos los diazótrofos poseen la nitrogenasa clásica (FeMo-nitrogenasa), existen algunos microorganismos que poseen además una o dos nitrogenasas alternativas (VFe y FeFe-nitrogenasa) (110). En las enzimas alternativas, el Mo del sitio catalítico es sustituido por V o Fe. Con el objetivo de adaptar el proceso de FBN a las restricciones fisiológicas, las bacterias diazótrofas presentan varios mecanismos para censar múltiples señales ambientales como la disponibilidad de N fijado (principalmente amonio y glutamina) o de O_2. La regulación del proceso es principalmente a nivel transcripcional y en algunos casos, se ha reportado la regulación postraduccional de la nitrogenasa mediante una inactivación reversible (121).

1.3.4.3. Captación de hierro

El hierro participa en un gran número de reacciones metabólicas y es un componente esencial de distintas macromoléculas. Este elemento es abundante en el suelo pero se encuentra en forma de hidróxidos insolubles por lo cual es poco biodisponible. En condiciones limitantes de este metal, las plantas excretan quelantes y/o fitosideróforos que unen el Fe_{3+} y lo transportan a la superficie de la raíz donde es reducido y absorbido por la planta. Por su parte las bacterias, en las mismas condiciones, producen metabolitos de bajo peso molecular con alta afinidad por el Fe llamados sideróforos (143). El complejo Fe_{3+} sideróforo es reconocido específicamente por receptores de membrana externa e internalizado (13). Los sideróforos bacterianos tienen mayor afinidad por el Fe en comparación con los fitosideróforos de las plantas (88). Las

bacterias que presentan estos sistemas, afectan la comunidad rizosferica al ser buenos competidores por el Fe nutricional (88).

1.4. Genómica de bacterias promotoras del crecimiento vegetal

Las BPCV asociadas a cultivos de interés agronómico como el sorgo dulce (*Sorgum bicolor*) (102), el maíz (*Zea mays*) (135), el arroz (*Oryza sativa*) o la caña de azúcar (*Saccharum officiarum* L.) tienen el potencial biotecnológico de mejorar la sustentabilidad de la explotación de dichos cultivos (136; 167). En este sentido el análisis genómico de las BPCV ha aportado información sobre las características genéticas que dominan los mecanismos moleculares imperantes en la interacción (23; 82; 98; 60; 91; 135). La detección de las capacidades PCV en bacterias rizosféricas o endofiticas se estudian comúnmente mediante técnicas que miden la actividad fisiológica de una enzima o una vía metabólica. Sin embargo, esta aproximación muchas veces tiene la desventaja de que algunos genes no son expresados en las condiciones de laboratorio en las que se hace el ensayo. Mediante el conocimiento de la secuencia y el análisis del genoma de la bacteria en estudio, a priori, esta problemática podría solucionarse, al poder analizarse si el genoma codifica para los genes correspondientes. Del mismo modo, es posible estudiar aquellos genes que podrían estar involucrados en la colonización e interacción con la planta. Sin embargo, una vez identificados en el genoma, es necesario realizar estudios fisiológicos y moleculares que confirmen su expresión y su rol en la interacción o PCV. Mediante el análisis de la secuencia genómica se han caracterizado genomas de cepas PCV endófitas modelo como: *Gluconacetobacter diazothropicus* PA15 (23), *Azoarcus* sp. BH72 (82) y *Azospirillum brasilense* Az39 (135). A partir del análisis de la secuencia genómica de *Azoarcus* sp. BH72 por ejemplo, se reveló que esta cepa endófita-diazótrofa, no posee los sistemas de secreción de tipo III presentes en bacterias asociadas a plantas, así como los genes que codifican para enzimas degradadoras de la pared vegetal, ni genes del sistema de *quorum sensing* del tipo homocerin lactona. Por su parte, el análisis del genoma del endófito-diazótrofo de caña de azúcar *G. diazothropicus* PA15 arrojó información valiosa para comprender su historia evolutiva y ecológica, destacándose su gran plasticidad, ya que posee 20 islas genómicas adquiridas

por transferencia horizontal. En cuanto a la cepa *A. brasilense* Az39, aislada de plantas de maíz (*Zea maize*), el análisis de su genoma permitió realizar un estudio comparativo con respecto a otras cepas de la misma especie, definiéndose un conjunto de genes comunes a todos los genomas, lo que se conoce como el genoma núcleo de la especie (178). En su conjunto este tipo de estudios han sido cruciales para profundizar en el entendimiento de la fisiología, ecología y evolución de estas bacterias y su interacción con plantas.

1.5. Tecnologías empleadas para la secuenciación de ADN

La primera metodología de secuenciación de ADN empleada, la cual prevaleció por 30 años, se basaba en la utilización de nucléotidos marcados que terminaban la transcripción (141). Esta aproximación fue utilizada en el desarrollo del Proyecto Genoma Humano, el cual se extendió durante 15 años a pesar de ser la colaboración internacional de mayor envergadura para un proyecto biológico. Debido a esto surgió la necesidad de mejorar las plataformas de secuenciación por lo cual el *National Human Genome Research Institute* de los Estados Unidos lanzo un programa cuyo fin era reducir el costo de secuenciar un genoma humano a mil dolares americanos (145). Este programa tuvo como consecuencia un gran impulso del área, lo que resultó en el desarrollo de las tecnologías de secuenciación de ADN de segunda generación.

1.5.1. Plataformas de secuenciación de segunda generación

Los secuenciadores de segunda generación producen una gran cantidad de lecturas cortas a un costo mucho menor, que las lecturas obtenidas mediante la secuenciación de Sanger. La plataforma Roche 454 (103) fue la que dio inicio a esta generación al cambiar de paradigma introduciendo la capacidad de paralelizar en gran medida las reacciones de secuenciación, estrategia que ha sido adoptado por otras plataformas. A su vez, la preparación de las muestras de ADN para su secuenciación, es menos laboriosa al no ser necesario clonar las secuencias en vectores biológicos como plásmidos o BACs (*bacterial artifial chromosomes*). Si bien cada plataforma tiene sus características particulares,

todas comparten los principios básicos incluyendo: 1- el fragmentado al azar del ADN y la construcción de la librería; 2- la ligación de los adaptadores en los extremos de los fragmentos mediante la enzima ligasa; 3- la hibridación de los fragmentos y posterior amplificación mediante la reacción en cadena de la polimerasa (PCR, por sus siglas en ingles) con el fin de generar una población de ADN molde clonal a partir de una sola molécula; 4- la secuenciación de forma paralela y cíclica del ADN molde. El hecho de contar con muchas copias de cada fragmento es lo que proporciona una señal lo suficientemente fuerte como para ser detectada por sobre el ruido de base (169).

1.5.2. Ion Torrent

Ion Torrent, a diferencia de otras plataformas, utiliza circuitos integrados que detectan el cambio de pH que ocurre cuando se libera un protón H^+ durante la polimerización de una base (138). El método de detección de las bases repercute en el bajo costo de manufactura de los dispositivos respecto a otras tecnologías. A su vez la detección de los iones permite utilizar nucleótidos y una polimerasa de ADN sin modificar, disminuyendo el costo por base secuenciada. Para el empleo de esta tecnología, en una primera instancia las muestras de ADN se fragmentan, los adaptadores se añaden en los extremos de estos y luego los fragmentos son capturados en perlas de Sepharosa y finalmente amplificados utilizando PCR en emulsión (emPCR) (131). Posteriormente los fragmentos amplificados son colocados en microporos de forma que se espera una perla por microporo (Figura 1.2 paso 1). Cada ciclo de secuenciación está compuesto de tres pasos: 1- inyección de un único tipo de nucleótido (A, C, G o T), 2- medida de la señal eléctrica generada en los microporos que presentan polimerización, y 3- lavado de los nucleótidos excedentes que quedaron en el microporo que no hayan sido polimerizados (Figura 1.2). Si la base en el fragmento molde es complementaria a la base inyectada ocurre una incorporación, por lo que si el molde presenta un repetido, la señal obtenida es proporcional al largo del repetido (138).

El perfil de errores de esta plataforma se asemeja mucho a la plataforma Roche 454 ya que el principio básico es similar. Para Ion Torrent PGM se reportó una tasa de error de 1.0 % aproximadamente, siendo las inserciones más comunes que las deleciones (154). Otro error caracterizado en Ion Torrent son los indels con una alta frecuencia en algunos en regiones particulares

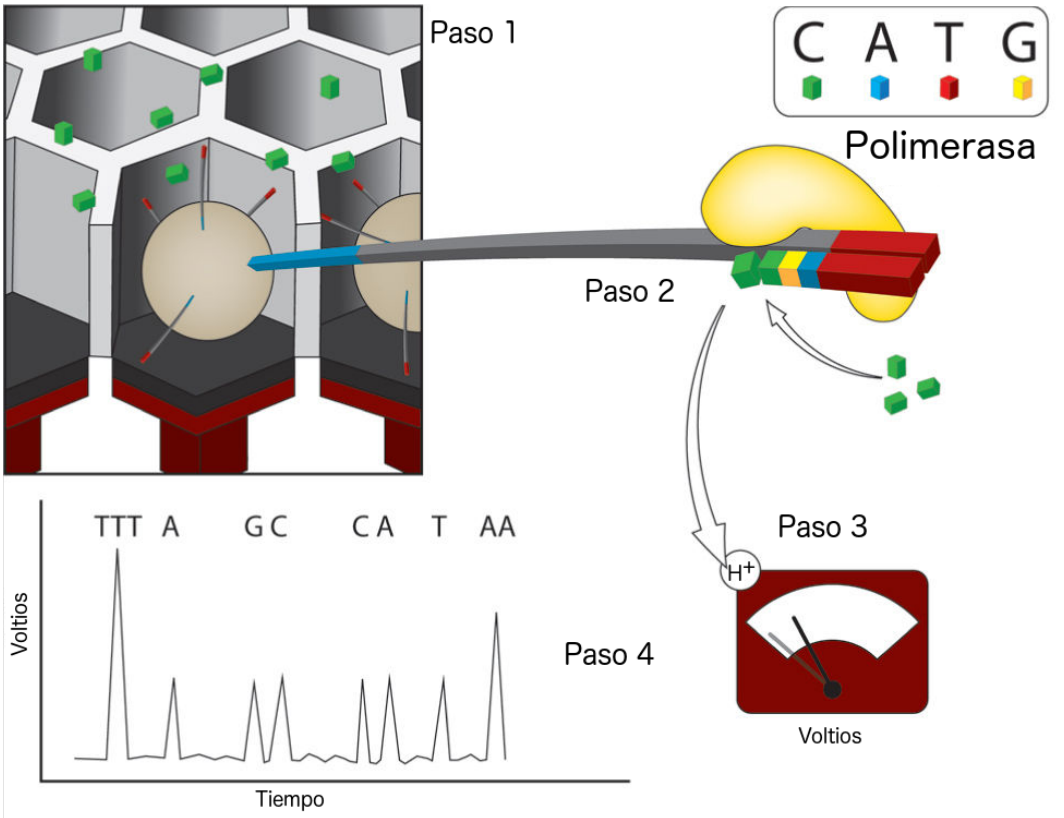

Figura 1.2: Esquema del proceso de secuenciación de ADN realizado por la plataforma Ion Torrent. El proceso esta divido en 4 etapas. Paso 1) luego de la amplificación del ADN molde se colocan las perlas con el ADN clonal en una placa de titulación donde cada pocillo tiene un sensor iónico. Paso 2) se suministran cantidades limitantes de un dNTP que son incorporadas al ADN molde de las perlas por la ADN polimerasa. Paso 3) con cada base incorporada se libera al medio un protón H^+. Paso 4) un detector capta el protón H^+ y el dispositivo traduce la señal en la base correspondiente. Tomado con modificaciones de Escalante y colaboradores (2014).

del genoma, lo que puede derivar en errores en la interpretación de los datos (28). Por otra parte, la precisión del Ion Torrent disminuye sobre el final de las lectura debido a la acumulación de desechos a lo largo de ciclado (28). Asimismo esta tecnología es menos precisa que la de Roche 454 ante la presencia de homopolímeros. Para homopolímeros de largo 2 el error más frecuente son las inserciones, mientras que para homopolímeros de largo mayor a 14 el dispositivo no es capaz de leerlo (94).

1.6. Corrección de errores de secuenciación del ADN

Diversos estudios han demostrado el impacto negativo de los errores de secuenciación en distintas aplicaciones (140; 154). Una aproximación sencilla para lograr mejorar la calidad de los datos proveniente de un dispositivo de secuenciación de ADN de alto rendimiento es el recortado o *trimming* de los datos en función de la calidad de las bases, lo que lleva a una perdida de información.

Uno de los principales problemas que presentan las plataformas de secuenciación de segunda generación es la alta tasa de error (sustituciones, inserciones y deleciones), entre 1 % y 1.8 % (3). La exactitud de los datos de secuenciación es de vital importancia ya que entre individuos de una misma especie puede haber muy pocos cambios por lo que los errores pueden ser confundidos con variaciones biológicas (3). A su vez, durante el ensamblado de los datos obtenidos, la corrección de errores es un paso importante para lograr un ensamblado de calidad. Se ha reportado que la corrección de errores en los datos de plataformas de secuenciación pueden mejorar algunas métricas del ensamblado de un genoma bacteriano (140). Esta práctica es importante, habiéndose demostrado que el mapeo de lecturas contra una referencia, mejora si los datos son previamente corregidos. Otro caso es el estudio de los polimorfismos de nucleótido simple (SNP, por sus siglas en inglés), el cual se beneficia de esta corrección ya que un error puede confundirse con un polimorfismo (154). Es por lo antes expuesto que se han desarrollado múltiples herramientas para la corrección de datos de secuenciación masiva de ADN (148; 4; 87; 3; 85).

1.6.1. Estrategias de corrección de errores

Los algoritmos para la corrección de errores de datos para las plataformas de secuenciación de ADN fueron categorizados de acuerdo a su estrategia algorítmica, describiéndose 4 categorías: 1- sub-secuencias de largo k llamadas kmeros, 2- el alineamiento múltiple de secuencias (AMS), 3 - trie, y 4- modelos probabilísticos. A pesar de las diferencias entre las categorías, la mayoría de las herramientas parten de los mismos 3 supuestos (85):

1. Los errores por posición son raros con respecto a las bases correctas, asumiendo una cobertura lo suficientemente alta.

2. La cobertura es uniforme a lo largo de la muestra.
3. Los errores son independientes de la posición en el genoma.

1.6.1.1. Correctores de errores basados en kmeros

Las plataformas de secuenciación de ADN de segunda generación se caracterizan por producir una gran cantidad de lecturas a partir de una muestra, lo que redunda en una alta cobertura (94). Dicha característica puede ser utilizada para la corrección de los datos si las bases de las lecturas son alineadas con respecto a la secuencia de ADN de la que se originan. Luego, para cada uno de los nucleótidos es posible identificar las bases que divergen y evaluar si corresponde que sean corregidas. De lo contrario, si no se cuenta con una secuencia de referencia, es posible fraccionar las lecturas en subcadenas de largo k (kmeros) y luego obtener las frecuencias de los distintos kmeros. A partir de la población de kmeros para un determinado conjunto de datos se pueden clasificar los kmeros poco frecuentes como erróneos y corregirlos. Se ha descripto un algoritmo que utiliza la estrategia de kmeros mediante su clasificación en dos categorías: sólidos y débiles. Aquellos que aparecen al menos un número N de veces son considerados sólidos y los que no son considerados débiles (120). Posteriormente este algoritmo corrige los kmeros débiles mediante un numero mínimo de ediciones, hasta que la lectura quede compuesta únicamente por kmeros sólidos. Este algoritmo parte del supuesto de que los errores son al azar y poco frecuentes.

Diversos correctores han sido desarrollados basados en kmeros incluyendo: 1 - **EULER**: uno de los primeros programas desarrollado para ensamblar lecturas que incorpora la capacidad de corregir datos provenientes de plataformas de secuenciación masiva de ADN (120). Para la realización de la corrección, este programa guarda el total de ocurrencias de los kmeros para cada lectura, generando una distribución de frecuencias de kmeros por lectura. El programa clasifica los kmeros como erróneos si no pertenecen a M lecturas, donde M es un parámetro. Luego clasifica cada lectura como errónea si posee al menos un kmero que cae dentro de la categoría erróneo. Para la corrección se realizan cambios hasta que todos los kmeros erróneos son eliminados de cada lectura. Los cambios se introducen hasta lograr corregir la lectura o hasta que la distancia de edición supere un valor máximo n (parámetro).

2 - **BayesHammer**: es un corrector que utiliza un grafo donde los vértices

están representados por cada uno de los kmeros presentes en las lecturas y las aristas unen los vértices que tienen a lo sumo N diferencias, donde N es parámetro (111). Los kmeros definidos como centrales (con gran representación en el genoma), son considerados como correctos y los nodos que están conectados a estos son considerados erróneos. Una vez que el programa obtiene todos los kmeros centrales, recorre las lecturas con kmeros erróneos y utilizando los kmeros centrales va cambiando por consenso las bases erróneas hasta lograr que las lecturas queden libres de kmeros erróneos. Esta aproximación tiene algunos problemas, ya que puede clasificar como erróneos algunos kmeros que estén conectados a un nodo central por azar, así como regiones repetidas que tengan muy pocas variaciones entre ellas. BayesHammer construye el mismo grafo que Hammer pero luego lo refina construyendo subclusters. BayerHammer fue integrado como un componente del ensamblador Spades (15).

3 - **Karect**: Este programa considera cada lectura como una referencia, efectúa un alineamiento múltiple con las lecturas que presenten mayor similitud entre ellas y las guarda en un grafo parcialmente ordenado (POG por sus siglas en inglés, *partial order graph*) (4). Para generar el alineamiento múltiple de cada lectura el programa Karect considera como similares a las lecturas cuyos kmeros presenten a lo sumo d (parámetro) diferencias o que tengan sub-kmeros compartidos. Los alineamientos son insertados en un grafo POG y posteriormente el corrector extrae la versión corregida de las lecturas del mismo.

1.6.1.2. Correctores de errores basados en tries

Algunos correctores utilizan estructuras de datos tipo trie y derivados. Un trie es una estructura de datos tipo árbol. Su nombre proviene del inglés *re**tri**eve*, que significa recuperar. En un trie cada arista que sale de un nodo del árbol esta etiquetada con un símbolo distinto del alfabeto. A partir de un nodo raíz es posible navegar el trie en busca de palabras o de prefijos. Las búsquedas se realizan desde la raíz descendiendo por el camino formado por las aristas cuyo símbolo coincida con las letras de la búsqueda, en orden sucesivo de la palabra que se está buscando. Un árbol de sufijos de una cadena es un caso particular de un trie donde todos los sufijos de dicha cadena son utilizados como las ramas. Es posible generalizar un árbol de sufijos para un conjunto de cadenas, de forma que el conjunto de cadenas representan las lecturas obteni-

das de un genoma y cada arista esta etiquetada con un símbolo y la cantidad total de símbolos que la soportan (Figura 1.3). Las lecturas en un árbol de sufijos siguen un camino hasta que algún momento deban bifurcar el camino, si la bifurcación se debe a un error en la lectura las aristas del nuevo camino tendrán un soporte muy bajo. Es en las bifurcaciones que los programas de corrección deben decidir si corregir una base. En el ejemplo de la figura 1.3 se puede observar que la lectura $TA\mathbf{A}A\$$ tiene un error en la tercera base, esto ocasiona que la rama $TA\mathbf{A}A\$$ tenga una sub-rama con poco sustento ($AA\$$). En este ejemplo, corrigiendo $TA\mathbf{A}A\$$ por $TA\mathbf{G}A\$$ en la lectura $TA\mathbf{A}A\$$ se elimina la sub-rama con poco sustento. Es importante recordar que parte del supuesto de que los errores son poco frecuentes, por lo que la rama errónea en el árbol tendría pocos representantes.

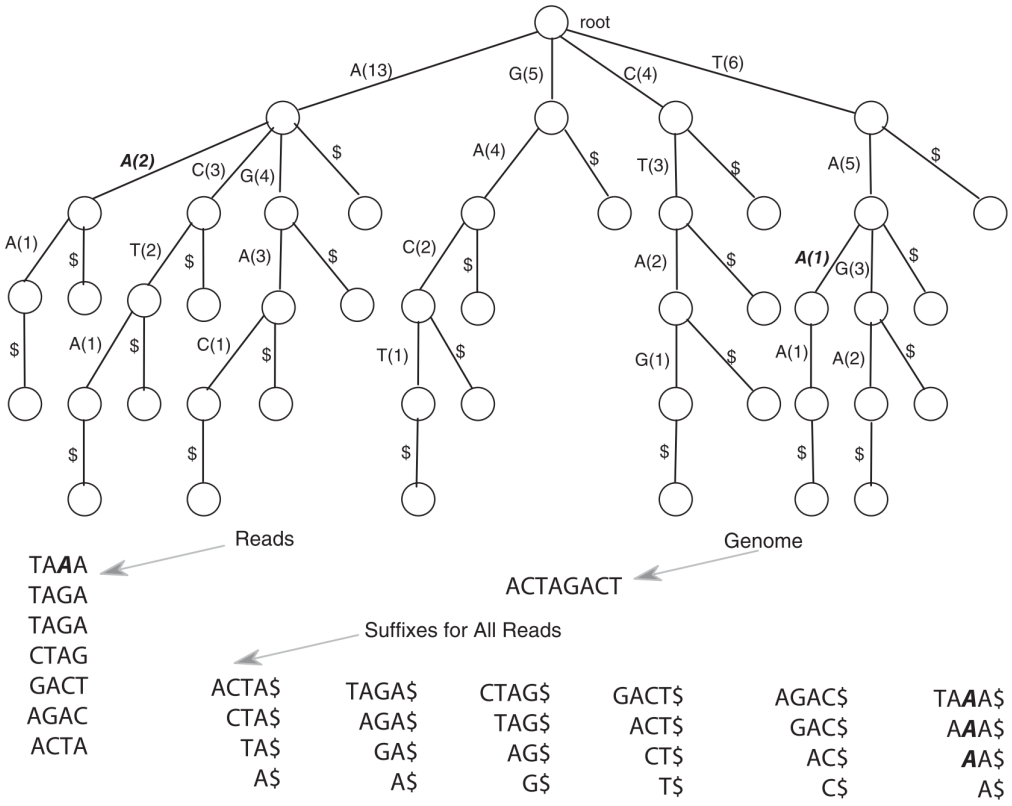

Figura 1.3: Ejemplo de un árbol de sufijos para una secuencia ejemplo donde la lectura $TAAA$ tiene un error en la tercera posición. Tomado de Alic y colaboradores (2016).

Como ejemplos de este grupo de correctores podemos mencionar:

1 - **SHREC**: este programa construye primero un trie a partir de las lecturas donde cada nodo es un kmero y el peso de cada nodo es la cantidad de

lecturas que presentan dicho kmero. Posteriormente intenta corregir los nodos cuyo peso es menor que un valor predefinido. La corrección se basa en la conversión del nodo erróneo en un nodo hermano de forma que el árbol por debajo del nodo corregido se ajuste exactamente al sub-árbol que existe por debajo del nodo hermano (147). Por otro lado, el programa **Hybrid-SHREC** (139), un programa derivado de SHREC, plantea la detecciióon de nodos erróneos de forma idéntica pero en lugar de convertir el nodo erróneo, este programa inserta o remueve bases de forma que los árboles debajo de los nodos se ajusten.

2 - **Fiona**: este programa utiliza un árbol de sufijos como SHREC y Hybrid-SHREC pero optimizado para la detección de superposiciones entre lecturas y capaz de corregir errores de indels. Este programa construye árboles de sufijos parciales para las lecturas y las lecturas invertidas, las cuales luego recorre en busca de errores. Los errores se detectan calculando para cada kmero el logaritmo de la razón de la probabilidad condicional de que tenga errores dada la cobertura del kmero, dividida por la probabilidad de que cualquier kmero tenga errores. Si este coeficiente es positivo, entonces el kmero se marca como erróneo. Luego de identificado el error, el programa intenta correcciones en el kmero considerando los subárboles correctos (148).

1.6.1.3. Correctores de errores basados en modelos probabilísticos

La mayoría de los programas para la corrección de errores carecen de un modelado estadístico de los datos. Asimismo, muchos programas no consideran los errores específicos o características particulares de la plataforma, como la caída de la calidad de las lecturas sobre el final. Sin embargo, algunos programas utilizan un modelo para describir las sustituciones y/o indels que ocurren durante la secuenciación del ADN. Entre este tipo de programas se encuentran:

1 - El programa **RECOUNT** se basa en el algoritmo de EM para determinar la base correcta en cada posición mediante el calculo de la verosimilitud de las variantes existentes para una posición dada (176).

2 - **Fiona** emplea un modelo estadístico para la detección de errores (ver sección 1.6.1.2).

3 - **DUDE-Seq**: este programa utiliza el algoritmo ***D****iscrete* ***U****niversal* ***DE****noiser* (DUDE) (174) para la corrección de datos de NGS orientados a la secuenciación de amplicones (87). En este trabajo se discutirá la implementación del algoritmo DUDE para la corrección de datos de secuenciación de ADN

generados por la plataforma Ion Torrent. Por este motivo que a continuación se describirá en mayor profundidad las características de este algoritmo.

1.6.2. Discrete Universal DEnoiser

En este trabajo exploramos la aplicación del algoritmo ***D**iscrete **U**niversal **DE**noiser* (DUDE) (174) para la corrección de datos de secuenciación de ADN provenientes de la plataforma Ion Torrent PGM. Este algoritmo fue desarrollado para la reconstrucción de secuencias con un alfabeto finito que fueron corrompidas por un canal que procesa cada símbolo de forma independiente y estadísticamente idéntica. Bajo estas hipótesis DUDE ofrece garantías teóricas de rendimiento en la capacidad de corregir, incluso cuando no se hacen supuestos sobre el modelo estocástico para los datos limpios subyacentes, utilizando únicamente el modelo con el que introduce errores el canal. Esta formulación semi-estocástica se ajusta al problema de la corrección de datos provenientes de la secuenciación del ADN. En la secuenciación del ADN es difícil establecer modelos estocásticos precisos para secuencias de ADN limpias, pero es simple y bastante realista suponer modelos de ruido (es decir, matrices de confusión) para dispositivos de secuenciación específicos. Adaptar el DUDE al problema de la corrección de datos de secuenciación de ADN requiere algunos ajustes, es por eso que se describe más adelante en este trabajo.

1.6.2.1. Definición formal de DUDE

En esta sección se presenta formalmente el algoritmo DUDE. En la figura 1.4 se esquematiza el problema de corrección de errores para secuencias discretas.

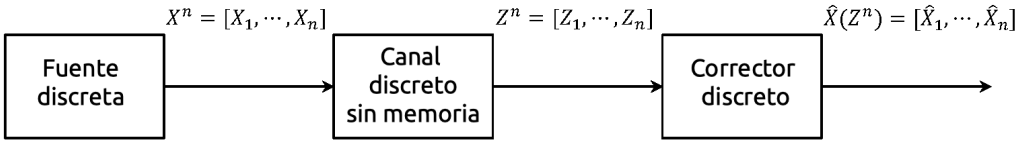

Figura 1.4: La configuración general para la corrección de errores para secuencias discretas. Tomado con modificaciones de Lee y colaboradores (2017).

Para la formalización del DUDE se utiliza la siguiente notación. La secuencia original (no corrupta) se denota como X^n, donde n es el largo de la secuencia y el símbolo en la posición i, $1 \leqslant i \leqslant n$, se denota como x_i. Por otro

lado, la secuencia corrupta por el canal se denota como Z^n y cada símbolo como z_i. Finalmente, la secuencia reconstruida por el DUDE se denota como $\hat{X}^n$, donde cada elemento de esta secuencia se denota $\hat{x}_i$. Los símbolos x_i pertenecen a un alfabeto finito $\mathcal{X}$ mientras que los símbolos z_i pertenecen a un alfabeto finito $\mathcal{Z}$.

El tipo de canal considerado en este modelo introduce errores en la secuencia de forma independiente y estadísticamente idéntica para cada símbolo. Esta clase de canales son denominados como "canales discretos sin memoria" o DMC por sus siglas en ingles ("Discrete Memoryless Channel"). Un DMC esta completamente caracterizado por la matriz de transición de estados del canal, $\Pi \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{Z}|}$, donde $\Pi(x, z)$ representa la probabilidad condicional de que el símbolo ruidoso sea z dado que el original es x. Un corrector de errores discreto opera sobre la secuencia corrupta Z^n e intenta reconstruir la cadena original a través de una estimación $\hat{X}^n$. Es por esto que para cada símbolo de la cadena ruidosa, z_i, el corrector intenta determinar una estimación del símbolo correcto, $\hat{x}_i = \hat{x}_i(Z^n)$, de forma que $\hat{x}^n = (\hat{x}_1(Z^n), \ldots, \hat{X}_n(Z^n))$. El modelo asume una función de pérdida $\Lambda : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ que está representada por la matriz $\Lambda = \{\Lambda(i, j)\}_{i,j} \in \mathcal{X}$, de forma que la calidad de la corrección de errores de $\hat{X}^n$ se mide con la pérdida promedio

$$L_{\hat{X}^n}(X^n, Z^n) = \frac{1}{n}\Sigma_{i=1}^{n}\Lambda(x_i, \hat{X}_i(Z^n)), \tag{1.1}$$

donde $\Lambda(x, x_i)$ es el costo de estimar x con x_i. Usualmente se cumple $\Lambda(x, x_i) = 0$ si $x_i = x$ y $\Lambda(x, x_i) > 0$ en otro caso.

Para corregir, DUDE hace dos recorridas por los datos. En la primera pasada colecta las estadísticas de la secuencia ruidosa Z^n. Específicamente, para cada $a \in \mathcal{Z}$, para cada contexto $l^k \in \mathcal{Z}^k, r^k \in \mathcal{Z}^k$ y $z_i^j = z_i, z_{j+i}, ..., z_j$ calcula

$$\mathbf{m}(z^n, l^k, r^k)[a] = |\{i : k + 1 \leq i \leq n - k, z_{i-k}^{i+k} = l^k a r^k\}|, \tag{1.2}$$

de forma que $m(z^n, l^k, r^k)$ es un vector columna de tamaño $|\mathcal{Z}|$ que cuenta las ocurrencias de cada símbolo a lo largo de la secuencia ruidosa en el contexto izquierdo l^k y derecho r^k. Cabe destacar que k es el único parámetro del DUDE y determina el largo de contexto considerado. Una vez que el vector m fue completado, DUDE recorre la secuencia ruidosa una segunda vez y para cada símbolo aplica la siguiente regla para cada $i, k + 1 \leq i \leq n - k$,

$$\hat{X}_i(Z^n) = \underset{\hat{x}\in\mathcal{X}}{\text{argmín}}\, m^T(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k})\Pi^{-1}[\lambda_{\hat{x}} \bigodot \pi_{z_i}], \tag{1.3}$$

donde π_{z_i} es la columna z_i-ésima de la matriz del canal Π, $\lambda_{\hat{x}}$ es la columna $\hat{x}$-ésima de la matriz de pérdida Λ, $\bigodot$ denota la multiplicación punto a punto entre vectores y m^T denota la transposición de m. El algoritmo DUDE no aplica un regla de decisión por mayoría simple basada en $\mathbf{m}$, sino que incorpora el conocimiento del canal a través de la matriz Π y la función de perdida Λ para la estimación precisa del símbolo correcto.

DUDE garantiza un rendimiento asintóticamente óptimo para correctores de ventana-deslizante que utilicen el mismo tamaño de ventana (174). Esto significa que se puede demostrar que el DUDE logra, de forma asintótica y sin acceso a ninguna información sobre las estadísticas de la señal limpia, el mismo rendimiento que el mejor corrector que tiene acceso a esta información.

1.7. Aportes de este trabajo

Estudios recientes han mostrado que la corrección de errores de secuenciación de ADN permite lograr secuencias genómicas de mayor calidad sin aumentar el costo económico. En este sentido, el empleo de estas herramientas mejora el ensamblado de genomas bacterianos y los resultados de los análisis posteriores.

En este trabajo se presentan los resultados del desarrollo de una aplicación de corrección de errores de secuenciación de ADN basada en el algoritmo DUDE. El desarrollo y la evaluación experimental de esta herramienta se presentan en el capítulo 2, donde se observa que mejora algunas de las métricas de calidad evaluadas para datos provenientes del secuenciador de ADN Ion Torrent PGM. Sin embargo, otras herramientas disponibles en la literatura superan a la herramienta desarrollada en este trabajo. En particular, el programa Fiona se destaca en la capacidad de corrección por sobre las otras herramientas evaluadas. Utilizando el programa Fiona es posible obtener ensamblados de mayor calidad, maximizando la información disponible para el estudio de la biología de los organismos. En función de estos resultados se empleó dicho programa para la corrección de los datos de secuenciación de las cepas endófitas de caña de azúcar *Kosakonia* sp. UYSO10 y *Rhizobium* sp. UYSO24.

