# CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS DEL INSTITUTO POLITÉCNICO NACIONAL

UNIDAD ZACATENCO

DEPARTAMENTO DE BIOTECNOLOGÍA Y BIOINGENIERÍA

**"Ensamblaje y anotación genómica de *Texcoconibacillus texcoconensis* gen. nov., sp. nov., así como la comprobación de genes de resistencia a arsénico, cadmio, cobalto, cobre y zinc"**

TESIS

Que presenta:

**ISA. Mónica Carolina Benítez Cárdenas**

Para obtener el grado de

**MAESTRA EN CIENCIAS**

**EN LA ESPECIALIDAD DE BIOTECNOLOGÍA**

Directores de la Tesis:

Dr. Luc Dendooven

Dr. Reynold Ramón Farrera Rebollo

**México, D.F.** **Agosto, 2015**

## Agradecimientos y dedicatorias

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT), por el apoyo económico brindado para realizar mis estudios de maestría.

Agradezco al Dr. Luc Dendooven, por permitirme ser parte de su equipo de trabajo y confiar en que podía iniciar con un tema totalmente nuevo.

Al Dr. Pablo Cruz, por su valiosa colaboración, sin la cual no hubiera podido lograr ensamblar el genoma del microorganismo con el que trabajé.

Al M. en C. Víctor Flores, por enseñarme bases de bioinformática, ayudarme en el ensamblaje y en la resolución de las múltiples dudas que tenía, y que aún tengo.

Al Dr. Francisco Barona, por abrirme generosamente las puertas del laboratorio de Evolución de la Diversidad Metabólica en el Langebio.

Al Dr. Reynold Ramón Farrera Rebollo, por sus consejos, confianza, dedicación e interés en el desarrollo de mi trabajo.

A mis asesores, la Dra. María Eugenia Hidalgo Lara y el Dr. Luis Bernardo Flores Cotera, por su colaboración en el desarrollo de mi tesis, el apoyo brindado y los consejos para enriquecer mi trabajo.

A mis padres, por siempre estar a mi lado apoyándome en las decisiones que he tomado y en cada proyecto que inicio.

Agradezco al Dr. Marco Luna y Blanca Ramírez, por la orientación y ayuda para trabajar en el laboratorio.

Al Dr. Santiago Domínguez por enseñarme sobre computadoras y programas y por su enorme paciencia conmigo.

Agradezco al personal del Departamento de Biotecnología y Bioingeniería del CINVESTAV por el apoyo brindado para el desarrollo de mi maestría.

A mis compañeros de generación y del laboratorio, por hacer más grato el trabajo.

**Resumen**

En el suelo del ex Lago de Texcoco hay gran diversidad de microorganismos extremófilos, entre ellos, los halófilos. *Texcoconibacillus texcoconensis* gen. nov., sp. nov., cepa 13CC$^{T}$ es una bacteria, Gram-positiva halófila (20% w/v de NaCl), aislada de dicho suelo (Ruiz-Romero *et al.*, 2013). Para determinar las bases genéticas de su capacidad para sobrevivir en un ambiente extremo, se secuenció, ensambló y anotó su genoma.

La secuenciación del genoma se hizo mediante Illumina y el ensamblaje se realizó con Velvet, obteniendo un genoma de 3, 395, 935 pb de longitud, con un total de 60 contigs y un estimado de N´s de 5917. Con este ensamblaje se partió para hacer la anotación, la cual se hizo en el servidor RAST, donde se lograron identificar 1925 posibles genes, de los cuales 11 genes son de resistencia a metales pesados. Después se procedió a hacer la comprobación de los genes anotados mediante pruebas de resistencia a arsénico, cadmio, cobalto, cobre y zinc, y además se determinó la Concentración Mínima Inhibitoria (MIC) para cada uno de estos metales, las cuales fueron de 625 mM para de arseniato, 7 mM para arsenito, 3.8 x$10^{-4}$ mM para cadmio, 6 mM para cobalto y 0.5 mM para zinc.

Es así como se corroboró que los genes anotados por RAST están activos, es decir, las proteínas que codifican realizan su función.

**Abstract**

There is a great variety of extremophile microorganisms in the soil of the Ex-Lake of Texcoco, among them, halophiles. *Texcoconibacillus texcoconensis* gen. nov., sp. nov., strain 13CC$^{T}$ is a halophile (20% w/v NaCl), Gram-positive bacterium, isolated from that soil (Ruíz-Romero *et al.*, 2013). To determine the genetic basis of its ability to survive in an extreme environment its genome was sequenced, assembled, and annotated.

The genome sequencing was performed with Illumina. The assembly was made using Velvet and a genome of 3, 395, 935 bp of length, with a total of 60 contigs and an estimate of 5917 N´s was obtained. Regarding annotation, 1925 possible genes were identified, of which 11 are related to heavy-metal resistance. Thus, tests for resistance to arsenic, cadmium, cobalt, copper and zinc were made, and Minimum Inhibitory Concentrations (MICs) were determined for each of these metals. The minimum inhibitory concentrations were 625 mM of arsenate, 7 mM of arsenite, $3.8x10^{-4}$ mM of cadmium, 6 mM of cobalt and 0.5 mM of zinc.

This way, it was corroborated that RAST-annotated genes are active, i.e. the encoded proteins are performing their function.

## Índice

## Índice de tablas



## Índice de figuras

## 1. Introducción

Hoy en día, es conocido que ambientes que hasta hace poco eran considerados inhabitables por el hombre, son colonizados por organismos capaces de adaptarse a ambientes extremos; estos organismos son llamados extremófilos. El descubrimiento de microorganismos que habitan en ambientes con temperaturas extremas, pH extremos, altas presiones y alta salinidad, ha despertado interés de estudiarlos desde el punto de vista biotecnológico, debido a las características de estos microorganismos, puesto que sus biomoléculas son resistentes a las condiciones agresivas de su entorno, por lo que se estudian además con la perspectiva del desarrollo de potenciales aplicaciones industriales, puesto que pueden resistir condiciones y concentraciones de compuestos a los que otros microorganismos no podrían sobrevivir (Oren, 2002b; Ramírez *et al.*, 2006).

En cuanto a los ambientes hipersalinos, estos suelen localizarse en zonas calientes y secas, en el caso de México, se tienen suelos y lagos salinos, como es el caso del ex Lago de Texcoco (Ramírez *et al.*, 2006).

### 1.1 *Texcoconibacillus texcoconensis* cepa 13CC

*Texcoconibacillus texcoconensis* es una bacteria Gram-positiva, con forma de bacilo, que esporula, fue aislada del suelo salino-alcalino del exLago de Texcoco. Es una bacteria aerobia, con actividad catalasa positiva, que crece a 37°C y a un pH 8.9, después de 1–2 días de incubación, y que además puede crecer hasta una salinidad del 20% (w/v) de NaCl. El género *Texcoconibacillus* pertenece a la clase *Bacilli* y a la familia *Bacillaceae*. Por tanto, la especie es *Texcoconibacillus texcoconensis*, siendo asi el primer microorganismos de un género nuevo (Ruiz-Romero *et al.*, 2013).

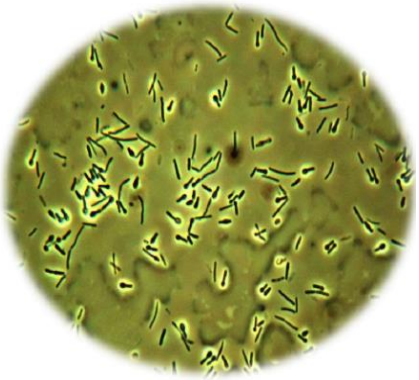


**Figura 1.** Microfotografía de *Texcoconibacillus texcoconensis* cepa 13CC (Ruiz-Romero *et al.*, 2013)

La capacidad de *Texcoconibacillus texcoconensis* para crecer a una alta concentración de cloruro de sodio (20% (w/v) NaCl), lo hace de interés, ya que se sabe que al presentar una resistencia a la alta salinidad, los microorganismos halófilos, pueden tener características que pueden ser de importancia biotecnológica, puesto que se ha visto que estos microorganismos tienen una gran versatilidad en su metabolismo y en la forma que toleran el estrés osmótico causado por la altas concentraciones de sales en su medio (Oren, 2002a; Ruiz-Romero *et al.*, 2013). Su posible aplicación biotecnológica está, por ejemplo, en la producción de biopolímeros, de enzimas y de solutos compatibles, así como a que presentan resistencia a metales pesados y otros contaminantes; o bien pueden emplearse en la biodegradación de compuestos tóxicos (Oren, 2002a; Ventosa & Nieto, 1995).

En la actualidad, con las nuevas tecnologías desarrolladas, la búsqueda de microorganismos halófilos con una posible aplicación podría ser mejor entendida mediante el ensamblaje y la anotación de su genoma, pues ayudaría a entender e identificar el rol que pueden tener sus genes en la adaptacion a esos medios extremos, además de que esta clase de trabajos podrían ayudar a identificar otros genes, que, como se mencionó antes, pueden ser de interés (Oren, 2002a).

## 1.2 Genoma

Un genoma es el conjunto de genes codificados dentro del ADN de un organismo, el cual contiene las instrucciones genéticas necesarias para desarrollar y dirigir sus actividades. Las moléculas del ADN están conformadas por dos hélices emparejadas y las coforman cuatro bases nucleótidas, adenina (A), timina (T), guanina (G) y citosina (C). ( National Human Genome Research Institute).

## 1.3 Genómica

Es la subdisciplina de la genética interesada en la descripción y análisis molecular de genomas, suele subdividirse en genómica estructural y funcional; la primera se ocupa de la caracterización física de los genomas, y la segunda de ubicar todos los elementos funcionales, tanto los componentes de un genoma (las proteínas codificadas, los elementos reguladores, estructurales, etc.) como determinar su papel en el organismo (Rodríguez, 2004).

## 1.4 Secuenciación

Secuenciar un genoma significa determinar el orden exacto de los pares de bases en un segmento de ADN (National Human Genome Research Institute).

La secuenciación es el núcleo de la genómica. Entre sus usos se tiene la determinación de la secuencia del genoma de una nueva especie o de un individuo dentro de una población, la secuenciación de las moléculas de ARN de una muestra en particular o el uso de la secuencia de ADN como un ensayo de lectura en técnicas de biología molecular, entre otras (Zerbino & Birney, 2008).

### 1.4.1 Tecnologías de secuenciación

#### 1.4.1.1 La secuenciación por el Método de Sanger.

Actualmente esta tecnología se emplea con algunas modificaciones. Es conocida como el método de los terminadores de cadena. La clave de esta técnica consiste en adicionar, ademas de los reactivos específicos para ésta, una pequeña cantidad de nucleótidos modificados que son conocidos como dideoxinucleótidos (ddA, ddG, ddC o ddT), que se incorporan en la cadena que se está elongando, deteniendo su polimerización (Sanger *et al.,* 1977). En el método original de Sanger, la detección de las moléculas de ADN en la reacción de secuenciación se realizaba utilizando un ADN iniciador marcado radioactivamente (Rodríguez, 2004).

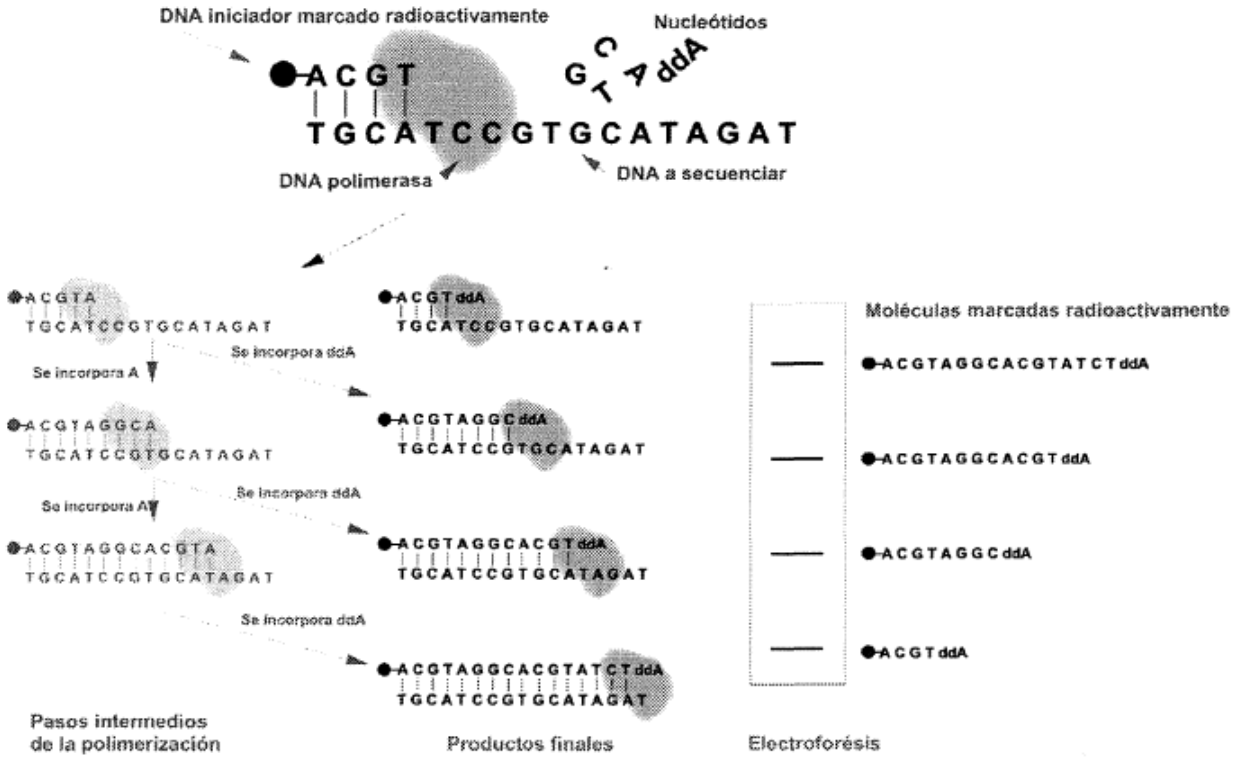



**Figura 2.** Secuenciación del ADN mediante el método de Sanger. Esquema de la reacción correspondiente a la determinación de la posición en la secuencia del ADN de una base, en este caso la adenina (A). La incorporación de ddA en lugar de A detiene el proceso de polimerización. Esta incorporación ocurre de forma aleatoria durante la polimerización de las moléculas de ADN de tal forma que una fracción de las moléculas elongadas se habrá detenido en cada posición en que A debiera incorporarse a la molécula. Las moléculas marcadas radioactivamente se detectan después de haberlas separado por su tamaño mediante electroforesis en geles de poliacrilamida. El tamaño de las moléculas detectadas nos indica en qué posición de la secuencia se encuentra el nucleótido A (Rodríguez, 2004).