Por otro lado, posteriormente se realizó el estudio genómico de las cepas UYSO10 y UYSO24, endófitas y promotoras del crecimiento vegetal de plantas de caña de azúcar, el cual se presenta en el capítulo 3. Los datos de secuenciación de estas cepas fueron corregidos con el programa Fiona y luego ensamblados y anotados. A partir de las anotaciones de estos, se caracterizó y se realizó una evaluación de la presencia de genes relacionados con la interacción planta-bacteria, así como con la PCV. En primer lugar, se determinó la afiliación filogenética de las cepas en estudio aprovechando los datos provenientes de las secuencias genómicas, resultando que la cepa UYSO10 pertenece al género *Kosakonia* y a la especie *radicincitans*; mientras que la cepa UYSO24 pertenece al género *Rhizobium*. Por otro lado, se determinó que ambas cepas poseen diversos genes relacionados a la interacción planta bacteria incluyendo sensores, reguladores, estructuras flagelares y pili entre otros; así como genes relacionados a la PCV tales como la nitrogenasa, así como la producción y regulación de fitohormonas. Estos resultados se correlacionan con resultados de las caracterizaciones *in vitro* previamente obtenidos.

En el capítulo final de esta tesis (capítulo 4), se presentan conclusiones y perspectivas a futuro de este trabajo.

1.8. Objetivos

1.8.1. Objetivo general

Desarrollar una herramienta informática para la corrección de errores de secuenciación, capaz de ser aplicada a un conjunto de datos genómicos de dos cepas bacterianas modelos, proveniente de la plataforma Ion Torrent PGM.

1.8.2. Objetivos particulares

Como objetivos particulares de la tesis se plantean:

1. Desarrollar una herramienta que permita estimar los errores que comete el dispositivo de secuenciación de ADN Ion Torrent PGM.
2. Desarrollar e implementar una herramienta de reducción de errores en secuencias de ADN basada en el algoritmo DUDE.
3. Corregir, ensamblar y anotar un conjunto de datos provenientes de la secuenciación de dos genomas bacterianos (cepas UYSO10 y UYSO24), secuenciados con la plataforma Ion Torrent PGM.
4. Realizar la caracterización genómica de las cepas en estudio poniendo énfasis en las características relacionadas a la interacción planta-endófito y la promoción del crecimiento vegetal.

Capítulo 2

Corrección de errores de secuenciación en la plataforma Ion Torrent

En este capítulo exploramos la aplicación una herramienta basada en el algoritmo DUDE para la reducción de errores de secuenciación del ADN en la plataforma Ion Torrent PGM. En primer lugar, se desarrolló una herramienta que permite estimar los errores que comete el dispositivo de secuenciación de ADN Ion Torrent PGM. En segundo lugar, se implementó una herramienta de reducción de errores en secuencias de ADN basada en el algoritmo DUDE.

2.1. Materiales y métodos

La formulación original del algoritmo DUDE se ajusta muy bien a la introducción de sustituciones durante la secuenciación de ADN. En este contexto definimos $X = Z = \{A, C, G, T\}$. Como función de pérdida utilizamos la distancia Hamming, $\Lambda(x, \hat{x}) = 0$ si $x = \hat{x}$ y $\Lambda(x, \hat{x}) = 1$ de lo contrario. Para los errores de indels el alfabeto X es definido a partir de la codificación de las secuencias de ADN en base al largo del homopolímero en cada posición. Específicamente, tomamos $X = \{A_1, C_1, G_1, T_1, A_2, C_2, G_2, T_2, ..., A_{10}, C_{10}, G_{10}, T_{10}\}$ siendo A_1 la base A repetida 1 vez, A_2 es la base A repetida 2 veces, y así sucesivamente hasta 10. A esta codificación la denominaremos "codificación por largo de corrida"(CLC) en el contexto de este trabajo y será representada de la forma $[A_1...T_{10}]$.

El algoritmo DUDE asume en primer lugar que la matriz del canal es conocida, pero en la práctica esta información no está disponible por lo que es necesario estimarla. Dicha estimación se realizó de forma empírica alineando las lecturas obtenidas de un secuenciador contra una referencia libre de errores. A partir de dicho alineamiento se calculó la relación de bases sustituidas e indels. En segundo lugar, DUDE asume que el canal no tiene memoria y por lo tanto la tasa de error es independiente de la posición de la base dentro de la secuencia. Esta hipótesis no se cumple para el dispositivo Ion Torrent PGM donde la tasa de error tiende a aumentar sobre el final de la lectura (28). Atendiendo a esta situación utilizamos tres aproximaciones:

1. empleo de una matriz única para el canal formada por el promedio de la tasa de error (IonDUDE),
2. empleo de matrices de canal dependientes de la posición de la base dentro de la lectura (IonDUDE M),
3. empleo de matrices de canal dependientes de la calidad q de cada base leída (IonDUDE Q).

A continuación se detalla la implementación del programa para la estimación de errores, así como las tres aproximaciones para la implementación del algoritmo DUDE para la corrección de errores de secuenciación.

2.1.1. Estimación del error del canal

Se desarrolló un programa para la estimación de la tasa de errores de secuenciación de ADN. La estimación se realiza mediante la comparación de las lecturas mapeadas contra la secuencia de referencia. Se asume que las sustituciones o indels encontrados en las lecturas mapeadas son errores introducidos durante el proceso de secuenciación y no son producto de variaciones biológicas. Esta estrategia ha sido empleada antes para evaluar la tasa de error de las plataformas Illumina y 454 (94). Este programa toma como entrada un archivo con el mapeo de las lecturas contra una referencia en formato .sam o .bam (90). Para la lectura de los archivos .sam y .bam se emplea la librería PySAM (https://github.com/pysam-developers/pysam). El resultado del programa es un conjunto de matrices de canal, que pueden ser utilizadas por IonDUDE para la corrección de datos.

2.1.2. IonDUDE

IonDUDE es la implementación del algoritmo DUDE para la corrección de errores de secuenciación del dispositivo IonTorrent PGM. IonDUDE utiliza una matriz única para el canal. Los errores de sustitución son corregidos por un módulo, mientras que los errores de indels son corregidos por otro módulo. El primer tipo de errores son corregidos según el algoritmo 1, el cual recibe como entradas el conjunto de lecturas a corregir, el largo de contexto que debe emplearse y la matriz del canal.

Entrada: Lecturas D, Matriz del canal Π, Largo del contexto k
Resultado: Lecturas corregidas $\hat{D}$
Datos: $m(D, l^k, r^k) \in \mathbb{N}^4$ para todo $(l^k, r^k) \in \{A, C, G, T\}^{2k}$

$m(D, l^k, r^k) \leftarrow [0, 0, 0, 0]^{2k}$
para cada lectura d en D **hacer**
  $n \leftarrow LargoSecuencia(d)$
  para $i \leftarrow k+1, \ldots, n-k$ **hacer**
    Incrementar $m(D, d_{i-k}^{i-i}, d_{i+1}^{i+k})[d_i]$
  fin
fin
para cada lectura d en D **hacer**
  $\hat{d} \leftarrow Copiar(d)$
  $n \leftarrow LargoSecuencia(d)$
  para $i \leftarrow k, \ldots, n-k+1$ **hacer**
    si $q_i \geqslant 20$ **entonces** `/* ` q_i `es la calidad de` d_i `*/`
      $\hat{d}_i \leftarrow d_i$
    sinó
      `/* ` $\lambda_{\hat{X}}$ `corresponde a la distancia de Hamming */`
      $\hat{d}_i \leftarrow \underset{\hat{X} \in \{A,C,G,T\}}{\text{argmín}} \; m(D, d_{i-k}^{i-1}, d_{i+1}^{i+k}) \Pi^{-1} [\lambda_{\hat{X}} \bigodot \pi_{d_i}]$
    fin
  fin
  Agregar $\hat{d}$ a $\hat{D}$
fin

Algoritmo 1: IonDUDE para sustituciones

El algoritmo 1 no modifica las bases que cumple alguna de las siguientes condiciones:

1. Su valor de calidad supera el umbral predefinido de $q \geqslant 20$ (línea 12). El valor de calidad $q = 20$ fue tomado de las prácticas comunes de control de calidad de las secuencias de ADN (103).
2. Su contexto derecho o izquierdo es de largo menor al parámetro k (línea 11). Dichos largos de contexto afectan los extremos de las lecturas.

Por otro lado, para la corrección de errores de indels, el primer paso es codificar la secuencia tomando en cuenta el largo de repetición de una misma base de forma consecutiva. Esta transformación permite utilizar el algoritmo DUDE en su configuración original para este tipo de aplicaciones. Los homopolímeros cuyo largo es mayor a 10 bases no son tenidos en cuenta al momento de la corrección de datos. Esta decisión se tomó considerando que la frecuencia de homopolímeros de mayor longitud es baja (28). Los errores de indels son corregidos de acuerdo al algoritmo 2.

Entrada: Lecturas D, Matriz del canal Π, Largo del contexto k
Resultado: Lecturas corregidas $\hat{D}$
Datos: Secuencia codificada en CLC D_c,
$m(D_c, l^k, r^k) \in \mathbb{N}^4$ para todo $(l^k, r^k) \in \{\mathrm{A}_1 ... T_{10}\}^{2k}$

$m(D_c, l^k, r^k) \leftarrow [0, 0, 0, 0, \ldots, 0, 0, 0, 0]^{2k}$
para cada lectura d en D **hacer**
  Agregar $CodificarSecuenciaCLC(d)$ a Dc
fin
para cada lectura d en D_c **hacer**
  $n \leftarrow LargoSecuencia(d)$
  para $i \leftarrow k+1, \ldots, n-k$ **hacer**
    Incrementar $m(D_c, d_{i-k}^{i-i}, d_{i+1}^{i+k})[d_i]$
  fin
fin
para cada lectura d en D_c **hacer**
  $n \leftarrow LargoSecuencia(d)$
  $\hat{d} \leftarrow Copiar(d)$
  para $i \leftarrow k, \ldots, n-k+1$ **hacer**
    $\hat{d}_i \leftarrow \underset{\hat{X} \in \{A_1 ... T_{10}\}}{\text{argmín}} \; m(Dc, d_{i-k}^{i-1}, d_{i+1}^{i+k}) \Pi^{-1} [\lambda_{\hat{X}} \bigodot \pi_{d_i}]$
  fin
  Agregar $\hat{d}$ a $\hat{D}c$
fin
para cada lectura $\hat{d}$ en $\hat{D}c$ **hacer**
  Agregar $DecodificarSecuenciaCLC(\hat{d})$ a $\hat{D}$
fin

Algoritmo 2: IonDUDE para indels

2.1.2.1. IonDUDE M

Como se ha expuesto anteriormente, en el caso del secuenciador Ion Torrent PGM la tasa de error tiende a aumentar sobre el final de las lecturas (Figura 2.1). Para ajustar el DUDE a este comportamiento se modificó el estimador de errores y el IonDUDE de forma tal de incorporar dicha información al modelo. La versión modificada se denomina IonDUDE M. En esta versión el modelo del canal fue segmentado en intervalos discretos de forma de capturar el descenso de calidad observada. Se construyen matrices de canal diferentes para los intervalos generados a partir de la segmentación de las lecturas, es decir que si se consideran lecturas con un largo de 100 pares de bases e intervalos de 10 pares de bases, se construirían 10 matrices.

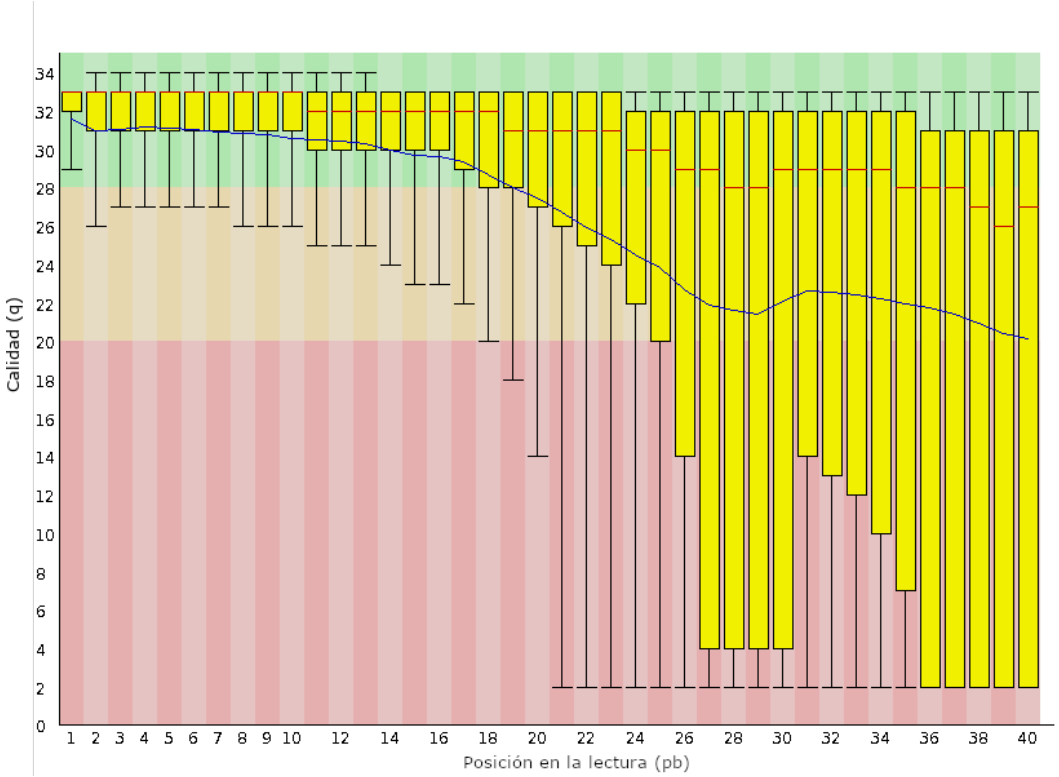

Figura 2.1: Ejemplo de calidad promedio de una base según la posición que ocupa en una lectura. Es notorio el decaimiento de la calidad de las bases sobre el final de la lectura. Las cajas amarillas representan los intercuartiles (25-75 %) y la línea central de las mismas representa la mediana, los bigotes superiores e inferiores representan los puntos 10 % y 90 % mientras que la línea azul representa el promedio del valor de calidad. Figura confeccionada con el programa FastQC (6).

2.1.2.2. IonDUDE Q

El formato FASTQ es utilizado para almacenar en un archivo de texto planto el resultado de una secuenciación de ADN y los valores de calidad. A pesar de ser el estándar *de facto*, no es un formato que haya sido sometido a un proceso de estandarización (35). Dicho formato surgió como sucesor natural del formato FASTA, ya que permite guardar los valores numéricos de calidad de cada base en un mismo archivo. Los valores de calidad son asignados durante el proceso de secuenciación por el programa de asignación de bases PHRED (47), este programa asigna los valores de calidad en términos de la probabilidad estimada p de que una base sea incorrecta:

$$q = -10 \cdot \log_{10}(p). \tag{2.1}$$

Esta versión del IonDUDE incorpora la calidad de las bases en el proceso de corrección de errores. Para hacerlo se modificaron las matrices del canal y el IonDUDE M. La probabilidad p de que una determinada base sea incorrecta en una lectura se codifica en el valor q (47). Utilizando el valor de calidad q es posible obtener p a partir de (2.1) como

$$p = 10^{-q/10}. \tag{2.2}$$

A su vez la probabilidad de error p se pondera por el tipo de error ya que las indels representan el 80 % de los errores que comete el equipo. Para esto se construyeron una serie de matrices de canal para sustituciones e indels donde los valores de la misma fueron calculados utilizando la fórmula (2.2).

2.1.3. Evaluación de IonDUDE

La evaluación de los resultados de la corrección de errores de secuenciación por las distintas versiones de IonDUDE está compuesta por dos elementos:

1. Mapeo de las lecturas contra una referencia y comparación entre lecturas corregidas y no corregidas.
2. Ensamblado *de novo* de las lectura y evaluación de parámetros del ensamblado.

Para llevar a cabo la evaluación se tomaron datos de genomas de especies muy bien caracterizadas (Tablas 2.1 y 2.2) y que han sido utilizados en

otros estudios de corrección de errores. En este sentido: 1- los datos de *Escherichia coli* O104:H4 fueron utilizados en un estudio de evaluación de errores de las plataformas de secuenciación de segunda generación (94), 2- los datos de *Escherichia coli* DH10B, *Staphylococcus aureus* LGA251, *Bordetella pertussis* 18323 y *Plasmodium falciparum* 3D7 fueron utilizados para evaluar la capacidad de corrección del programa Fiona (148).

Tabla 2.1: Descripción de los datos experimentales utilizados para la evaluación de la corrección de errores.

Organismo	N^o de acceso	LPL*	Total de lecturas	Cobertura
Escherichia coli O104:H4	SRR254209	178	977.971	32X
Escherichia coli DH10B	ERR039477	92	390.976	8X
Staphylococcus aureus LGA251	ERR236069	227	1.338.465	109X
Bordetella pertussis 18323	ERR161541	142	2.464.690	85X
Plasmodium falciparum 3D7	ERR161543	177	1.959.564	13X

*LPL: largo promedio de las lecturas.

Tabla 2.2: Identificación y fuente de los datos utilizados como referencia.

Organismo	N^o de acceso	TG*	%GC
Escherichia coli O104:H4	GCA_000299455.1	5.43	50.7
Escherichia coli DH10B	GCA_000019425.1	4.68	50.8
Staphylococcus aureus LGA251	GCA_000237265.1	2.73	32.9
Bordetella pertussis 18323	GCA_000306945.1	4.04	67.7
Plasmodium falciparum 3D7	GCA_000002765.1	23.2	20.0

*TG: tamaño del genoma en Mb.

Asimismo, los resultados de IonDUDE, IonDUDE M y IonDUDE Q fueron comparados con otros programas de corrección de errores de secuenciación de ADN. Las herramientas seleccionadas para la comparación fueron: Fiona 0.2.9 (148), Pollux 1.0 (104), Karect 1.0 (4) y IonHammer (15). Todas las herramientas utilizadas en la comparación fueron desarrolladas recientemente y son empleadas para la corrección de datos de secuenciación en el ensamblado de genomas.

2.1.3.1. Mapeo de las lecturas

La primera evaluación se realizó utilizando el programa para alinear lecturas *Burrows-Wheeler Aligner* (BWA) (89) y el programa de control de calidad de alineamientos Qualimap 2 (113). Para esto, en primera instancia se alinearon las lecturas contra los genomas de referencia utilizando el programa BWA MEM versión 0.7.12-r1039 empleando los parámetros por defecto. Posteriormente, el archivo de mapeo fue procesado con el módulo "BAM QC" del programa Qualimap 2. El análisis de "BAM QC" utiliza el tag *Concise Idiosyncratic Gapped Alignment Report* (CIGAR) de las entradas del archivo SAM (90) para obtener el número de sustituciones e indels en una secuencia mapeada con respecto a una referencia. A partir del análisis "BAM QC" se obtuvieron las siguientes métricas:

1. LM (Lecturas mapeadas): total de lecturas que fueron mapeadas exitosamente contra la referencia.
2. TE (Tasa de error): es la relación entre el total de bases incorrectas y el total de bases mapeadas.
3. LMI (Lecturas mapeadas con inserciones): total de lecturas mapeadas exitosamente contra la referencia que tienen al menos una base insertada con respecto a la referencia.
4. LMD (Lecturas mapeadas con deleciones): total de lecturas mapeadas exitosamente contra la referencia que tienen al menos una base borrada con respecto a la referencia.
5. Indels HP (Indels en homopolímeros): total de indels en regiones homopoliméricas.

2.1.3.2. Métricas del ensamblado de los genomas

En segundo lugar se evaluó el efecto de la corrección de errores sobre el ensamblado *de novo*. Los datos fueron ensamblados con el programa Spades 3.11.1 (15) empleando los parámetros por defecto, salvo que el programa se ejecutó en modo ensamblador únicamente (–only-assembler), utilizando como entrada datos provenientes de Ion Torrent (–iontorrent) y con los largos de kmeros igual a $k \in \{21, 33, 55, 77\}$. Posteriormente, los resultados del ensamblado fueron procesados con el programa de control de calidad QUAST 4.5 (61) de forma de obtener un conjunto de métricas para la comparación. Del

total de contigs obtenidos se tomaron en cuenta aquellos mayores a 500 pb, para los que se reportaron los valores:

1. #C: cantidad de contigs en el ensamblaje.
2. CML: longitud del contig más largo en el ensamblaje.
3. N50: es la longitud del contig tal que el uso de contigs de longitud mayor o igual a N50 produce la mitad (50 %) de las bases del ensamblado.
4. NG50: es la longitud del contig tal que el uso de contigs de longitud mayor o igual a NG50 produce al menos la mitad (50 %) de las bases del genoma de referencia.
5. #MA: cantidad de posiciones en los contigs ensamblados que sufrieron translocaciónes o inversiones.
6. FG: fracción del genoma de la referencia cubierta por al menos una base de un contig.
7. LTA: largo total alineado.
8. AML: ensamblado de mayor longuitud.

2.2. Resultados

A continuación se describen los resultados obtenidos con los programas estimador y IonDUDE.

2.2.1. Estimación del error del canal

El algoritmo DUDE parte de la base de que la forma en la cual el canal introduce los errores es conocida, es decir que la matriz del canal Π es conocida. Dicha matriz no es provista por los fabricantes de los secuenciadores de ADN por lo que debe ser estimada. El programa *estimador* de matrices fue diseñado para estimar Π empíricamente. La estimación de Π para IonDUDE se realizó a partir de un conjunto de datos que el fabricante de Ion Torrent PGM proporciona en su web, que pertenecen a un experimento de re-secuenciación de cepas de referencia de *Escherichia coli* DH10B. En la tablas 2.3 y 2.4 se observan las matrices de sustituciones e indels, respectivamente, para el IonDUDE donde cada fila corresponde a una base correcta, x, y cada columna corresponde a la base observada, z. Estos valores representan la probabilidad condicional $P(z|x)$. Las matrices de confusión generadas con esta herramienta fueron utilizadas en los experimentos que se describen en la sección 2.2.2.

Tabla 2.3: Ejemplo de una matriz de canal utilizada por IonDUDE para la corrección de errores de sustituciones. Cada fila corresponde a la base correcta y cada columna corresponde a la base observada.

-	A	C	G	T
A	0.9988	0.0003	0.0006	0.0003
C	0.0003	0.999	0.0002	0.0004
G	0.0005	0.0002	0.9990	0.0003
T	0.0003	0.0006	0.0003	0.9988

Tabla 2.4: Ejemplo de una matriz del canal utilizada por el IonDUDE para la corrección de errores de indels. Cada fila corresponde a la longitud correcta del homopolímero y cada columna corresponde a la longitud observada del homopolímero.

-	1	2	3	4	5	6	7	8	9	10
1	0.9970	0.0030	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0048	0.9900	0.0048	0.0002	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0003	0.0133	0.9754	0.0103	0.0005	4.8132e-05	0.0	0.0	0.0	0.0
4	2.4905e-05	0.0005	0.0370	0.9354	0.0256	0.0011	0.0001	6.5999e-05	0.0	0.0
5	3.7456e-06	5.6185e-05	0.0010	0.0739	0.8795	0.0424	0.0024	0.0003	0.0001	0.0001
6	0.0	0.0	0.0002	0.0033	0.1410	0.7779	0.0689	0.0067	0.00136	0.0003
7	0.0	0.0	0.0	0.0011	0.0110	0.2734	0.5566	0.1256	0.02782	0.0042
8	0.0	0.0	0.0	0.0005	0.0041	0.0348	0.2627	0.3512	0.2998	0.0468
9	0.0	0.0	0.0	0.0	0.0	0.0163	0.0521	0.0619	0.5114	0.3583
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0909	0.3636	0.5454

2.2.2. Evaluación de IonDUDE

En la siguiente sección se describen los resultados de la evaluación de la capacidad de corrección de errores de los programas IonDUDE, IonDUDE M y IonDUDE Q.

2.2.2.1. Mapeo de las lecturas

La primera evaluación de la capacidad de corrección de datos de secuenciación masiva de las implementaciones de IonDUDE fue el mapeo de las lecturas corregidas contra una referencia y la posterior contabilización de errores. En primer lugar se determinó el efecto del largo de contexto en las distintas implementaciones de IonDUDE. Este análisis consistió en comparar distintos valores para los parámetros k y h, que corresponden a los largos de contexto considerados para sustituciones e indels, para IonDUDE, IonDUDE M y IonDUDE Q. Los resultados de las muestras ERR161541 y SRR254209 se muestran en las tablas 2.5 y 2.6, respectivamente. A su vez, en las tablas 1.1, 1.2 y 1.3 del Apéndice 1 se muestran los resultados de las muestras ERR039477, ERR236069 y ERR161543, respectivamente. Estas tablas presentan resultados similares a los de las tablas 2.5 y 2.6. Para cada experimento se evaluaron un conjunto de métricas que se detallan en la sección 2.1.3.1.

En los resultados que se muestran en las tablas 2.5, 2.6 se puede observar una mejora de los parámetros evaluados para algunas de las implementaciones del IonDUDE respecto a los datos sin corregir. Los datos para la muestra ERR161541 (Tabla 2.5) muestran que el programa IonDUDE Q corriendo en

Tabla 2.5: Resultados de "BAM QC" para la muestra ERR161541 procesados con IonDUDE, IonDUDE M y IonDUDE

Programa	Contexto	Modo	LM (%)	TE	Mismatches	LMI (%)	LMD (%)	Indels HP (%)
Sin corregir	NA	NA	96.78	0.0155	1325788	28.69	43.01	42.53
IonDUDE	k5	Subs	96.78	0.0156	1325285	28.69	43	42.52
IonDUDE	k8	Subs	96.78	0.0155	1325916	28.69	43.01	42.52
IonDUDE	k10	Subs	96.8	0.0155	1325843	28.69	43.01	42.52
IonDUDE	h3	Indels	96.83	0.0155	1245113	27.55	43.87	41.84
IonDUDE	h4	Indels	96.86	0.0152	1190683	26.62	43.22	41.36
IonDUDE	h5	Indels	96.85	0.0152	1223385	26.98	42.87	41.57
IonDUDE M	k5	Subs	96.75	0.0158	1314708	28.6	42.93	42.48
IonDUDE M	k8	Subs	96.78	0.0155	1325896	28.69	43	42.52
IonDUDE M	k10	Subs	96.8	0.0155	1325901	28.7	43	42.52
IonDUDE M	h3	Indels	96.84	0.0155	1199756	26.79	44.56	41.65
IonDUDE M	h4	Indels	96.89	0.0151	1121627	25.42	43.3	40.53
IonDUDE M	h5	Indels	96.89	0.0149	1149964	25.66	42.64	**40.3**
IonDUDE Q	k5	Subs	96.75	0.0156	1316236	28.62	42.93	42.35
IonDUDE Q	k8	Subs	96.78	0.0155	1325474	28.69	43	42.51
IonDUDE Q	k10	Subs	96.8	0.0155	1326017	28.7	43.01	42.52
IonDUDE Q	h3	Indels	96.78	0.0158	1079617	25.28	45.29	40.61
IonDUDE Q	h4	Indels	96.87	0.015	**981528**	23.43	43.55	41.07
IonDUDE Q	h5	Indels	**97**	**0.0142**	994033	**23.09**	**41.41**	42.87

Contexto, k y h representa el largo de contexto utilizado por lo que por ejemplo $k10$ significa k de largo 10, $h3$ significa h de largo 3; Modo, *Subs* significa que se corrió el programa para corregir solamente errores de sustituciones (en este modo solo se considera k); $Indels$ significa que se corrió el para corregir solamente errores de indels en homopolímeros (en este modo solo se considera h); LM, Lecturas Mapeadas; TE, Tasa de error; LMI, Lecturas mapeadas con inserciones; LMD, Lecturas mapeadas con deleciones; Indels HP, indels en homopolímeros. El mejor resultado para cada métrica esta resaltado en negrita.

Tabla 2.6: Resultados de "BAM QC" para la muestra SRR254209 procesados con IonDUDE, IonDUDE M y IonDUDE Q

Programa	Contexto	Modo	LM (%)	TE	Mismatches	LMI (%)	LMD (%)	Indels HP (%)
Sin corregir	NA	NA	96.59	0.0203	964830	41.88	38.27	46.93
IonDUDE	k5	Subs	96.59	0.0203	966235	41.87	38.27	46.93
IonDUDE	k8	Subs	96.59	0.0203	964842	41.88	38.27	46.93
IonDUDE	k10	Subs	96.59	0.0203	964833	41.88	38.27	46.93
IonDUDE	h3	Indels	96.63	0.0201	993629	42.11	38.74	46.29
IonDUDE	h4	Indels	96.63	0.0201	990592	42.1	38.61	46.31
IonDUDE	h5	Indels	96.63	0.0201	988963	42.2	38.55	46.26
IonDUDE M	k5	Subs	96.56	0.0206	998039	41.74	38.24	46.86
IonDUDE M	k8	Subs	96.59	0.0203	964978	41.87	38.27	46.93
IonDUDE M	k10	Subs	96.59	0.0203	964819	41.88	38.27	46.93
IonDUDE M	h3	Indels	96.66	0.0201	1001930	41.84	38.92	45.97
IonDUDE M	h4	Indels	96.66	0.0199	993093	41.78	38.53	45.75
IonDUDE M	h5	Indels	96.64	0.0199	987126	41.98	38.39	45.86
IonDUDE Q	k5	Subs	96.57	0.0204	979306	41.65	38.2	46.89
IonDUDE Q	k8	Subs	96.59	0.0203	964894	41.87	38.27	46.93
IonDUDE Q	k10	Subs	96.59	0.0203	**964800**	41.88	38.27	46.93
IonDUDE Q	h3	Indels	96.65	0.0203	1114764	38.54	42.35	**45.66**
IonDUDE Q	h4	Indels	96.71	0.0196	1062992	**37.46**	40.01	46.54
IonDUDE Q	h5	Indels	**96.73**	**0.0193**	983693	39.29	**37.84**	47.54

Contexto, k y h representa el largo de contexto utilizado por lo que por ejemplo $k10$ significa k de largo 10, $h3$ significa h de largo 3; Modo, *Subs* significa que se corrió el programa para corregir solamente errores de sustituciones (en este modo solo se considera k); $Indels$ significa que se corrió el para corregir solamente errores de indels en homopolímeros (en este modo solo se considera h); LM, Lecturas Mapeadas; TE, Tasa de error; LMI, Lecturas mapeadas con inserciones; LMD, Lecturas mapeadas con deleciones; Indels HP, indels en homopolímeros. El mejor resultado para cada métrica esta resaltado en negrita.

el modo indels con un contexto h de largo cinco (IonDUDE Q h5 Indels), logra aumentar el porcentaje de lecturas mapeadas exitosamente, disminuir la tasa de error, disminuir el porcentaje de lecturas mapeadas con inserciones y con deleciones. Asimismo, para la misma muestra el programa IonDUDE M h5 Indels redujo la cantidad de indels en homopolímeros y el programa IonDUDE Q h4 Indels muestra una disminución de las bases incorrectas. Por otro lado, los resultados para la muestra SRR254209 (Tabla 2.6) muestran que la corrección con el programa IonDUDE Q h5 Indels aumento el porcentaje de lecturas mapeadas exitosamente, disminuyo la tasa de error y el porcentaje de lecturas con deleciones. Mientras que para la misma muestra el programa IonDUDE Q k10 Subs redujo la cantidad de mismatches y el programa IonDUDE Q h3 redujo el porcentaje de homopolímeros con indels. Sin embargo, algunas de las métricas evaluadas muestran que la corrección de las lecturas con el programa IonDUDE o sus variantes ocasionan pérdida de la calidad en los ensamblajes. Un ejemplo de lo antes mencionado se puede observar en la tabla 2.6 donde la corrección programa IonDUDE Q h3 Indels ocasionó el aumento de la cantidad de bases incorrectas y de la proporción de lecturas mapeadas con deleciones. En vista de algunos resultados negativos con las distintas versiones de IonDUDE se efectuó el siguiente análisis para estudiar el comportamiento de IonDUDE. Partiendo de la formula 1.3 y suponiendo que $z_i = A$

$$m^T \cdot \Pi^{-1} = [n_a, n_c, n_g, n_t]^T \cdot \Pi^{-1}, \tag{2.3}$$

y consideramos los valores n_C, n_G y n_C fijos, estudiamos para qué valores de n_A la regla de decisión de DUDE determina $\hat{x} = A$. Los costos E asociadas a cada base que son objeto de minimización en la regla de decisión de DUDE quedan de la forma

$$\begin{aligned} E_A &= -A_A n_A + K_A, \\ E_C &= A_C n_A + K_C, \\ E_G &= A_G n_A + K_G, \\ E_T &= A_T n_A + K_T, \end{aligned} \tag{2.4}$$

donde los términos K_A, K_C, K_G, K_T y los coeficientes positivos A_A, A_C, A_G, A_T dependen de la matriz Π y de los valores n_C, n_G y n_C (que son fijos para este caso). Si se compara el costo E_A contra E_C, la diferencia entre ellos es

$$E_C - E_A = (A_C + A_A)n_A + K_C - K_A. \tag{2.5}$$

Igualando la diferencia a cero es posible obtener el punto de quiebre de n_A tal que a partir de ese valor E_A es menor y antes de alcanzar ese valor E_C es el menor,

$$n_A = \frac{K_A - K_C}{A_C + A_A}. \tag{2.6}$$

El termino $K_A - K_C$ es de la forma

$$K_A - K_C = \alpha_C n_C + \alpha_G n_G + \alpha_T n_T, \tag{2.7}$$

para ciertos cocientes α_C,α_G y α_T. Dividiendo (2.7) entre la cantidad de veces que ocurre el contexto, queda todo expresado en términos de la probabilidades empíricas de cada símbolo en el contexto $\hat{p}_A$, $\hat{p}_C$, $\hat{p}_G$, $\hat{p}_T$ de forma que

$$\hat{p}_A = \frac{\alpha_C \hat{p}_C + \alpha_G \hat{p}_G + \alpha_T \hat{p}_T}{A_C + A_A}, \tag{2.8}$$

Supongamos que el símbolo correcto es A, es decir, no hubo un error en la posición analizada. En (2.8) se observa que para obtener el umbral más alto, es decir, la situación más exigente para el DUDE, es necesario otorgarle todo el peso que no tenga $\hat{p}_A$ al símbolo C, G o T que tenga el valor más grande de α_C, α_G, α_T y cero al resto. Tomando como ejemplo el símbolo C, los valores quedan de la forma $\hat{p}_G = \hat{p}_T = 0$ y $\hat{p}_C = 1 - \hat{p}_A$. A partir de lo anterior y de la ecuación (2.8) se obtiene que $\hat{p}_A$ debe que ser como mínimo el valor que satisface

$$\frac{\hat{p}_A}{1 - \hat{p}_A} = \frac{\alpha_C}{A_C - A_A}, \tag{2.9}$$

A partir de la ecuación (2.9) se puede obtener $\hat{p}_A$ y repitiendo este análisis para $E_G - E_A$ y $E_T - E_A$ es posible llegar a un valor mínimo de $\hat{p}_A$ que tiene que alcanzar un contexto dado para que el DUDE elija A dado que observó una A.

A partir de un análisis análogo el precedente, tomando el caso particular de sustituciones de A por C, determinamos el valor mínimo de p_A de forma que dada una base observada C siendo la correcta A el DUDE realizara

correctamente la corrección correspondiente (Figura 2.2).

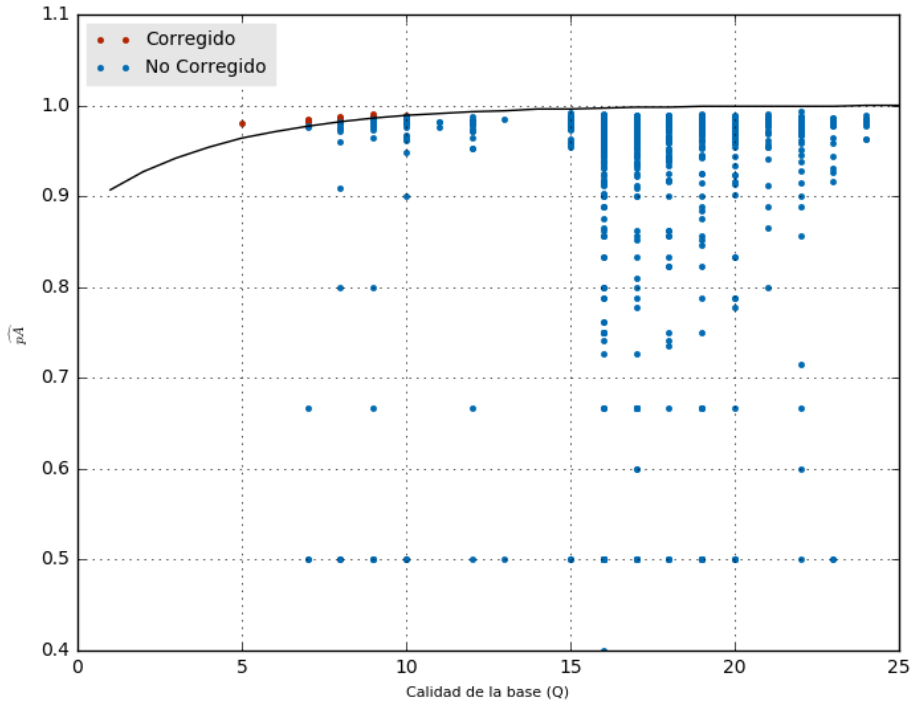

Figura 2.2: $\hat{p}A$ para cada valor de calidad Q para la muestra ERR039477 corregida con IonDUDE Q. La línea representa el valor mínimo de $\hat{p}A$ (Formula 2.9) necesario para que el DUDE realice la corrección de C por A. Los puntos azules representan las bases que no fueron corregidas, y los puntos rojos las bases corregidas.

En la figura 2.2 se aprecia que los datos corregidos con el programa IonDUDE Q h3 para la muestra ERR039477 la enorme mayoría de los valores caen por debajo del umbral, por lo cual no con corregidos. Esto explica por qué el IonDUDE no es capaz de corregir la mayoría de los errores.

Evaluación

Se compararon las implementaciones de IonDUDE respecto a los programas de corrección de datos Fiona, Pollux, Karect e IonHammer. Los resultados se pueden observar en las tablas 2.7, 2.8 para las muestras ERR161541 y SRR254209, respectivamente. Las tablas 1.4, 1.5 y 1.6 del apéndice A, muestran los resultados de las muestras ERR039477, ERR161543 y ERR236069, respectivamente. Estos resultados son similares a los de las muestras ERR161541 y SRR254209 por lo que fueron colocados en el apéndice A.

Tabla 2.7: Resultados de "BAM QC" para la muestra ERR161541

Programa	LM (%)	TE	Mismatches	LMI (%)	LMD (%)	Indels HP (%)
Sin corregir	96.78	0.0155	1325788	28.69	43.01	42.53
Pollux	**99.79**	**0.0052**	197628	7.2	17.39	53.87
Karect	96.77	0.0146	1232736	26.22	41.42	43.33
Ionhammer	97.59	0.0069	**122821**	**3.26**	**8.98**	56.72
Fiona	97.62	0.0056	343160	8.7	22.7	51.38
IonDUDE k5 Subs	96.78	0.0156	1325285	28.69	43	42.52
IonDUDE h3 Indels	96.83	0.0155	1245113	27.55	43.87	41.84
IonDUDE M k5 Subs	96.75	0.0158	1314708	28.6	42.93	42.48
IonDUDE M h3 Indels	96.84	0.0155	1199756	26.79	44.56	41.65
IonDUDE Q k5 Subs	96.75	0.0156	1316236	28.62	42.93	42.35
IonDUDE Q h3 Indels	96.78	0.0158	1079617	25.28	45.29	**40.61**

LM, Lecturas Mapeadas; TE, Tasa de error; LMI, Lecturas mapeadas con inserciones; LMD, Lecturas mapeadas con deleciones; Indels HP, indels en homopolímeros;. El mejor resultado para cada métrica esta resaltado en negrita.

Tabla 2.8: Resultados de "BAM QC"para la muestra SRR254209.

Programa	LM (%)	TE	Mismatches	LMI (%)	LMD (%)	Indels HP (%)
Sin corregir	96.59	0.0203	964830	41.88	38.27	46.93
Pollux	**99.82**	**0.0049**	**221415**	**6.72**	**14.14**	61.51
Karect	96.89	0.0088	499710	18.51	24.42	53.74
IonHammer	96.67	0.0193	961410	37.77	34.95	47.18
Fiona	97.58	0.0074	374670	15.23	16.37	48.44
IonDUDE k5 Subs	96.59	0.0203	966235	41.87	38.27	46.93
IonDUDE h3 Indels	96.63	0.0201	993629	42.11	38.74	46.29
IonDUDE M k5 Subs	96.56	0.0206	998039	41.74	38.24	46.86
IonDUDE M h3 Indels	96.66	0.0201	1001930	41.84	38.92	45.97
IonDUDE Q k5 Sub	96.57	0.0204	979306	41.65	38.2	46.89
IonDUDE Q h3 Indels	96.65	0.0203	1114764	38.54	42.35	**45.66**

LM, Lecturas Mapeadas; TE, Tasa de error; LMI, Lecturas mapeadas con inserciones; LMD, Lecturas mapeadas con deleciones; Indels HP, indels en homopolímeros;. El mejor resultado para cada métrica esta resaltado en negrita.

De los programas correctores de errores evaluadas, Pollux, Fiona y Karect mostraron la capacidad de mejorar las métricas evaluadas en todas las muestras utilizadas. Por otro lado, el programa IonHammer logro mejorar todas las métricas en la muestra ERR161541 salvo por la proporción de homopolímeros con indels que se incrementó pasando de 36.37 % a 47.18 % (Tabla 2.7). El programa Pollux logra destacarse por encima del resto en al menos una métrica en todos los conjuntos de datos ensayados. En la muestra SRR254209 Pollux logra los mejores resultados en todas las métricas evaluadas, salvo en la cantidad de indels en homopolímeros (Tabla 2.8). Por otra parte, los resultados de del programa Karect para la muestra ERR039477 muestran que logra mejorar cuatro de las métricas por encima de los otros programas evaluados (Tabla 2.7). El programa IonDUDE Q h3 Indels logro disminuir la cantidad de indels en regiones homopoliméricas, con respecto a los otros programas evaluados, para las muestras ERR161541 (Tabla 2.7) y SRR254209 (Tabla 2.8).

2.2.2.2. Evaluación del impacto de la corrección sobre el ensamblado de genomas

En esta sección se describen los resultados de la corrección de errores como paso previo al ensamblado *de novo* de genomas. Una vez corregidos los datos, se ensamblaron con el programa Spades y los contigs resultantes fueron procesados con el programa QUAST. Para los contigs obtenidos del paso anterior se reportan las métricas mencionadas en la sección 2.1.3.2.

En la tablas 2.9, 2.10, 2.11, 2.12 y 2.13 se observan las métricas obtenidas para los ensamblados.

Los resultados obtenidos permitieron determinar que el programa Fiona logra los mejores resultados en los conjuntos de datos ERR161541, SRR254209 y ERR039477 (Tablas 2.9, 2.10 y 2.11). A su vez, en la muestra ERR161541 el programa Fiona logra el contig de mayor longitud, un mayor porcentaje del genoma de referencia cubierto y el ensalmado de mayor longitud (Tabla 2.9). Por otra parte, el programa Pollux logró mejorar el valor de N50 en la muestra ERR161543 (Tabla 2.12) y que el ensamblado resultara en el contig más largo en el conjunto ERR236069 (Tabla 2.13). A su vez, la corrección de los datos empleando el programa Karect disminuyó la cantidad de ensamblajes incorrectos en las muestras ERR161541, ERR03947 y ERR236069 (Tablas 2.9, 2.11 y 2.13). Por otra parte, los resultados obtenidos con el programa IonHammer

Tabla 2.9: Resultados del ensamblado de la muestra ERR161541.

Programa	#C	CML	N50	NG50	#MA	FG (%)	AML	LTA
Sin corregir	766	22864	4335	2843	6	67.566	22832	2809470
Pollux	527	49155	10130	8948	2	89.068	49101	3688262
Fiona	**467**	**49162**	**11364**	**10341**	2	**89.737**	**49116**	**3711061**
IonHammer	478	49127	10632	9687	**1**	89.23	49104	3690458
Karect	1925	28496	3037	3180	-	-	-	-
IonDUDE k5 Subs	683	36973	7841	6926	4	88.731	36969	3683921
IonDUDE h3 Indels	722	25931	7586	6513	2	89.23	25826	3705486
IonDUDE M k5 Subs	708	39150	7346	6515	2	88.677	39150	3682760
IonDUDE M h3 Indels	971	15229	3220	2099	5	64.624	15146	2688837
IonDUDE Q k5 Subs	770	22864	4333	2825	6	67.532	22832	2807556
IonDUDE Q h3 Indels	713	39108	6065	4290	6	76.106	39100	3154090

#C, Cantidad de contigs; CML, Contig más largo; #MA, Cantidad de missassemblies; FG, Fracción del genoma cubierto; AML, Alineamiento mas largo; LTA, Largo total alineado. El mejor resultado para cada métrica esta resaltado en negrita.

Tabla 2.10: Resultados del ensamblado de la muestra SRR254209.

Programa	#C	CML	N50	NG50	#MA	FG (%)	AML	LTA
Sin corregir	593	42839	9741	7866	23	79.276	42751	4319965
Pollux	553	31098	4216	-	37	39.34	30925	2138558
Fiona	**481**	**78879**	**18377**	**17715**	27	**92.105**	62387	**5015767**
IonHammer	659	30671	8148	6652	**15**	77.085	30586	4199443
Karect	519	72980	12827	11762	40	86.142	**72707**	4693751
IonDUDE k5 Subs	588	42839	10023	8148	25	80.092	42751	4363665
IonDUDE h3 Indels	672	50700	10073	8580	26	85.898	50597	4685391
IonDUDE M k5 Subs	619	50162	9538	8094	23	80.758	49834	4401756
IonDUDE M h3 Indels	681	48002	9872	8178	28	85.017	47875	4638126
IonDUDE Q k5 Subs	584	51981	9798	8009	23	78.582	51764	4281904
IonDUDE Q h3 Indels	860	29005	6400	5302	22	78.491	28918	4281619

#C, Cantidad de contigs; CML, Contig más largo; #MA, Cantidad de missassemblies; FG, Fracción del genoma cubierto; AML, Alineamiento mas largo; LTA, Largo total alineado. El mejor resultado para cada métrica esta resaltado en negrita.

Tabla 2.11: Resultados del ensamblado de la muestra ERR039477.

Programa	#C	CML	N50	NG50	#MA	FG (%)	AML	LTA
Sin corregir	**1890**	39574	3154	2863	26	91.826	39574	4325251
Pollux	2921	39558	1551	1309	25	82.711	39558	3888517
Fiona	1859	**39575**	**3281**	**3001**	24	**92.112**	**39575**	**4338671**
IonHammer	1891	39572	3139	2851	33	91.64	39572	4316227
Karect	1925	28496	3037	2795	**20**	91.94	28496	4332806
IonDUDE k5 Subs	1890	39574	3154	2863	26	91.826	39574	4325251
IonDUDE h3 Indels	1895	39574	3142	2863	27	91.853	39574	4326187
IonDUDE M k5 Subs	1892	39574	3142	2857	26	91.826	39574	4325300
IonDUDE M h3 Indels	1892	39573	3168	2886	25	91.845	39573	4325059
IonDUDE Q k5 Subs	1895	39574	3139	2858	26	91.822	39574	4325119
IonDUDE Q h3 Indels	1913	39573	3101	2831	26	91.724	39573	4320455

#C, Cantidad de contigs; CML, Contig más largo; #MA, Cantidad de missassemblies; FG, Fracción del genoma cubierto; AML, Alineamiento mas largo; LTA, Largo total alineado. El mejor resultado para cada métrica esta resaltado en negrita.

Tabla 2.12: Resultados del ensamblado de la muestra ERR161543.

Programa	#C	CML	N50	NG50	#MA	FG (%)	AML	LTA
Sin corregir	**101**	6044	2345	–	**7**	0.426	3820	98946
Pollux	139	9746	**2801**	–	13	0.404	5507	94245
Fiona	485	**12113**	2656	–	45	**2.758**	**7750**	**642366**
IonHammer	88	6044	2283	–	9	0.449	3338	104506
Karect	786	19846	2319	–	30	4.644	7219	1080152
IonDUDE k5 Subs	342	9453	2253	–	18	2.438	7289	566890
IonDUDE h3 Indels	443	10174	2389	–	34	2.102	7287	488965
IonDUDE M k5 Subs	330	9520	2322	–	25	1.343	5971	312617
IonDUDE M h3 Indels	239	7827	2213	–	22	1.369	6745	318236
IonDUDE Q k5 Subs	180	6187	2265	–	19	1.073	6064	249713
IonDUDE Q h3 Indels	439	8163	1900	–	38	2.487	4591	578328

#C, Cantidad de contigs; CML, Contig más largo; #MA, Cantidad de missassemblies; FG, Fracción del genoma cubierto; AML, Alineamiento mas largo; LTA, Largo total alineado. El mejor resultado para cada métrica esta resaltado en negrita.

Tabla 2.13: Resultados del ensamblado de la muestra ERR236069.

Programa	#C	CML	N50	NG50	#MA	FG (%)	AML	LTA
Sin corregir	109	168847	**2784608**	63067	27	87.813	154829	2421991
Pollux	94	**350889**	2780678	74286	30	87.988	179836	2423683
Fiona	93	251427	2780587	93816	26	88.086	187236	2426790
IonHammer	92	188170	2782103	86830	**25**	**88.277**	187552	2432270
Karect	**78**	257204	2784233	**108604**	29	88.106	**199382**	**2428189**
IonDUDE k5 Subs	109	168847	2784608	63067	27	87.813	154829	2421991
Iondude h3 Indels	118	194102	2784445	48937	28	87.773	159716	2424113
IonDUDE M k5 Subs	148	84585	2615853	30293	**25**	83.292	83834	2295792
IonDUDE M h3 Indels	111	215153	2783101	63063	27	87.815	159716	2423646
IonDUDE Q k5 Subs	109	168847	2784470	63067	26	87.813	154829	2421975
IonDUDE Q h3 Indels	124	193999	2788746	48817	28	87.669	165146	2421401

#C, Cantidad de contigs; CML, Contig más largo; #MA, Cantidad de missassemblies; FG, Fracción del genoma cubierto; AML, Alineamiento mas largo; LTA, Largo total alineado. El mejor resultado para cada métrica esta resaltado en negrita.

muestran que este programa logró una disminución de los ensamblajes incorrectos en las muestras ERR161541, SRR254209 y ERR236069 (Tablas 2.9. 2.10 y 2.13); y a su vez mejoro el total del genoma cubierto en la muestra ERR236069 (Tabla 2.13). Por último, las implementaciones de IonDUDE logran mejorar algunas de las métricas evaluadas en algunas de las muestras. Para el conjunto de datos ERR161541 (Tabla 2.9) el programa IonDUDE con un largo de contexto k igual a 5 y en modo sustituciones (k5 Subs) Subs, IonDUDE k3 Indels, IonDUDE M k5 Subs e IonDUDE h3 Indels logran mejorar todas las métricas evaluadas. Sin embargo, para el conjunto de datos antes mencionado los programas Pollux, Fiona e IonHammer obtienen mejores resultados en todas las métricas. A su vez, el conjunto de datos SRR254209 (Tabla 2.10) muestra que algunas de las métricas evaluadas mejoraron luego de la corrección con IonDUDE k5 Subs, IonDUDE Mk5 Subs, IonDUDE h3 Indels e IonDUDE Q k5 Subs. Cabe destacar que para el conjunto de datos antes mencionado el programa Fiona logra mejores resultados. Es importante destacar que algunas de los resultados obtenidos con algunas de las variantes del IonDUDE ocasionan que los ensamblajes muestren un deterioro de las métricas evaluadas. Un ejemplo de lo antes mencionado se puede observar para los resultados del programa IonDUDE Qh3 Indels para la muestra SRR254209 (Tabla 2.10). Otro ejemplo del deterioro de los ensamblados, se puede obser-

var en los resultados obtenidos con el programa IonDUDE M h3 Indels para la muestra ERR161541 (Tabla 2.9), donde todas las métricas evaluadas son inferiores respecto a los datos sin corregir salvo el número de ensamblados incorrectos.

2.3. Discusión

En este trabajo se introduce y evalúa una aproximación a la corrección de errores de secuenciación de ADN con la plataforma Ion Torrent PGM. Se adaptó un algoritmo de reducción de errores, DUDE, a la corrección de errores de secuenciación de ADN.

2.3.1. Estimación de errores

El DUDE, es su formulación original (174), esta diseñado para corregir errores introducidos en una secuencia que es transmitida por un canal que la corrompe. El canal introduce cambios en los símbolos de la secuencia de forma independiente y estadísticamente idéntica. En este trabajo los símbolos de la secuencia son los nucleótidos de una secuencia de ADN y el secuenciador de ADN representa el canal por el cual se transmiten. La primera tarea en la implementación del DUDE al problema de la secuenciación de ADN fue la estimación de la matriz del canal, dicha matriz de transición de estados describe por completo al canal (183). La implementación del estimador de la matriz del canal de este trabajo es independiente de la plataforma. El único requerimiento para el funcionamiento del estimador es contar con un conjunto de datos de la secuenciación de una muestra y la secuencia de la muestra libre de errores, típicamente datos de la re-secuenciación de un genoma. Utilizando la información del mapeo de las lecturas contra la referencia se estima la matriz del canal.

En este trabajo de tesis el secuenciador utilizado para llevar a cabo los experimentos es el Ion Torrent PGM. Este secuenciador tiene un perfil de errores característico, donde la mayoría de los errores son inserciones o deleciones en regiones homopoliméricas (94; 28). Las matrices obtenidas en este trabajo para datos secuenciados con Ion Torrent coinciden con la clase de errores reportados para esta plataforma (28).

2.3.2. Evaluación de IonDUDE

En este trabajo se implementó el algoritmo DUDE utilizando tres aproximaciones diferentes llamadas IonDUDE, IonDUDE M y IonDUDE Q (sección 2.1.2). La evaluación de la capacidad de corrección de errores de cada una de las versiones de IonDUDE se llevó a cabo empleando dos estrategias. La primera

estrategia consistió en evaluar la calidad del mapeo de las lecturas contra una referencia antes y después ser procesadas con las variantes de IonDUDE. En los resultados de la evaluación antes mencionada se destacó la versión IonDUDE Q, obteniendo mejores resultados en la enorme mayoría de los experimentos (Tablas 2.5, 2.6 y Tablas 1.1, 1.2 y 1.3 del apéndice A). Este resultado siguiere que la incorporación de matrices dependientes de la posición y de los valores de calidad q a las matrices del canal permitieron ajustar mejor el modelo de IonDUDE. A pesar de lo anterior, es importante resaltar que los resultados de algunas versiones de IonDUDE ocasionaron un deterioro de las métricas evaluadas, por lo se llevó a cabo un análisis en mayor en profundidad del IonDUDE. El análisis del comportamiento de IonDUDE Q mostró que la regla de decisión del DUDE no es capaz de corregir la mayoría de los errores. Este resultado sugiere que un grupo de datos con mayor cobertura podría ser corregido con mayor éxito. En este sentido existe una herramienta de corrección de datos de secuenciación de ADN basada en el algoritmo DUDE denominada DUDE-Seq (87). El programa DUDE-Seq fue diseñado para la corrección de errores en datos de secuenciación de ADN provenientes de la amplificación de un único gen. Está metodología de trabajo permite estudiar la población de un gen en una muestra, en el caso de estudio de diversidad bacteriana el gen usualmente amplificado es *16S ARNr* (31; 30; 126; 17). Los datos de secuenciación de amplicones poseen una alta cobertura ya que se secuencia el mismo gen en los distintos organismos presentes en la muestra. Esto sugiere que la alta cobertura de las secuencias con las que el DUDE-Seq trabaja permiten que el algoritmo DUDE pueda realizar los cambios correspondientes en las bases erróneas. Sería interesante utilizar el DUDE-Seq para corregir los datos utilizados en este trabajo. Sin embargo llevar esta clase de experimentos seria desafiante por que el DUDE-Seq utiliza los espectrogramas que producen los secuenciadores y no los archivos .fasta o .fastq. Por otra parte, sería interesante utilizar IonDUDE para la corrección de datos de secuenciación de amplicones.

La segunda evaluación de las variantes del IonDUDE se realizó utilizando la misma metodología antes mencionada pero incorporando herramientas de corrección disponibles en la literatura como comparativa. Las herramientas seleccionadas para la comparación fueron Pollux (104), Karect (4), IonHammer (15) y Fiona (148). Los resultados de la evaluación muestran los programas IonDUDE, IonDUDE M y IonDUDE Q no logran el nivel de corrección de las herramientas Pollux, Karect y Fiona. En este sentido, el programa Pollux logra

destacarse en dos de los cinco conjuntos de datos, donde obtiene los mejores resultados en cuatro de cinco métricas evaluadas.

2.3.3. Efecto de la corrección sobre el ensamblado de genomas

El programa para ensamblado Spades construye un grafo a partir de las lecturas, dicho grafo representa los kmeros que ocurren en las lecturas y los solapamientos entre ellas (15). La secuencia genómica de la que se originan las lecturas está representada por un camino en el grafo antes mencionado. Los errores en las lecturas pueden dificultar la construcción del grafo ya que estos generan kmeros con conexiones quiméricas entre nodos, caminos paralelos o caminos sin salida. A pesar de esto, una gran parte de los errores son inocuos para el ensamblado ya que los ensambladores poseen mecanismos para la corrección de los artefactos antes mencionados (67). En este sentido la disminución de la tasa de error medida a partir de las lecturas mapeadas contra una referencia no lleva necesariamente a la mejora del ensamblado de los datos. Sin embargo, los errores que ocasionan que un kmero se transforme en otro ya existente en el genoma pueden ocasionar problemas para los ensambladores. Este tipo de errores ocasiona ensamblados incorrectos (translocaciónes o inversiones) o contigs de menor tamaño, por lo cual la corrección de estos errores es un paso importante en el ensamblado de genomas (140; 67). Teniendo en cuenta lo anterior y tomando como ejemplo los resultados obtenidos para la muestra SRR254209, el programa Pollux logra los mejores resultados en las métricas de "BAM QC" pero el ensamblado resultante está más fragmentado con respecto al obtenido post-corrección con Fiona. Este resultado siguiere que Fiona corrige un mayor número de errores problemáticos para el ensamblado, lo cual redunda en un ensamblado menos fragmentado.

Los datos procesados con el IonDUDE lograron leves mejoras en las métricas de los ensamblados. Sin embargo, el programa Fiona se destaca mejorando la mayoría de las métricas de los ensamblados en cuatro de las cinco muestras. Fiona logró principalmente disminuir la fragmentación del ensamblado, lo que permitirá lograr mejores resultados en los análisis posteriores. Este resultado llevo a la decisión de utilizar este programa para el ensamblado de los datos de los genomas de las cepas *Enterobacter* sp. UYSO10 y *Shinella* sp. UYSO24.

Capítulo 3

Genómica de bacterias endófitas asociadas al cultivo de caña de azúcar *Saccharum officinarum*

En este capítulo se llevo a cabo el estudio genómico de las cepas endófitas modelo *Enterobacter* sp. UYSO10 y *Shinella* sp. UYSO24. En primer lugar, se corrigieron, ensamblaron y anotaron los datos de secuenciación de ADN de las cepas UYSO10 y UYSO24. En segundo lugar, se realizo la caracterización genómica de las cepas en estudio poniendo énfasis en las características relacionadas a la interacción planta-endófito y la promoción del crecimiento vegetal.

3.1. Materiales y métodos

Este trabajo se enmarcó en una de las líneas de investigación del Departamento focalizada en el estudio de bacterias endófitas promotoras del crecimiento vegetal (PCV) asociadas a cultivos de interés agronómico. Previamente se construyó y caracterizó (bioquímica, fisiológica y fenotípicamente), una colección de probables endófitos bacterianos a partir de variedades de caña de azúcar cultivadas en Uruguay (167). Como parte de la misma, posteriormente se demostró que los aislamientos *Enterobacter* sp. UYSO10 y *Shinella* sp. UYSO24, son específicamente PCV de la variedad de caña LCP85-384 y son endófitos verdaderos (166). Teniendo en cuenta las características mencionadas, dichas cepas son empleadas como modelos de estudio en el laboratorio. En este contexto, los genomas de las cepas *Enterobacter* sp. UYSO10 y *Shinella* sp. UYSO24 fueron secuenciados en la Plataforma de Secuenciación Masiva del Instituto de Investigaciones Biológicas "Clemente Estable" utilizando la plataforma *Ion Torrent PGM*.

3.1.1. Secuenciación, control de calidad, ensamblado y anotación de los genomas

Los aislamientos *Enterobacter* sp. UYSO10 y *Shinella* sp. UYSO24 fueron crecidos en tubos de ensayo conteniendo 10 ml de medio de cultivo líquido rico TSB a 28° en agitación, hasta alcanzar una densidad óptica a 620nm de 1.0 ($D.O._{620} = 1.0$). Dicho cultivo fue centrifugado a 10000x g durante 5 min, y el pellet re-suspendido en un 1ml de NaCl 0.9 %. A partir de la suspensión obtenida se extrajo el ADN utilizando el kit Quick-DNA^{TM}-Fungal/Bacterial Miniprep Kit (Zymo Research, EEUU) siguiendo las indicaciones del fabricante. La calidad del ADN (integridad, presencia de proteínas y ARN), fue verificada en un espectrómetro Nanodrop (Thermo Fisher Scientific, Waltham, EEUU), así como en un Bioanalyzer 2100 (Agilent Genomics, Santa Clara, EEUU). Posteriormente el ADN fue secuenciado utilizando la estrategia de *shotgun sequencing* en la plataforma Ion Torrent PGM (Thermo Fisher Scientific, Waltham, EEUU). La calidad de las lecturas fue evaluada con el programa FastQC (6) y posteriormente procesadas con el programa Sickle 1.33 (75) para remover las lecturas de baja calidad o longitud incorrecta. Luego, las lecturas fueron corregidas con el programa Fiona 0.2.9 empleándose los parámetros por

defecto salvo el valor del parámetro **g**. En esté parámetro, para la cepa UYSO10 se empleó un valor de **g** = 5800000, mientras que para la cepa UYSO24 un valor **g** = 6000000. En ambos casos el valor corresponde aproximadamente al tamaño total ensamblado de los datos sin corregir. Después del control de calidad y corrección de los datos, estos fueron ensamblados *de novo* con el programa Spades 3.11.1 (15) empleando los parámetros [—only-assembler] [-–iontorrent] y [k = 21,33,55,71]. La calidad de los datos ensamblados fue evaluada con el programa QUAST 4.5 (61) (Figura 3.1). Asimismo, se empleo el programa plasmidSPAdes (8) para la reconstrucción de plásmidos en los datos, siendo los parámetros utilizados: [–only-assembler] [–iontorrent] [–plasmid] y [k = 21,33,55,71]. Los contigs resultantes del ensamblado se ordenaron y orientaron (scaffolding) utilizando los programas Mauve (134) y Medusa (27) por separado, ejecutándose ambas herramientas con los parámetros por defecto. Posteriormente, los scaffolds fueron anotados utilizando el servicio web *Rapid Annotation using Subsystem Technology* (RAST) (10), el programa Prokka (149) y el servicio web *KEGG Automatic Annotation Server* KAAS (108). Un esquema del proceso antes mencionado se puede observar en la figura 3.1.

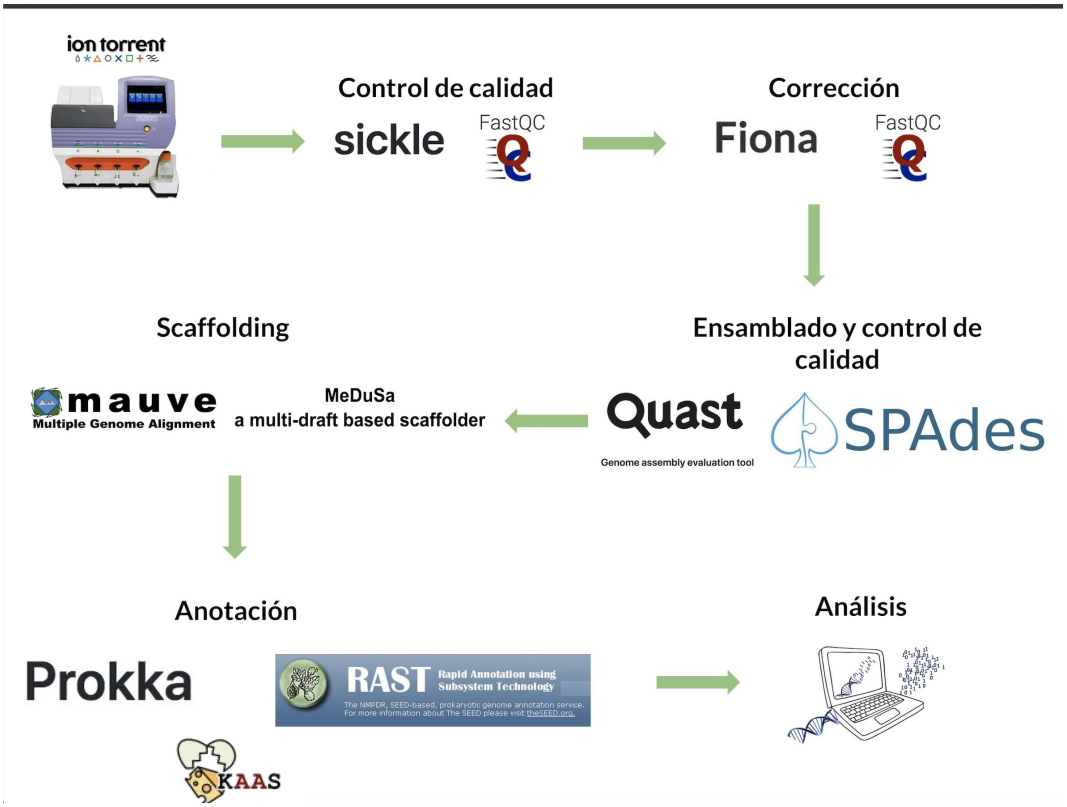

Figura 3.1: Esquema del proceso de control de calidad, ensamblado y anotación de los datos genómicos de las cepas UYSO10 y UYSO24.