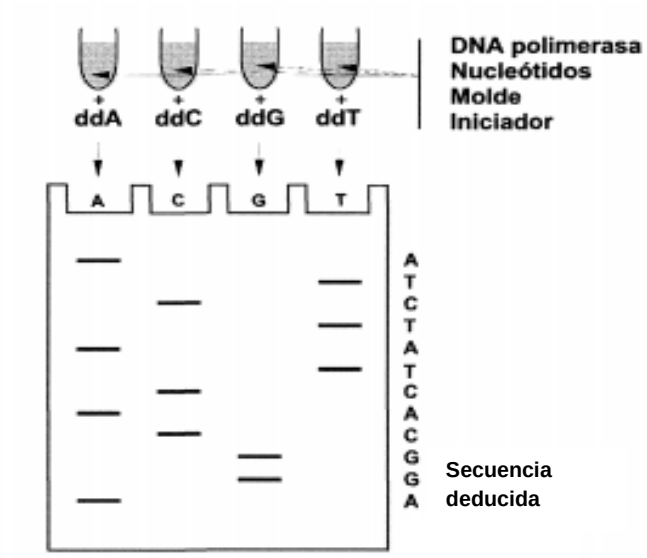



**Figura 3.** Secuenciación del ADN mediante el método de Sanger. Experimento completo de secuenciación del ADN. Se realizan reacciones como las descritas en la Figura 2, para cada uno de los nucleótidos y la secuencia completa del ADN se deduce de la posición en que aparecen las moléculas marcadas radioactivamente (Rodríguez, 2004).

### 1.4.1.2 La pirosecuenciación (Secuenciación 454).

Esta técnica se basa en la "secuenciación por síntesis", que depende de la liberación de pirofosfato en la incorporación de nucleótidos (Pareek *et al.*, 2011). Con esta tecnología se producen lecturas cortas. Equipos como el GS FLX Titanium pueden generar 500 Mpb de ADN en lecturas de un tamaño aproximado de 400 pb por lectura, que es aproximadamente 1.2 millones de lecturas por corrida, mientras que los instrumentos de la generación anterior (como GS FLX), generaban hasta 100Mpb de ADN secuenciado en lecturas de un tamaño de 250 pb, lo que es aproximadamente 400,000 lecturas por corrida (Nagarajan & Pop, 2010). La técnica se basa en la detección quimioluminiscente de pirofosfato liberado durante la incorporación de desoxinucleótido trifosfato (dNTP) mediada por polimerasa (Nyren *et al.*, 1993; Nyren, 2007) y en la secuenciación de ADN en tiempo real, utilizando esta detección de la liberación de pirofosfato (Pareek *et al.*, 2011; Ronaghi *et al.*, 1998).

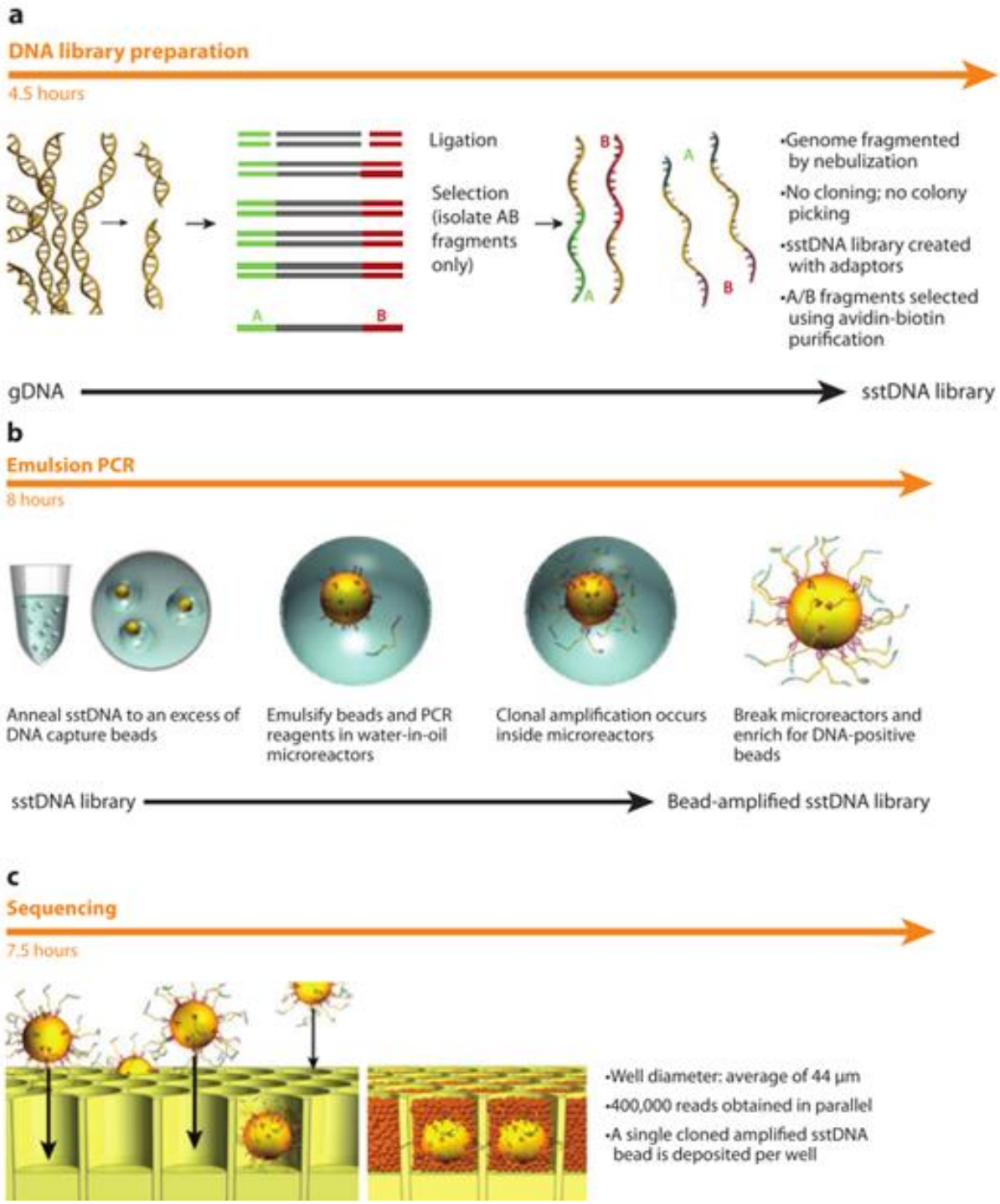



**Figura 4.** Método usado por el secuenciador Roche/454 para amplificar copias de ADN de cadena sencilla desde una librería de fragmentos en perlas de agarosa. Una mezcla de fragmentos de ADN con cuentas o perlas de agarosa que contienen oligonucleótidos complementarios a los adaptadores en los extremos de los fragmentos, se mezcla en una proporción aproximada de 1:1 .Las mezcla es encapsulada por agitación vigorosa en vortex en micelas acuosas que contienen reactivos para PCR rodeados por aceite, y pipeteados en placas de microtitulación con 96 pozos para amplificación PCR. Las perlas resultantes son impregnadas con aproximadamente 1 millón de copias del segmento de la cadena sencilla original, lo cual provee suficiente fuerza de señal durante la reacción de pirosecuenciación que sigue, para detectar y registrar eventos de incorporación de nucleótido (Mardis, 2008).

#### 1.4.1.3 Illumina-Solexa

Es una plataforma de secuenciación de segunda generación, conocida como Solexa. Con este tipo de tecnología se obtienen lecturas de entre 25-35 pb (Nagarajan & Pop, 2010; Mardis, 2008) y de entre 50-100 según lo reportado por Miller *et al.*, (2010) y Paszkiewicz & Studholme, (2010), siendo estas lecturas cortas de alto rendimiento que son menores que las producidas por Sanger, presentando tasas de error menores al 1% (Nagarajan & Pop, 2010).

La secuenciación por Illumina se inicia con una biblioteca adaptadora específica para Illumina, que consta de una celda de flujo con ocho canales, la cual amplifica fragmentos en su superficie y usa ADN polimerasa para producir múltiples copias de ADN, o clústeres, cada uno de los cuales representa a la molécula que inicia la amplificación del clúster, y es ejecutada por un dispositivo automático llamado Cluster Station. Cada clúster contiene aproximadamente un millón de copias del fragmento original, lo que es suficiente para reportar las bases incorporadas con la intensidad de señal requerida para la detección durante la secuenciación. El sistema Illumina utiliza una estrategia de secuenciación por síntesis, en el cual, los cuatro nucleótidos son agregados simultáneamente a la celda de flujo, junto con ADN polimerasa, para su incorporación en los fragmentos iniciadores de los oligos del clúster (Figura 4), particularmente, los nucleótidos tienen una etiqueta fluorescente específica para cada base y tienen el grupo 3´-OH químicamente bloqueado de tal manera que cada incorporación de un nucleótido es un evento único (Figura 5). A cada paso de incorporación de una base le sigue un paso de generación de imagen, en el cual, de cada carril de la celda de flujo se obtienen 100 fracciones de imagen repetidas tres veces, por la óptica del instrumento, a una densidad de 30,000

clústeres por fracción de imagen. Después de cada paso de generación de imagen, el grupo bloqueador en el enlace 3´ es químicamente removido para preparar cada filamento para la siguiente incorporación por la ADN polimerasa (Figura 6). Esta serie de pasos continúa por un número específico de ciclos según lo defina el usuario, lo que permite longitudes de lecturas discretas de 25 a 35 pb. Un algoritmo de asignación de bases asigna las secuencias y valores de calidad asociados a cada lectura y un canal de chequeo de calidad evalúa los datos de Illumina de cada corrida, removiendo secuencias con poca calidad (Mardis, 2008).

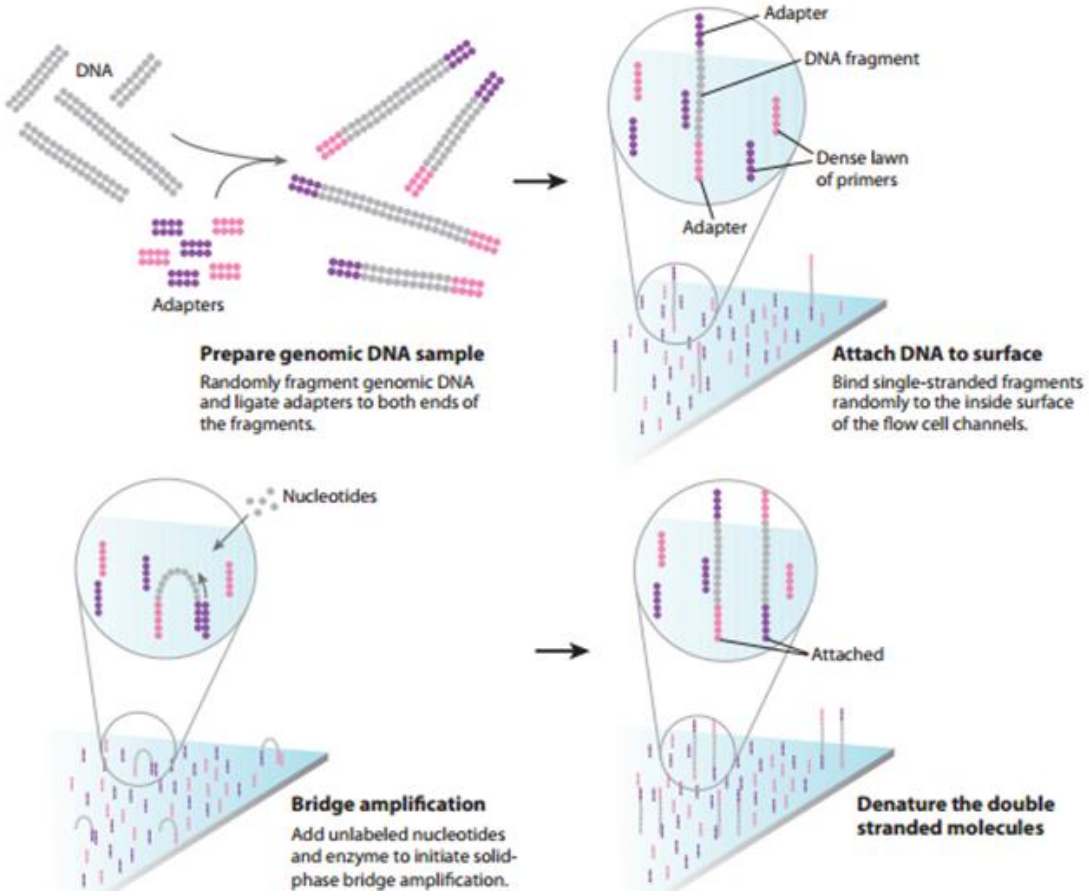



**Figura 5.** Secuenciación por síntesis de Illumina. A los filamentos del clúster obtenidos por amplificación en puente se les agregan iniciadores y los cuatro son marcados fluorescentemente. Nucleótidos bloqueados en el extremo 3´-OH son agregados a la celda de flujo con ADN polimerasa. Los filamentos del clúster son extendidos con sólo un nucleótido. Siguiendo este paso de incorporación, los nucleótidos no usados y la ADN polimerasa son extraídos por lavado; un buffer de escaneo se agrega a la celda de flujo y el sistema óptico escanea cada carril de la celda de flujo en unidades de imagen denominadas “fracciones” (tiles) (Mardis, 2008).

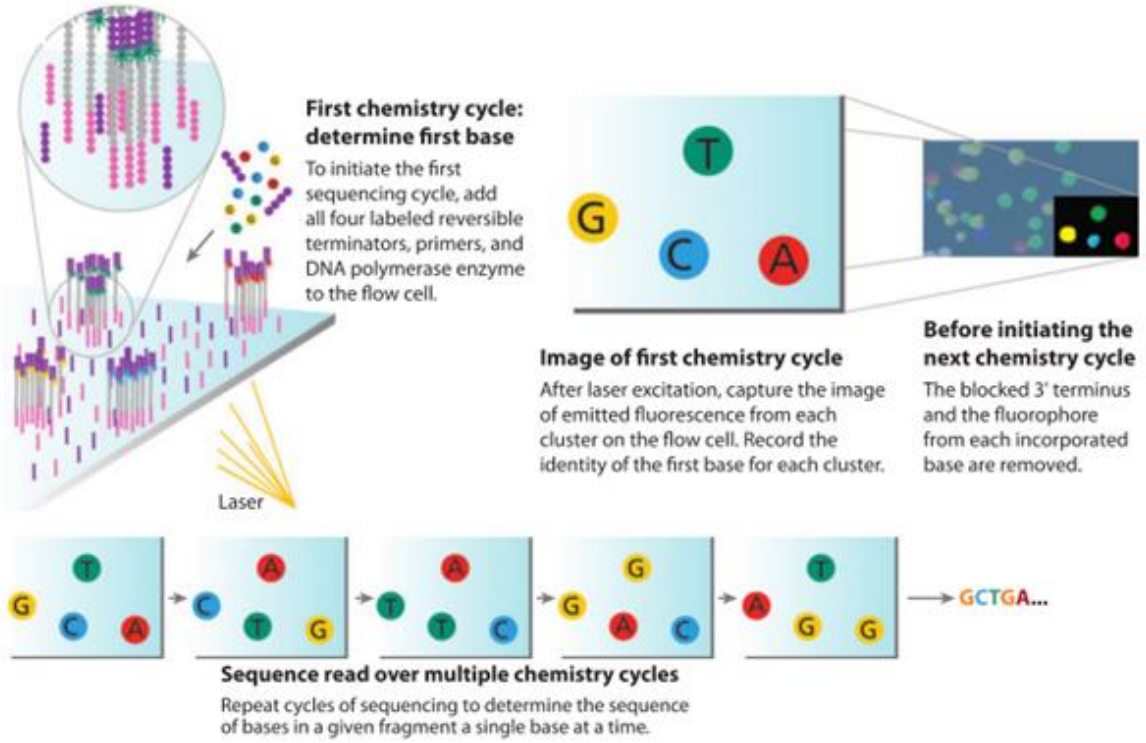



**Figura 6.** Ciclo químico de la secuenciación por síntesis de Illumina. Una vez que el proceso de generación de imagen se termina, se agregan a la celda de flujo agentes químicos que efectúan la separación de las etiquetas fluorescentes y los grupos que bloquen los extremos 3´-OH, lo que prepara los filamentos del clúster para otra ronda de incorporación de nucleótido fluorescente (Mardis, 2008).

### 1.5 De la secuenciación al ensamblaje de un genoma

En la actualidad, con las diferentes tecnologías de secuenciación de alto rendimiento (High-throughput sequencing technologies: HT-NGS), se obtiene más información cualitativa y cuantitativa sobre los genomas, puesto que generan desde cientos de millones hasta miles de millones de bases en una sola corrida, lo que antes no era técnicamente posible o tenía un costo muy elevado (Nagarajan & Pop, 2010; Pareek *et al.*, 2011). En el caso de Illumina, se producen lecturas con una amplia cobertura (coverage), que se obtienen a expensas de una longitud muy corta de las lecturas (~35pb, Mardis, 2008; Nagarajan & Pop, 2010), generando millones de estas con la misma longitud.

La habilidad de secuenciar y ensamblar eficientemente genomas bacterianos completos tiene implicaciones muy significativas para propósitos evolutivos, (Smith *et al.*, 2006; Mwangi *et al.,* 2007) y metagenómicos (Handelsman *et al.*, 1998; Eisen, 2007), principalmente.

### 1.5.1 Ensamblaje

Un ensamblaje es una estructura jerárquica de datos que se obtiene como resultado de la reconstrucción del genoma a partir de la información de un grupo de secuencias producto de un procedimiento de secuenciación (Miller *et al.*, 2010).

Durante el ensamblaje, se agrupan las lecturas o “reads”, en “contigs” y los contigs en andamios o “scaffolds”. Los contigs proporcionan un alineamiento múltiple de las lecturas más la secuencia consenso, que es la secuencia de nucleótidos más comunes en una región definida del ADN. Los scaffolds, a veces llamados supercontigs o metacontigs, definen el orden, la orientación y el tamaño de los contigs y los espacios entre ellos. Un scaffold puede ser de dos tipos: el que presente un camino simple, es decir, una única opción de acomodo de las secuencias o bien una red, que significa que puede haber más de una opción en el orden de las secuencias. La mayoría de los programas ensambladores generan un grupo de lecturas no ensambladas o parcialmente ensambladas. En cuanto al formato del archivo generado, el más ampliamente aceptado es *fasta*, donde la secuencia consenso del contig puede ser representada por los caracteres A, C, G y T, más, posiblemente otros caracteres con un significado especial. Los ensamblajes se miden por el tamaño y la precisión de sus contigs y scaffolds. El tamaño de ensamblaje es dado usualmente por estadística, que incluye la longitud máxima, la

longitud media, la longitud total y el N50. El contig N50 se obtiene ordenando los contigs de mayor a menor longitud hasta obtener un conjunto con el cual se tenga representado el 50% del tamaño total del genoma, entonces el contig N50 será el contig más pequeño que nos permite llegar a ese porcentaje (Miller *et al.*, 2010). Las estadísticas N50 para diferentes ensamblajes no son comparables a menos que estén calculadas usando el mismo valor de la longitud combinada total, por esta razón la precisión del ensamblaje es difícil de medir. El alineamiento con secuencias de referencia es útil cuando existen secuencias de referencia de confianza (Miller *et al.*, 2010; Paszkiewicz & Studholme, 2010).

### 1.5.2 Tecnologías para el ensamblaje

Las tecnologías de secuenciación de ADN comparten la limitación fundamental de que la longitud de las lecturas generadas, son mucho más cortas que aún los más pequeños genomas (Miller *et al.*, 2010). Debido a esto, se han desarrollado diferentes programas para ensamblar las lecturas que se generan al secuenciar, y así poder reconstruir los genomas (Hernandez *et al.*, 2008).

#### 1.5.2.1 Programas disponibles para el ensamblaje

Dependiendo del tipo de tecnología de secuenciación usada, se elige el programa para efectuar el ensamblaje. Estos programas se basan en gráficas de ensamblaje (Assembly graphs), las cuales son un conjunto de nodos o vértices, que pueden visualizarse como esferas, y con un conjunto de flechas (filos) o arcos entre los nodos. Si las flechas o filos solo pueden ser recorridas en una sola dirección, el gráfico es directo (Miller *et al.*, 2010).

De esta manera los ensambladores pueden clasificarse dentro de tres categorías, según el tipo de gráfica que usan:

- Los métodos Overlap/Layout/Consensus (OLC), dependen de una gráfica de superposición (overlap graph).
- Los métodos De Bruijn Graph (DBG), usan alguna forma de gráficas de K-mer (K-mer graph).
- Los algoritmos de gráficas voraces (o ávidos), pueden usar tanto OLC como DBG.

#### 1.5.2.1.1 Ensambladores tipo OLC (Overlap/Layout/Consensus).

El método de los ensambladores OLC es típico para lecturas producidas por Sanger y 454. Usan una gráfica de superposición, que representa las lecturas de secuenciación y sus sobreposiciones (Myers, 1995), las cuales deben ser precomputadas por una serie de alineamientos de secuencia por apareamiento que son computacionalmente demandantes. Conceptualmente, los nodos representan las lecturas y las flechas las superposiciones entre esas lecturas. (Miller *et al.*, 2010).

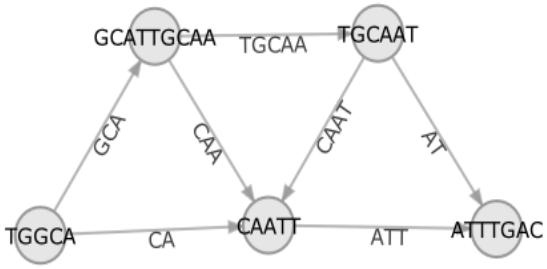



**Figura 7.** Gráfica de ensamblaje OLC. Los nodos son lecturas completas y las flechas conectan las lecturas que se sobreponen. Note que en una gráfica de ensamblaje OLC real, las lecturas y las superposiciones serían mucho más grandes. En esta figura se acortan para mayor claridad.

La operación de este tipo de ensambladores se da en tres fases, primero descubre las superposiciones, comparando las lecturas por apareamiento de todas contra

todas, esto lo hace definiendo K-mers, que son una representación compacta basada en palabras cortas ideal para conjuntos de datos, a lo largo de todas las lecturas y selecciona los candidatos para superposición que comparten K-mers. Después construye y manipula una gráfica de superposición, produciendo una disposición de las lecturas aproximada y por ultimo hace una alineación de secuencias múltiples (Multiple sequence alignment o MSA), para determinar la disposición precisa y la secuencia consenso (Miller *et al.*, 2010; Wang *et al.*, 1994).

A continuación se presentan tres ensambladores que funcionan con este método.

- **Newbler**

Newbler es un software ampliamente usado por 454 Life Sciences (Margulies *et al.*, 2005), se usa para construir scaffolds a partir de limitaciones de lecturas apareadas en los extremos (paired end). Newbler explota la cobertura, si es posible, para superar los errores de asignación de bases. El programa calcula los consensos de contig y unitig (o contig de alta confianza) en el "espacio de flujo" (Myers *et al.*, 2000), usando la fuerza de la señal suministrada por la plataforma, asociada con cada flujo de un nucleótido particular. La señal normalizada es correlacionada proporcionalmente al número de repeticiones directas de ese nucleótido en esa posición de la lectura (Miller *et al.*, 2010).

- **Celera**

Es un ensamblador usado para manejar datos de 454. Funciona con el software CABOG (Celera Assembler with the Best Overlap Graph), descubriendo las superposiciones usando iniciadores (seeds) comprimidos. CABOG reduce las corridas de homopolímero, esto es, repeticiones de letras sencillas, a bases simple, para superar la incertidumbre de longitud de corrida en los datos y construye unitigs

iniciales excluyendo lecturas que son subsecuencias de otras lecturas, debido a que estas lecturas son más susceptibles a superposiciones falsas inducidas por repeticiones (Miller *et al.*, 2010).

- **Edena (Exact DE Novo Assembler).**

Es un software usado para el ensamblaje de *novo* de contigs precisos de conjuntos de datos con lecturas muy cortas de la misma longitud. De igual manera que otros ensambladores, Edena estructura las superposiciones en una gráfica, produciendo contigs exactos de varias kilobases que cubren la mayoría del genoma que está siendo ensamblado (Hernandez *et al.*, 2008).

El software Edena descarta las lecturas duplicadas y encuentra todas las superposiciones perfectas y libres de error. Remueve superposiciones individuales que son redundantes con pares de otras superposiciones, lo que es una aplicación del algoritmo de reducción de superposiciones transitivas (Myers *et al.*, 2000). Edena poda puntas y burbujas y con el software Shorty ataca el caso especial en que unas pocas lecturas largas están disponibles para actuar como iniciadores para incorporar lecturas cortas y sus parejas acompañantes. De esta manera, trabajando por medio de iteraciones, Shorty usa contigs como iniciadores (seeds) para obtener contigs más grandes (Miller *et al.*, 2010).

#### 1.5.2.1.2 La gráfica de De Bruijn

Es una representación compacta basada en palabras cortas llamadas K-mers que es ideal para conjuntos de datos con lecturas muy cortas (25-50 bp) de alta cobertura. Los nodos representan todas las cadenas de letras posibles de longitud fija. Las flechas representan superposiciones perfectas que abarcan desde el prefijo hasta el sufijo de las letras en los nodos (Miller *et al.*, 2010; Zerbino & Birney, 2008).

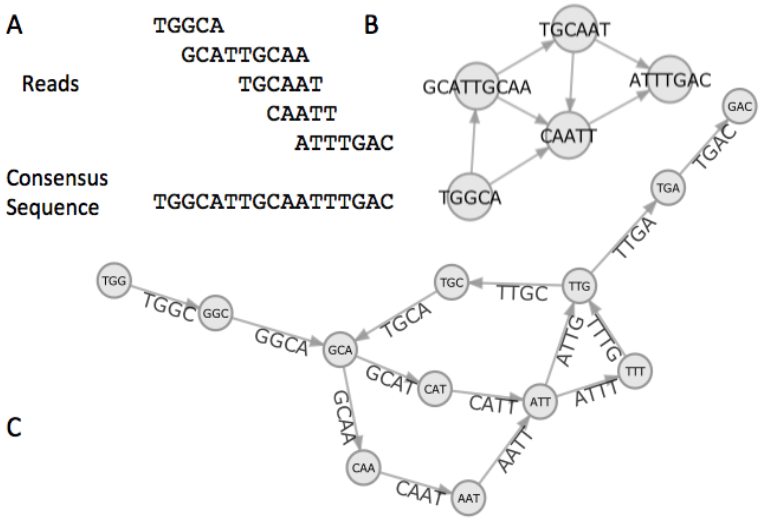



**Figura 8.** Lecturas y dos posibles gráficas de ensamblaje. A: Lecturas hipotéticas alineadas a la secuencia consenso. B: Gráfica OLC creada con las lecturas; las flechas representan superposiciones de dos o más nucleótidos. El proceso de ensamblaje requiere pasar por todos los nodos. C: Gráfica de ensamble de De Bruijn creada de las mismas lecturas. Las flechas representan segmentos de 4 nucleótidos con dos de ellos superpuestos entre los nodos El proceso de ensamble requiere recorrer todos las flechas.