El servicio web RAST provee anotaciones de genes individuales y de subsistemas a partir de los datos de secuenciación. Un subsistema es un conjunto de roles funcionales que fueron agrupados por un anotador experto y es utilizado por el servicio de anotación RAST (116). Un rol funcional por su parte es una función metabólica abstracta como por ejemplo *aspartokinasa*.

3.1.2. Clasificación taxonómica y estructura de los genomas en estudio

3.1.2.1. Clasificación taxonómica de las cepas

Con el fin de verificar la correcta identificación a nivel de género, se utilizaron herramientas que emplean el total de los datos genómicos. El cálculo de la identidad nucleótidica promedio (ANI por sus siglas en inglés *Average Nucleotide Identity*) (58; 80), consiste en el alineamiento de las secuencias genómicas de las cepas a comparar y el posterior conteo de las bases homólogas. Dicha técnica fue propuesta como estándar para la determinación de las relaciones

taxonómicas entre procariotas, en reemplazo de la técnica de hibridización ADN-ADN (HAA), teniendo en cuenta la gran cantidad de datos genómicos disponibles en los últimos años (132). A su vez, se utilizo un índice basado en la frecuencia de tetra-nucleótidos (TETRA) presentes en una secuencia genómica (132). A diferencia del calculo de ANI, en este caso no es necesario alinear los genomas sino que el índice de correlación TETRA se obtiene para una secuencia genómica y es comparado contra el índice obtenido para otro genoma.

El calculo de ANI para las cepas UYSO10 y UYSO24, se realizó utilizando el servicio web JSpeciesWS (133), el cual también permite la comparación del índice TETRA de un genoma empleando la base de datos GenomesDB, mediante el análisis TCS (por sus siglas en ingles TETRA *correlation search*). Cabe mencionar que, a pesar de utilizar las mismas herramientas, las estrategias de identificación variaron entre los genomas. A continuación se detallan las estrategias utilizadas en cada caso.

Clasificación taxonómica de la cepa UYOS10

Partiendo de la secuencia del gen *16S ARNr* obtenida de la anotación realizada para este aislamiento (Sección 3.1.1), se confeccionó un árbol filogenético. Para esto, la secuencia del gen *16 ARNr* fue clasificada y alineada con el programa SINA v1.2.11 (124). La clasificación se realiza comparando la secuencia del gen *16S ARNr* contra las bases de datos de SILVA (125). Posteriormente, se seleccionaron aquellas secuencias que poseían una similitud mayor a 0.97 % de identidad nucleótidica del gen *16S ARNr* con respecto al aislamiento UYSO10 (Tabla 3.1). Dicho umbral de identidad es considerado el valor mínimo para clasificar dos organismos procariotas dentro de la misma especie (45). Finalmente, con las secuencias de la tabla 3.1 y el gen *16 ARNr* de UYSO10 se reconstruyó la filogenia de las mismas con el programa MEGA 7.0.26 y el algoritmo *Neighbour Joining* con 1000 replicas de *bootstrap* (83). En la reconstrucción del árbol filogenético, la cepa *Enterobacter cloacae subsp. cloacae* ATCC 13047 fue utilizada como grupo externo.

Tabla 3.1: Lista de los códigos de acceso de los genes *16S ARNr* tomadas del GenBank pertenecientes al género *Kosakonia*.

Organismo	N^o de acceso
Kosakonia radicincitans DSM 16656	CP018016
Kosakonia radicincitans GXGL-4A	CP015113
Kosakonia radicincitans UMEnt01	JDYJ01000053
Kosakonia radicincitans YD4	JSFC01000001
Kosakonia oryziphila REICA 142	JF795013
Kosakonia cowanii CIP 107300	AJ508303
Kosakonia oryzae Ola 51	EF488759
Kosakonia arachidis Ah-143	EU672801
Kosakonia sacchari SP1	ERR161543
Enterobacter cloacae subsp. cloacae ATCC 13047	ATCC13047

Posteriormente, se calculó el ANI con el servicio web JSpeciesWS, del genoma de la cepa UYSO10 con respecto a los genomas secuenciados disponibles en el servidor FTP de *National Center for Biotechnology Information* (NCBI) (ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/) pertenecientes a las cepas del género *Kosakonia*, las secuencias utilizadas se detallan en la tabla 3.2. El valor de ANI puede ser obtenido utilizando dos estrategias: ANIb y ANIm. Estas aproximaciones consisten en el fraccionamiento del genoma problema en secciones de 1020 pb, seguido del alineamiento de las mismas contra las referencias empleadas. ANIb utiliza el programa BLAST (5) para realizar el alineamiento de las secciones y ANIm hace uso del programa de alineamiento de secuencias MUMMer (84). A su vez, se comparó el valor de índice TETRA del genoma de UYSO10 con respecto al índice de los genomas de la tabla 3.2.

Tabla 3.2: Lista de los códigos de acceso de los genomas tomados del GenBank pertenecientes al género *Kosakonia*.

Organismo	N^o de acceso
Kosakonia radicincitans DSM1665	GCA_000280495.1
Kosakonia radicinticans GXGL-4A	GCA_001887675.1
Kosakonia radicincitans UMEnt01	GCA_000691205.1
Kosakonia radicincitans YD4	GCA_000877295.1
Kosakonia oryzendophytica REICA 082	GCA_900185945.1
Kosakonia oryzae Ola 51	GCA_001658025.1
Kosakonia sacchari SP1	GCA_000300455.4
Kosakonia sacchari CGMCC 1.12102	GCA_900100995.1
Kosakonia oryzae CGMCC 1.7012	GCA_900112145.1
Kosakonia sacchari BO-1	GCA_001683395.1
Kosakonia arachidis Ah-143	GCA_900116535.1
Kosakonia sp. S29	GCA_900112785.1

Clasificación taxonómica de la cepa UYSO24

En primera instancia para, la secuencia del gen *16S ARNr* se alineó y clasificó con el programa SINA v1.2.11 (124) utilizando el mismo criterio que para la cepa UYSO10. Debido a que esta aproximación no fue concluyente y sólo fue capaz de colocar al aislamiento dentro de la familia *Rhizobiaceae* se realizo la reconstrucción filogenética de la familia *Rhizobiaceae* con el programa PhyloPhlAn (150). El programa PhyloPhlAl fue ejecutado en modalidad *de novo* la cual reconstruye la filogenia en base a un acervo de 400 proteínas las cuales son detectadas en los genomas provistos y alineadas para finalmente reconstruir la filogenia. A partir de la filogenia antes mencionada se seleccionaron los genomas filogenéticamente cercanos a la cepa UYSO24 para el análisis de TETRA, ANIm y ANIb.

3.1.2.2. Estructura de los genomas de las cepas UYSO10 y UYSO24

Una vez finalizado el ensamblado y la anotación de los datos, se prosiguió a determinar la estructura genómica de ambos genomas en estudio. Este análisis se realizó para determinar el número de replicones que cada aislamiento posee, así como para estudiar la distribución de los genes a lo largo de los mismos.

Asimismo, se estudió la estabilidad genómica de ambos aislamientos mediante la detección de secuencias de inserción (SI) e islas genómicas. Para esto, en primer lugar, se utilizó el servicio web ISsaga para la anotación de las SI (171). Por último, la anotación de las islas genómicas se realizó con el servicio web IslandViewer 4 (24).

Por otro lado se confeccionaron mapas de los replicones de cada aislamiento en estudio. Para el caso de la cepa UYSO10, los contigs obtenidos del ensamblado con el programa Spades, fueron procesados con el programa Medusa 1.6 (27), empleando los parámetros por defecto, con el fin de obtener un conjunto de *scaffolds* a partir de la combinación de los mismos. Posteriormente, cada uno de los genes anotados del replicón fueron clasificados en sus correspondientes categorías *clusters of orthologous groups* COG (164), mediante el programa RPS-BLAST versión 2.7.1 con las matrices de puntuación especificas de posición obtenidas de COG. Por último, mediante la aplicación web GView Server (https://server.gview.ca), se confeccionó un mapa circular del replicón. Para el caso del genoma UYSO24, en primera instancia se ordenaron y orientaron los contigs, obtenidos con el programa Spades, con el programa Mauve *snapshot_2015-02-25* empleando los parámetros por defecto. Posteriormente se clasificaron cada uno de los genes anotados en las correspondientes categorías de la base de datos COG utilizando el programa RPS-BLAST 2.7.1 con las matrices de puntuación especificas de posición obtenidas de COG. Por último, empleando la aplicación web GView Server (https://server.gview.ca) se confeccionó un mapa del replicón.

3.1.3. Principales características genómicas relacionadas a la interacción planta-bacteria

Los resultados de la anotación de los genomas fueron utilizados para extraer las principales características de los mismos. En este trabajo se hizo énfasis en aquellas características relacionadas con la capacidad de colonizar los tejidos vegetales y prosperar en ese ambiente, así como en las características relacionadas con la PCV. Considerando lo antes mencionado y teniendo en cuenta la bibliografía, se buscaron en el genoma genes, operones y regulones relacionados a la fijación biológica de nitrógeno (FBN), producción de fitohormonas, producción de sideróforos, motilidad, quimiotaxis, *pillus* y sistemas de secreción. Particularmente y teniendo en cuenta la capacidad de FBN

de la cepa *Enterobacter* sp. UYSO10 (167; 166) se buscaron, en el genoma de esta cepa, posibles genes codificantes para la enzima nitrogenasa. Los genes antes mencionados fueron comparados con los genes homólogos presentes en los genomas de cepas bacterianas del mismo género y de la cepa de referencia *Gluconacetobacter diazotrophicus* PA15 (23). Por otra parte, se presume para la cepa *Shinella* sp. UYSO24 que el rol de la formación de biopelículas es de gran relevancia en la capacidad PCV observada. Esta hipótesis se basa en el hecho de que está cepa presenta una profusa colonización de la superficie radicular en forma de biopelícula, pero una discreta colonización de los tejidos internos de la planta (166). Teniendo en cuenta estas características, en primera instancia, se buscaron en el genoma posibles genes relacionados con la producción de exopolisacáridos (EPS). Los genes u operones relacionados con esta función fueron comparados con los homólogos presentes en las cepas *Neorhizobium galegae* HAMBI540 y *Sinorhizobium meliloti* 1021 (51).

Una vez obtenidos los conjuntos de posibles genes codificantes para la enzima nitrogenasa, así como para la producción de exopolisacáridos, en las cepas UYSO10 y UYSO24 respectivamente; se obtuvieron las regiones de similitud entre las secuencias con el programa TBLASTX versión 2.6.0+ (5) con un valor de *e-value* de 0.001. Finalmente, se construyeron visualizaciones de los conjuntos de genes a partir de los alineamientos con el programa Easyfig versión 2.2.2 (161).

3.2. Resultados

A continuación se describen los resultados obtenidos a partir del estudio de las principales características de los genomas de las cepas de UYSO10 y UYSO24.

3.2.1. Secuenciación, control de calidad, ensamblado y anotación de los genomas

Para la secuenciación del genoma de la cepa UYSO10 se realizaron dos experimentos en la plataforma Ion Torrent PGM. Como resultado y luego del control de calidad se obtuvieron 658.146 lecturas. El ensamblado de las lecturas resulto en 99 contigs, a partir de los cuales se obtuvieron 12 scaffolds con un contendido GC de 54 %, un valor de N50 de 346.125, siendo el contig de mayor longitud obtenido de 678.970 pb. Luego del proceso de *scaffolding*, se obtuvo un único scaffold de 5.790.203 pb el cual tiene 25 gaps. El genoma predicho presenta 5948 secuencias codificantes para proteínas, una copia del gen *16S ARNr*. Entre los genes predichos el 33 % (1908) comprenden 405 subsistemas del sistema SEED.

Por otra parte, para la secuenciación de genoma de la cepa *Shinella* sp. UYSO24, se realizaron también dos experimentos de secuenciación en la plataforma Ion Torrent PGM. Como resultado y luego del control de calidad se obtuvieron 787.334 lecturas. El ensamblado de las mismas resultó en 135 contigs y un contenido GC de 61.1 %, con valor de N50 de 269.086 pb y siendo el contig de mayor longitud obtenido de 588.718 pb. El análisis del genoma predicho presentó 8093 secuencias codificantes para proteínas, las cuales comprenden 365 subsistemas del sistema SEED y una copia del gen *16S ARNr*. Mediante la aproximación empleada, para el aislamiento en cuestión no fue posible obtener un conjunto scaffolds. Asimismo, para esta cepa se logró ensamblar un plásmido con un longitud de 136,818 pb, compuest por 24 contigs y con un contenido GC de 58.7 %. La anotación de dicho plásmido con el servicio web RAST muestra que presenta 191 CDS, de las cuales el 14 % fueron clasificadas dentro de algún subsistema de sistema SEED.

3.2.2. Identificación y estructura genómica

3.2.2.1. Clasificación taxonómica de las cepas en estudio

Previamente la cepa UYSO10 fue identificada como relacionada al género *Enterobacter* mediante la secuenciación parcial del gen *16S ARNr* (167). A partir de la secuencia completa de dicho gen obtenida en este trabajo, se realizó la reconstrucción filogenética con el fin de confirmar su asignación filogenética (Figura 3.2). Los resultados sugieren que la cepa UYSO10 pertenece al género *Kosakonia* y a la especie *radicincitans*.

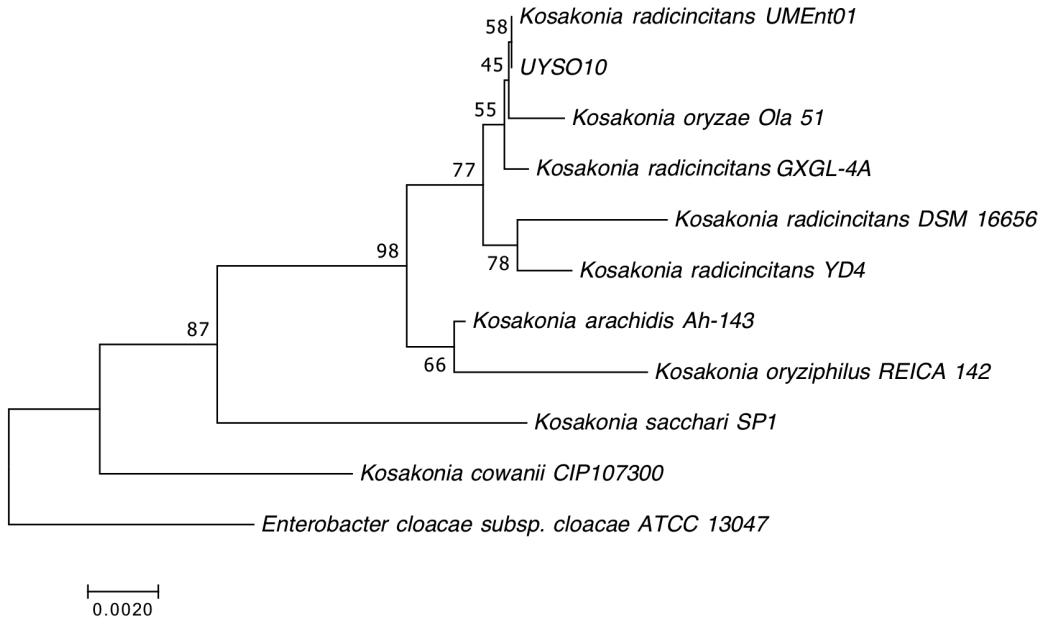

Figura 3.2: Reconstrucción filogenética basada en el gen *16S ARNr* utilizado para la asignación taxonómica del aislamiento *Enterobacter* sp. UYSO10. Se construyo utilizando el método de *neighbor-joining* (1000 replicas de *bootstrap*). Solamente se muestran los valores de *bootstrap* < 100.

A partir de los genomas disponibles en las bases de datos para cepas del género *Kosakonia* (Tabla 3.2), se calcularon los valores de ANIb, ANIm y TETRA con respecto a la cepa UYSO10 (Tabla 3.3). Por lo general, un valor de 96 % de ANI y 0.99 de TETRA son considerados una medida robusta de similitud genómica entre cepas (132; 78). Cabe destacar que los valores de ANIb, ANIm y TETRA superan los valores umbrales para las cepas *K. radicincitans* UMEnt01, *K. radicincitans* YD4, *K. radicincitans* DSM1665, *K. radicincitans* GXGL-4A, *K. oryzae* CGMCC1.7012, *K. sacchari* SP1 y *K. oryzae* Ola 51.

Tabla 3.3: Resultados de ANIb, ANIm y TETRA obtenidos con el servicio JSpeciesWS, con respecto a la secuencia genómica de la cepa UYSO10.

Organismo	ANIb	ANIm	Tetra
***Kosakonia radicincitans* DSM 16656**	**99**	**99**	**0.99**
***Kosakonia radicinticans* GXGL-4A**	**99**	**99**	**0.99**
***Kosakonia radicincitans* UMEnt01**	**99**	**99**	**0.99**
Kosakonia radicincitans YD4	95	96	0.99
Kosakonia oryzendophytica REICA 082	82	85	0.98
***Kosakonia oryzae* Ola 51**	**96**	**96**	**0.99**
Kosakonia sacchari SP1	95	96	0.99
Kosakonia sacchari CGMCC 1.12102	83	86	0.99
***Kosakonia oryzae* CGMCC 1.7012**	**96**	**96**	**0.99**
Kosakonia sacchari BO-1	83	86	0.99
Kosakonia arachidis Ah-143	93	93	0.99
Kosakonia sp. S29	79	84	0.96

Las cepas que superan los valores umbrales para los valores de ANIb, ANIm y Tetra están resaltados en negrita.

Por otro lado, análisis preliminares utilizando el gen marcador *16S ARNr* mostraron que la cepa UYSO24 esta filogenéticamente relacionada con bacterias de la familia *Rhizobiaceae* (166). El estudio antes mencionado no permitió clasificar a la cepa UYSO24 a nivel de especie debido a la baja homología del gen marcador *16S ARNr* con secuencias depositadas en la base de datos SILVA. Es por esto último que se realizo la reconstrucción filogenética con el programa PhyloPhlAn de la familia *Rhizobiaceae* junto con la cepa UYSO24 (https://github.com/mberacochea/tesis-master). Los resultados mostraron que la cepa UYSO24 está relacionada a un grupo de rizobios aislados de *Arabidopsis thaliana* (11) y al clado de *Neorhizobium galegae* (Tabla 3.4)).

Tabla 3.4: Resultados de ANIb, ANIm y TETRA obtenidos con el servicio JSpeciesWS, con respecto a la secuencia genómica de la cepa UYSO24.

Organismo	Estado*	N^o de acceso	ANIb	ANIm	Tetra
Rhizobium sp. Leaf 306	Contig	GCA_001423425.1	76	85	0.98
Rhizobium sp. Leaf 321	Scaffold	GCA_001423215.1	76	85	0.98
Rhizobium sp. NFR12	Contig	GCA_900108545.1	76	85	0.98
Rhizobium sp. LCM 4573	Contig	GCA_001854865.1	75	84	0.95
Rhizobium oryzae B4P	Contig	GCA_900177415.1	77	85	0.92
Rhizobium sp. YS-1r	Scaffold	GCA_000757525.1	76	84	0.96
Rhizobium sp. CF080	Contig	GCA_000282095.2	76	84	0.94
Rhizobium sp. LC145	Contig	GCA_001005825.1	75	84	0.95
Rhizobium sp. NFR07	Contig	GCA_900111905.1	76	85	0.97
Neorhizobium galegae HAMBI 1145	Contig	GCA_000985915.1	76	84	0.94
Neorhizobium galegae HAMBI 2427	Contig	GCF_000986035.1	76	84	0.94
Neorhizobium galegae HAMBI 540	Complete	GCA_000731315.1	76	84	0.94
Neorhizobium galegae HAMBI 1141	Complete	GCA_000731295.1	76	84	0.94
Neorhizobium vignae CCBAU 05176	Contig	GCA_000732195.1	76	84	0.94

*Estado del proyecto de secuenciación en la base de datos del NCBI.

3.2.2.2. Estructura genómica de las cepas en estudio

A partir del scaffold obtenido para la cepa UYSO10, se confeccionó un mapa genómico de la cepa UYSO10 (Figura 3.3).

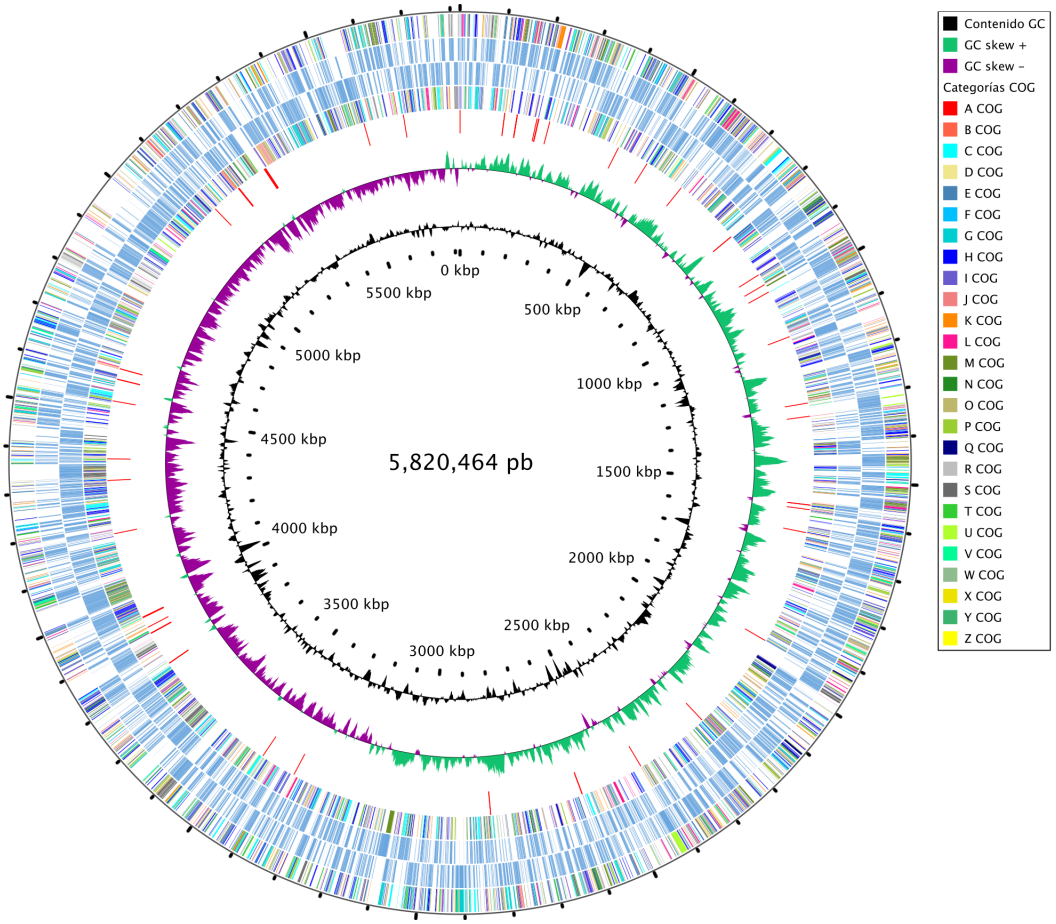

Figura 3.3: Mapa circular del cromosoma de *Kosakoni radicincitansa* UYSO10. Desde afuera hacia adentro: CDS en la hebra sentido coloreados de acuerdo a su categoría COG, CDS en la hebra sentido, CDS en la hebra anti-sentido, CDS en la hebra anti-sentido coloreados de acuerdo a su categoría COG, tRNA y rRNA, GC skew y contendido GC %.

Por otro lado, los resultados obtenidos con en el servicio web IslandViewer muestra que el genoma de la cepa UYSO10 presenta 30 posibles islas genómicas (Figura 3.4). A su vez, este genoma presenta cuatro probables secuencias de inserción, las cuales pertenecen a las familias IS3 e IS30.

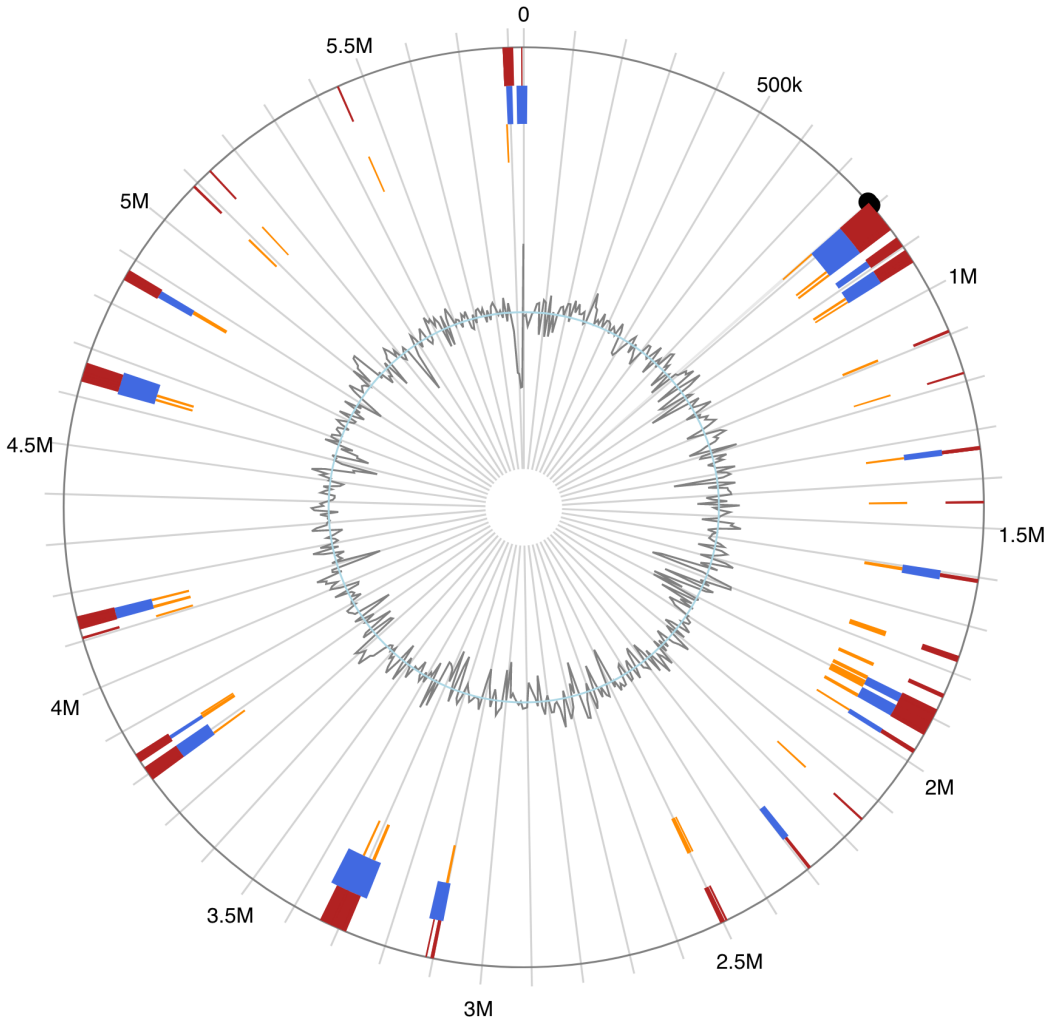

Figura 3.4: Mapa del genoma de la cepa *Kosakonia radicincitans* UYSO10. De afuera hacia adentro: marcador de posición, islas genómicas detectadas por: al menos un método (rojo), IslandPath-DIMOB (azul), y SIGI-HMM (amarillo); contenido GC.

Por otro lado, a partir de los contigs obtenidos para la cepa UYSO24 se confeccionó un mapa genómico con el servicio web GView (Figura 3.5), incluyendo la comparación el subconjunto de los genomas de la tabla 3.4.

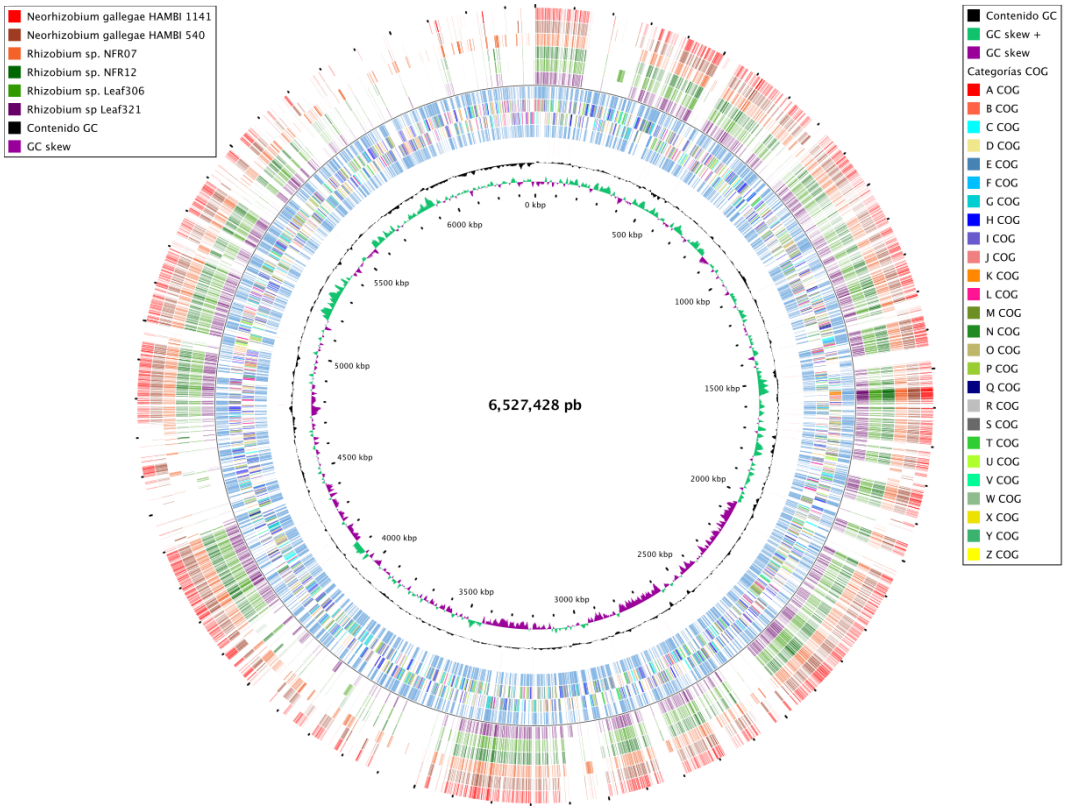

Figura 3.5: Mapa circular del cromosoma de *Rhizobium* sp. UYSO24. Desde afuera hacia adentro: Comparación basada en BLAST+ de los genomas *Neorhizobium gallegae* HAMBI 1141, *Neorhizobium gallegae* HAMBI 540, *Rhizobium* sp. NFR07, *Rhizobium* sp. NFR12, *Rhzibium* sp. Leaf306 y *Rhizobium* sp. Leaf321; CDS en la hebra sentido coloreados de acuerdo a su categoría COG, CDS en la hebra sentido, CDS en la hebra anti-sentido, CDS en la hebra anti-sentido coloreados de acuerdo a su categoría COG, tRNA y rRNA, GC skew y contendido GC %.

Los resultados del ensamblado de las lecturas con el programa plasmidSpades mostraron que la cepa UYSO24 presenta un probable plásmido de 136,818 pb (Figura 3.6). El análisis del mismo mostró posee un probable sistema de replicación *repABC* (uyso24.8317, uyso24.8318, uyso24.8319).