Este tipo de ensamblaje es ampliamente aplicado a las lecturas cortas que se obtienen de las plataformas Solexa y SOLID. Se basa en gráficas de K-mer, que son una forma de gráfica de De Brujin. Sus nodos representan todas las posibles subsecuenciaciones de longitud fija, y las flechas representan todas las superposiciones de longitud fija entre las subsecuenciaciones. Se tiene una flecha para cada K-mer que comienza en cada base, (excluyendo las últimas K-1 bases). Los nodos representan superposiciones de K-1 bases. Este tipo de gráfica no requiere detección de superposiciones de todo contra todo, no almacena necesariamente las lecturas individuales o sus superposiciones y comprime las secuencias redundantes. Por el contrario, la gráfica de K-mer contiene la secuencia efectiva y por ello puede agotar la memoria disponible en genomas grandes. Los enfoques de memoria distribuida pueden aliviar esta limitante (Miller *et al.*, 2010).

La forma en que este tipo de gráfica opera es con un enfoque de gráfica de De Bruijn (DBG) o un enfoque Euleriano, basado en un escenario ideal, es decir, dado un conjunto de datos perfecto, o sea K-mers libres de errores que provean cobertura total y abarquen cada repetición, la gráfica de K-mer sería una gráfica de De Bruijn y contendría una trayectoria Euleriana, esto es, una vía que atraviese cada flecha de la gráfica exactamente una vez. La trayectoria sería evidente de encontrar, haciendo el problema del ensamblaje trivial por extensión, pero las gráficas de K-mer construidas a partir de datos de secuenciación reales son más complicadas. La fase de construcción de la gráfica avanza rápidamente usando una búsqueda en tabla hash, que es una estructura de datos que asocia claves con valores, a tiempo constante, de la existencia de cada K-mer en la corriente de datos.

A continuación se presentan algunos ensambladores basados en gráficas de De Brujin.

- **Ensamblador EULER**

En este ensamblador, en la representación de los datos, los elementos no están organizados alrededor de lecturas, sino en torno a palabras de K nucleótidos o K-mers (Chaisson *et al.*, 2004 y 2008). Las lecturas son mapeadas como trayectorias a través de la gráfica, yendo de una palabra a la siguiente en un orden determinado (Jiang *et al.* 2008; Miller *et al.*, 2010;). El software euleriano fue desarrollado para lecturas Sanger y fue subsecuentemente modificado para lecturas cortas de 454 GS20, para lecturas no apareadas Illumina/Solexa aún más cortas y para lecturas Solexa apareadas en los extremos (paired end).

Euler aplica un filtro a las lecturas antes de construir su gráfica. El filtro detecta la asignación errónea de bases tomando nota de K-mers de baja frecuencia, es decir,

se basa en la redundancia de las lecturas: muchos de los K-mers verdaderos deberían ser repetidos en varias lecturas y la mayoría de los K-mers erróneos deberían ser únicos, por lo que se implementa con una lista de K-mers y sus frecuencias observadas en las lecturas y el filtro excluye o corrige K-mers de baja frecuencia. Sin embargo, esta corrección reduce el número total de K-mers y con ello el conteo de nodos en la gráfica, además genera el riesgo de enmascarar un polimorfismo verdadero y puede corromper K-mers válidos que tienen, de casualidad, baja cobertura. (Miller *et al.*, 2010).

- **Velvet**

Es una aplicación que está basada en un conjunto de algoritmos desarrollado para manipular gráficas de De Bruijn, en particular gráficas de K-mer, para el ensamblaje de la secuencia genómica (Pevzner *et al.*, 2001; Zerbino & Birney, 2008). Velvet hace uso extensivo de la simplificación de gráficas para reducir a nodos sencillos las trayectorias que no se intersectan. La simplificación comprime la gráfica sin pérdida de información. Velvet usa este paso de simplificación durante la construcción de la gráfica y de nuevo varias veces durante el proceso de ensamblaje. Este proceso de eliminación de casos únicos, es análoga a la formación de unitigs en gráficas de superposición y en ensambladores OLC (Zerbino & Birney, 2008).

Velvet corta la gráfica de K-mer removiendo puntas (spurs) iterativamente. Su algoritmo de remoción de puntas es similar al procedimiento de filtración de Euler. Tiene un parámetro para el mínimo número de ocurrencias de las lecturas para que un K-mer pueda calificar como un nodo de gráfica. Velvet reduce la complejidad de la gráfica con una búsqueda acotada de burbujas en la gráfica, donde las trayectorias candidatas son atravesadas por pasos, moviéndose hacia adelante un nodo en todas

las trayectorias, por cada iteración, hasta que la longitud de la trayectoria exceda un umbral, además, estrecha las candidatas a burbujas hasta aquellas con un requerimiento de similitud de secuencia en las trayectorias alternas. Una vez que ha encontrado una burbuja, remueve la trayectoria que representa menores lecturas y, trabajando fuera de la gráfica, realinea las lecturas desde la trayectoria removida hasta la trayectoria remanente. Debido a que la multiplicidad de lectura más alta determina la trayectoria objetivo, el realineador asigna las bases de consenso con un algoritmo de votación por columnas. Velvet aún reduce más la complejidad de la gráfica por enhebrado o ensartado de lecturas. Remueve trayectorias que representan menores lecturas que un umbral. Esta operación genera el riesgo de remover secuencias de baja cobertura pero se supone que remueve mayormente conexiones falsas inducidas por errores de secuenciación convergentes. Velvet explota las lecturas largas, si alguna fue proporcionada, por medio de un algoritmo que llama Banda de Rock (RockBand), éste, forma nodos a partir de trayectorias que son confirmadas por dos o más lecturas largas, siempre y cuando no haya otras dos lecturas largas que proporcionen una contradicción consistente. Además, puede ser ejecutado varias veces por cada conjunto de datos para optimizar la selección de tres parámetros críticos: la longitud de los K-mers, que está limitada a ser impar para excluir nodos que representen repeticiones palindrómicas, la frecuencia mínima esperada de K-mers en las lecturas determina qué K-mers serán podados *a priori* y la cobertura esperada del genoma en las lecturas controla rupturas de conexión falsa (Zerbino & Birney, 2008). El software se enfoca en el ensamblado *de novo* para lecturas cortas con extremos apareados (paired end) de la plataforma Solexa. Una

extensión permite ensamblar conjuntos de datos compuestos solamente de lecturas SOLID (Miller *et al.*, 2010).

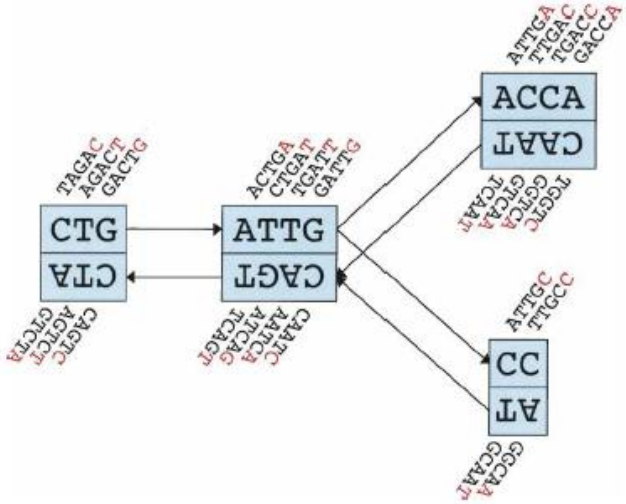



**Figura 9**. Representación esquemática de la implementación de Velvet de una gráfica de De Bruijn. Cada nodo, representado por un único rectángulo, representa a su vez una serie de K-mers que se superponen (en este caso K=5). Listados directamente ya sea arriba o abajo (rojo), se muestra el último nucleótido de cada K-mer. La secuencia de esos nucleótidos finales, copiado en letras grandes en el rectángulo, es la secuencia del nodo. El nodo gemelo, directamente pegado al nodo ya sea arriba o abajo, representa la serie reversa de K-mers complementos reversos. Los arcos o filos son representados por flechas entre los nodos. El último K-mer de un origen de arco se superpone con el primero de su destino. Cada arco tiene un arco simétrico. Note que los dos nodos en la izquierda podrían fusionarse en uno sin pérdida de información, debido a que forman una cadena (Zerbino & Birney, 2008).

- **Ensamblador ABySS**

En este programa, el ensamblaje se lleva a cabo en dos etapas: primero, sin usar la información de extremos apareados, los contigs se extienden hasta que ya no pueden seguir extendiéndose sin ambigüedades o hasta que llegan a un extremo romo (blunt-end) debido a falta de cobertura. En el segundo paso, la información de extremos apareados (paired end) se usa para resolver las ambigüedades y fusionar los contigs. ABySS distribuye la gráfica K-mer y las computaciones de la gráfica a lo largo de una retícula de cómputo con una vasta memoria. ABySS particiona el ensamblaje en el nivel de grado de disgregación del nodo de gráfica individual (cada

nodo del gráfico es procesado separadamente, por eficiencia, muchos nodos de la gráfica son asignados a cada CPU). La asignación de un nodo de gráfica al CPU se logra convirtiendo el K-mer a un entero (Miller *et al.*, 2010).

#### 1.5.2.1.3 Algoritmos basados en gráficas voraces (o ávidas).

Los algoritmos ávidos aplican una operación básica: dada cualquier lectura o contig, agregan una lectura o contig más. La operación básica se repite hasta que no sean posibles más operaciones. Cada operación usa la siguiente superposición con más alta puntuación para hacer el siguiente empalme. Así, los contigs crecen por extensión ávida o ambiciosa, siempre escogiendo la lectura que es encontrada siguiendo la superposición con el más alto puntaje. Como todos los ensambladores, los algoritmos ávidos necesitan mecanismos para evitar incorporar superposiciones falsas positivas en los contigs. Las superposiciones inducidas por secuencias repetitivas pueden tener más altos puntajes que las superposiciones inducidas por una posición común de origen.

A continuación se presentan algunos ensambladores basados en gráficas ávidas.

- **SSAKE ( Ensamblaje de Secuencias Cortas por Búsqueda de K-mer y Extensión de Lectura 3' o Short Sequence Assembly by K-mer search and 3' read Extension)**

Fue el primer ensamblador de lecturas cortas y fue diseñado para lecturas cortas no apareadas de longitud uniforme. Estaba basado en la noción de que una alta cobertura proveería un conjunto de lecturas libres de error si las lecturas erróneas pudieran ser evitadas. SSAKE no usa una gráfica explícitamente, usa una tabla de búsqueda de lecturas indexada por sus prefijos, donde busca iterativamente lecturas que se sobreponen en un extremo del contig, de esta manera ensambla

agresivamente millones de secuencias cortas de nucleótidos por búsqueda progresiva de K-mers perfectos situados en el punto más extremo 3' de las secuencias (perfect 3'-most k-mers) usando un árbol de prefijos de ADN (Warren *et al.*, 2007). Sus lecturas candidatas deben tener una sobreposición idéntica de prefijo a sufijo cuya longitud esté arriba de un umbral, por lo que escoge cuidadosamente entre lecturas múltiples con superposiciones igualmente largas. SSAKE prefiere lecturas con confirmación de extremo a extremo en otras lecturas, lo que favorece lecturas libres de error, además el software detecta cuando el conjunto de candidatos presenta extensiones múltiples, en particular, detecta cuando los sufijos de las lecturas candidatas exhiben diferencias que son cada una confirmadas en otras lecturas. En este punto, el software termina la extensión del contig. Cuando ninguna lectura satisfaga el umbral mínimo inicial, el programa disminuye el umbral hasta que un segundo mínimo es alcanzado (Miller *et al.,* 2010).

- **SHARCGS**

Es un ensamblador que prefiltra las lecturas de acuerdo con sus valores de calidad y su redundancia en el conjunto de datos. (Dohm *et al.*, 2007). Este ensamblador también opera con lecturas cortas no apareadas de alta cobertura y longitud uniforme. Agrega funcionalidades pre y post-procesador al algoritmo básico de SSAKE. El preprocesador filtra lecturas erróneas al requerir un número mínimo de correspondencias exactas a todo lo largo, en otras lecturas. Un filtro aún más restrictivo es opcional, requiriendo que los valores de calidad (QVs) de lecturas que coincidan excedan un umbral mínimo. SHARCGS filtra el conjunto de lecturas crudas tres veces, cada una con un conjunto de restricciones diferente, para generar tres conjuntos filtrados diferentes. Después ensambla cada conjunto separadamente por

extensión iterativa de los contigs. Entonces, en el postproceso, fusiona los tres conjuntos de contigs usando alineamiento de secuencia. La fusión tiene el propósito de extender los contigs desde lecturas altamente confirmadas por medio de la integración de contigs más largos provenientes de filtros de menor restricción. (Miller *et al.*, 2010).

### 1.5.3 Aspectos adicionales en el ensamblaje.

- **Factores que complican la aplicación de las gráficas de K-mer al ensamblaje de secuencias de ADN.**

1. Con la doble cadena del ADN, la secuencia hacia adelante (forward) de cualquier lectura puede sobreponerse a la secuencia complementaria hacia adelante o hacia atrás (reverse) de otras lecturas. Una ejecución de la gráfica de K-mer contiene los nodos y las flechas para ambas cadenas, cuidando evitar la producción del ensamblaje entero dos veces. Otra ejecución almacena las secuencias forward y reverse como semi-nodos con la misma rotulación. Otra ejecución más representa cadenas alternas en un nodo único con dos lados, limitando las trayectorias a que entren y abandonen lados opuestos (Miller *et al.*, 2010).