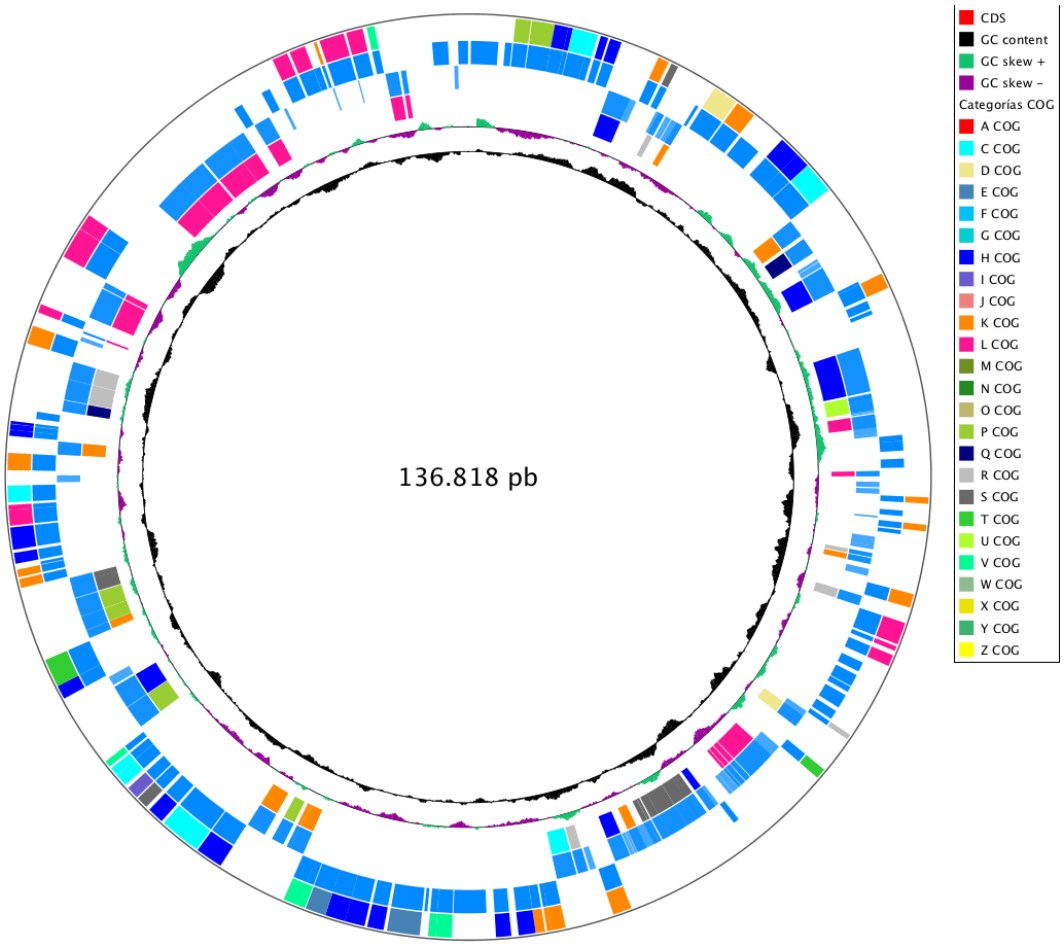

Figura 3.6: Mapa circular del plásmido de *Rhizobium* sp. UYSO24. Desde el exterior al centro: CDS en la hebra sentido coloreados de acuerdo a su categoría COG, CDS en la hebra sentido, CDS en la hebra anti-sentido, CDS en la hebra anti-sentido coloreados de acuerdo a su categoría COG, GC skew y contendido GC %.

Asimismo, los resultados del análisis de los elementos móviles en el genoma de la cepa UYSO24 mostraron que el mismo presenta 45 secuencias de inserción, 31 islas genómicas (Figura 3.7). El mismo análisis se realizó para los genomas de las cepas *Neorhizobium* HAMBI 540 y *Neorhizobium* HAMBI 1141, identificándose 11 islas genómicas en ambos (datos no mostrados).

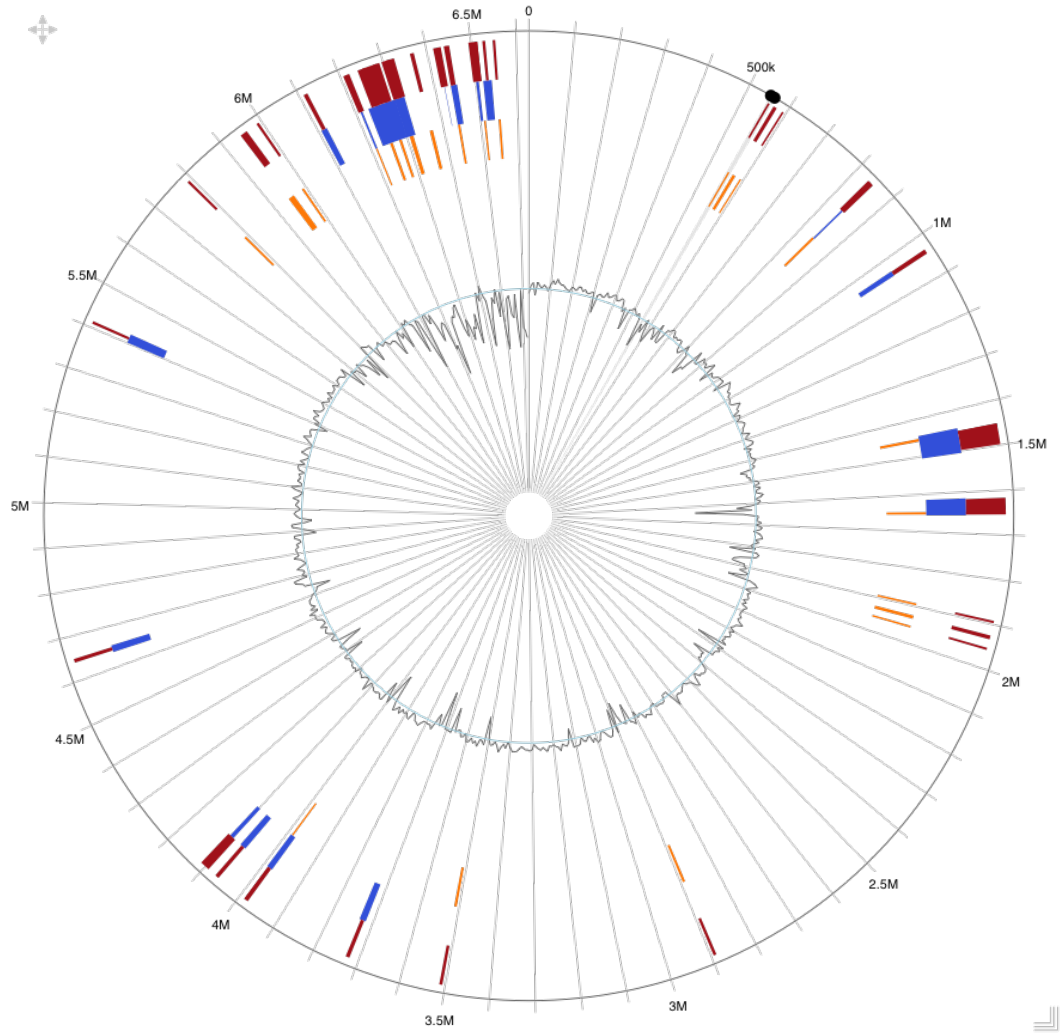

Figura 3.7: Mapa del genoma de la cepa *Rhizobium* sp. UYSO24 mostrando las islas genómicas. De afuera hacia adentro: marcador de posición, islas genómicas detectadas por: al menos un método (rojo), IslandPath-DIMOB (azul), y SIGI-HMM (amarillo); contenido GC.

3.2.3. Principales características genómicas relacionadas a la interacción planta-bacteria

A continuación se detallan las principales características identificadas en los genomas de las cepas *Kosakonia radicincitans* UYSO10 y *Rhizobium* sp. UYSO24, relacionadas con la interacción planta-microorganismo y la promoción del crecimiento vegetal.

3.2.3.1. Sistemas de secreción

Los resultados de los análisis mostraron que la cepa UYSO10 presenta probables genes codificantes para los sistemas de secreción del tipo I, IV, V y VI (Tabla 3.5). Por su lado, la cepa UYSO24 presenta probables genes para los sistemas II, IV y V (Tabla 3.6).

Tabla 3.5: Genes pertenecientes a sistemas de secreción presentes en el genoma de la cepa *Kosakonia radicincitans* UYSO10.

Tipo de sistema de secreción	Número del gen	Nombre del gen
SSTI	uyso10.226	*lapC*
	uyso10.2263	*lapB*
	uyso10.2264	*lapE*
	uyso10.513	*raxB*
	uyso10.514	*raxA*
	uyso10.3257	*raxB*
	uyso10.3258	*raxA*
SSTIV	uyso10.1312	*traL*
	uyso10.1313	*traE*
	uyso10.1314	*traK*
	uyso10.1315	*traB*
	uyso10.1320	*traV*
	uyso10.1322	*virb4*
	uyso10.1324	*traW*
	uyso10.1330	*traU*
	uyso10.1331	*trbC*
	uyso10.1332	*traN*
	uyso10.1333	*traF*
	uyso10.1338	*traH*

Continuación de la tabla 3.5 de la página anterior.

Tipo de sistema de secreción	Número del gen	Nombre del gen
	uyso10.1339	*traG*
	uyso10.1343	*t4cp2*
SSTV	uyso10.2029	-
	uyso10.316	*misL*
	uyso10.1196	-
	uyso10.2433	-
SSTVI	uyso10.4823	*clpB*
	uyso10.4824	
	uyso10.3698	*icmF*
	uyso10.3633	*impA*
	uyso10.3696	
	uyso10.4382	
	uyso10.3622	*impB*
	uyso10.3695	
	uyso10.4398	
	uyso10.3623	*impC*
	uyso10.3694	
	uyso10.4397	
	uyso10.3690	*impD*
	uyso10.4393	
	uyso10.3685	*impE*
	uyso10.3584	*impF*
	uyso10.3626	*impG*
	uyso10.3627	
	uyso10.3682	
	uyso10.3683	
	uyso10.4381	
	uyso10.3628	*impH*
	uyso10.3629	
	uyso10.3681	
	uyso10.4380	
	uyso10.3688	*impI*
	uyso10.3700	*impJ*
	uyso10.3701	

Continuación de la tabla 3.5 de la página anterior.

Tipo de sistema de secreción	Número del gen	Nombre del gen
	uyso10.4396	
	uyso10.3697	*impM*
	uyso10.3702	*vasD*
	uyso10.298	
	uyso10.3630	
	uyso10.3677	
	uyso10.415	*vgrG*
	uyso10.416	
	uyso10.4389	
	uyso10.5617	

Tabla 3.6: Genes pertenecientes a sistemas de secreción presentes en el genoma de la cepa *Rhizobium* sp. UYSO24.

Tipo de sistema de secreción	Número del gen	Nombre del gen
SSTII	uyso24.7366	*flp*
	uyso24.7369	*rcpA*
	uyso24.7372	*tadA*
	uyso24.7373	*tadB*
	uyso24.7376	*tadC*
	uyso24.7367	*tadV*
	uyso24.7371	*tadZ*
SSTIV	uyso24.3566	*mobA*
	uyso24.3563	*t4cp1*
	uyso24.3562	*t4cp2*
	uyso24.3663	*virB1*
	uyso24.3676	*virB10*
	uyso24.3607	*virB11*
	uyso24.3606	*virB2*
	uyso24.3664	*virB2*
	uyso24.3605	*virB3*
	uyso24.3665	*virB3*
	uyso24.3600	*virB5*
	uyso24.3668	*virB5*
	uyso24.3599	*virB6*
	uyso24.3670	*virB6*
	uyso24.3598	*virB8*
	uyso24.3672	*virB8*
	uyso24.3597	*virB9*
SSTV	uyso24.6383	*sadA*

3.2.3.2. Quimioreceptores, transducción de señales ambientales, motilidad y formación de biopelículas

En las tablas 3.7 e 3.8 se presentan los probables genes codificantes para los quimiorreceptores del tipo *Methyl accepting chemotaxis protein* (MCPs), identificados en los genomas de las cepas UYSO10 y UYSO24 respectivamente.

Tabla 3.7: Genes pertenecientes a MCPs presentes en el genoma de la cepa *Kosakonia radicincitans* UYSO10

Número del gen	Función predicha
uyso10.865, uyso10.943 uyso10.1030, uyso10.1050 uyso10.1630, uyso10.2027 uyso10.2072, uyso10.2115 uyso10.2431, uyso10.2756 uyso10.3086, uyso10.3099 uyso10.3281, uyso10.3611 uyso10.4126	*Methyl-accepting chemotaxis sensor/transducer protein*
uyso10.821	*Methyl-accepting chemotaxis citrate transducer*
uyso10.1012, uyso10.1013 uyso10.3514, uyso10.4427 uyso10.5134, uyso10.5221	*Methyl-accepting chemotaxis protein*
uyso10.1147, uyso10.1279	*Methyl-accepting chemotaxis protein I (serine chemoreceptor protein)*
uyso10.3815	*Methyl-accepting chemotaxis protein II (aspartate chemoreceptor protein)*
uyso10.3374	*Methyl-accepting chemotaxis protein III (ribose and galactose chemoreceptor protein))*
uyso10.3814	*Methyl-accepting chemotaxis protein IV (dipeptide chemoreceptor protein)*

Tabla 3.8: Genes pertenecientes a MCPs presentes en el genoma de la cepa *Rhizobium* sp. UYSO24

Número del gen	Función predicha
uyso24.885, uyso24.1365, uyso24.1366, uyso24.1729, uyso24.2957, uyso24.3318, uyso24.3625, uyso24.3638, uyso24.3639, uyso24.3702, uyso24.3703, uyso24.4057, uyso24.4954, uyso24.4990, uyso24.4991, uyso24.5045, uyso24.5169, uyso24.5338, uyso24.5391, uyso24.6309, uyso24.6310, uyso24.6372, uyso24.6578, uyso24.6700, uyso24.6701, uyso24.6966, uyso24.8265	*Methyl-accepting chemotaxis protein*
uyso24.3534, uyso24.5091, uyso24.6579, uyso24.7273	*Methyl-accepting chemotaxis protein I (serine chemoreceptor protein)*
uyso24.5099, uyso24.4572, uyso24.5092, uyso24.5093, uyso24.5773, uyso24.6035, uyso24.8196, uyso24.8198, uyso24.3317	*Methyl-accepting chemotaxis sensor/transducer protein*

Con respecto a la transducción de señales, en este trabajo se identificaron en el genoma de la cepa UYSO10 (Tabla 3.10) y en el de la cepa UYSO24 (Tabla 3.11), 20 y 13 sistemas de dos componentes (SDC), respectivamente; así como 6 SDC en común: *CheAB*, *EnvZ-OmpR*, *KdpDE*, *QseCB*, *PhoRB* y *FixJL* (Tabla 3.9).

Tabla 3.9: Sistemas de dos componentes encontrados en los genomas de las cepas *Kosakonia radicincitans* UYSO10 y *Rhiozobium* sp. UYSO24.

Complejo	Número de gen	Descripción
CheAB	uyso10.1086 y uyso.1081 uyso24.1360. y uyso24.1354	Quimiotaxis (163)
EnvZ-OmpR	uyso10.2981 y uyso10.2980 uyso24.4099 y uyso24.4841	Respuesta al estrés osmótico (158)
KdpED	uyso10.4618 y uyso10.4617 uyso24.1245 y uyso24.1241	Transporte de potasio (173)
QseBC	uyso10.2015 y uyso10.5157 uyso24.7849 y uyso24.5327	Quorum sensing (156)
PhoBR	uyso10.1786 y uyso10.1787 uyso24.1450 y uyso24.1459	Homeostasis de fosfato (162)
FixLJ	uyso10.3191 y uyso10.3191 uyso24.8231 y uyso24.4272	Relacionado a la FBN y el sensado de la concentración de oxigeno (180)

Tabla 3.10: Sistemas de dos componentes encontrados en el genoma de la cepa *Kosakonia radicincitans* UYSO10.

Complejo	Número de gen	Descripción
ArcAB	uyso10.1376 y uyso10.5337	Control anoxigénico redox (54)
BaeSR	uyso10.4112 y uyso10.4113	Resistencia a múltiples drogas bacterianas (16)
BasSR	uyso10.4541 y uyso10.4542	Sensor de Hierro y Zinc (63; 112)
BarAY	uyso10.1590 y uyso10.980	Metabolismo central de carbono (33)
CitAB	uyso10.424 y uyso10.423	Regulación de la fermentación de nitrato (144)
CpxAR	uyso10.55 y uyso10.56	Detección de cobre (181)
DcuSR	uyso10.3665 y uyso10.3666	Transporte de C4-dicarboxilato (56)
GlrKR	uyso10.4791 y uyso10.4789	Metabolismo de amino azucares (128)
NarXL	uyso10.3509 y uyso10.3510	Respiración de nitrato y nitrito (127)
PhoQP	uyso10.2524 y uyso10.2525	Adaptación a concentraciones bajas de Mg_{2+} (59)
RstBA	uyso10.2950 y uyso10.2949	Relacionado a la capacidad de formar biopelículas (92)
RcsBD	uyso10.4268 y uyso10.4267	Regulación de los genes de síntesis de los polisacáridos de cápsula (100)
NtrBC	uyso10.13 y uyso10.12	Relacionado a la FBN y el sensado de la concentración de amonio (130)
TctED	uyso10.424 y uyso10.423	Transporte de ácido tricarboxílico (175)

Tabla 3.11: Sistemas de dos componentes encontrados en el genoma de la cepa *Rhiozobium* sp. UYSO24.

Complejo	Número de gen	Descripción
ChvGI	uyso24.434. y uyso24.436	Sensor de acidez del medio (34)
DctBD	uyso24.7856 y uyso24.7857	Transporte del C4-dicarboxilato (71)
CckA-CtrA/CpdR	uyso24.6652 y uyso24.5639	Regulación del ciclo celular (72)
GlnGL	uyso24.1000 y uyso24.994	Regulación del metabolismo de nitrógeno (160)
NtrYX	uyso24.998 y uyso24.999	Regulación del metabolismo de nitrógeno (69)
PleCD	uyso24.2871 y uyso24.2461	Regulación Ciclo celular (119)
RegAB	uyso24.5530 y uyso24.5528	Involucrado en los procesos redox de la célula, entre ellos la FBN (177)

Con respecto a la presencia de los probables genes codificantes para un aparato flagelar (involucrado en la movilidad del tipo *swimming*), en los genomas de las cepas en estudio, los resultados de los análisis permitieron identificar en el genoma de la cepa UYSO10 todos los genes necesarios para la síntesis y ensamblado de un aparato flagelar funcional (Figura 3.8). Sin embargo, los resultados de los análisis mostraron que el genoma de la cepa UYSO24 presenta un aparto flagelar incompleto (Figura 3.9).

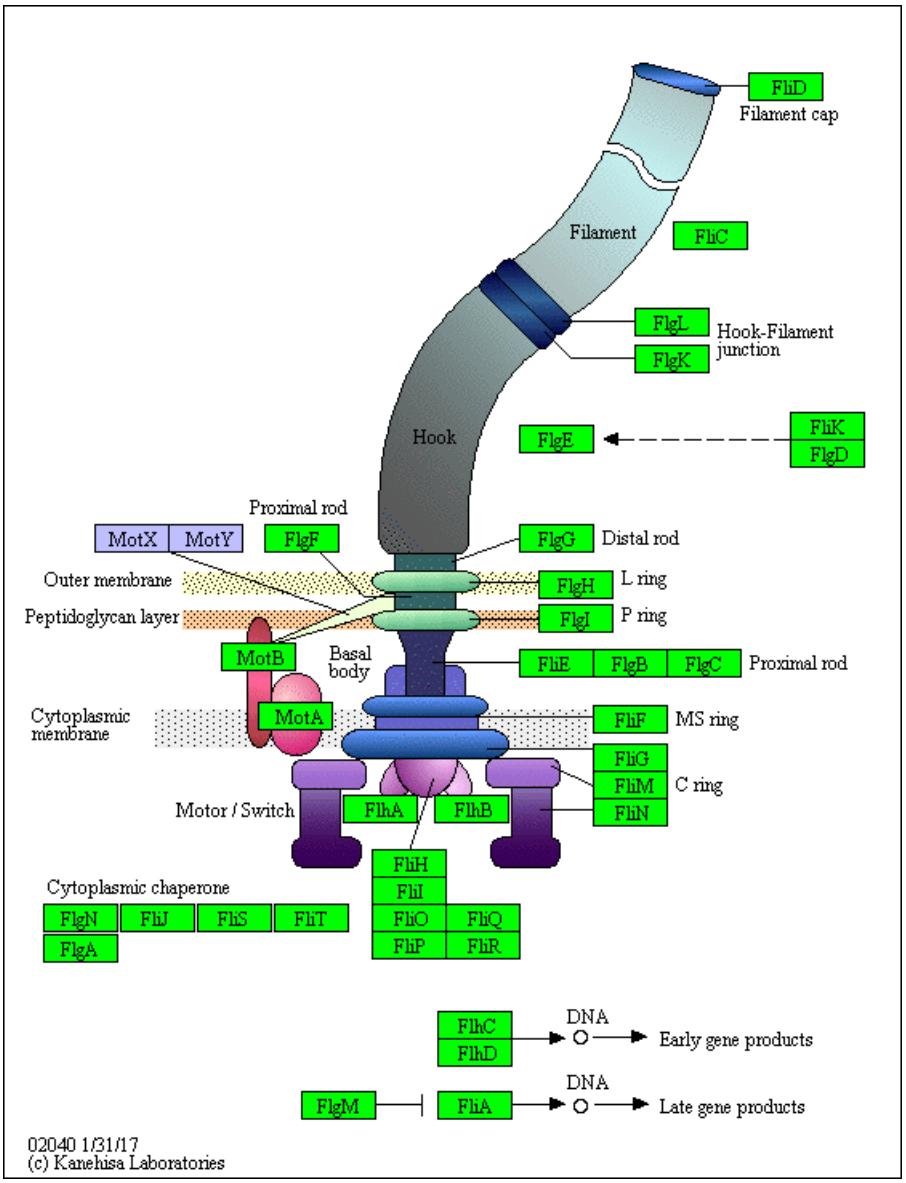

Figura 3.8: Modelo esquemático del flagelo externo bacteriano y los genes involucrados en el ensamblado del mismo. Rectángulos verdes y lilas: genes presentes o ausentes en el genoma de la cepa de *Koskonia* sp. UYSO10 respectivamente. Imagen reconstruida utilizando la anotación realizada con el servicio web KAAS y el servicio KEGG empleando el genoma de la cepa de *Koskonia* sp. UYSO10

.

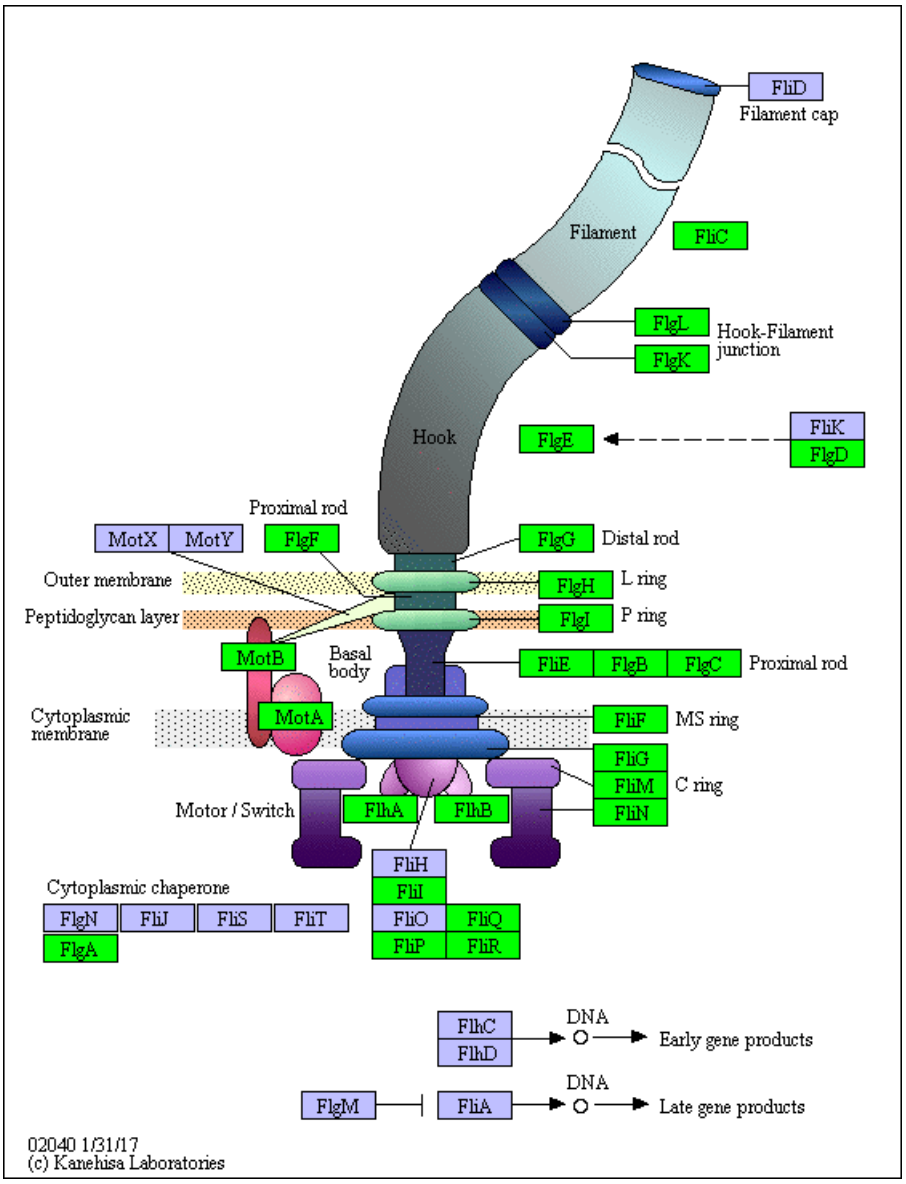

Figura 3.9: Modelo esquemático del flagelo bacteriano y los genes involucrados en el ensamblado del mismo. Rectángulos verdes y lilas: genes presentes o ausentes en el genoma de la cepa *Rhizobium* sp. UYSO24 respectivamente. Imagen reconstruida utilizando la anotación del servicio web KAAS y el servicio web KEGG para el genoma *Rhizobium* sp. UYSO24.

.

Por otra parte, los análisis mostraron que el genoma de la cepa UYSO10 presenta los probables genes para la biosíntesis del *pillus* tipo IV responsable del movimiento del tipo *twitching* (Tabla 3.12). Sin embargo, no se identificaron en el genoma de la cepa UYSO24 genes relacionados con el *pillus* tipo IV.

Tabla 3.12: Genes pertenecientes al *pillus* tipo IV presentes en el genoma de la cepa *Kosakonia radicincitans* UYSO10

Número del gen	Nombre del gen	Función predicha
uyso10.1473	*pilA*	*Type IV pilin PilA*
uyso10.1472	*pilB*	*Type IV fimbrial assembly, ATPase PilB*
uyso10.1471	*pilC*	*Type IV fimbrial assembly protein PilC*
uyso10.1820, uyso10.5430	*pilD*	*Leader peptidase (Prepilin peptidase) (EC 3.4.23.43)*
uysp10.447	*pilT*	*Twitching motility protein PilT*
uyso10.5482	*aroB*	*3-dehydroquinate synthase (EC 4.2.3.4)*
uyso10.5484	*pilQ*	*Type IV pilus biogenesis protein PilQ*
uyso10.5488	*pilM*	*Type IV pilus biogenesis protein PilM*
uyso10.5487	*pilN*	*Type IV pilus biogenesis protein PilN*
uyso10.5486	*pilO*	*Type IV pilus biogenesis protein PilO*
uyso10.5485	*pilP*	*Type IV pilus biogenesis protein PilP*
uyso10.1510, uyso10.5489	*mrcB*	*Multlimodular transpeptidase-transglycosylase (EC 2.4.1.129) (EC 3.4.-.-)*

A partir del estudio de las anotaciones del genoma de la cepa UYSO24 se identificó el operón *exo*, responsable de la producción de exopolisacáridos. Asimismo, se realizó la comparación del operón *exo* de la cepa UYSO24 con respecto a el operón ortólogo de las cepas *Neorhizobium galegae* HAMBI540 y *Sinorhizobium meliloti* 1021 (Figura 3.10).

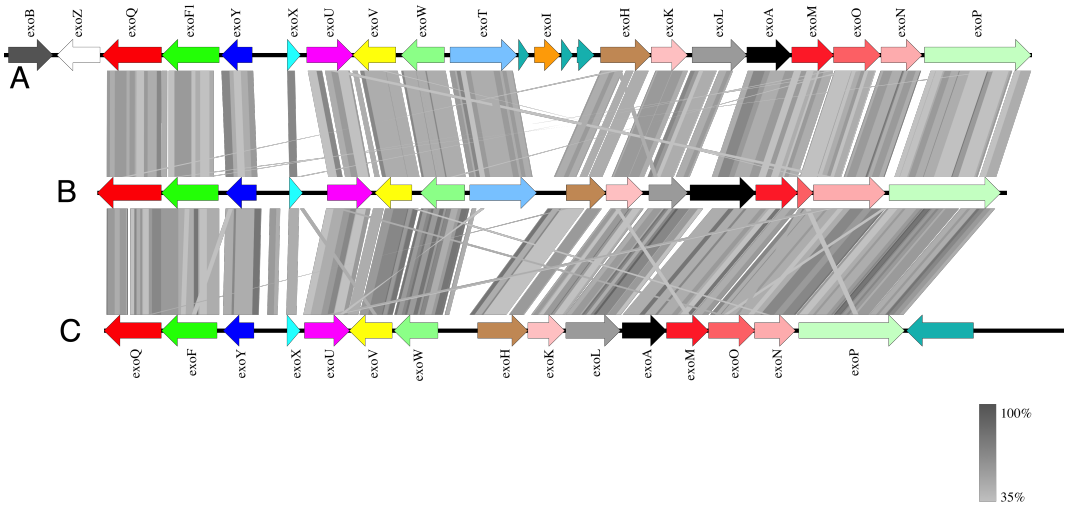

Figura 3.10: Comparación del operón *exo* entre las cepas (A) *Sinorhizobium meliloti* 1021, (B) *Rhizobium* sp. UYSO24 y (C) *Neorhizobium galegae* HAMBI540. La barra muestra el grado de identidad aminoácídica entre los genes.

El operón *exo* en la cepa UYSO24 posee una organización genómica similar al presente en el genoma de la cepa *S. meliloti* 1021. Sin embargo, en la cepa UYSO24 no se encuentran presentes los genes *exoZ* y *exoI*, mientras que el gen *exoB* se encuentra distante del resto de los genes del grupo. Por otra parte, en la cepa *N. galegae* HAMBI540 la organización del operón *exo* es similar al presente en el genoma de la cepa UYSO24. No obstante, la cepa HAMBI540 no presenta el gen *exoT*, el cual si está presente en la cepa UYSO24. Es de destacar, que al igual que en la cepa HAMBI540 el probable gen *exoB* (uyso24.779) se encuentra distante del resto de los genes del operón en la cepa UYSO24. Finalmente, es importante resaltar que la cepa UYSO10 no presenta los genes que codifican para el operón *exo*.

3.2.3.3. Principales características genómicas relacionadas a la promoción del crecimiento vegetal

Previamente las cepas UYSO10 y UYSO24 fueron reportadas como PCV de plantas de caña de azúcar (166). En este trabajo se analizaron algunas características PCV presentes en los genomas de las cepas en estudio incluyendo la producción y modulación de fitohormonas, producción de sideróforos y FBN.

Producción y modulación de fitohormonas

Estudios previos demostraron que las cepas en estudio tienen la capacidad de producir ácido indol acético (AIA) *in vitro* de forma dependiente de triptófano (167). Un gen marcador clave de esta vía es el que codifica la enzima indol–3–piruvuato decarboxilasa (*ipdC*). En ese sentido, los resultados del análisis del genoma de la cepa UYSO10 mostraron la ausencia del mencionado gen. Sin embargo, utilizando el programa BLAST con un *e-value* de 0.001 y el gen de *ipdC* de la cepa *Azospirillum brasilense* SP7 como referencia, se identificó el gen uyso10.4406 como un posible gen *ipdC* en el genoma de la cepa UYSO10. Por otro lado, y utilizando la misma estrategia, no fue posible identificar el gen *ipdC* en el genoma de la cepa UYSO24.

Con respecto a la presencia en los genomas en estudio del gen *acdS* codificante para la ACC desaminasa, enzima responsable de modular los niveles de la fitohormona etileno en la planta (55). El resultado de los análisis mostraron la presencia de un posible gen *acdS* (uyso24.7096) en en el genoma de la cepa UYSO24 y no en el genoma de la cepa UYSO10.

Sistemas de adquisición de hierro

Los resultados de los análisis mostraron que el genoma de la cepa UYSO10 presenta probables genes para la biosíntesis de sideróforos del tipo *enterobactina* (Tabla 3.13) y *vibrioferrina* (Tabla 3.14). Sin embargo, ambos genomas presentan los genes necesarios para la internalización de sideróforos del tipo hidroxamato, incluyendo los genes que codifican para un posible operón *fhuABCD*, así como los genes para el complejo TonB (10 y 4 para el genoma de la cepa UYSO10 y UYSO24 respectivamente).