2. Los genomas reales presentan estructuras repetidas complejas que incluyen repeticiones (repeats) en tándem, repeticiones invertidas, repeticiones imperfectas y repeticiones insertadas en repeticiones. Las repeticiones más grandes que el valor K producen gráficas de K-mer enredadas que complican el ensamblaje. Repeticiones perfectas de longitud K o mayores colapsan dentro de la gráfica, dejando una estructura de gráfica local que se asemeja a una cuerda con extremos deshilachados; las trayectorias convergen a lo largo de la repetición y después divergen. El ensamblaje exitoso requiere la separación de la trayectoria convergente,

que representa una repetición colapsada. Los ensambladores típicamente consultan las lecturas y posiblemente los compañeros de pareja (mate-pairs), en intentos de resolver estas regiones (Miller *et al.*, 2010).

3. Un palíndromo es una secuencia de ADN que es su propio complemento reverso. Los palíndromos inducen trayectorias que se doblan y regresan sobre sí mismas. Velvet requiere que los valores de K, que son las longitudes de un K-mer, sean números impares. Un K-mer de tamaño impar no puede corresponder a su propio complemento reverso (Miller *et al.*, 2010).

4. Los datos reales incluyen errores de secuenciación o por mal procesamiento de la muestra. Los ensambladores DBG para reducir este problema pre-procesan las lecturas para remover errores, después ponderan los filos de las gráficas por el número de lecturas que las soportan y luego filtran las trayectorias ligeramente soportadas. Por último, convierten trayectorias en secuencias y usan algoritmos de alineamiento de secuencias para colapsar trayectorias casi idénticas. Estos errores, en el caso de Velvet, se enfocan en características topológicas, donde los datos erróneos pueden crear tres tipos de estructuras: “puntas” (tips) debidas a errores en los bordes de las lecturas, salientes o protuberancias (bulges) debidas a errores de lectura internos o a conexiones entre puntas cercanas, y conexiones erróneas debidas a errores de clonado o a puntas distantes que se fusionan entre sí. Los tres rasgos son removidos consecutivamente (Zerbino & Birney, 2008). Las gráficas de K-mer son más sensibles al error de secuenciación, ya que cada base mal asignada introduce hasta k nodos erróneos; sin embargo, para mitigar esto, los enfoques OLC y DBG, ambos, emplean etapas de preproceso para filtrar o corregir porciones no

confirmadas de las lecturas, así como post-procesos para reparar gráficas por filtración, suavización y enhebramiento (Miller *et al.*, 2010).

- **Evaluando la calidad de las lecturas: FastQC**

Es importante mencionar que antes de hacer el ensamblaje del genoma, se debe verificar la calidad de las lecturas generadas por el secuenciador y antes de analizar la secuenciación para obtener conclusiones biológicas, se debería llevar a cabo el control de calidad, para asegurar que los datos crudos están bien. Muchos secuenciadores generarán reportes de control de calidad como parte de su canalización (pipeline) de análisis, pero esos reportes están usualmente enfocados en identificar problemas que son generados por el propio secuenciador. FastQC se enfoca en proveer un reporte de control de calidad (QC report) que puede poner en evidencia problemas que se originan ya sea en el secuenciador o en material de la biblioteca o librería de inicio. El análisis en FastQC se lleva a cabo por una serie de módulos de análisis de los datos de secuenciación. En el reporte de calidad se muestran un resumen de los módulos que fueron ejecutados y una evaluación rápida, que muestra si los resultados del módulo parecen enteramente normales (marca verde), ligeramente anormales (marca naranja) o muy inusuales (marca roja) (Figura 10). Estas evaluaciones deben ser tomadas en el contexto de lo que uno espera de su biblioteca o librería, ya que la información ayuda a encontrar orientaciones específicas sobre cómo interpretar los resultados de cada módulo. (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

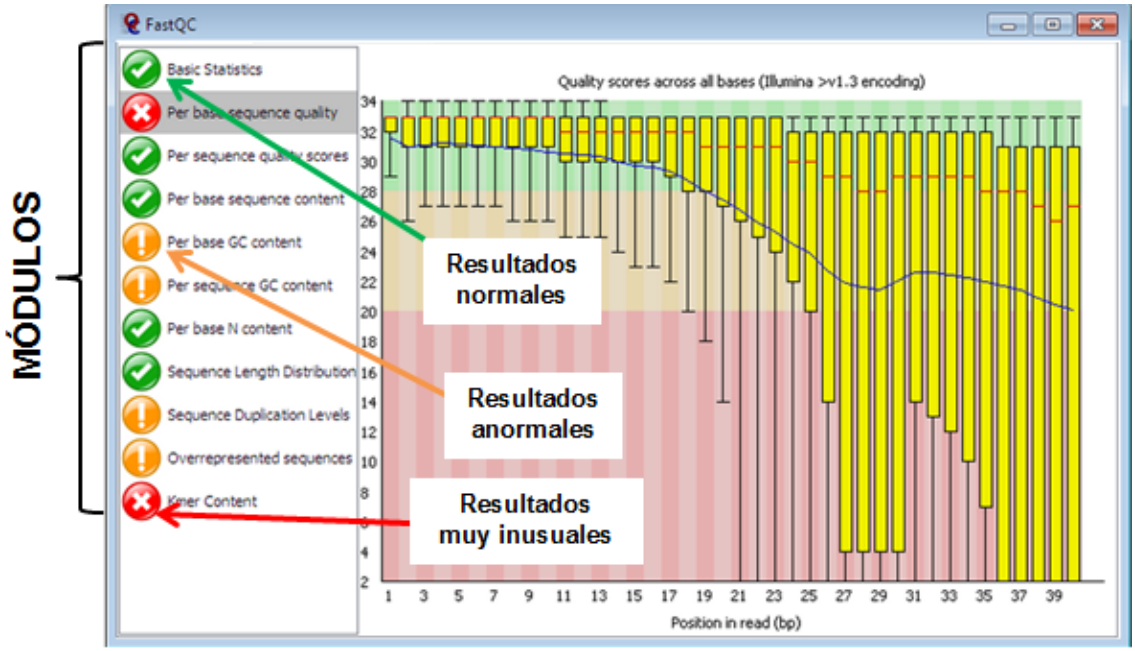



**Figura 10.** Visualización del reporte generado por FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

### 1.6 Otros programas

Existen otros programas que sirven de apoyo cuando se está trabajando en un ensamblaje.

#### 1.6.1 Mauve

Mauve permite alinear genomas para compararlos. Este programa representa un sistema de alineamiento que integra análisis de eventos evolutivos de gran escala con alineamiento de secuencias múltiples tradicionales. Mauve usa el anclaje como una heurística para incrementar la rapidez del alineamiento e identifica y alinea regiones de colinealidad local llamados bloques localmente colineales (locally collinear blocks o LCBs). Cada LCB es una región homóloga de secuencia compartida por dos o más de los genomas bajo estudio y no contiene ningún rearreglo de la secuencia homóloga (Darling *et al.*, 2004).

El “display” de alineamiento está organizado en un “panel” horizontal por cada secuencia de genoma introducida. Cada panel de genoma contiene el nombre de la secuencia de genoma, una escala que muestra las coordenadas de secuencia para ese genoma y una línea horizontal central única. Pueden aparecer siluetas de bloques coloreados arriba y posiblemente abajo de la línea central, donde, cada uno de estos trazos de bloques rodea una región de la secuencia del genoma que se alineó aparte de otro genoma y es presumiblemente homólogo e internamente libre de rearreglo genómico. Cuando un bloque se encuentra arriba de la línea central, la región alineada está en la dirección hacia adelante (forward), en relación a la secuencia del primer genoma. Los bloques debajo de la línea central indican regiones que se alinean en la orientación reversa (reverse). En el esquema de color estándar, la región de secuencia cubierta por un bloque coloreado es enteramente colineal y homóloga entre los genomas  Las límites de los bloques coloreados usualmente indican los puntos de interrupción del rearreglo del genoma. Las áreas que son completamente blancas no fueron alineadas y probablemente contienen elementos de secuencia específicos para el genoma particular. La altura del perfil de similitud corresponde al nivel promedio de conservación en esa región de la secuencia del genoma, está calculada para que sea inversamente proporcional a la entropía de la columna de alineamiento promedio sobre una región del alineamiento (Darling *et al.*, 2004; http://darlinglab.org/mauve/user-guide/viewer.html).

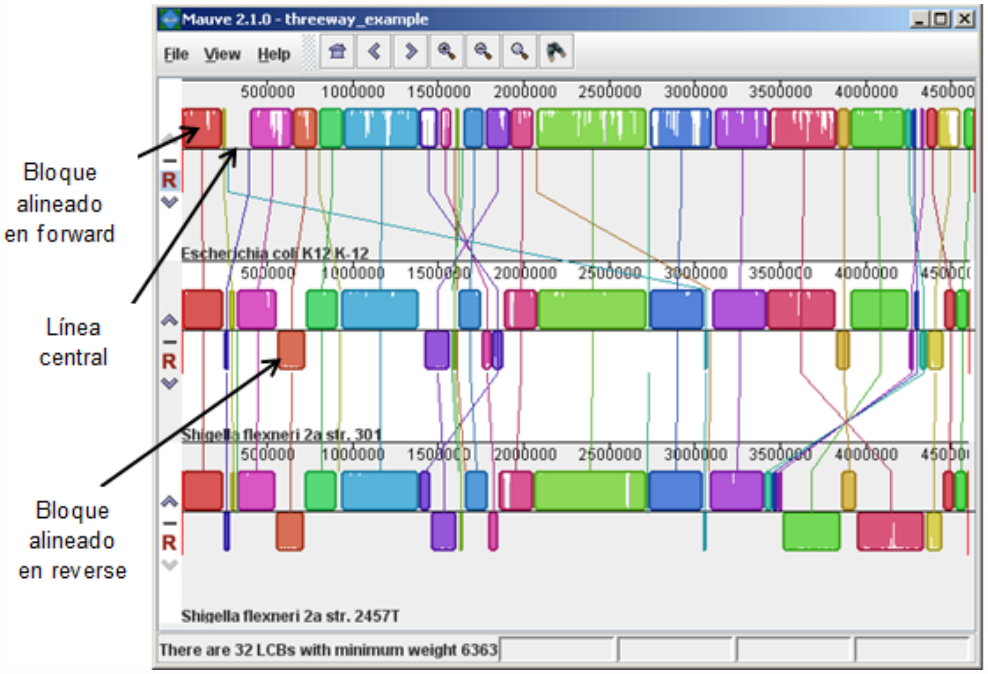



**Figura 11.** Alineamiento de *E. coli* K12 MG1655, *S. flexneri* 2a 301, y *S. flexneri* 2457. Las regiones invertidas en el genoma de *S. flexneri* están claramente dibujadas como bloques debajo de la línea central de cada genoma. Estos tres genomas fueron tomados del sitio NCBI FTP y alineados con Progressive Mauve usando parámetros de default. En esta figura, los bloques coloreados del primer genoma son conectados por líneas a bloques similarmente coloreados en el segundo y el tercer genoma. Estas líneas indican cuales regiones en cada genoma son homólogas. Note el patrón de cruce en X de las líneas que tienden a ocurrir en la vecindad del origen y término de la replicación predichos en estos organismos (http://darlinglab.org/mauve/user-guide/viewer.html).

Cuando un alineamiento ha sido computado con Progressive Mauve, que es la porción del programa que permite la visualización, está disponible un modo de display que colorea regiones conservadas entre todos los genomas, de modo diferente a las regiones conservadas entre subgrupos de los genomas. A las regiones conservadas entre todos los genomas se les denomina como “columna vertebral” y son dibujadas en color Malva (Mauve, en inglés) (http://darlinglab.org/mauve/user-guide/viewer.html).

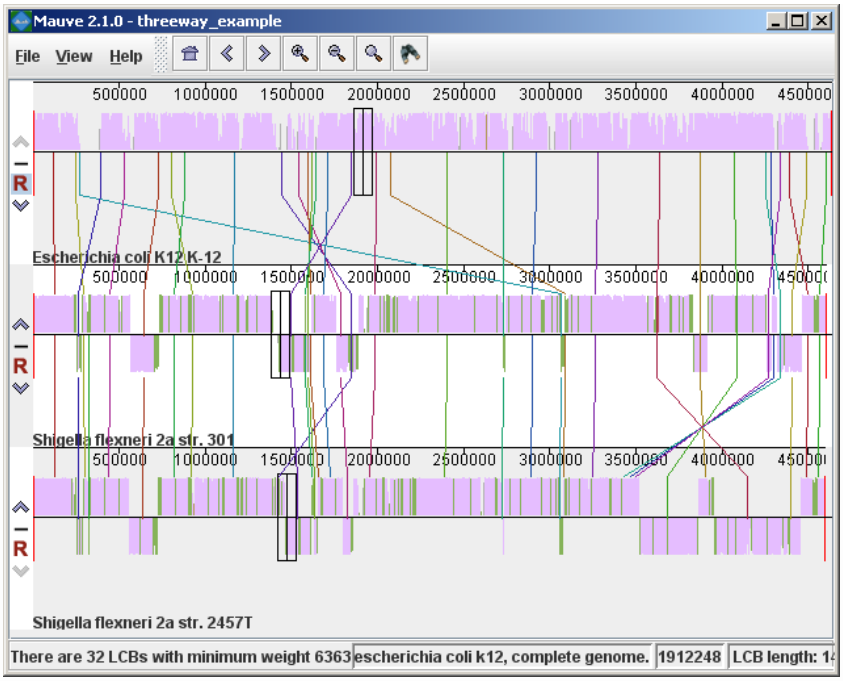



**Figura 12.** Alineamiento de *E. coli* K12 MG1655,*S. flexneri* 2a 301, y *S. flexneri* 2457T con diferentes settings de estilo. Las partes del despliegue de similitud que están en color Malva están conservadas entre los tres genomas. Mientras que las porciones en color verde son segmentos conservados solo entre *Shigella flexneri*. La caja negra sigue el cursor del mouse y resalta el sitio homólogo en cada genoma (http://darlinglab.org/mauve/user-guide/viewer.html).