Tabla 3.13: Genes involucrados en la síntesis y transporte del sideróforo *enterobactina* presentes en el genoma de la cepa *Kosakonia radicincitans* UYSO10

Número del gen	Nombre del gen	Función predicha
uyso10.2123	*entC*	*Isochorismate synthase (EC 5.4.4.2) [enterobactin] siderophore*
uyso10.2134	*entB*	*Isochorismatase (EC 3.3.2.1) [enterobactin] siderophore*
uyso10.2134	*entB2*	*Apo-aryl carrier domain of EntB*
uyso10.2135	*entA*	*2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase (EC 1.3.1.28) [enterobactin] siderophore*
uyso10.2133	*entE*	*2,3-dihydroxybenzoate-AMP ligase (EC 2.7.7.58) [enterobactin] siderophore*
uyso10.2120	*entD*	*4'-phosphopantetheinyl transferase (EC 2.7.8.-) [enterobactin] siderophore*
uyso10.2125	*entF*	*Enterobactin synthetase component F, serine activating enzyme (EC 2.7.7.-)*
uyso10.2136	*entH*	*Proofreading thioesterase in enterobactin biosynthesis EntH*
uyso10.2124	*ybdZ*	*Putative cytoplasmic protein YbdZ in enterobactin biosynthesis operon*
uyso10.2129, uyso10.2130	*entS*	*Enterobactin exporter EntS*
uyso10.2123	*fes*	*Enterobactin esterase*
uyso10.2131	*fepB*	*Ferric enterobactin-binding periplasmic protein FepB (TC 3.A.1.14.2)*
uyso10.2126	*fepC*	*Ferric enterobactin transport ATP-binding protein FepC (TC 3.A.1.14.2)*
uyso10.2128	*fepD*	*Ferric enterobactin transport system permease protein FepD (TC 3.A.1.14.2)*
uyso10.2127	*fepG*	*Ferric enterobactin transport system permease protein FepG (TC 3.A.1.14.2)*
uyso10.441	*fepE*	*Ferric enterobactin uptake protein FepE*
uyso10.2041, uyso10.2121, uyso10.2122	*fepA*	*Outer membrane receptor for ferric enterobactin and colicins B, D*

Tabla 3.14: Genes involucrados en la síntesis y transporte del sideróforo *vibrioferrina* presentes en el genoma de la cepa *Kosakonia radicincitans* UYSO10

Número del gen	Nombre del gen	Función predicha
uyso10.1178	*pvsA*	*Vibrioferrin ligase/carboxylase protein PvsA*
uyso10.1177	*pvsB*	*Vibrioferrin amide bond forming protein PvsB*
uyso10.1176	*pvsC*	*Vibrioferrin membrane-spanning transport protein PvsC*
uyso10.1175	*pvsD*	*Vibrioferrin amide bond forming protein PvsD*
uyso10.1174	*pvsE*	*Vibrioferrin decarboxylase protein PvsE*
uyso10.1181	*pvuA*	*Vibrioferrin receptor PvuA*

Por último, cabe destacar que la cepa UYSO10 presenta el gen uyso10.4625 el cual tiene una alta homología con el gen *fur*, el cual es regulador del metabolismo central de hierro a nivel celular. No se encontraron homólogos al gen *fur* en el genoma de la cepa UYSO24.

Fijación biológica de nitrógeno

El análisis del genoma de la cepa UYSO24 mostró la ausencia de genes correspondientes a la enzima nitrogenasa. Por su parte, el genoma de la cepa UYSO10 presenta un conjunto de probables genes para dos tipos de nitrogenasas: la "clásica" FeMo-nitrogensa, así como la "alternativa" FeFe-nitrogenasa. La nitrogenasa del tipo FeMo es codificada por el regulón *nif* que está compuesto por los genes *nifHDK*. Mientras que la nitrogenasa alternativa del tipo FeFe esta codificada por el operón *anf*, el cual esta compuesto por los genes *anfHDHL*. A continuación se describen los resultados del estudio de la organización y composición de regulones con respecto a los presentes en cepas de referencia. En primer lugar, se comparó la organización e identidad de los genes que componen la nitrogenasa FeMo de la cepa UYSO10 con el regulón ortólogo en las cepas *Kosakonia radicincitans* DSM16656 (179) y *Gluconacetobacter diazotrophicus* PA15 (23) (Figura 3.11). Los resultados mostraron un alto grado de identidad, tanto a nivel de secuencia como en la organización de los genes, entre los regulones *nif* de la cepas UYSO10 y DSM16656. Particular-

mente el gen uyso.3199 (*nifH*) comparte un 100 % de identidad aminoacídica entre las cepas antes mencionadas, mientras que para los otros genes del replicón la identidad varia entre 99.2 y 99.8 %. Por otra parte, los resultados mostraron que el regulón *nif* de las cepas UYSO10 y PA15, está organizados de igual manera pero con un bajo grado de identidad aminoacídica. A su vez, el gen *nifL* no está presente en *G. diazhotrophicus* PA15.

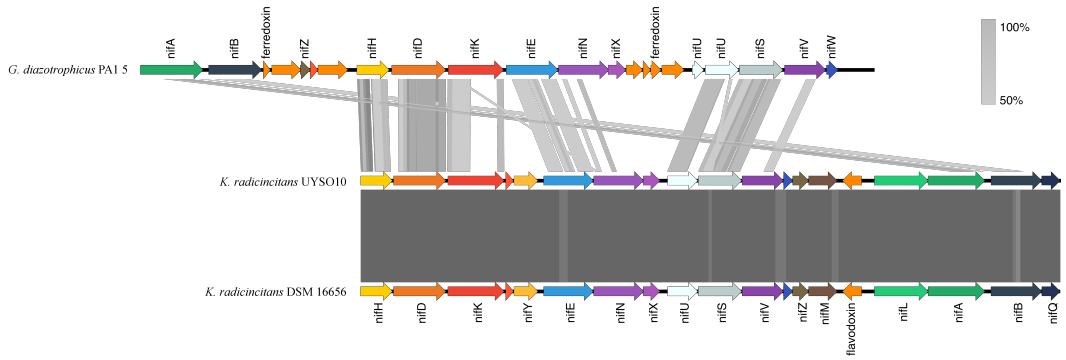

Figura 3.11: Comparación del regulón *nif* entre las cepas *Gluconacetobacter diazothrophicus* PAL 5, *Kosakonia* radicincitans. UYSO10 y *Kosakonia radicincitans* DSM 16656.La barra muestra el grado de identidad aminoácidica entre los genes.

En segundo lugar, se analizó el operón *anf* de la cepa UYSO10,con respecto a las cepas *Kosakonia radicincitans* DSM 16656 y *Azotobacter vinelandii* DJ (152) (Figura 3.12). Los resultados mostraron que la organización de los genes es la misma y que la identidad de los todos genes es mayor al 99 % entre la cepa UYSO10 y DSM16656. Por otro lado, al compararse las cepas UYSO10 y *A. vinelandii* DJ, se observó que los genes presentan la misma organización pero con un bajo grado de identidad aminoacídica (Figura 3.12).

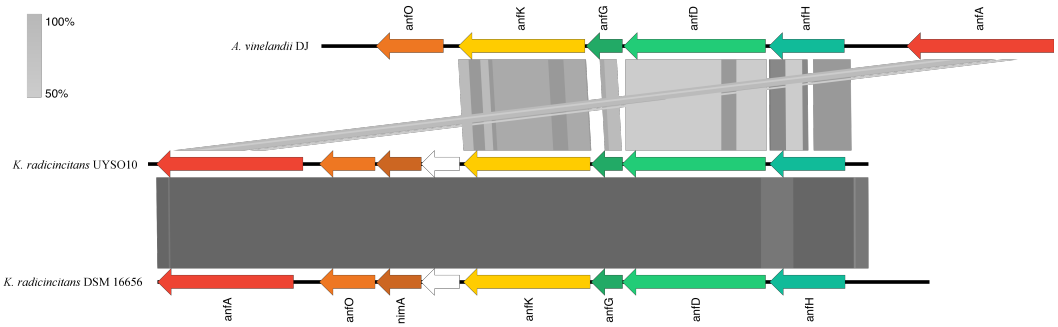

Figura 3.12: Comparación del operón *anf* entre las cepas *Azotobacter vinelandii* DJ, *Kosakonia radicincitans* UYSO10 y *Kosakonia radicincitans* DSM 16656. La barra muestra el grado de identidad a.a entre los genes.

3.3. Discusión

3.3.1. Análisis genómico de las cepas UYSO10 y UYSO24

Las cepas *Kosakonia radicincitans* UYSO10 y *Rhizobium* sp. UYSO24 fueron aisladas de los tejidos internos de variedades de caña de azúcar cultivadas en Uruguay (165), las mismas tienen un gran potencial biotecnológico ya que demostraron ser endófitos verdaderos y PCV de plantas de caña de azúcar en condiciones *in vitro* y controladas de invernáculo (167; 166). Los resultados presentados en este trabajo, se complementan con los resultados de las caracterizaciones fisiológicas, bioquímicas y proteómicas realizadas para las cepas UYSO10 y UYSO24 (166).

3.3.1.1. Análisis filogenético

Los resultados del análisis del gen uyso10.5346 de la cepa UYSO10 , el cual codifica para un probable gen *16S ARNr*, permitieron clasificarla dentro del género *Kosakonia*. Esto concuerda con los resultados de ANI, los cuales siguieren que la misma pertenece a la especie *Kosakonia radicincitans*, ampliamente reportada por su asociación a diversas especies vegetales (179; 107; 18; 21). Asimismo, la reconstrucción filogenética realizada en base a la secuencia completa del gen *16S ARNr* de la cepa *Kosakonia radicincitans* UYSO10 mostró que esta cepa, conforma un nodo monofilético con las distintas cepas de *Kosakonia radicincitans* aisladas de tejidos internos de vegetales al igual que la cepa UYSO10. Por otro lado, la secuencia del gen uyso24.4234 de la cepa UYSO24, el cual codifica para un probable gen *16S ARNr*, presenta una identidad menor al 97 %, valor mínimo aceptado para definir especies, con respecto a secuencias disponibles en bases de datos. A partir de la reconstrucción filogenética de la familia *Rhizobaceae* se observo que a cepa UYSO24 esta relacionada a un grupo de rizobios aislados de *Arabidopsis thaliana* (11) (*Rhizobium* sp. Leaf306, *Rhizobium* sp. Leaf321) y de *Panicum virgatum* (*Rhizobium* sp. NFR07, *Rhizobium* sp. NFR12). Asimismo, los resultados de los análisis ANIb, ANIm y TETRA de la cepa UYSO24 con respecto a las cepas *Rhizobium* sp. Leaf306, *Rhizobium* sp. Leaf321, *Rhizobium* sp. NFR07 y *Rhizobium* sp. NFR12 se encuentran por debajo del valor 95-96 % de ANI, el cual se considera un umbral robusto de similitud entre genomas (79). Este resultado siguiere que la cepa

UYSO24 pertenece al género *Rhizobium* pero no a la misma especie que las cepas antes mencionadas. El género *Rhizobium* ha sido ampliamente reportado en asociaciones benéficas planta-bacteria, principalmente a plantas de leguminosas (114). Sin embargo, en los últimos años se han reportado varias cepas no noduladoras pertenecientes a este género. Dichas cepas han sido aisladas de tejidos de diferentes plantas esterilizados en su superficie, y en algunos casos como PCV de plantas de arroz, sorgo, maíz, trigo y tomate (102; 62; 53). Teniendo en cuenta los resultados antes mencionados, se necesitan realizar nuevos experimentos y análisis complementarios para poder determinar la especie de esta cepa.

3.3.1.2. Estructura de los genomas de las cepas UYSO10 y UYSO24

Los resultados del ensamblaje y *scaffolding* de los datos de secuenciación siguieren que la cepa UYSO10 posee un único replicón. Esto difiere de la cepa de referencia *Kosakonia radicintans* DSM 16656, la cual está compuesta por tres replicones: un cromosoma y dos plásmidos (18). Teniendo en cuenta el alto grado de homología que presentan ambas cepas, este resultado siguiere que la cepa UYSO10 podría tener otros replicones, pero que con la estrategia de trabajo empelada, no fue posible identificarlos. En este sentido, se ha reportado que el programa plasmidSpades no es capaz de ensamblar todos los plásmidos en algunos conjuntos de datos (9). Por otro lado, los análisis realizados demostraron que el genoma de esta cepa posee un bajo número de elementos móviles, indicando un bajo grado de transferencia horizontal de genes. La misma observación fue reportada para la cepa *Azoarcus* sp. BH72 (82), proponiéndose que esa característica es debido a que la cepa esta adaptada a un ambiente estable, como los tejidos internos de las plantas. Por otra parte, los resultados del ensamblado del borrador del genoma de UYSO24 sugieren que el mismo está formado por dos replicones: un cromosoma y un plásmido. La secuencia del plásmido ensamblada con el programa plasmidSpades posee un sistema completo de replicación del tipo *repABC*, resultado que coincide la presencia del plásmido mencionado. Es interesante destacar que los genomas de las cepas *Neorhizobium gallegae*, relacionadas filogenéticamente a UYSO24, están formados por varios replicones (115). En los genomas de rizobios es común la presencia de elementos extra-cromosomales de gran tamaño en donde en muchos de los casos se encuentran los genes relacionados a la nodulación y

la FBN (96). Es necesario realizar más experimentos de secuenciado de la cepa UYSO24 para identificar los elementos extra-cromosomales y la estructura completa del genoma.

3.3.2. Los genomas de las cepas en estudio codifican para características probablemente involucradas en la interacción planta-bacteria

En estudios previos se reportó que las cepas en estudio son PCV de variedades de caña de azúcar en condiciones *in vitro* y de invernáculo, así como endófitos verdaderos de caña de azúcar (166). En este sentido y con la finalidad de profundizar en la caracterización de ambas cepas, en este trabajo, se analizaron los genomas buscando en primera instancia características genómicas posiblemente involucradas en el proceso de interacción con la planta hospedera figuras 3.13 y 3.14.

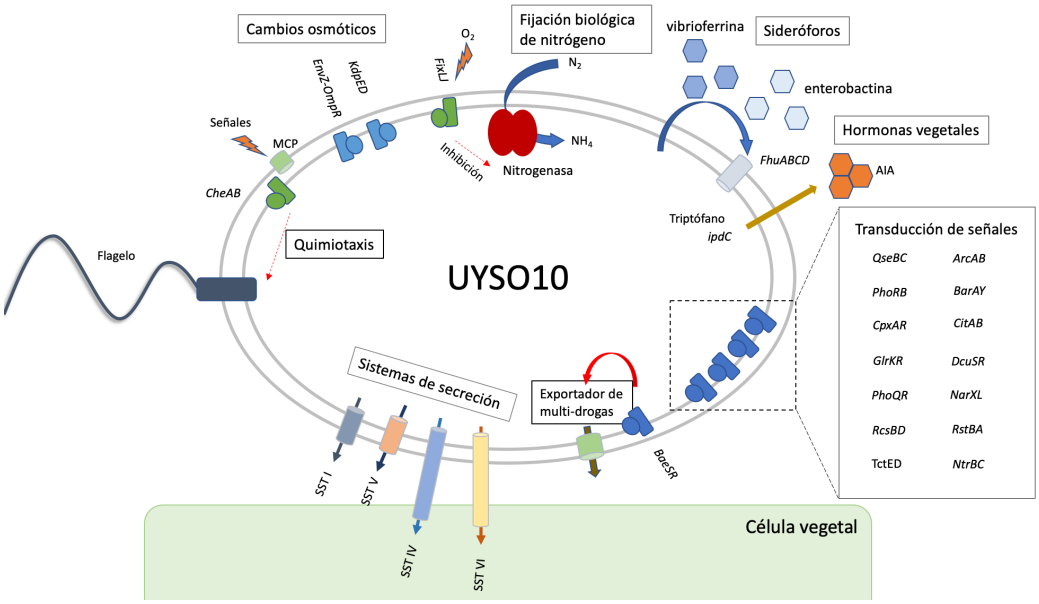

Figura 3.13: Esquema mostrando los principales mecanismos posiblemente involucrados a la interacción planta-bacteria presentes en el genoma de la cepa *Kosakonia radicincitans* UYSO10

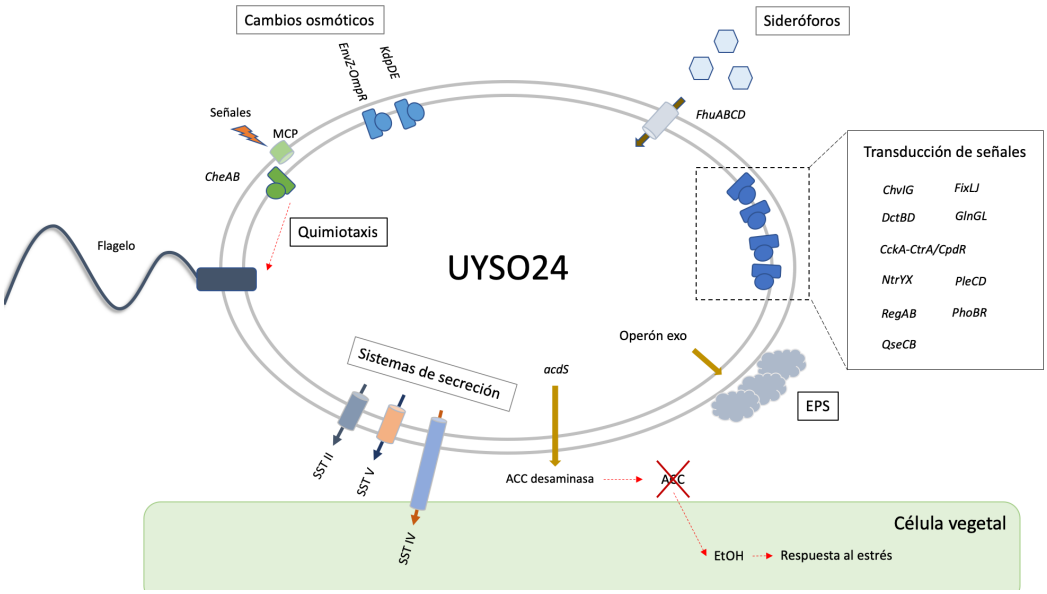

Figura 3.14: Esquema mostrando los principales mecanismos posiblemente involucrados a la interacción planta-bacteria presentes en el genoma de la *Rhizobium* sp. UYSO24

3.3.2.1. Respuesta de las bacterias al ambiente e inicio de la interacción planta-bacteria

La mayoría de los endófitos bacterianos cumplen parte de su ciclo de vida en el suelo en donde compiten con otros microorganismos por nichos y nutrientes, debiendo ser capaces de prevalecer en esas condiciones, con el fin de lograr una posterior colonización e infección efectiva de los tejidos internos de sus plantas huéspedes. En ese nicho ecológico, los procesos de censado y transducción de señales son fundamentales para poder adaptarse a los cambios ambientales. En este sentido, los cambios en los parámetros fisicoquímicos en bacteria son detectados mediante quimioreceptores conocidos como *Methyl-accepting chemotaxis protein* (MCPs) (172). Asimismo, uno de los mecanismos de transducción de señales más extendido en bacterias, es el basado en sistemas de dos componentes (SDC). Los SDC están formados por una kinasa sensor, que responde a una señal específica, la cual fosforila un regulador de la respuesta (172). Este tipo de sistemas juegan un papel importante en un amplio rango de mecanismos adaptativos tales como la virulencia, la quimiotaxis, el metabolismo y la motilidad (2). Teniendo en cuenta lo anterior, se buscaron e identificaron en los genomas de las cepas en estudio un amplio repertorio de genes codificantes para MCPs y SDC. El genoma de la cepa UYSO10 presenta

27 probables MCPs incluyendo receptores para: citrato, ribosa y galactosa; serina, aspartato, dipéptidos así como varios sin clasificar. Por su parte el genoma de la cepa UYSO24 presenta 40 probables MCPs incluyendo receptores para serina, polipéptidos así como varios sin clasificar. Estos resultados demuestran la capacidad de las cepas en estudio de poder censar y traducir un conjunto de estímulos diversos presentes en el ambiente, incluso aquellos que podría provenir de la planta y que están involucrados en los primeros pasos de la interacción. En este sentido, las plantas exudan a través de sus raíces una serie de compuestos de bajo peso molecular al suelo rizosférico, que actúan como quimioatrayentes para las bacterias rizosféricas (65; 36; 106). Dichos exudados inducen como respuesta una quimiotaxis dependiente de flagelo en las bacterias del suelo, las cuales se mueven como consecuencia hacia las raíces. En los genomas de las cepas UYSO10 y UYSO24 se identificó el SDC *CheAB*, el cual ha sido reportado como uno de los mecanismos centrales de la regulación de la quimiotaxis (163). Cepas mutantes de *Pseudomonas fluorescens* en el gen *cheA* pierden la capacidad quimiotáctica dependiente de flagelo, pero mantienen la capacidad motil. Esta pérdida de la capacidad quimiotáctica ocasiona que dichas cepas pierdan a su vez, la capacidad de colonizar las raíces de plantas de tomate remarcando su rol clave en la interacción planta-bacteria (40). Por otra parte, ambos genomas analizados presentaron los genes estructurales del aparato flagelar, coincidiendo con los previamente reportado (166). Sin embargo, el genoma de la cepa UYSO24 no presenta algunos de los genes del aparato flagelar. Teniendo en cuenta los antecedentes así como el conjunto de resultados obtenidos, se puede hipotetizar que las cepas en estudio podrían censar diversos compuestos presentes en los exudados radiculares y que como consecuencia se mueven hacia las raíces siendo la quimiotaxis mediada por flagelos uno de los mecanismo claves involucrados durante el proceso de colonización de las raíces de caña de azúcar por parte de las cepas UYSO10 y UYSO24. Asimismo, otro mecanismo reportado como involucrado en la supervivencia de las bacterias en la rizósfera, es el constituido por los transportadores multidrogas (25). En este sentido el análisis de los genomas de las cepas en estudio mostró la presencia de probables genes para dos posibles transportadores de multidrogas del tipo *AcrAB-TolC* y *MdtABC-TolC*. Este tipo de transportadores pueden exportar un gran rango de sustratos incluyendo antibióticos, metales pesados, compuestos vegetales como las auxinas, señales de *quorum sensing* y metabolitos bacterianos entre otros (25)). Particularmente se ha reportado en las

cepas diazótrofas PCV *Azospirillum lipoferum* 4B y *Hebaspirullum seropedicae* SmR1, la expresión de este tipo de transportadores durante la interacción con diferentes cultivares de arroz y plantas de maíz, respectivamente (43; 14). En este contexto es interesante hipotetizar sobre la posible participación de estos transportadores en la interacción de las cepas en estudio con la planta huésped, por lo que estos genes son muy interesantes como candidatos para ser mutados y evaluados sus fenotipos *in planta*.

3.3.3. Anclaje y colonización de los tejidos vegetales

Como se describió previamente, una vez que las bacterias se han aproximado a las raíces de las plantas, comienza el proceso de colonización de las diferentes zonas de la raíz. En dicho proceso participan diversas estructuras celulares tales como los flagelos, las fimbrias, los polisacáridos de superficie y el *pillus* de tipo IV. Estudios previos en las cepas *Kosakonia radicincitans* UYSO10 y *Rhizobium* sp. UYSO24 revelaron patrones de colonización de los tejidos vegetales característicos para las cepas en estudio (166). La cepa UYSO10 forma pequeñas biopelículas esféricas en la bifurcación entre la raíz principal y las raíces laterales, así como entre las raíces laterales y los pelos radiculares. Por su lado, la cepa UYSO24 forma una profusa biopelícula laminar sobre el tejido de las raíces principales y en la zona de pelos radiculares (166). Una biopelícula es un agregado de microorganismos en el cual las células están embebidas en una matriz de polímeros extracelulares, unida o adherida a una superficie (49; 39). Las biopelícula están formadas mayoritariamente por agua y un matriz de exopolisacáridos (EPS), fimbrias, proteínas de adherencia, ADN exógeno (eDNA) y *pillus*, así como diversos productos procedentes de la lisis de las bacterias (49). En este sentido, el fenotipo observado en la cepa UYSO24 llevó al estudio en profundidad de las características genómicas probablemente relacionadas a la producción de biopelículas. Como producto del análisis se identificó un probable conjunto de genes *exo* en el genoma de la cepa UYSO24. Este conjunto de genes presenta una alta homología aminoácidica y colinearidad con respecto a los genes *exo* presentes en el genoma de la cepa *Sinorhizobium meliloti* 1021, sugiriendo que ambos operones son ortólogos. En la cepa *S. meliloti* 1021, mutantes en los genes *exo* son deficientes en la producción de EPS y muestran una capacidad reducida de formar nódulos en *Medicago truncatula* (74). Asimismo, esta reportado para la cepa endófita

diazótrofa *Gluconacetobacter diazotrophicus* Pal5, que mutantes deficientes en la producción de EPS pierden la capacidad de colonizar plantas de arroz (105). Este fenotipo no-colonizador se revirtió al añadirse exógenamente EPS purificados al medio de cultivo, evidenciando claramente el rol de los EPS en la colonización. La regulación de la síntesis de EPS en *S. meliloti* 1021 está mediada por el SDC (*ChvGI*), el cual regula positivamente la transcripción del gen *exoS* (182). El genoma de la cepa UYSO24 presenta los genes uyso24.434 y uyso24.436 que codifican para un probable SDC del tipo *ChvGI* lo que sugiere que la regulación de la producción de EPS en esta cepa podría estar mediada por el producto de este gen. En su conjunto los resultados obtenidos siguieren que ambas cepas producirían EPS probablemente involucrados en la formación de biopelículas durante la interacción con la planta huésped, lo cual concuerda con lo previamente reportado (166). Teniendo en cuenta la bibliografía consultada y los resultados obtenidos se puede especular que los genes exo en la cepa UYSO24 podrían estar relacionados con la colonización de la superficie radicular reportada. Por último, cabe destacar que la cepa UYSO10 presenta probables genes que codificarían a un *pillus* tipo IV, así como un probable gen *pilT* (uyso10.447), asociado a la retracción de este. Esta estructura está reportada por su relación con la etapa de anclaje de las células bacterianas a la superficie de los tejidos vegetales. Cepas mutantes de *Azoarcus* sp. BH72 incapaces de sintetizar *pillus* del tipo IV son incapaces de formar colonias en la superficie de la raíz y no pueden colonizar tejidos internos de dichas plantas de arroz (41; 26). Los resultados obtenidos sugieren que el *pillus* tipo IV presente en el genoma de la cepa UYSO10, participaría en el proceso de colonización de plantas de caña de azúcar.

3.3.4. Mecanismos bacterianos de supervivencia

Siguiendo con el proceso de interacción planta-endófito, una vez que las bacterias colonizaron la superficie de los tejidos vegetales, éstas deben competir por este hábitat con otros organismos y superar la respuesta inmune del huésped con el fin de poder infectar los tejidos internos (64). En este proceso, los sistemas de secreción cumplen un rol importante (18). Particularmente los SSTVI cumplen un doble rol: 1- permiten a las bacterias mitigar la respuesta inmune de la planta y 2- permiten inyectar efectores tóxicos en las células bacterianas vecinas, compitiendo mejor por la colonización de los diferentes nichos

(22). Los SSTVI se asociaron inicialmente a patógenos pero en los últimos años se ha descripto que su presencia en bacterias endófitas es bastante extendida (151; 106; 18). Cepas mutantes de *Azoarcus* sp. BH72 para el SSTVI colonizan más agresivamente las raíces de plantas de arroz, sugiriéndose que la pérdida de este sistema mejora la colonización de las plantas (153). Por otro lado, el genoma de la cepa *Kosakonia radicincitans* DSM 166656 codifica para tres SSTVI, postulándose que esta característica le confiere una ventaja adaptativa en condiciones de competencia durante la colonización de las raíces (18). En los genomas en estudio solo fue posible detectar la presencia de probables genes codificantes para al menos un SSTVI en la cepa UYSO10. Teniendo en cuenta las características de colonización e infección por parte de esta cepa, en la cual se observa una discreta colonización de la superficie radicular pero una profusa colonización de los tejidos internos (166), es de sumo interés poder realizar experimentos conducentes a la caracterización de este tipo de sistemas en la interacción caña de azúcar-UYSO10. Asimismo, los genomas analizados presentaron también probables genes que codifican para un SSTIV. Los SSTIV están evolutivamente relacionados a los sistemas de conjugación de ADN y son capaces de secretar moléculas variadas, desde proteínas simples a complejos proteína-proteína o ADN-proteína, a través de las membranas internas y externas (97). Una vez que las bacterias endófitas logran infectar los tejidos internos de las plantas, las mismas deben adaptarse a las nuevas condiciones. Los tejidos internos de caña de azúcar presentan una elevada concentración de sacarosa, lo que ocasiona un estrés osmótico para las bacterias. Los genomas de las cepas en estudio presentan probables genes que codifican para los SDC del tipo *KdpDE* y *EnvZ-OmpR*, los cuales han sido descriptos por su relación a la osmotolerancia (12; 81). A su vez, en *Escherichia coli* el SDC *EnvZ-OmpR* sensa la osmolaridad del medio y regula la expresión de las porinas *OmpF* y *OmpC* (81); mientras que el SDC *KdpDE* regula el flujo de potasio en respuesta a la osmolaridad del medio (12). Estos antecedentes resaltan el probable rol de este tipo de sistemas en las bacterias en estudio.

3.3.5. Los genomas de las cepas en estudio codifican para mecanismos probablemente involucrados en la promoción del crecimiento vegetal

Como se mencionó, los mecanismos de PCV presentes en bacterias endófitas pueden ser directos o indirectos. Previamente se reportó para las cepas en estudio la presencia de diferentes mecanismos PCV *in vitro* incluyendo la solubilización de fosfato, la FBN, la producción de sideróforos así como la producción de hormonas vegetales. Teniendo en cuenta estos resultados previos, en este trabajo se buscaron en los genomas en estudio genes que codifiquen para estos probables mecanismos identificados *in vitro*.

3.3.5.1. Hormonas vegetales

Previamente se demostró que las cepas *Kosakonia radicincitans* UYSO10 y *Rhizobium* sp. UYSO24 tienen la capacidad de producir AIA de forma dependiente de triptófano (165). La producción de AIA dependiente de triptófano en bacterias de los géneros *Azosporillum* y *Enterobacter* se da mediante la vía metabólica del ácido indol–3–piruvico (*IPyA*) (142). Un gen clave de esta vía es el *ipdC* que codifica la enzima indol–3–piruvuato decarboxilasa. Teniendo en cuenta lo anterior, se buscó en los genomas de las cepas en estudio el gen *ipdC*, identificándose el mismo sólo en la cepa UYSO10. Este resultado siguiere que la cepa UYSO10 podría utilizar esta vía para la producción de AIA, lo cual coincide con los resultados obtenidos *in vitro* (166). Por otro lado, la ausencia del gen *ipdC* en la cepa UYSO24, sugiere que está cepa posee otra vía para la producción de AIA. Por otro lado, se analizo la presencia de los genes codificantes para la enzima ACC desaminasa en los genomas de las cepas en estudio. Esta enzima modula la concentración de la fitohormona etileno, producida por la planta en condiciones de estrés (55). En este sentido, el genoma de la cepa UYSO24 presenta el gen uyso24.7096 que es una probable ACC desaminasa (*acdS*), coincidiendo con los datos reportados de los estudios *in vitro* (166). Sin embargo, el genoma de la cepa UYSO10 no presenta el gen *acdS*, lo cual no concuerda con los resultados obtenidos en laboratorio en condiciones controladas. Este resultado sugiere que el gen se encuentra en una región no secuenciada o que el ensayo *in vitro* no logra reproducir las condiciones necesarias para determinar la actividad de la enzima.

3.3.5.2. Sistemas de adquisición de hierro de alta afinidad mediado por sideróforos

El hierro es un elemento esencial para los microorganismos, sin embargo es escaso en los suelos debido a su poca solubilidad. Las bacterias han desarrollado diferentes estrategias para adquirir este nutriente, dentro de los cuales se encuentran 1- el contacto directo entre la bacteria y el hierro del medio, 2- la síntesis y secreción de moléculas (sideróforos) con alta afinidad por este metal. La producción de sideróforos se ha propuesto como un mecanismo indirecto de PCV, donde las bacterias desplazan a los fitopatógenos al competir por este metal (65). Los sideróforos pueden unirse a diversas formas de hierro en el medio y el complejo formado es posteriormente reconocido por un receptor de membrana externa e internalizado por un sistema proteico del tipo *TonB-ExbB-ExbD*. Los genomas de las cepas UYSO10 y UYSO24 presentan, 10 y 4 probables genes que codifican para receptores de membrana *TonB* dependientes, respectivamente. Por su parte, a modo de comparación, el genoma de la cepa *Gluconacetobacter diazotrophicus* Pal5 codifica para 6 receptores TonB dependientes, la cepa *Herbaspirillum seropedicae* 17 y la cepa endófita-diazótrofa *Azoarcus* sp. BH72. También cabe destacar que ambas cepas en estudio presentan un conjunto de probables genes ortólogos al operón *fhuABCD*. Este operón fue descripto y caracterizado en *Escherichia coli* y es responsable de la captación e internalización de sideróforos del tipo ferricromo (7). Por último, la cepa UYSO10 presenta los genes necesarios para la síntesis de los sideróforos *vibrioferrina* y *enterobactina* concordando con lo previamente reportado a partir de ensayos *in vitro* (167). Sin embargo, el genoma de la cepa UYSO24 no presenta genes codificantes para la síntesis de sideróforos, lo cual no concuerda con los resultados reportados del análisis *in vitro* para esta cepa. Este resultado sugiere que la cepa UYSO24 tiene la capacidad de producir sideróforos pero los genes responsables no fueron secuenciados o anotados a partir de la metodología de este trabajo. En su conjunto los resultados sugieren que las cepas UYSO10 y UYSO24 poseen un gran potencial de competitividad en ambientes limitados de hierro, al poseer diferentes receptores de membrana externa posiblemente involucrados en la internalización del complejo fe-sideróforo lo cuales permitirían competir eficazmente con otras bacterias rizosféricas.

3.3.5.3. Fijación Biológica de Nitrógeno

La capacidad de fijar el N_2 atmosférico por bacterias es uno de los procesos biológicos de mayor interés biotecnológico en la actualidad. Bacterias del género *Rhizobium* han sido extensivamente estudiadas y utilizadas como inoculantes en cultivos de leguminosas debido a la elevada capacidad PCV que presentan, en gran medida gracias a la capacidad de FBN que poseen. Esta característica PCV a sido reportada en muchas bacterias asociadas a cultivos de interés agronómico. En caña de azúcar, por ejemplo, esta reportado que la FBN es uno de los mecanismos activos responsables de la PCV por parte de la cepa *Gluconacetobacter diazotrophicus* PAL5 (23). Las cepas UYSO10 y UYSO24 son capaces de realizar la FBN en *in vitro* y se postula que la FBN es uno de los mecanismos probablemente involucrados en la PCV observada (166). El genoma de la cepa *Kosakonia radicincitans* UYSO10 presenta genes para dos tipos de nitrogenasas distintas, la nitrogenasa clásica *FeMo* codificada por los genes *nifHDK* y la nitrogenasa alternativa *FeFe* codificada por los genes *anfHDGK*. Mientras que los genes *nifHDK* se encuentran en todas las bacterias diazótrofas, los genes *anfHDGK* solo se encuentran en un pequeño grupo de diazótrofos (46). Es interesante destacar que las cepas filogenéticamente relacionadas *K. radicincitans* DSM 16656 y UMEnt01, *Kosakonia* sp. NN145S y NN143E, *K. oryzae* Ola51 y YD4 poseen también ambos tipos de nitrogenasas (46; 91; 93); y que otras enterobacterias filogenéticamente cercanas como *Klebsiella* y *Enterobacter* no poseen la nitrogenasa alternativa FeFe (18). Está reportado que la FBN es un mecanismo activo de PCV en algunas cepas de del género *Kosakonia* (99; 184). Particularmente las cepas *Kosakonia* sp. NN145S y NN143E, son capaces de PCV de plantas de caña de azúcar y la FBN es uno de los mecanismos involucrados en el proceso (93). Asimismo, la cepa *Kosakonia* sp. R4-368, aislada de plantas de Jatropa, fue capaz de reducir acetileno y de PCV de su planta huésped (98). La cepa mencionada presenta únicamente el regulón *nif* y mutantes en los genes *nifH*, *nifD* y *nifK* ocasionaron la pérdida de la actividad de la nitrogenasa así como de la capacidad PCV por la cepa (99). Con respecto a la cepa UYSO10, se ha demostrado que las 2 nitrogenasas presentes son funcionales, y que la nitrogenasa *FeFe* no es regulada por la concentración de Mo o V (Taulé en revisión 2018). Esta característica es diferente a la cepa *Azotobacter vinelandii* donde concentraciones de 1μM o 1nM inhiben la expresión de las nitrogenasas alternativas

(73; 146), sugiriendo que la cepa UYSO10 utiliza sistemas de regulación de la FBN diferentes a *A. vinelandii*. En ensayos *in vitro* en los que se evaluaron mutantes de la cepa UYSO10 para los genes *nifH* y *anfH*, se ha demostrado que la FBN es uno de los mecanismos activos de PCV en plantas de caña de azúcar (Taulé 2018 revisión). Asimismo, se demostró que ambas nitrogenasas tienen el mismo efecto de PCV, lo cual contrasta con la cepa *Kosakonia radicincitans* DSM16656, donde mutantes en la nitrogenasa *FeMo* no son capaces de realizar la FBN (46). Este resultado es interesante ya que a pesar de la alta similitud reportada entre las nitrogenasas de ambas cepas, ($> 99\,\%$ a nivel aa), las mismas se comportan de forma distinta. Esta diferencia se podría explicar por la posible presencia de mutaciones puntuales en la nitrogenasa FeFe o por diferencias en las condiciones que se realizaron los experimentos. La FBN se regula a nivel transcripcional en respuesta a los niveles ambientales de oxígeno y amonio. Debido a que los componentes de la nitrogenasa son lábiles al oxígeno, es favorable para las bacterias reprimir su transcripción cuando los niveles de oxígeno son altos. También es favorable reprimir su expresión cuando el nivel celular de nitrógeno es suficientemente alto ya que la síntesis y actividad de la enzima es un proceso metabólico muy costoso. El genoma de la cepa UYSO10 codifica para dos SDC posiblemente encargados de la regulación de la transcripción de los genes *nif*, los SDC *NtrBC* y *FixJL*. El SDC *FixJL* regula la FBN en respuesta la concentración de oxígeno, controlando que la transcripción de los genes *nif* solo suceda en condiciones de microaerofilia (180). Por otro lado, está reportado que el SDC *NtrBC* activa la transcripción de los genes *nif* en condiciones de poca disponibilidad de amonio. Estos resultados muestran que la cepa *Kosakonia radicincitans* UYSO10 tiene potencial como organismo modelo para el estudio de la PCV mediado por la FBN.

Por otra pare, el análisis del genoma de cepa *Rhizobium* sp. UYSO24 mostró la ausencia de genes responsables de la síntesis de ambas nitrogenasas. Este resultado no coincide con los resultados obtenidos *in vitro* (166). Es importante mencionar que las cepas pertenecientes al nodo monofilogenético que pertenece UYSO24 no presentan los genes que codifican para la nitrogenasa. Este resultado siguiere que el análisis *in vitro* de reducción del acetileno que se realizó reporto un resultado falso positivo para la cepa UYSO24, hecho ya reportado para el método empleado (157). Otro aspecto interesante con respecto a la FBN en esta cepa, es que el genoma codifica para los probables genes *fixJL*. Teniendo en cuenta lo anterior, otra posibilidad es que los genes

que codifican para la nitrogenasa se encuentren en alguna de las regiones que no fueron secuenciadas.

Capítulo 4

Conclusiones y perspectivas

4.1. Conclusiones

Los resultados obtenidos en el transcurso de este trabajo de tesis coinciden con los reportados en la literatura sobre el impacto positivo de la corrección de errores de secuenciación sobre la calidad de los datos. En particular el IonDUDE es capaz de corregir de errores de secuenciación de datos obtenidos con el secuenciador Ion Torrent, pero su rendimiento es superado por herramientas disponibles en la literatura. De las herramientas ensayadas en este trabajo, el programa Fiona se destacó sobre el resto por su capacidad de corrección de errores. Es importante destacar que la evaluación de parámetros de calidad basados en el mapeo de las lecturas no debe tomarse como una medida definitiva de la capacidad de corrección de una herramienta. En el caso de genomas bacterianos, los resultados de la evaluación de la calidad de los ensamblados puede ser un indicador más relevante sobre la capacidad de corrección de una herramienta.

En vista de los resultados obtenidos con el programa Fiona, se decidió utilizar este programa para el procesamiento de los datos de las cepas *Kosakonia* sp. UYSO10 y *Rhizobium* sp. UYSO24. El análisis genómico de las cepas UYSO10 y UYSO24 permitió determinar que presentan genes reportados por su asociación con mecanismos de PCV. A su vez, algunos de los genes encontrados han sido reportados en otras cepas relacionadas de caña de azúcar. En particular los genes involucrados en la tolerancia al estrés osmótico, lo cual probablemente represente un mecánico de adaptación particular de este huésped. Es interesante destacar que contar con la secuencia genómica de las cepas en

estudio permitió la clasificación taxonómica de la cepa *Rhizobium* sp. UYSO24. Mientras que la cepa UYSO10 fue reclasificada dentro del género *Kosakonia*.

4.2. Perspectivas

En vista de los resultados obtenidos proponemos:

1 - Aplicar el modelo del DUDE para canales con memoria.

2 - Aplicar el IonDUDE a otras tecnologías de secuenciación de ADN.

3 - Con el objetivo de determinar la estructura completa de los genomas de las cepas en estudio se plantea secuenciar los genomas empleando tecnologías de secuenciado de tercera generación.

4 - Realizar estudios de genómica comparativa con otras cepas PCV, empleando la vasta disponibilidad de genomas secuenciados en bases de datos públicas. Se plantea particularmente para la cepa UYSO24 dada la afiliación filogenética de la cepa con un grupo de rizobios poco caracterizado.

Referencias bibliográficas

[1] 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., y Abecasis, G. R. A global reference for human genetic variation. *Nature 526*, 7571 (2015), 68–74.

[2] Abby, S. S., Cury, J., Guglielmini, J., Néron, B., Touchon, M., y Rocha, E. P. C. Identification of protein secretion systems in bacterial genomes. *Scientific Reports 6*, 1 (2016), 23080.

[3] Alic, A. S., Ruzafa, D., Dopazo, J., y Blanquer, I. Objective review of de novo stand-alone error correction methods for NGS data. *Wiley Interdisciplinary Reviews: Computational Molecular Science 6*, 2 (2016), 111–146.

[4] Allam, A., Kalnis, P., y Solovyev, V. Karect: Accurate correction of substitution, insertion and deletion errors for next-generation sequencing data. *Bioinformatics 31*, 21 (2015), 3421–3428.

[5] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., y Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology 215*, 3 (1990), 403–10.

[6] Andrews, S. FastQC: A quality control tool for high throughput sequence data, 2010.

[7] Andrews, S. C., Robinson, A. K., y Rodríguez-Quiñones, F. Bacterial iron homeostasis. *FEMS microbiology reviews 27*, 2-3 (2003), 215–37.

[8] Antipov, D., Hartwick, N., Shen, M., Raiko, M., Lapidus, A., y Pevzner, P. A. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics 32*, 22 (2016), btw493.

[9] Arredondo-Alonso, S., Willems, R. J., van Schaik, W., y Schürch, A. C. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microbial Genomics 3*, 10 (2017), 1–8.

[10] Aziz, R. K., Bartels, D., Best, A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., y Zagnitko, O. The RAST Server: Rapid annotations using subsystems technology. *BMC Genomics 9* (2008), 1–15.

[11] Bai, Y., Müller, D. B., Srinivas, G., Garrido-Oter, R., Potthoff, E., Rott, M., Dombrowski, N., Münch, P. C., Spaepen, S., Remus-Emsermann, M., Hüttel, B., McHardy, A. C., Vorholt, J. A., y Schulze-Lefert, P. Functional overlap of the Arabidopsis leaf and root microbiota. *Nature 528*, 7582 (2015), 364–369.

[12] Ballal, A., Basu, B., y Apte, S. K. The Kdp-ATPase system and its regulation. *Journal of Biosciences 32*, 3 (2007), 559–568.

[13] Bally, R., y Elmerich, C. Biocontrol of plant diseases by associative and endophytic nitrogen-fixing bbacteria. In *Associative and Endophytic Nitrogen-fixing Bacteria and Cyanobacterial Associations*, C. Elmerich and W. E. Newton, Eds., vol. 5 of *Nitrogen Fixation: Origins, Applications, and Research Progress*. Springer Netherlands, Dordrecht, 2007, pp. 171–190.

[14] Balsanelli, E., Tadra-Sfeir, M. Z., Faoro, H., Pankievicz, V. C., de Baura, V. A., Pedrosa, F. O., de Souza, E. M., Dixon, R., y Monteiro, R. A. Molecular adaptations of Herbaspirillum seropedicae during colonization of the maize rhizosphere. *Environmental microbiology 18*, 8 (2016), 2343–56.

[15] Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., y Pevzner, P. A. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology 19*, 5 (2012), 455–477.

[16] Baranova, N., y Nikaido, H. The BaeSR Two-Component Regulatory System Activates Transcription of the yegMNOB (mdtABCD) Transporter Gene Cluster in Escherichia coli and Increases Its Resistance to Novobiocin and Deoxycholate. *Journal of Bacteriology 184*, 15 (2002), 4168–4176.

[17] Barb, J. J., Oler, A. J., Kim, H.-S., Chalmers, N., Wallen, G. R., Cashion, A., Munson, P. J., y Ames, N. J. Development of an Analysis Pipeline Characterizing Multiple Hypervariable Regions of 16S rRNA Using Mock Samples. *PLoS ONE 11*, 2 (2016), e0148047.

[18] Becker, M., Patz, S., Becker, Y., Berger, B., Drungowski, M., Bunk, B., Overmann, J., Spröer, C., Reetz, J., Tchuisseu Tchakounte, G. V., y Ruppel, S. Comparative Genomics Reveal a Flagellar System, a Type VI Secretion System and Plant Growth-Promoting Gene Clusters Unique to the Endophytic Bacterium Kosakonia radicincitans. *Frontiers in Microbiology 9* (2018), 1–22.

[19] Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi,

C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E. Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., VandeVondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D.,

Durbin, R., y Smith, A. J. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature 456*, 7218 (2008), 53–59.

[20] Berendsen, R. L., Pieterse, C. M., y Bakker, P. A. The rhizosphere microbiome and plant health. *Trends in Plant Science 17*, 8 (2012), 478–486.

[21] Berger, B., Wiesner, M., Brock, A. K., Schreiner, M., y Ruppel, S. K. radicincitans, a beneficial bacteria that promotes radish growth under field conditions. *Agronomy for Sustainable Development 35*, 4 (2015), 1521–1528.

[22] Bernal, P., Llamas, M. A., y Filloux, A. Type VI secretion systems in plant-associated bacteria. *Environmental microbiology 20*, 1 (2018), 1–15.

[23] Bertalan, M., Albano, R., de Pádua, V., Rouws, L., Rojas, C., Hemerly, A., Teixeira, K., Schwab, S., Araujo, J., Oliveira, A., França, L., Magalhães, V., Alquéres, S., Cardoso, A., Almeida, W., Loureiro, M. M., Nogueira, E., Cidade, D., Oliveira, D., Simão, T., Macedo, J. J., Valadão, A., Dreschsel, M., Freitas, F., Vidal, M., Guedes, H., Rodrigues, E., Meneses, C., Brioso, P., Pozzer, L., Figueiredo, D., Montano, H., Junior, J., de Souza Filho, G., Martin Quintana Flores, V., Ferreira, B., Branco, A., Gonzalez, P., Guillobel, H., Lemos, M., Seibel, L., Alves-Ferreira, M., Sachetto-Martins, G., Coelho, A., Santos, E., Amaral, G., Neves, A., Pacheco, A. B., Carvalho, D., Lery, L., Bisch, P., Rössle, S. C., Urményi, T., Rael Pereira, A., Silva, R., Rondinelli, E., von Krüger, W., Martins, O., Baldani, J. I., y Ferreira, P. C. G. Complete genome sequence of the sugarcane nitrogen-fixing endophyte Gluconacetobacter diazotrophicus Pal5. *BMC genomics 10* (2009), 450.

[24] Bertelli, C., Laird, M. R., Williams, K. P., Lau, B. Y., Hoad, G., Winsor, G. L., y Brinkman, F. S. IslandViewer 4: Expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Research 45*, W1 (2017), W30–W35.

[25] Blanco, P., Hernando-Amado, S., Reales-Calderon, J., Corona, F., Lira, F., Alcalde-Rico, M., Bernardini, A., Sanchez, M., y Martinez, J. Bacterial Multidrug Efflux Pumps: Much More Than Antibiotic Resistance Determinants. *Microorganisms 4*, 1 (2016), 14.

[26] Böhm, M., Hurek, T., y Reinhold-Hurek, B. Twitching motility is essential for endophytic rice colonization by the N2-fixing endophyte Azoarcus sp. strain BH72. *Molecular plant-microbe interactions : MPMI 20*, 5 (2007), 526–33.

[27] Bosi, E., Donati, B., Galardini, M., Brunetti, S., Sagot, M.-F., Lió, P., Crescenzi, P., Fani, R., y Fondi, M. MeDuSa: a multi-draft based scaffolder. *Bioinformatics 31*, 15 (2015), 2443–2451.

[28] Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P., y Tyson, G. W. Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. *PLoS Computational Biology 9*, 4 (2013).

[29] Burdman, S., Dulguerova, G., Okon, Y., y Jurkevitch, E. Purification of the major outer membrane protein of Azospirillum brasilense, its affinity to plant roots, and its involvement in cell aggregation. *Molecular plant-microbe interactions : MPMI 14*, 4 (2001), 555–61.

[30] Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., Mcdonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., y Knight, R. QIIME allows analysis of high- throughput community sequencing data Intensity normalization improves color calling in SOLiD sequencing. *Nature Publishing Group 7*, 5 (2010), 335–336.

[31] Caporaso, J. G., Lauber, C. L., Walters, W. a., Berg-Lyons, D., Lozupone, C. a., Turnbaugh, P. J., Fierer, N., y Knight, R. Global patterns of 16S rRNA diversity at a depth of millions of

sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America 108 Suppl* (2011), 4516–22.

[32] Carvalho, T. L. G., Balsemao-Pires, E., Saraiva, R. M., Ferreira, P. C. G., y Hemerly, A. S. Nitrogen signalling in plant interactions with associative and endophytic diazotrophic bacteria. *Journal of Experimental Botany 65*, 19 (2014), 5631–5642.

[33] Chavez, R. G., Alvarez, A. F., Romeo, T., y Georgellis, D. The physiological stimulus for the BarA sensor kinase. *Journal of Bacteriology 192*, 7 (2010), 2009–2012.

[34] Chen, E. J., Fisher, R. F., Perovich, V. M., Sabio, E. A., y Long, S. R. Identification of direct transcriptional target genes of ExoS/ChvI two-component signaling in Sinorhizobium meliloti. *Journal of Bacteriology 191*, 22 (2009), 6833–6842.

[35] Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., y Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research 38*, 6 (2009), 1767–1771.

[36] Compant, S., Clément, C., y Sessitsch, A. Plant growth-promoting bacteria in the rhizo- and endosphere of plants: Their role, colonization, mechanisms involved and prospects for utilization. *Soil Biology and Biochemistry 42*, 5 (2010), 669–678.

[37] Compant, S., Duffy, B., Nowak, J., Clément, C., y Ait Barka, E. Use of Plant Growth-Promoting Bacteria for Biocontrol of Plant Diseases: Principles, Mechanisms of Action, and Future Prospects. *Applied and Environmental Microbiology 71*, 9 (2005), 4951–4959.

[38] Croucher, N. J., Finkelstein, J. A., Pelton, S. I., Mitchell, P. K., Lee, G. M., Parkhill, J., Bentley, S. D., Hanage, W. P., y Lipsitch, M. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nature Genetics 45*, 6 (2013), 656–663.

[39] Danhorn, T., y Fuqua, C. Biofilm formation by plant-associated bacteria. *Annual review of microbiology 61*, 1 (2007), 401–22.

[40] de Weert, S., Vermeiren, H., Mulders, I. H. M., Kuiper, I., Hendrickx, N., Bloemberg, G. V., Vanderleyden, J., De Mot, R., y Lugtenberg, B. J. J. Flagella-driven chemotaxis towards exudate components is an important trait for tomato root colonization by Pseudomonas fluorescens. *Molecular plant-microbe interactions : MPMI 15*, 11 (2002), 1173–80.

[41] Dörr, J., Hurek, T., y Reinhold-Hurek, B. Type IV pili are involved in plant-microbe and fungus-microbe interactions. *Molecular microbiology 30*, 1 (1998), 7–17.

[42] Downie, J. A. The roles of extracellular proteins, polysaccharides and signals in the interactions of rhizobia with legume roots. *FEMS Microbiology Reviews 34*, 2 (2010), 150–170.

[43] Drogue, B., Sanguin, H., Borland, S., Prigent-Combaret, C., y Wisniewski-Dyé, F. Genome wide profiling of Azospirillum lipoferum 4B gene expression during interaction with rice roots. *FEMS Microbiology Ecology 87*, 2 (2014), 543–555.

[44] Duca, D., Lorv, J., Patten, C. L., Rose, D., y Glick, B. R. Indole-3-acetic acid in plant–microbe interactions. *Antonie van Leeuwenhoek 106*, 1 (2014), 85–125.

[45] Edgar, R. C. Updating the 97 % identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics 34*, 14 (2018), 2371–2375.

[46] Ekandjo, L. K., Ruppel, S., Remus, R., Witzel, K., Patz, S., y Becker, Y. Site-directed mutagenesis to deactivate two nitrogenase isozymes of Kosakonia radicincitans DSM16656 T. *Canadian Journal of Microbiology 64*, 2 (2018), 97–106.

[47] Ewing, B., Hillier, L., Wendl, M. C., y Green, P. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Research 8*, 3 (1998), 175–185.

[48] Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B.,

Merrick, J., y Al., E. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science 269*, 5223 (1995), 496–512.

[49] Flemming, H.-C., y Wingender, J. The biofilm matrix. *Nature Reviews Microbiology 8*, 9 (2010), 623–633.

[50] Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, P.-C., y Lucier, T. S. The Minimal Gene Complement of Mycoplasma genitalium. *Science 270*, 5235 (1995), 397–404.

[51] Galibert, F. The Composite Genome of the Legume Symbiont Sinorhizobium meliloti. *Science 293*, 5530 (2001), 668–672.

[52] Gamalero, E., y Glick, B. R. Bacterial Modulation of Plant Ethylene Levels. *Plant Physiology 169*, 1 (2015), 13–22.

[53] García-Fraile, P., Menéndez, E., y Rivas, R. Role of bacterial biofertilizers in agriculture and forestry. *AIMS Bioengineering 2*, 3 (2015), 183–205.

[54] Georgellis, D., Kwon, O., y Lin, E. C. Quinones as the redox signal for the arc two-component system of bacteria. *Science 292*, 5525 (2001), 2314–6.

[55] Glick, B. R., Cheng, Z., Czarny, J., y Duan, J. Promotion of plant growth by ACC deaminase-producing soil bacteria. *New Perspectives and Approaches in Plant Growth-Promoting Rhizobacteria Research* (2007), 329–339.

[56] Golby, P., Davies, S., Kelly, D. J., Guest, J. R., y Andrews, S. C. Identification and Characterization of a Controlling Gene Expression in Response to C 4 -Dicarboxylates in Escherichia coli Identification

and Characterization of a Two-Component Sensor-Kinase and Response-Regulator System (DcuS-DcuR) Controlling Gene E. *J. Bacteriol. 181*, 4 (1999), 1238–1248.

[57] Goodwin, S., McPherson, J. D., y McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics 17*, 6 (2016), 333–351.

[58] Goris, J., Klappenbach, J. A., Vandamme, P., Coenye, T., Konstantinidis, K. T., Tiedje, J. M., Goris, J., Vandamme, P., Coenye, T., Konstantinidis, K. T., y Tiedje, J. M. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology 57*, 1 (2007), 81–91.

[59] Groisman, E. A. The Pleiotropic Two-Component Regulatory System PhoP-PhoQ MINIREVIEW The Pleiotropic Two-Component Regulatory System PhoP-PhoQ. *Journal of Bacteriology 183*, 6 (2001), 1835–1842.

[60] Guizelini, D., Saizaki, P. M., Coimbra, N. A. R., Weiss, V. A., Faoro, H., Sfeir, M. Z. T., Baura, V. A., Monteiro, R. A., Chubatsu, L. S., Souza, E. M., Cruz, L. M., Pedrosa, F. O., Raittz, R. T., Marchaukoski, J. N., y Steffens, M. B. R. Complete Genome Sequence of Herbaspirillum hiltneri N3 (DSM 17495), Isolated from Surface-Sterilized Wheat Roots. *Genome Announcements 3*, 5 (2015), 01288–15.

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

[62] Gutiérrez-Zamora, M. L., y Martínez-Romero, E. Natural endophytic association between Rhizobium etli and maize (Zea mays L.). *Journal of Biotechnology 91*, 2-3 (2001), 117–126.

[63] Hagiwara, D., Yamashino, T., y Mizuno, T. A Genome-wide view of the Escherichia coli BasS-BasR two-component system implicated in iron-responses. *Bioscience, biotechnology, and biochemistry 68*, 8 (2004), 1758–1767.

[64] Hardoim, P. R., van Overbeek, L. S., Berg, G., Pirttilä, A. M., Compant, S., Campisano, A., Döring, M., y Sessitsch, A. The Hidden World within Plants: Ecological and Evolutionary Considerations for Defining Functioning of Microbial Endophytes. *Microbiology and Molecular Biology Reviews 79*, 3 (2015), 293–320.

[65] Hardoim, P. R., van Overbeek, L. S., y Elsas, J. D. V. Properties of bacterial endophytes and their proposed role in plant growth. *Trends in microbiology 16*, 10 (2008), 463–71.

[66] Harshey, R. M. Bacterial motility on a surface: many ways to a common goal. *Annual review of microbiology 57*, 1 (2003), 249–73.

[67] Heydari, M., Miclotte, G., Demeester, P., Van de Peer, Y., y Fostier, J. Evaluation of the impact of Illumina error correction tools on de novo genome assembly. *BMC bioinformatics 18*, 1 (2017), 374.

[68] Hurek, T., y Reinhold-Hurek, B. Azoarcus sp. strain BH72 as a model for nitrogen-fixing grass endophytes. *Journal of Biotechnology 106* (2003), 169–178.

[69] Ishida, M. L., Assumpção, M. C., Machado, H. B., Benelli, E. M., Souza, E. M., y Pedrosa, F. O. Identification and characterization of the two-component NtrY/NtrX regulatory system in Azospirillum brasilense. *Brazilian Journal of Medical and Biological Research 35*, 6 (2002), 651–661.

[70] Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., y Akeson, M. Improved data analysis for the MinION nanopore sequencer. *Nature methods 12*, 4 (2015), 351–6.

[71] Janausch, I. G., Zientz, E., Tran, Q. H., Kröger, A., y Unden, G. C4-dicarboxylate carriers and sensors in bacteria. *Biochimica et biophysica acta 1553*, 1-2 (2002), 39–56.

[72] Jenal, U. The role of proteolysis in the Caulobacter crescentus cell cycle and development. *Research in Microbiology 160*, 9 (2009), 687–695.

[73] Joerger, R. D., Loveless, T. M., Pau, R. N., Mitchenall, L. A., Simon, B. H., y Bishop, P. E. Nucleotide sequences and mutational analysis of the structural genes for nitrogenase 2 of Azotobacter vinelandii. *Journal of bacteriology 172*, 6 (1990), 3400–8.

[74] Jones, K. M., Sharopova, N., Lohar, D. P., Zhang, J. Q., VandenBosch, K. A., y Walker, G. C. Differential response of the plant Medicago truncatula to its symbiont Sinorhizobium meliloti or an exopolysaccharide-deficient mutant. *Proceedings of the National Academy of Sciences 105*, 2 (2008), 704–709.

[75] Joshi, N. A., y Fass, J. N. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files, 2011.

[76] Jünemann, S., Prior, K., Szczepanowski, R., Harks, I., Ehmke, B., Goesmann, A., Stoye, J., y Harmsen, D. Bacterial Community Shift in Treated Periodontitis Patients Revealed by Ion Torrent 16S rRNA Gene Amplicon Sequencing. *PLoS ONE 7*, 8 (2012), e41606.

[77] Kandel, S., Joubert, P., y Doty, S. Bacterial Endophyte Colonization and Distribution within Plants. *Microorganisms 5*, 4 (2017), 77.

[78] Kim, D., Kim, W.-Y., Lee, S.-Y., Lee, S.-Y., Yun, H., Shin, S.-Y., Lee, J., Hong, Y., Won, Y., Kim, S.-J., Lee, Y. S., y Ahn, S.-M. Revising a Personal Genome by Comparing and Combining Data from Two Different Sequencing Platforms. *PLoS ONE 8*, 4 (2013), e60585.

[79] Kim, M., Oh, H.-S., Park, S.-C., y Chun, J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International journal of systematic and evolutionary microbiology 64*, 2 (2014), 346–51.

[80] Konstantinidis, K. T., y Tiedje, J. M. Towards a Genome-Based Taxonomy for Prokaryotes. *Journal of Bacteriology 187*, 18 (2005), 6258–6264.

[81] Krämer, R. Bacterial stimulus perception and signal transduction: Response to osmotic stress. *The Chemical Record 10*, 4 (2010), 217–229.

[82] Krause, A., Ramakumar, A., Bartels, D., Battistoni, F., Bekel, T., Boch, J., Friedrich, F., Hurek, T., Krause, L., Linke, B., Mchardy, A. C., Sarkar, A., Schneiker, S., Syed, A. A., Thauer, R., Reinhold-hurek, B., Kaiser, O., Goesmann, A., Weidner, S., Pu, A., Böhm, M., Friedrich, F., Hurek, T., Krause, L., Linke, B., Mchardy, A. C., Sarkar, A., Schneiker, S., Syed, A. A., Thauer, R., Vorhölter, F.-J., Weidner, S., Pühler, A., Reinhold-hurek, B., Kaiser, O., y Goesmann, A. Complete genome of the mutualistic, N2-fixing grass endophyte Azoarcus sp. strain BH72. *Nature biotechnology 24*, 11 (2006), 1385–91.

[83] Kumar, S., Stecher, G., y Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution 33*, 7 (2016), 1870–1874.

[84] Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., y Salzberg, S. L. Versatile and open software for comparing large genomes. *Genome biology 5*, 2 (2004), R12.

[85] Laehnemann, D., Borkhardt, A., y McHardy, A. C. Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. *Briefings in Bioinformatics 17*, 1 (2016), 154–179.

[86] Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., Poudel, S., y Ussery, D. W. Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics 15*, 2 (2015), 141–161.

[87] Lee, B., Moon, T., Yoon, S., y Weissman, T. DUDE-Seq: Fast, flexible, and robust denoising for targeted amplicon sequencing. *PLOS ONE 12*, 7 (2017), e0181463.

[88] Lemanceau, P., Bauer, P., Kraemer, S., y Briat, J. Iron dynamics in the rhizosphere as a case study for analyzing interactions between soils, plants and microbes. *Plant and Soil 321*, 1-2 (2009), 513–535.

[89] Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Bioinformatics 28*, 14 (2013), 1838–1844.

[90] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., y Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics 25*, 16 (2009), 2078–2079.

[91] Li, Y., Li, S., Chen, M., Peng, G., Tan, Z., y An, Q. Complete genome sequence of Kosakonia oryzae type strain Ola 51T. *Standards in Genomic Sciences 12*, 1 (2017), 8–11.

[92] Li, Y.-C., Chang, C.-k., Chang, C.-F., Cheng, Y.-H., Fang, P.-J., Yu, T., Chen, S.-C., Li, Y.-C., Hsiao, C.-D., y Huang, T.-h. Structural dynamics of the two-component response regulator RstA in recognition of promoter DNA element. *Nucleic Acids Research 42*, 13 (2014), 8777–8788.

[93] Lin, L., Li, Z., Hu, C., Zhang, X., Chang, S., Yang, L., Li, Y., y An, Q. Plant Growth-Promoting Nitrogen-Fixing Enterobacteria Are in Association with Sugarcane Plants Growing in Guangxi, China. *Microbes and environments 27*, 4 (2012), 391–398.

[94] Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., y Pallen, M. J. Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology 30*, 5 (2012), 434–439.

[95] Loman, N. J., y Pallen, M. J. Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology 13*, 12 (2015), 787–794.

[96] López-Guerrero, M. G., Ormeño-Orrillo, E., Acosta, J. L., Mendoza-Vargas, A., Rogel, M. A., Ramírez, M. A., Rosenblueth, M., Martínez-Romero, J., y Martínez-Romero, E. Rhizobial extrachromosomal replicon variability, stability and expression in natural niches. *Plasmid 68*, 3 (2012), 149–58.

[97] Low, H. H., Gubellini, F., Rivera-Calzada, A., Braun, N., Connery, S., Dujeancourt, A., Lu, F., Redzej, A., Fronzes, R., Orlova, E. V., y Waksman, G. Structure of a type IV secretion system. *Nature 508*, 7497 (2014), 550–553.

[98] Madhaiyan, M., Peng, N., y Ji, L. Complete Genome Sequence of Enterobacter sp. Strain R4-368, an Endophytic N-Fixing Gammaproteobacterium Isolated from Surface-Sterilized Roots of Jatropha curcas L. *Genome Announcements 1*, 4 (2013), 00544–13.

[99] Madhaiyan, M., Peng, N., Te, N., Hsin I, C., Lin, C., Lin, F., Reddy, C., Yan, H., y Ji, L. Improvement of plant growth and seed yield in Jatropha curcas by a novel nitrogen-fixing root associated Enterobacter species. *Biotechnology for Biofuels 6*, 1 (2013), 140.

[100] Majdalani, N., y Gottesman, S. The RCS Phosphorelay: A Complex Signal Transduction System. *Annual Review of Microbiology 59*, 1 (2005), 379–405.

[101] Malfanova, N., Lugtenberg, B. J. J., y Berg, G. Bacterial Endophytes: Who and Where, and What are they doing there? In *Molecular Microbial Ecology of the Rhizosphere*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2013, pp. 391–403.

[102] Mareque, C., Taulé, C., Beracochea, M., y Battistoni, F. Isolation, characterization and plant growth promotion effects of putative bacterial endophytes associated with sweet sorghum (Sorghum bicolor (L) Moench). *Annals of Microbiology 65*, 2 (2015), 1057–1067.