### 1.6.2. Tablet

Tablet es un visualizador usado para el ensamblaje y alineamientos de secuencias. Provee visualizaciones de alta calidad mostrando los datos en vistas empacadas o apiladas, permitiendo acceso instantáneo y navegación a cualquier región de interés, así como visión de conjunto de contigs y resúmenes de datos. Tablet puede a la vez trabajar con multi-cores y es eficiente en el uso de la memoria, por lo que es capaz de manejar ensamblajes que contengan millones de lecturas, aún en una máquina de escritorio de 32 bits. Puede importar datos de los formatos de ensamblaje ACE, AFG, MAQ y SOAP (con un soporte preliminar para SAM), y puede manejar datos de 454 y Solexa. Sus visualizaciones se dividen en varias áreas; el “display” principal provee una visión de un contig único a la vez, con lecturas alineadas contra su secuencia

consenso. Las lecturas se colorean de acuerdo al tipo de nucleótido y un uso de los gradientes y del color permiten que la estructura visual se mantenga aun cuando se hagan máximos acercamientos (Milne *et al.*, 2009).

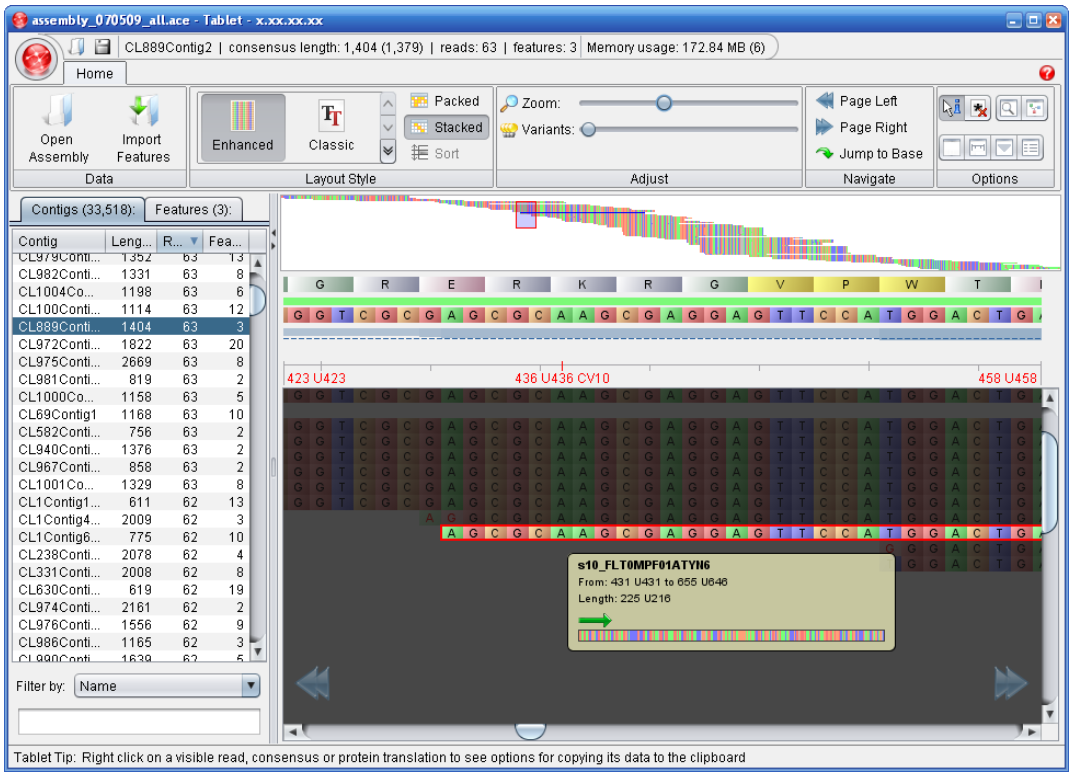



**Figura 13.** Visualización de cada uno de los nucleótidos contenidos en un contig seleccionado. Cada uno de ellos se presenta en un color diferente para hacer la distinción entre ellos (Milne *et al.*, 2009).

Tablet puede configurar los datos en dos formatos: empacados (mostrando tantas lecturas por línea como sea posible sin superponerlas) o apilados (mostrando una lectura por línea), y permite al usuario pasar instantáneamente de uno a otro. Una lista ordenable que contiene todos los contigs disponibles, muestra tanto la longitud de los contigs como los números de lecturas y aspectos de la anotación, y puede ser dinámicamente filtrada por cualquiera de sus campos. El acercamiento continuo del contig en tiempo real esta soportado por medio de un control deslizante, y hay también una opción para variar el contraste entre nucleótidos variantes y no variantes, la cual ajusta la brillantez para distinguir las bases de las lecturas que

difieren del consenso. La traducción de proteínas se proporciona opcionalmente para todos de los seis recuadros de lectura de la secuencia consenso (Milne *et al.*, 2009).

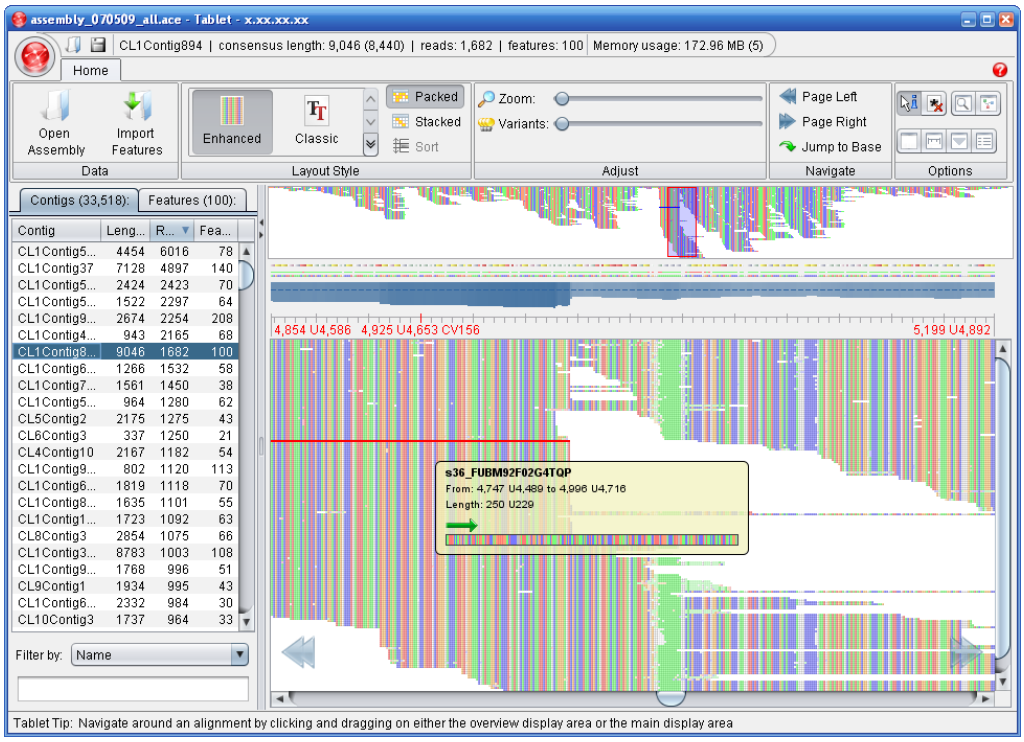



**Figura 14.** Contig pequeño generado con Illumina en un tipo de archivo BAM visto en Tablet. La parte oscura son las lecturas donde se pueden ver según la posición de sus nucleótidos. Se destaca además un resumen de la información de la lectura seleccionada, y se hace una representación gráfica de la secuencia (Milne *et al.*, 2009).

### 1.6.3. BWA (Burrows-Wheeler Aligner)

BWA es un paquete de software para mapear secuencias de baja diferencia contra un genoma de referencia. Consiste de tres algoritmos, denominados BWA-backtrack, BWA-SW y BWA-MEM. El primer algoritmo está diseñado para lecturas de secuencia de Illumina, hasta de 100 pb, mientras que los restantes dos lo están para secuencias más largas que van de 70 pb hasta 1 Mbp (http://bio-bwa.sourceforge.net/). BWA-MEM y BWA-SW comparten características similares tales como soporte para lecturas largas y alineamientos divididos, pero BWA-MEM, el cuál es el más reciente, está generalmente recomendado para búsquedas de alta

calidad, ya que es más rápido y más preciso. BWA-MEM también tiene un mejor desempeño que BWA-backtrack para lecturas de 70-100 bp de Illumina (Li & Durbin, 2009).

### 1.7 Anotación

Una vez que se ha ensamblado un genoma, el siguiente paso es hacer una anotación del mismo, es decir, hacer un marcado de los genes y otras características biológicas en la secuencia de ADN, por medio de un software que permite localizar e identificar los genes y sus potenciales funciones. La anotación de un genoma es un proceso iterativo que utiliza diversas bases de datos (Figura 14). La existencia de una ruta de reacción conocida provee información muy valiosa que soporta la inferencia de una función a través de un proceso de eliminación sistemático (Overbeek *et al.*, 2004).

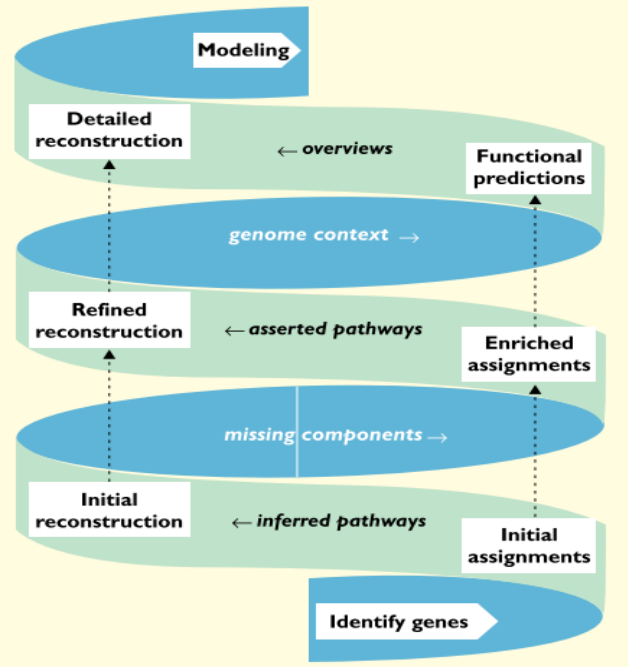



**Figura 15.** El proceso iterativo de anotación del genoma (Overbeek *et al.*, 2004).

Los métodos computacionales para determinar la función de genes se basan en similitudes de las secuencias de genes cuya función ha sido determinada

experimentalmente. En cuanto a los métodos de predicción de funciones, se pueden ampliar usando enfoques de análisis de contexto de genes, tales como examinar la conservación de clústeres de genes cromosomales, eventos de fusión de genes y perfiles de co-ocurrencia a través de genomas. El análisis de contexto se basa en la observación de que los genes funcionalmente relacionados tienen frecuentemente un contexto genético similar y descansan en la identificación de tales eventos a través de una colección de genomas filogenéticamente diversos. De esta manera, dependiendo del nivel de similitud de secuencia entre genes recién secuenciados y genes con función determinada por métodos experimentales, las funciones predichas con tales métodos van desde asignación de categorías funcionales “gruesas”, por ejemplo, “familia de proteínas aminotransferasas”, hasta descripciones muy precisas de la función, por ejemplo, “alanina aminotransferasa”. Los genes sin similitud con genes caracterizados funcionalmente se dice que tienen una función desconocida y son frecuentemente rotulados como “proteínas hipotéticas” (Mavromatis *et al.*, 2009).

### 1.7.1 Algunos programas para hacer la anotación

Tal y como se presentan diferentes opciones para ensamblar un genoma, también las hay para hacer la anotación. A continuación se mencionan algunas de las herramientas usadas para la anotación de genomas.

#### 1.7.1.1 RAST (Rapid Annotations using Subsystems Technology)

RAST es un servicio automatizado para la anotación de genomas completos, o casi completos de bacterias y de arqueas. Identifica genes que codifican para proteínas, rRNS y tARN; asigna funciones a los genes, predice cuáles subsistemas están representados en el genoma, usa esta información para reconstruir la red metabólica

y hace que el resultado sea fácilmente descargable por el usuario. Además, el genoma anotado puede ser visto y analizado desde el ambiente SEED (http://pubseed.theseed.org; Aziz *et al.*, 2008), el cual es una integración de datos de genomas constantemente actualizada, con una base de datos de genomas. El ambiente SEED también aloja subsistemas (acopios de familias de proteínas relacionadas funcionalmente) y sus FIGfams (familias de proteínas) derivadas, las cuales representan el núcleo del motor de anotación de RAST (Figura 16) (Overbeek *et al.*, 2014). Cuando un nuevo genoma es sometido a RAST, los genes son asignados y sus anotaciones son hechas por comparación con la colección de FIGfams. Si el genoma se hace público, será alojado dentro de SEED y sus proteínas pasan a formar parte de la colección de FIGfams (Aziz *et al.*, 2008; Overbeek *et al.*, 2014).