[103] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-j., Chen, Z., Dewell, B., Du, L., Fierro, J. M., Gomes, X. V., Goodwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Maria, L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-b., Knight, J. R., Lanza, R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Kenton, L., Lu, H., Makhijani, V. B., Mcdade, K. E., Mckenna, M. P., Myers, W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Greg, A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Richard, F., y Rothberg, J. M. Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. *Nature biotechnology 437*, 7057 (2006), 376–380.

[104] Marinier, E., Brown, D. G., y McConkey, B. J. Pollux: platform independent error correction of single and mixed genomes. *BMC bioinformatics 16*, 1 (2015), 10.

[105] Meneses, C. H. S. G., Rouws, L. F. M., Simoes-Araujo, J. L., Vidal, M. S., y Baldani, J. I. Exopolysaccharide production is required for biofilm formation and plant colonization by the nitrogen-fixing endophyte Gluconacetobacter diazotrophicus. *Molecular plant-microbe interactions : MPMI 24*, 12 (2011), 1448–58.

[106] Mitter, B., Petric, A., Shin, M. W., Chain, P. S. G., Hauberg-Lotte, L., Reinhold-Hurek, B., Nowak, J., y Sessitsch, A. Comparative genome analysis of Burkholderia phytofirmans PsJN reveals a wide spectrum of endophytic lifestyles based on interaction strategies with host plants. *Frontiers in plant science 4* (2013), 120.

[107] Mohd Suhaimi, N. S., Yap, K. P., Ajam, N., y Thong, K. L. Genome sequence of Kosakonia radicincitans UMEnt01/12, a bacterium associated with bacterial wilt diseased banana plant. *FEMS Microbiology Letters 358*, 1 (2014), 11–13.

[108] Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., y Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research 35*, Web Server issue (2007), 182–5.

[109] Myer, P. R., Kim, M., Freetly, H. C., y Smith, T. P. Evaluation of 16S rRNA amplicon sequencing using two next-generation sequencing technologies for phylogenetic analysis of the rumen bacterial community in steers. *Journal of Microbiological Methods 127* (2016), 132–140.

[110] Newton, W. E. Recent advances in understanding nitrogenases and How they work. In *Biological Nitrogen Fixation*, vol. 1-2. John Wiley & Sons, Inc, Hoboken, NJ, USA, 2015, pp. 5–20.

[111] Nikolenko, S. I., Korobeynikov, A. I., y Alekseyev, M. A. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics 14*, Suppl 1 (2013), S7.

[112] Ogasawara, H., Shinohara, S., Yamamoto, K., y Ishihama, A. Novel regulation targets of the metal-response BasS-BasR two-component system of Escherichia coli. *Microbiology 158* (2012), 1482–1492.

[113] Okonechnikov, K., Conesa, A., y García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics 32*, 2 (2015).

[114] Ormeño-Orrillo, E., Servín-Garcidueñas, L. E., Rogel, M. A., González, V., Peralta, H., Mora, J., Martínez-Romero, J., y Martínez-Romero, E. Taxonomy of rhizobia and agrobacteria from the Rhizobiaceae family in light of genomics. *Systematic and Applied Microbiology 38*, 4 (2015), 287–291.

[115] Österman, J., Marsh, J., Laine, P. K., Zeng, Z., Alatalo, E., Sullivan, J. T., Young, J. P. W., Thomas-Oates, J., Paulin, L., y Lindström, K. Genome sequencing of two Neorhizobium galegae strains reveals a noeT gene responsible for the unusual acetylation of the nodulation factors. *BMC Genomics 15*, 1 (2014), 500.

[116] Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., y Stevens, R. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic acids research 42*, Database issue (2014), 206–14.

[117] Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A. K., Chapman, J., Feltus, F. A., Gowik, U., Grigoriev, I. V., Lyons, E., Maher, C. a., Martis, M., Narechania, A., Otillar, R. P., Penning, B. W., Salamov, A. a., Wang, Y., Zhang, L., Carpita, N. C., Freeling, M., Gingle, A. R., Hash, C. T., Keller, B., Klein, P., Kresovich, S., McCann, M. C., Ming, R., Peterson, D. G., Mehboob-ur-Rahman, Ware, D., Westhoff, P., Mayer, K.

F. X., Messing, J., y Rokhsar, D. S. The Sorghum bicolor genome and the diversification of grasses. *Nature 457*, 7229 (2009), 551–556.

[118] Patten, C., y Glick, B. Bacterial biosynthesis of indole-3-acetic acid. *Canadian Journal of Micorbiology 42* (1996), 207–220.

[119] Paul, R., Jaeger, T., Abel, S., Wiederkehr, I., Folcher, M., Biondi, E. G., Laub, M. T., y Jenal, U. Allosteric Regulation of Histidine Kinases by Their Cognate Response Regulator Determines Cell Fate. *Cell 133*, 3 (2008), 452–461.

[120] Pevzner, P. A., Tang, H., y Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences 98*, 17 (2001), 9748–9753.

[121] Poza-Carrión, C., Echavarri-Erasun, C., y Rubio, L. M. Regulation of nif gene expression in Azotobacter vinelandii. In *Biological Nitrogen Fixation*, vol. 1-2. John Wiley & Sons, Inc, Hoboken, NJ, USA, 2015, pp. 99–108.

[122] Prat, A., y Perou, C. M. Mammary development meets cancer genomics. *Nature Medicine 15*, 8 (2009), 842–844.

[123] Prober, S. M., Leff, J. W., Bates, S. T., Borer, E. T., Firn, J., Harpole, W. S., Lind, E. M., Seabloom, E. W., Adler, P. B., Bakker, J. D., Cleland, E. E., DeCrappeo, N. M., DeLorenze, E., Hagenah, N., Hautier, Y., Hofmockel, K. S., Kirkman, K. P., Knops, J. M. H., La Pierre, K. J., MacDougall, A. S., McCulley, R. L., Mitchell, C. E., Risch, A. C., Schuetz, M., Stevens, C. J., Williams, R. J., y Fierer, N. Plant diversity predicts beta but not alpha diversity of soil microbes across grasslands worldwide. *Ecology letters 18*, 1 (2015), 85–95.

[124] Pruesse, E., Peplies, J., y Glöckner, F. O. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics 28*, 14 (2012), 1823–1829.

[125] Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Jorg, P., y Glockner, F. O. SILVA: a comprehensive online

resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic acids research 35*, 21 (2007), 7188–7196.

[126] Pylro, V. S., Roesch, L. F. W., Morais, D. K., Clark, I. M., Hirsch, P. R., y Tótola, M. R. Data analysis for 16S microbial profiling from different benchtop sequencing platforms. *Journal of microbiological methods 107* (2014), 30–7.

[127] Rabin, R. S., y Stewart, V. Dual response regulators (NarL and NarP) interact with dual sensors (NarX and NarQ) to control nitrate- and nitrite-regulated gene expression in Escherichia coli K-12. *Journal of bacteriology 175*, 11 (1993), 3259–68.

[128] Reichenbach, B., Göpel, Y., y Görke, B. Dual control by perfectly overlapping σ54- and σ70-promoters adjusts small RNA GlmY expression to different environmental signals. *Molecular Microbiology 74*, 5 (2009), 1054–1070.

[129] Reinhold-Hurek, B., y Hurek, T. Living inside plants: bacterial endophytes. *Current opinion in plant biology 14*, 4 (2011), 435–443.

[130] Reitzer, L. Nitrogen Assimilation and Global Regulation in Escherichia coli. *Annual Review of Microbiology 57*, 1 (2003), 155–176.

[131] Reuter, J. A., Spacek, D. V., y Snyder, M. P. Review High-Throughput Sequencing Technologies. *Molecular Cell 58*, 4 (2015), 586–597.

[132] Richter, M., y Rossello-Mora, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences 106*, 45 (2009), 19126–19131.

[133] Richter, M., Rosselló-Móra, R., Oliver Glöckner, F., y Peplies, J. JSpeciesWS: A web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics 32*, 6 (2015), 929–931.

[134] Rissman, A. I., Mau, B., Biehl, B. S., Darling, A. E., Glasner, J. D., y Perna, N. T. Reordering contigs of draft genomes using the Mauve Aligner. *Bioinformatics 25*, 16 (2009), 2071–2073.

[135] Rivera, D., Revale, S., Molina, R., Gualpa, J., Puente, M., Maroniche, G., Paris, G., Baker, D., Clavijo, B., McLay, K., Spaepen, S., Perticari, A., Vazquez, M., Wisniewski-Dye, F., Watkins, C., Martinez-Abarca, F., Vanderleyden, J., y Cassan, F. Complete Genome Sequence of the Model Rhizosphere Strain Azospirillum brasilense Az39, Successfully Applied in Agriculture. *Genome Announcements 2*, 4 (2014), 00683–14.

[136] Rodrigues, E. P., Rodrigues, L. S., de Oliveira, A. L. M., Baldani, V. L. D., Teixeira, K. R. D. S., Urquiaga, S., y Reis, V. M. Azospirillum amazonense inoculation: effects on growth, yield and N2 fixation of rice (Oryza sativa L.). *Plant and Soil 302*, 1-2 (2008), 249–261.

[137] Rosenblueth, M., y Martínez-Romero, E. Bacterial Endophytes and Their Interactions with Hosts. *Molecular Plant-Microbe Interactions 19*, 8 (2006), 827–837.

[138] Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., McKernan, K. J., Williams, A., Roth, G. T., y Bustillo, J. An integrated semiconductor device enabling non-optical genome sequencing. *Nature 475*, 7356 (2011), 348–352.

[139] Salmela, L. Correction of sequencing errors in a mixed set of reads. *Bioinformatics 26*, 10 (2010), 1284–1290.

[140] Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T. J., Schatz, M. C., Delcher, A. L., Roberts, M., Marçais, G., Pop, M., y Yorke, J. a. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome research 22*, 3 (2012), 557–67.

[141] Sanger, F., Nicklen, S., y Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences 74*, 12 (1977), 5463–5467.

[142] Sant'Anna, F. H., Almeida, L. G. P., Cecagno, R., Reolon, L. a., Siqueira, F. M., Machado, M. R. S., Vasconcelos, A. T. R., y Schrank, I. S. Genomic insights into the versatility of the plant growth-promoting bacterium Azospirillum amazonense. *BMC genomics 12* (2011), 409.

[143] Schalk, I. J., Hannauer, M., y Braud, A. New roles for bacterial siderophores in metal transport and tolerance. *Environmental microbiology 13*, 11 (2011), 2844–54.

[144] Scheu, P. D., Witan, J., Rauschmeier, M., Graf, S., Liao, Y. F., Ebert-Jung, A., Basché, T., Erker, W., y Unden, G. CitA/CitB two-component system regulating citrate fermentation in Escherichia coli and its relation to the DcuS/DcuR system In vivo. *Journal of Bacteriology 194*, 3 (2012), 636–645.

[145] Schloss, J. A. How to get genomes at one ten-thousandth the cost. *Nature Biotechnology 26*, 10 (2008), 1113–1115.

[146] Schneider, K., Müller, A., Schramm, U., y Klipp, W. Demonstration of a molybdenum- and vanadium-independent nitrogenase in a nifHDK-deletion mutant of Rhodobacter capsulatus. *European journal of biochemistry 195*, 3 (1991), 653–61.

[147] Schroder, J., Schroder, H., Puglisi, S. J., Sinha, R., y Schmidt, B. SHREC: a short-read error correction method. *Bioinformatics 25*, 17 (2009), 2157–2163.

[148] Schulz, M. H., Weese, D., Holtgrewe, M., Dimitrova, V., Niu, S., Reinert, K., y Richard, H. Fiona: A parallel and automatic strategy for read error correction. *Bioinformatics 30*, 17 (2014), i356–i363.

[149] Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics 30* (2014), 2068–2069.

[150] Segata, N., Börnigen, D., Morgan, X. C., y Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications 4* (2013).

[151] Sessitsch, A., Hardoim, P., Döring, J., Weilharter, A., Krause, A., Woyke, T., Mitter, B., Hauberg-Lotte, L., Friedrich, F., Rahalkar, M., Hurek, T., Sarkar, A., Bodrossy, L., van Overbeek, L., Brar, D., van Elsas, J. D., y Reinhold-Hurek, B. Functional characteristics of an endophyte community colonizing rice roots as revealed by metagenomic analysis. *Molecular plant-microbe interactions : MPMI 25*, 1 (2012), 28–36.

[152] Setubal, J. C., Moreira, L. M., y da Silva, A. C. R. Bacterial phytopathogens and genome science. *Current opinion in microbiology 8*, 5 (2005), 595–600.

[153] Shidore, T., Dinse, T., Öhrlein, J., Becker, A., y Reinhold-Hurek, B. Transcriptomic analysis of responses to exudates reveal genes required for rhizosphere competence of the endophyte Azoarcus sp. strain BH72. *Environmental Microbiology 14*, 10 (2012), 2775–2787.

[154] Song, L., Huang, W., Kang, J., Huang, Y., Ren, H., y Ding, K. Comparison of error correction algorithms for Ion Torrent PGM data: application to hepatitis B virus. *Scientific Reports 7*, 1 (2017), 8106.

[155] Spaepen, S. Plant Hormones Produced by Microbes. In *Principles of Plant-Microbe Interactions*, B. Lugtenberg, Ed. Springer International Publishing, Cham, 2015, pp. 247–256.

[156] Sperandio, V., Torres, A. G., y Kaper, J. B. Quorum sensing Escherichia coli regulators B and C (QseBC): a novel two-component regulatory system involved in the regulation of flagella and motility by quorum sensing in E. coli. *Molecular Microbiology 43*, 3 (2002), 809–821.

[157] Staal, M., Lintel-Hekkert, S. t., Harren, F., y Stal, L. Nitrogenase activity in cyanobacteria measured by the acetylene reduction assay: a comparison between batch incubation and on-line monitoring. *Environmental Microbiology 3*, 5 (2001), 343–351.

[158] Stock, A. M., Robinson, V. L., y Goudreau, P. N. Two-Component Signal Transduction. *Annual Review of Biochemistry 69*, 1 (2000), 183–215.

[159] Stolyar, S., Van Dien, S., Hillesland, K. L., Pinel, N., Lie, T. J., Leigh, J. A., y Stahl, D. A. Metabolic modeling of a mutualistic microbial community. *Molecular Systems Biology 3*, 1 (2007).

[160] Sukmarini, L., y Shimizu, K. Metabolic regulation of Escherichia coli and its glnG and zwf mutants under nitrogen limitation. *Biochemical Engineering Journal 48*, 2 (2010), 230–236.

[161] Sullivan, M. J., Petty, N. K., y Beatson, S. A. Easyfig: a genome comparison visualizer. *Bioinformatics 27*, 7 (2011), 1009–1010.

[162] Sultan, S. Z., Silva, A. J., y Benitez, J. A. The PhoB regulatory system modulates biofilm formation and stress response in El Tor biotype Vibrio cholerae. *FEMS Microbiology Letters 302*, 1 (2010), 22–31.

[163] Szurmant, H., Bunn, M. W., Cannistraro, V. J., y Ordal, G. W. Bacillus subtilis Hydrolyzes CheY-P at the Location of Its Action, the Flagellar Switch. *Journal of Biological Chemistry 278*, 49 (2003), 48611–48616.

[164] Tatusov, R. L. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research 28*, 1 (2000), 33–36.

[165] Taulé, C. *Bacterias promotoras del crecimiento vegetal asociadas a variedades de caña de azúcar en Uruguay: identificación, caracterización y estudios de interacción.* Tesis maestría, 2011.

[166] Taulé, C., Castillo, A., Villar, S., Olivares, F., y Battistoni, F. Endophytic colonization of sugarcane (Saccharum officinarum) by the novel diazotrophs Shinella sp. UYSO24 and Enterobacter sp. UYSO10. *Plant and Soil 403*, 1-2 (2016), 403–418.

[167] Taulé, C., Mareque, C., Barlocco, C., Hackembruch, F., Reis, V. M., Sicardi, M., y Battistoni, F. The contribution of nitrogen fixation to sugarcane (Saccharum officinarum L.), and the

identification and characterization of part of the associated diazotrophic bacterial community. *Plant and Soil 356*, 1-2 (2012), 35–49.

[168] UK10K Consortium, Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J. R. B., Xu, C., Futema, M., Lawson, D., Iotchkova, V., Schiffels, S., Hendricks, A. E., Danecek, P., Li, R., Floyd, J., Wain, L. V., Barroso, I., Humphries, S. E., Hurles, M. E., Zeggini, E., Barrett, J. C., Plagnol, V., Richards, J. B., Greenwood, C. M. T., Timpson, N. J., Durbin, R., y Soranzo, N. The UK10K project identifies rare variants in health and disease. *Nature 526*, 7571 (2015), 82–90.

[169] Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., y Thermes, C. Ten years of next-generation sequencing technology. *Trends in Genetics 30*, 9 (2014), 1–9.

[170] Vanbleu, E., y Vanderleyden, J. Molecular Genetics of Rhizosphere and Plant-Root Colonization. In *Associative and Endophytic Nitrogen-fixing Bacteria and Cyanobacterial Associations*, C. Elmerich and W. E. Newton, Eds. Springer Netherlands, Dordrecht, 2007, pp. 85–112.

[171] Varani, A. M., Siguier, P., Gourbeyre, E., Charneau, V., y Chandler, M. ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biology 12*, 3 (2011), R30.

[172] Wadhams, G. H., y Armitage, J. P. Making sense of it all: Bacterial chemotaxis. *Nature Reviews Molecular Cell Biology 5*, 12 (2004), 1024–1037.

[173] Walderhaug, M. O., Polarek, J. W., Voelkner, P., Daniel, J. M., Hesse, J. E., Altendorf, K., y Epstein, W. KdpD and KdpE, proteins that control expression of the kdpABC operon, are members of the two-component sensor-effector class of regulators. *Journal of Bacteriology 174*, 7 (1992), 2152–2159.

[174] Weissman, T., Ordentlich, E., Seroussi, G., Sergio, V., y Weinberger, M. J. Universal discrete denoising: Known channel. *IEEE Transactions on Information Theory 51*, 1 (2005), 5–28.

[175] Widenhorn, K. A., Somers, J. M., y Kay, W. W. Genetic regulation of the tricarboxylate transport operon (tctI) of Salmonella typhimurium. *Journal of bacteriology 171*, 8 (1989), 4436–4441.

[176] Wijaya, E., Frith, M. M. C., Suzuki, Y., y Horton, P. Recount: expectation maximization based error correction tool for next generation sequencing data. *Genome informatics. International Conference on Genome Informatics 23*, 1 (2009), 189–201.

[177] Willett, J. W., y Crosson, S. Atypical modes of bacterial histidine kinase signaling. *Molecular Microbiology 103*, 2 (2017), 197–202.

[178] Wisniewski-Dyé, F., Lozano, L., Acosta-Cruz, E., Borland, S., Drogue, B., Prigent-Combaret, C., Rouy, Z., Barbe, V., Herrera, A. M., González, V., y Mavingui, P. Genome Sequence of Azospirillum brasilense CBG497 and Comparative Analyses of Azospirillum Core and Accessory Genomes provide Insight into Niche Adaptation. *Genes 3*, 4 (2012), 576–602.

[179] Witzel, K., Gwinn-Giglio, M., Nadendla, S., Shefchek, K., y Ruppel, S. Genome sequence of Enterobacter radicincitans DSM16656T, a plant growth-promoting endophyte. *Journal of Bacteriology 194*, 19 (2012), 5469–5469.

[180] Wright, G. S. A., Saeki, A., Hikima, T., Nishizono, Y., Hisano, T., Kamaya, M., Nukina, K., Nishitani, H., Nakamura, H., Yamamoto, M., Antonyuk, S. V., Hasnain, S. S., Shiro, Y., y Sawai, H. Architecture of the complete oxygen-sensing FixL-FixJ two-component signal transduction system. *Science Signaling 11*, 525 (2018), eaaq0825.

[181] Yamamoto, K., y Ishihama, A. Transcriptional response of Escherichia coli to external copper. *Molecular microbiology 56*, 1 (2005), 215–27.

[182] Yao, S.-Y., Luo, L., Har, K. J., Becker, A., Rüberg, S., Yu, G.-Q., Zhu, J.-B., y Cheng, H.-P. Sinorhizobium meliloti ExoR and ExoS proteins regulate both succinoglycan and flagellum production. *Journal of bacteriology 186*, 18 (2004), 6042–9.

[183] Yeung, R. W. Discrete Memoryless Channels. In *Information Theory and Network Coding*. Springer US, Boston, MA, 2008, pp. 137–182.

[184] Zhu, B., Chen, M., Lin, L., Yang, L., Li, Y., y An, Q. Genome Sequence of Enterobacter sp. Strain SP1, an Endophytic Nitrogen-Fixing Bacterium Isolated from Sugarcane. *Journal of Bacteriology 194*, 24 (2012), 6963–6964.

ANEXOS

Anexo 1

Tablas

Tabla 1.1: Resultados de "BAM QC" para la muestra ERR039477 procesados con IonDUDE, IonDUDE M y IonDUDE Q

Programa	Contexto	Modo	LM (%)	TE	Mismatches	LMI (%)	LMD (%)	Indels HP (%)
Sin corregir	NA	NA	99.53	0.0076	173165	15.85	23.19	61.62
IonDUDE	k5	Subs	99.53	0.0076	173165	15.85	23.19	61.62
IonDUDE	k8	Subs	99.54	0.0076	173172	15.85	23.19	61.62
IonDUDE	k10	Subs	**99.58**	0.0076	173174	15.85	23.19	61.62
IonDUDE	h3	Indels	99.53	0.0076	177157	15.62	23.72	61.57
IonDUDE	h4	Indels	99.53	**0.0075**	174017	15.75	23.28	61.53
IonDUDE	h5	Indels	99.53	**0.0075**	173422	15.8	23.19	61.56
IonDUDE M	k5	Subs	99.53	0.0076	173441	15.85	23.19	61.62
IonDUDE M	k8	Subs	99.54	0.0076	173190	15.85	23.19	61.62
IonDUDE M	k10	Subs	**99.58**	0.0076	173185	15.85	23.19	61.62
IonDUDE M	h3	Indels	99.52	0.0077	181295	15.37	24.34	61.44
IonDUDE M	h4	Indels	99.53	**0.0075**	174028	15.58	23.28	**61.14**
IonDUDE M	h5	Indels	99.53	**0.0075**	**172711**	15.69	**23.08**	61.29
IonDUDE Q	k5	Subs	99.53	0.0076	174133	15.83	23.17	61.19
IonDUDE Q	k8	Subs	99.54	0.0076	173171	15.85	23.19	61.62
IonDUDE Q	k10	Subs	**99.58**	0.0076	173171	15.85	23.19	61.62
IonDUDE Q	h3	Indels	99.52	0.0077	186892	**14.77**	24.79	60.17
IonDUDE Q	h4	Indels	99.52	**0.0075**	177504	15.17	23.65	61.19
IonDUDE Q	h5	Indels	99.53	**0.0075**	173158	15.67	23.14	61.79

Contexto, k y h representa el largo de contexto utilizado por lo que por ejemplo $k10$ significa k de largo 10, $h3$ significa h de largo 3; Modo, *Subs* significa que se corrió el programa para corregir solamente errores de substituciones (en este modo solo se considera k); *Indels* significa que se corrió el para corregir solamente errores de indels en homopolímeros (en este modo solo se considera h); LM, Lecturas Mapeadas; TE, Tasa de error; LMI, Lecturas mapeadas con inserciones; LMD, Lecturas mapeadas con deleciones; Indels HP, indels en homopolímeros.

Tabla 1.2: Resultados de "BAM QC" para la muestra ERR161543 procesados con IonDUDE, IonDUDE M y IonDUDE

Programa	Contexto	Modo	LM (%)	TE	Mismatches	LMI (%)	LMD (%)	Indels HP (%)
Sin corregir	NA	NA	85.56	0.0217	1903447	43.03	34.44	65.6
IonDUDE	k5	Subs	85.56	0.0217	1907004	43.02	34.44	65.6
IonDUDE	k8	Subs	85.56	0.0217	1903396	43.03	34.44	65.6
IonDUDE	k10	Subs	85.58	0.0217	**1903224**	43.03	34.44	65.6
IonDUDE	h3	Indels	85.43	0.0218	1951604	42.6	34.8	65.26
IonDUDE	h4	Indels	85.47	0.0216	1933802	42.46	34.49	65.3
IonDUDE	h5	Indels	85.46	0.0216	1924077	42.41	34.29	65.36
IonDUDE M	k5	Subs	85.54	0.0219	1943475	42.99	34.43	65.6
IonDUDE M	k8	Subs	85.56	0.0217	1908877	43.02	34.44	65.59
IonDUDE M	k10	Subs	85.58	0.0217	1905163	43.03	34.44	65.59
IonDUDE M	h3	Indels	85.44	0.0218	1976574	42.29	35.09	64.92
IonDUDE M	h4	Indels	85.47	0.0215	1948823	41.83	34.37	64.66
IonDUDE M	h5	Indels	85.48	0.0214	1929190	41.77	33.96	64.79
IonDUDE Q	k5	Subs	85.55	0.0218	1915556	43	34.43	65.54
IonDUDE Q	k8	Subs	85.57	0.0217	1904690	43.02	34.44	65.59
IonDUDE Q	k10	Subs	85.58	0.0217	1903426	43.03	34.44	65.59
IonDUDE Q	h3	Indels	85.37	0.0227	2340473	39.29	41.89	**61.74**
IonDUDE Q	h4	Indels	85.53	0.0216	2150191	**38.87**	37.73	63.57
IonDUDE Q	h5	Indels	**85.65**	**0.0208**	1960110	39.32	**33.93**	65.79

Contexto, k y h representa el largo de contexto utilizado por lo que por ejemplo $k10$ significa k de largo 10, $h3$ significa h de largo 3; Modo, *Subs* significa que se corrió el programa para corregir solamente errores de substituciones (en este modo solo se considera k); *Indels* significa que se corrió el para corregir solamente errores de indels en homopolímeros (en este modo solo se considera h); LM, Lecturas Mapeadas; TE, Tasa de error; LMI, Lecturas mapeadas con inserciones; LMD, Lecturas mapeadas con deleciones; Indels HP, indels en homopolímeros.

Tabla 1.3: Resultados de "BAM QC" para la muestra ERR236069 procesados con IonDUDE, IonDUDE M y IonDUDE Q

Programa	Contexto	Modo	LM (%)	TE	Mismatches	LMI (%)	LMD (%)	Indels HP (%)
Sin corregir	NA	NA	84.13	0.0457	4748648	75.8	51.91	36.37
IonDUDE	k5	Subs	84.13	0.0457	4748677	75.8	51.91	36.37
IonDUDE	k8	Subs	84.13	0.0457	4748640	75.8	51.91	36.37
IonDUDE	k10	Subs	84.15	0.0457	4748577	75.8	51.91	36.37
IonDUDE	h3	Indels	84.15	0.0457	4765310	75.49	51.96	36.02
IonDUDE	h4	Indels	84.2	0.0455	4763130	75.15	51.8	35.58
IonDUDE	h5	Indels	84.2	0.0455	4752659	75.08	51.65	35.46
IonDUDE M	k5	Subs	84.12	0.0457	4750012	75.79	51.9	36.37
IonDUDE M	k8	Subs	84.13	0.0457	4748135	75.79	51.91	36.37
IonDUDE M	k10	Subs	84.15	0.0457	4748001	75.79	51.91	36.37
IonDUDE M	h3	Indels	84.21	0.0454	4786202	74.78	51.91	35.11
IonDUDE M	h4	Indels	84.31	0.0451	4777051	73.91	51.3	33.85
IonDUDE M	h5	Indels	84.33	0.0449	4747607	73.66	50.78	**33.35**
IonDUDE Q	k5	Subs	84.12	0.0457	4748467	75.78	51.89	36.35
IonDUDE Q	k8	Subs	84.13	0.0457	**4746188**	75.79	51.9	36.36
IonDUDE Q	k10	Subs	84.15	0.0457	4747223	75.79	51.91	36.37
IonDUDE Q	h3	Indels	84.12	0.0457	4957776	74	54.34	35.68
IonDUDE Q	h4	Indels	84.35	0.0448	4897660	72.31	51.99	35.75
IonDUDE Q	h5	Indels	**84.55**	**0.044**	4768443	**71.34**	**48.88**	36.33

Contexto, k y h representa el largo de contexto utilizado por lo que por ejemplo $k10$ significa k de largo 10, $h3$ significa h de largo 3; Modo, *Subs* significa que se corrió el programa para corregir solamente errores de substituciones (en este modo solo se considera k); $Indels$ significa que se corrió el para corregir solamente errores de indels en homopolímeros (en este modo solo se considera h); LM, Lecturas Mapeadas; TE, Tasa de error; LMI, Lecturas mapeadas con inserciones; LMD, Lecturas mapeadas con deleciones; Indels HP, indels en homopolímeros.

Tabla 1.4: Resultados de "BAM QC" para la muestra ERR039477

Programa	LM (%)	TE	Mismatches	LMI (%)	LMD (%)	Indels HP (%)
Sin corregir	99.53	0.0076	173165	15.85	23.19	61.62
Pollux	**99.73**	0.0037	89326	6.64	14.3	**57.19**
Karect	99.59	**0.0017**	**54715**	**1.34**	**9.97**	62.84
Ionhammer	99.53	0.0073	172348	14.42	21.86	61.75
Fiona	99.67	0.0018	55228	1.97	10.06	61.2
IonDUDE k5 Subs	99.53	0.0076	173165	15.85	23.19	61.62
IonDUDE h3 Indels	99.53	0.0076	177157	15.62	23.72	61.57
IonDUDE M k5 Subs	99.53	0.0076	173441	15.85	23.19	61.62
IonDUDE M h3 Indels	99.52	0.0077	181295	15.37	24.34	61.44
IonDUDE Q k5 Subs	99.53	0.0076	174133	15.83	23.17	61.19
IonDUDE Q h3 Indels	99.52	0.0077	186892	14.77	24.79	60.17

LM, Lecturas Mapeadas; TE, Tasa de error; LMI, Lecturas mapeadas con inserciones; LMD, Lecturas mapeadas con deleciones; Indels HP, indels en homopolímeros. El mejor resultado para cada métrica esta resaltado en negrita.

Tabla 1.5: Resultados de "BAM QC" para la muestra ERR161543

Programa	LM (%)	TE	Mismatches	LMI (%)	LMD (%)	Indels HP (%)
Sin corregir	85.56	0.0217	1903447	43.03	34.44	65.6
Pollux	**88.79**	**0.0073**	**559281**	**16.92**	**18.47**	72.13
Karect	85.99	0.0142	1540576	26.37	28.09	68.23
Ionhammer	85.68	0.0216	2015725	39.73	32.49	65.97
Fiona	86.08	0.0123	1303231	24.14	20.48	66.07
IonDUDE k5 Subs	85.56	0.0217	1907004	43.02	34.44	65.6
IonDUDE h3 Indels	85.43	0.0218	1951604	42.6	34.8	65.26
IonDUDE M k5 Subs	85.54	0.0219	1943475	42.99	34.43	65.6
IonDUDE M h3 Indels	85.44	0.0218	1976574	42.29	35.09	64.92
IonDUDE Q k5 Subs	85.55	0.0218	1915556	43	34.43	65.54
IonDUDE Q h3 Indels	85.37	0.0227	2340473	39.29	41.89	**61.74**

LM, Lecturas Mapeadas; TE, Tasa de error; LMI, Lecturas mapeadas con inserciones; LMD, Lecturas mapeadas con deleciones; Indels HP, indels en homopolímeros;. El mejor resultado para cada métrica esta resaltado en negrita.

Tabla 1.6: Resultados de "BAM QC" para la muestra ERR236069

Programa	LM (%)	TE	Mismatches	LMI (%)	LMD (%)	Indels HP (%)
Sin corregir	84.13	0.0457	4748648	75.8	51.91	36.37
Pollux	**88.59**	0.0219	**2543911**	20.12	12.88	**31.53**
Karect	86.3	0.0227	4550066	29.86	23.7	34.33
Ionhammer	86.29	**0.0208**	4095891	**7.1**	**6.74**	42.91
Fiona	85.91	0.0309	4206885	41.67	28.41	33.37
IonDUDE k5 Subs	84.13	0.0457	4748677	75.8	51.91	36.37
IonDUDE h3 Indels	84.15	0.0457	4765310	75.49	51.96	36.02
IonDUDE M k5 Subs	84.12	0.0457	4750012	75.79	51.9	36.37
IonDUDE M h3 Indels	84.21	0.0454	4786202	74.78	51.91	35.11
IonDUDE Q k5 Subs	84.12	0.0457	4748467	75.78	51.89	36.35
IonDUDE Q h3 Indels	84.12	0.0457	4957776	74	54.34	35.68

LM, Lecturas Mapeadas; TE, Tasa de error; LMI, Lecturas mapeadas con inserciones; LMD, Lecturas mapeadas con deleciones; Indels HP, indels en homopolímeros;. El mejor resultado para cada métrica esta resaltado en negrita.