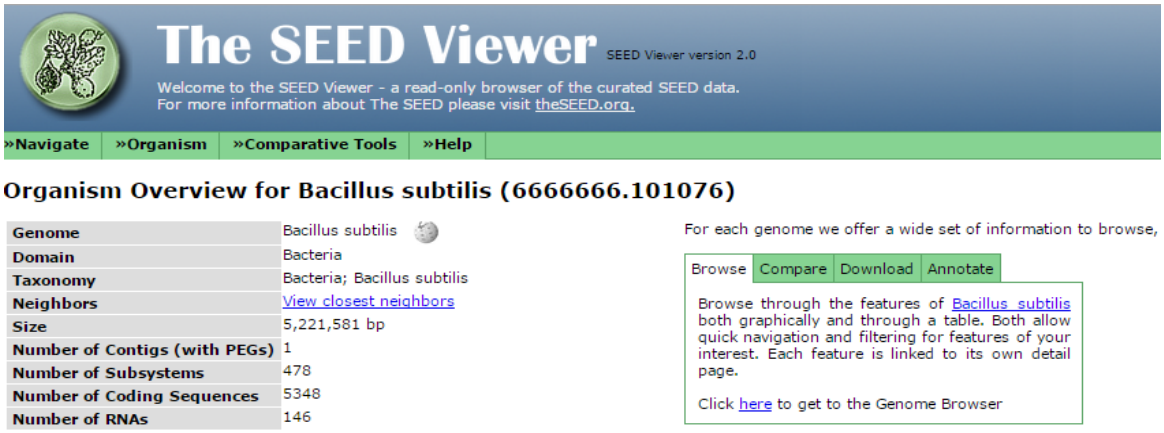



**Figura 16.** Resultados de una anotación realizada en RAST donde se puede visualizar con el ambiente SEED.

El servicio normalmente tiene disponible el genoma anotado después de 12-24 horas de haber sido sometido. RAST trata de producir evaluaciones de genes de alta calidad y una reconstrucción metabólica inicial, basándose en el uso de una librería creciente de subsistemas que son conservados manualmente y en familias de proteínas principalmente derivadas de esos subsistemas (FIGfams). De esta manera,

tenemos que RAST produce automáticamente dos clases de funciones declaradas de genes: las primeras, basadas en subsistemas, están basadas en el reconocimiento de las variantes funcionales de los subsistemas, y las no basadas en subsistemas, las cuales se llenan usando enfoques más comunes para la integración de la evidencia obtenida a partir de un cierto número de herramientas. El hecho de que RAST distinga entre estas dos clases de anotación y use la basada en subsistemas, que es relativamente más confiable, hace los resultados de RAST un esfuerzo integral de anotación excepcionalmente bueno. Además, este servicio provee de un ambiente para examinar el genoma anotado y compararlo con cientos de genomas mantenidos dentro de la integración SEED, a través de su visor de genomas, que además permite la determinación de genes que el genoma tiene en común con grupos específicos de otros genomas (o de genes que lo distinguen de esos grupos específicos), confiere la habilidad para desplegar el contexto genómico alrededor de genes específicos y la habilidad de descargar información relevante y anotaciones (Aziz *et al.*, 2008).

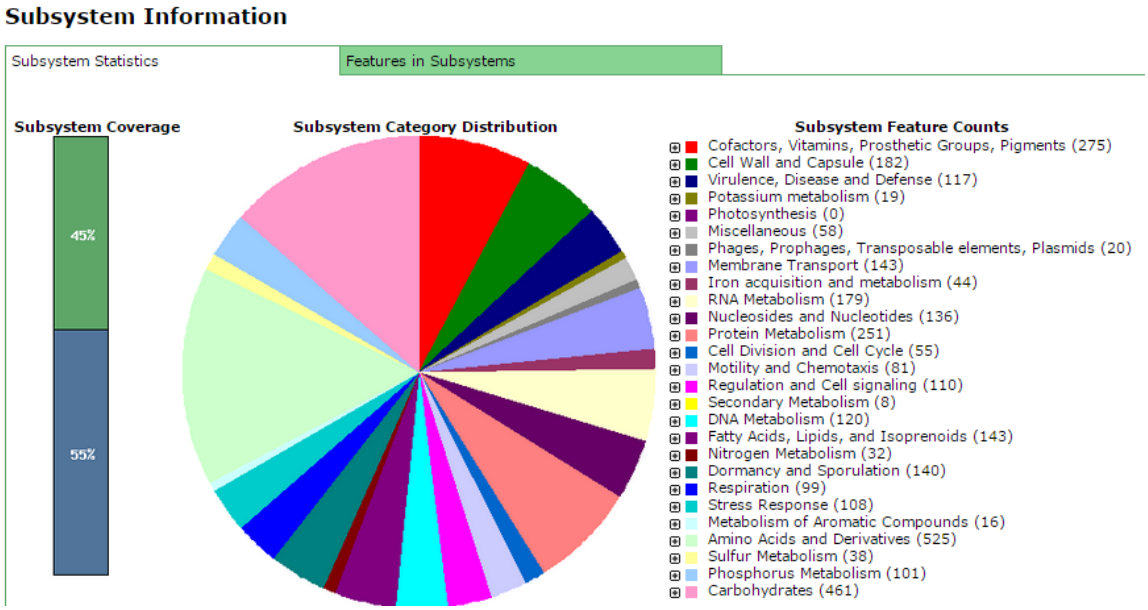



**Figura 17.** Genes conectados a subsistemas y su distribución en categorías diferentes. Las categorías se pueden expandir hasta el gen específico.

### 1.7.1.2 IMG (Integrated Microbial Genomes)

El Sistema Genomas Microbianos Integrados (Integrated Microbial Genomes o IMG) integra genomas, ya sean incompletos, así como genomas completos, de los tres dominios de la vida (Bacteria, Archaea y Eucarya) con un gran número de plásmidos y virus (Markowitz *et al.*, 2012). IMG emplea recursos de NCBI (National Center for Biotecnology Information) (Pruitt *et al.*, 2009) como su fuente principal de datos de secuencias de genomas y anotaciones "primarias" de genes predichos y sus proteínas producidas. Para cada genoma, IMG registra la información de la secuencia del genoma, incluyendo su organización en replicones cromosomales (para genomas terminados) y scaffolds y/o contigs (para genomas incompletos), junto con las secuencias codificantes de proteínas (protein-coding sequences o CDS) predichas, algunos genes codificantes de ARN y los nombres de los productos de proteínas que son proporcionados por los centros de secuencia de genomas (Figura 18) (Markowitz *et al.*, 2012). El pipeline (canalización) de integración de datos de IMG asocia cada genoma con metadatos de GOLD (Genomes OnLine Database) (Liolios *et al.*, 2010), y rellena cualquier información adicional potencialmente faltante de los archivos de referencia, tales como repeticiones CRISPR (Clustered Regularly Interspaced Palindromic Repeats) (Bland, *et al.*, 2007), señales de péptidos computadas usando SignalP (Emanuelsson *et al.*, 2007) y hélices transmembrana usando TMHMM (Transmembrane Helices Hidden Markov Models) (Moller *et al.*, 2001). Los ARN´s faltantes son identificados usando tRNAscan-SE-1.23 (Lowe & Eddy, 1997) para ARNs, HMMs (Hidden Markov Models) desarrollados de modo propio para ARNr (Lagesen *et al.*, 2007), y Rfam (Annotating non-coding RNAs in complete genomes) (Griffiths-Jones *et al.*, 2005), así como INFERNAL v1.0

(Inference of RNA Alignments) (Nawrocki *et al.*, 2009) para otros ARNs pequeños. Los genes son asociados con anotaciones funcionales “secundarias” y listas de genes relacionados (es decir homólogos, o parálogos). Las anotaciones generadas por IMG consisten en familias de proteínas y caracterizaciones de dominio basadas en clusters de grupos ortólogos (COGs) y en categorías funcionales (Tatusov *et al.*, 2003), Pfam (Protein Families Database) (Finn *et al.*, 2010), categorías de rol TIGRfam y TIGR (Selengut *et al.*, 2007), dominios InterPro (Hunter *et al.*, 2005), términos de Ontología de Genes (Genes Ontology o GO) (Gene Ontology Consortium, 2008) así como términos y trayectorias KEGG (Kyoto Encyclopedia of Genes and Genomes) Ortholog (KO) (Kanehisa *et al.*, 2010).

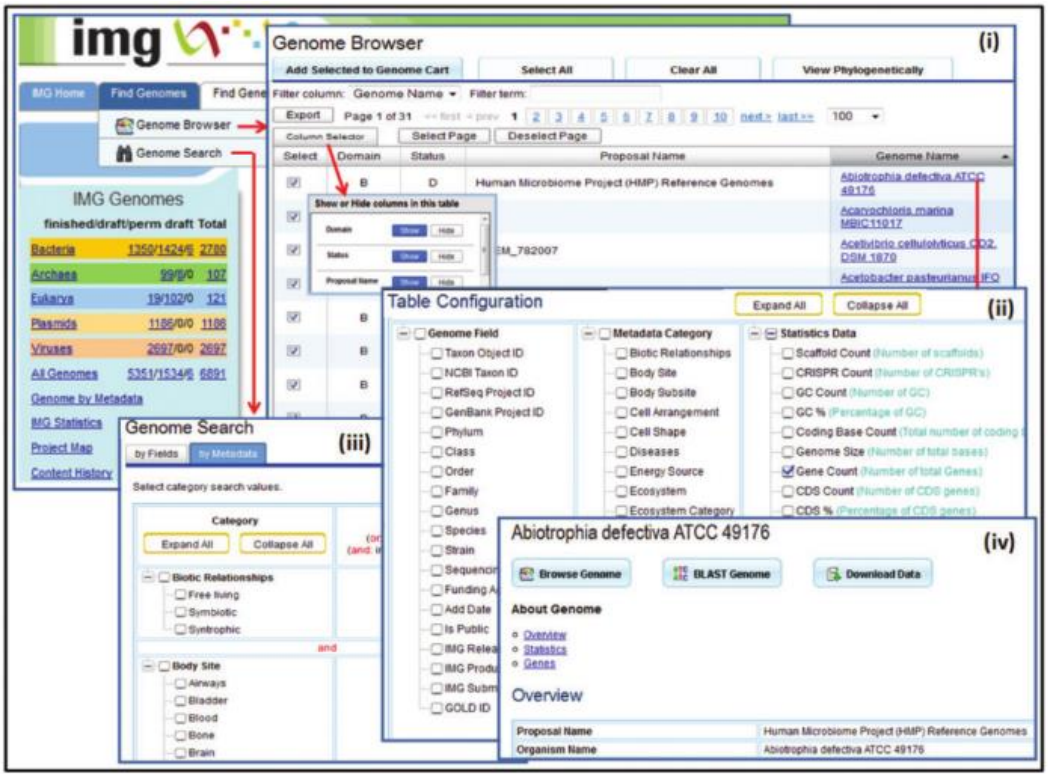



**Figura 18**. Herramienta exploradora de genomas y de búsqueda en IMG. El “Explorador de genomas” presenta los genomas dispuestos en un árbol filogenético o (i) en una lista tabular que puede ser configurada por (ii) el agregado o la remoción de columnas específicas de genomas, metadatos o anotaciones. (iii) El “buscador de genomas“ permite buscar genomas en campos específicos de genomas o metadatos (iv) Un genoma puede ser explorado usando una variedad de herramientas de exploración; examinado buscando la presencia de genes específicos usando BLAST o; descargado (Markowitz *et al.*, 2012).

Para cada gen, IMG provee listas de genes relacionados (esto es, candidatos homólogos, parálogos y ortólogos) que están basados en similitudes de secuencias computadas usando el BLASTp (Basic Local Alignment Search Tool for proteins) de NCBI (National Center for Biotechnology Information) para genes codificantes de proteínas y en BLASTn (Basic Local Alignment Search Tool for nucleotide)para genes de ARN. Tales listas de genes pueden ser filtradas usando el porcentaje de identidad, puntuación de bit (bit score) y valores E (E-Values) más estrictos.

### 1.8 Predicción de funciones en un microorganismo a partir de la anotación

Como se mencionó anteriormente, una vez que se han anotado los genes de un genoma, con esta información se puede hacer una predicción de las posibles funciones fisiológicas del microorganismo a analizar. Dentro de estas funciones, se puede analizar todo aquello de lo que presuntamente es capaz el microorganismo, por los genes que presenta, por ejemplo, si posee genes que codifican para la producción de xilanasas, amilasas, reductasas, deshidrogenasas, fijación de nitrógeno, o bien genes relacionados con la resistencia a compuestos tóxicos, como antibióticos y metales (Aziz *et al.*, 2008).

En cuanto a metales, algunos microorganismos pueden tener genes de resistencia a metales pesados, lo cual resulta interesante, pues estos componentes son un gran problema ambiental, ya que se bioacumulan en los organismos, lo que causa daño en las diferentes cadenas tróficas y eventualmente en los humanos, ya que no contamos con vías para su biodegradación (Bruins *et al.*, 2000; Choudhury & Srivastava, 2001).

## 1.9 Metales pesados y metaloides

En general, los metales juegan un papel integral en los procesos vitales de los microorganismos, ya que funcionan como catalizadores para reacciones bioquímicas, son estabilizadores de estructuras proteicas y paredes celulares bacterianas y sirven para mantener el balance osmótico. Por ejemplo, los metales de transición esenciales, como el hierro (Fe), el cobre (Cu) y el níquel (Ni), están involucrados en procesos redox. Otros metales esenciales, como el magnesio (Mg) y el zinc (Zn), estabilizan varias enzimas y al ADN a través de fuerzas electrostáticas. El hierro, el magnesio, el níquel y el cobalto son parte de moléculas complejas, teniendo una amplia gama de funciones, y el potasio y el sodio son requeridos para la regulación de la presión osmótica intracelular (Bruins *et al.*, 2000).

Es importante resaltar que existen otros metales que no tienen un papel biológico, como plata (Ag), aluminio (Al), cadmio (Cd), oro (Au), plomo (Pb) y mercurio (Hg) (Bruins *et al.*, 2000), debido a esto, tanto procariontes como eucariontes han desarrollado un amplio espectro de rutas tales como la oxidación/reducción, la compartamentalización, la exclusión y la inmovilización, como los mecanismos de defensa natural contra los compuestos tóxicos (Tsai *et al.*, 2009), como es el caso de los metales pesados.

### 1.9.1 Toxicidad de metales pesados

Los metales pesados son considerados compuestos tóxicos, se les llama pesados debido a que tienen la característica de presentar una alta densidad (superior a 5g/cm$^3$), y generalmente ejercen una acción inhibitoria en los microorganismos, bloqueando grupos funcionales esenciales, desplazando otros iones metálicos esenciales o modificando las conformaciones activas de moléculas biológicas; sin

embargo, en relativamente bajas concentraciones algunos metales pesados son esenciales para los organismos, (por ejemplo Co, Cu, Zn, Ni) ya que proveen cofactores vitales para metalo-proteínas y enzimas (Hassen *et al.*, 1998).

La toxicidad ocurre por el desplazamiento de metales esenciales de sus sitios de unión nativos o a través de interacciones de ligandos. Los metales no esenciales se unen con mayor afinidad a grupos que contienen tioles y a grupos con oxígeno, desplazando así a los metales esenciales. La toxicidad también resulta de alteraciones en la estructura conformacional de ácidos nucleicos y proteínas; así como por interferencia con la fosforilación oxidativa y el balance osmótico. Las bacterias se han adaptado a la presencia de metales pesados a través de una variedad de sistemas de resistencia mediados a nivel de cromosoma, de transposones y de plásmidos. (Bruins *et al.*, 2000).

Metales como el hierro (Fe), el cobre (Cu) el cromo (Cr), el vanadio (V) y el cobalto (Co) llevan a cabo reacciones cíclicas de óxido-reducción; por otro lado, para un segundo grupo de metales, mercurio (Hg), cadmio (Cd), y níquel (Ni); la ruta primaria para su toxicidad es el agotamiento del glutatión y la adhesión a grupos sulfhidrilos de las proteínas. El arsénico (As) se adhiere directamente a grupos tiol; sin embargo, han sido propuestos otros mecanismos, que involucran la formación de peróxido de hidrógeno bajo condiciones fisiológicas (Bruins *et al.*, 2000; Xiong & Jayaswal, 1998).

Los iones de metales pesados traza, tales como el cobalto (Co), el zinc (Zn), el cobre (Cu) y el níquel (Ni), juegan papeles importantes en las bacterias. Ellos regulan una amplia gama de funciones metabólicas como coenzimas, cofactores, catalizadores, ácidos de Lewis en enzimas y estabilizadores estructurales de enzimas y proteínas que se adhieren al ADN (Choudhury & Srivastava, 2001). Sin embargo, al aumentar

su concentración estos metales pueden producir daños a los microorganismos, de esta manera, éstos han desarrollado mecanismos para regular los procesos de entrada y salida de estos compuestos para mantener el nivel intracelular relativamente constante de los iones de metales pesados y así poder resistir su presencia (Xiong & Jayaswal, 1998).

### 1.9.2 Resistencia a metales pesados

Una célula puede desarrollar sistemas de resistencia a metales en un intento por proteger componentes celulares sensibles. El limitar el acceso del metal o la alteración de componentes celulares disminuye la sensibilidad a metales. Varios factores determinan la existencia de resistencia en un microorganismo: el tipo y número de mecanismos de captación del metal, el papel que cada metal juega en el metabolismo normal, y la presencia de genes localizados en plásmidos, cromosomas o transposones que controlan la resistencia al metal (Choudhury & Srivastava, 2001). Muchos microorganismos demuestran resistencia a metales en agua, suelo y desechos industriales. Los genes localizados en cromosomas, plásmidos y transposones codifican resistencia específica a una variedad de iones metálicos. (Bruins *et al.*, 2000).

Para entender la interacción metal-microorganismo, es necesario hacer una distinción entre dos tipos de metales pesados: metales pesados que son tóxicos *per se* y metales que son esenciales para el crecimiento y la manutención, pero que son tóxicos cuando están en exceso. Para la resistencia a metales que son fisiológicamente requeridos, la supervivencia se optimiza por cooperación entre el mecanismo de resistencia y el metabolismo celular normal del metal, permitiéndole a la célula acumular suficiente metal para el mantenimiento de actividades

dependientes a ese metal, mientras que se reacciona contra concentraciones del metal supra-óptimas (Choudhury & Srivastava, 2001).

Los mecanismos moleculares involucran un número de proteínas, tales como trasportadores de iones, reductasas, fitoquelatinas (cadistinas) asociadas a glutatión y metalotioneínas ricas en cisteína, así como ligandos para metal de bajo peso molecular ricos en cisteína. Estas proteínas exportan los iones metálicos fuera de las células o los detoxifican o secuestran de tal manera que las células puedan crecer en un ambiente que contienen altos niveles de metales tóxicos. Sin embargo, no hay un mecanismo común de resistencia a todos los iones de metales pesados. En las bacterias, los genes que codifican la resistencia a metales pesados se localizan ya sea en el cromosoma bacteriano, en los plásmidos o en ambos. (Xiong & Jayaswal, 1998). Hay seis mecanismos generalmente propuestos para la resistencia a metales pesados en las bacterias y otros microorganismos: (a) exclusión del metal por medio de una barrera de permeabilidad; (b) exclusión por bombas de evacuación con transporte activo; (c) secuestro intracelular físico por proteínas de unión u otros ligandos, para prevenir que dañe los sitios celulares sensibles al metal; (d) secuestro extracelular; (e) detoxificación enzimática del metal hasta una forma menos tóxica y (f) reducción de la sensibilidad de los sitios celulares sensibles a los iones metálicos. Estas estrategias se reflejan en el fenotipo de resistencia a uno o varios metales (Bruins *et al.*, 2000; Choudhury & Srivastava, 2001).

### 1.8.3 Características de algunos metales pesados

#### 1.8.3.1 Arsénico

El arsénico (As) es un metaloide tóxico que está presente en muchos ambientes y es liberado a través de procesos de origen natural o antropogénico. La OMS estableció

0.05 mg/l como límite permitido de arsénico en el agua para consumo humano. El arsénico existe en varios estados de oxidación, que incluyen al arsenato As (V), arsenito As (III), arsénico elemental As (0) y arsenuro As (-III) (Tsai *et al.*, 2009). Los dos estados de oxidación más comunes del arsénico soluble en la naturaleza son As (V) y As (III), presentes como los oxianiones: arsenato ($AsO_4^{3-}$) y arsenito ($AsO_3^{3-}$), respectivamente. Aunque el arsenato y el arsenito son ambos tóxicos para los sistemas biológicos, inducen diferentes tipos de daño celular. Debido a su analogía estructural al fosfato inorgánico, el arsenato puede entrar a la célula vía los sistemas de transporte de membrana para fosfato e interrumpir las reacciones metabólicas que requieren fosforilación. En contraste, el arsenito es transportado al interior de la célula por acuagliceroporinas, en bacterias, levaduras y mamíferos, y ejerce su toxicidad por adhesión a grupos tiol en proteínas, con lo que debilita su función.

El arsenito es mucho más toxico que el arsenato, sin embargo la resistencia al As (V) requiere su reducción a As (III), el cual es expulsado después. (Achour *et al*., 2007). Es posible que el sistema de eflujo de As (III) haya evolucionado primero bajo condiciones reductoras, y después fue acoplado con la reducción de As (V) para acomodarse a la toxicidad de As (V) una vez que la atmósfera de la tierra se volvió más oxidada (Tsai *et al.*, 2009). Por otro lado, la oxidación de arsenito puede servir como fuente de energía en microorganismos quimiolitotróficos. Las bacterias que pueden metabolizar elementos tóxicos representan por lo tanto una herramienta atractiva para restaurar sitios contaminados (Muller *et al.*, 2007).

Muchas bacterias Gram-positivas y Gram-negativas emplean mecanismos similares de resistencia, como puede ser la oxidación del arsenito o metilación hacia compuestos menos tóxicos, así como extrusión activa del arsenito de la célula. Los

genes que codifican la maquinaria de detoxificación del arsénico (genes *ars*) están ampliamente distribuidos en las bacterias y las arqueas y pueden ser hallados en plásmidos y cromosomas. Comúnmente consisten ya sea de tres (*arsRBC*) o cinco (*arsRDABC*) genes arreglados en una unidad transcripcional única (Tsai *et al.*, 2009).

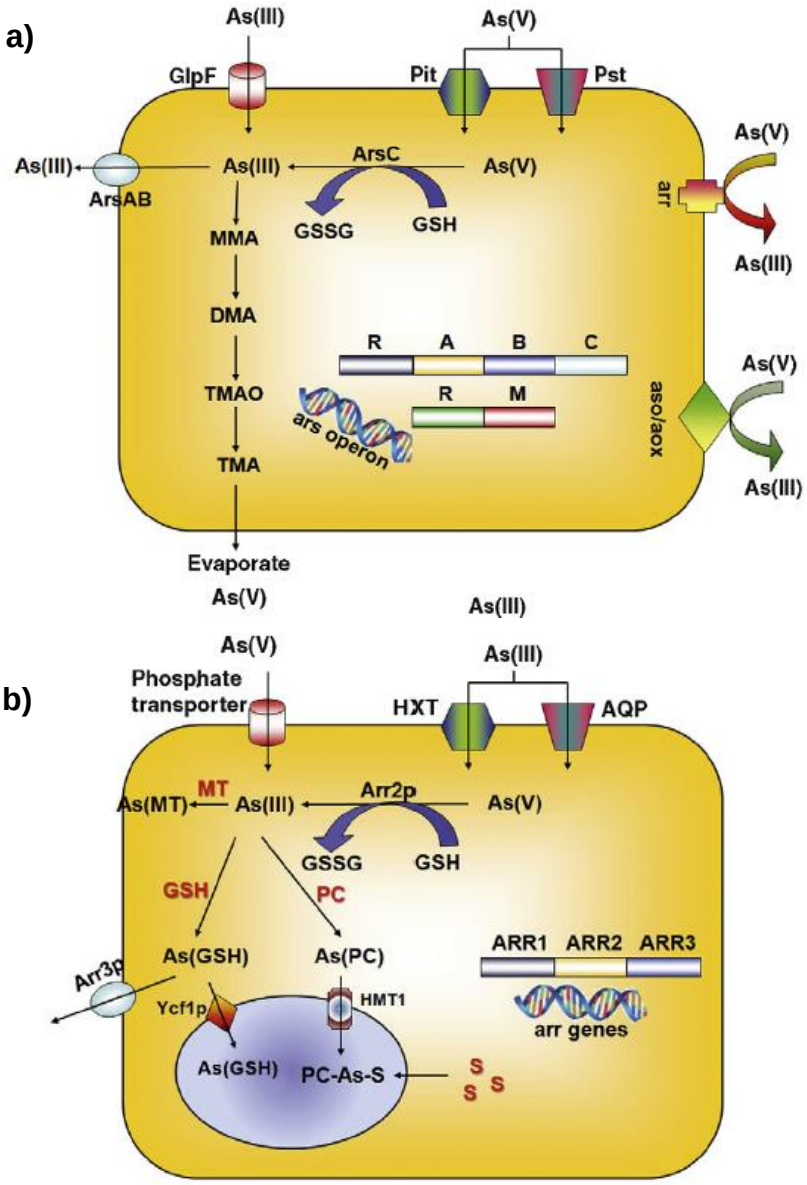



**Figura 19 .**Representación esquemática de los procesos involucrados en el metabolismo del arsénico en el ambiente de (a) procariontes y (b) eucariontes. En ambos casos, el arsénico entra a las células a través de transportadores. El arsenato es reducido a arsenito por una reductasa, y posteriormente es extruido fuera de la célula por una bomba de membrana específica. En eucariontes, el arsenito puede también ser detoxificado por formación de complejos con péptidos ricos en Cys tales como las fitoquelatinas y almacenados en la vacuola. Además, el arsenito puede servir como donador de electrones por oxidación a arsenato. El arsenato puede ser usado como el aceptor de electrones último durante la respiración y el arsénico inorgánico puede también ser transformado en especies orgánicas en una familia de metilación en cascada (Tsai *et al.*, 2009).

*ArsB* es una proteína integral de membrana que bombea el arsenito fuera de la célula, está frecuentemente asociada con una unidad de ATPasa; *ArsA,* la cual proporciona la energía necesaria a *ArsB,* vía hidrólisis de ATP. *ArsC* es una arsenato reductasa que convierte el arsenato a arsenito antes de expulsarlo. *ArsR* es un represor trans-actuante involucrado en la regulación basal del operón *ars*, mientras que *ArsD* es un segundo represor que controla los niveles altos de la expresión de los genes *ars* (Figura 19) (Achour *et al.*, 2007).

**1.8.3.2 Cadmio.**

El cadmio es un metal no esencial que es tóxico a baja concentración, la FDA ha determinado que los niveles de cadmio en el agua en botella no deben exceder 0.005 mg/L, por lo que representa un riesgo importante para la salud humana debido a que puede provocar estrés oxidativo, cambios en la actividad de muchas enzimas, interacciones con biomoléculas incluyendo ADN y ARN y los riesgos potenciales consecuentes; haciendo muy importante su oportuna detección. Este metal se encuentra comúnmente presente en la biósfera a concentraciones que se aproximan a 0.01 a 1.8 ppm y está frecuentemente asociado con minerales de zinc (Sochor *et al.*, 2011). En bacterias y otras células, el cadmio entra a través de sistemas de transporte de iones divalentes y puede ser co-transportado con magnesio. Los sistemas de transporte de iones divalentes son normalmente requeridos para transportar metales esenciales tales como magnesio, así como fosfato y sulfato. Una consecuencia adversa de esto es el co-transporte de otros cationes que pueden ser tóxicos al organismo (Sochor *et al.*, 2011). Dentro de la célula, el Cd(II) puede enlazarse a grupos sulfhidrilos en proteínas esenciales, interfiriendo con funciones celulares importantes y puede también causar ruptura de una hebra del ADN (Bruins,

2000). La ATPasa de resistencia a $Cd^{2+}$ de bacterias Gram-positivas (*Cad*A) es una bomba de membrana de cationes homóloga con otras ATPasas tipo P bacteriales, animales y de plantas. La *Cad*A provoca una incorporación dependiente de ATP, de $Cd^{2+}$ (y $Zn^{2+}$) por vesículas de membrana de adentro hacia afuera. El gen *czc* (y *cnr* así como *ncc*) es otro gen de bacterias Gram negativas que confieren resistencias a $Cd^{2+}$, $Zn^{2+}$, $Co^{2+}$ y $Ni^{2+}$. Estos iones son bombeados por un complejo de tres polipéptidos de membrana, que no es una ATPasa, pero que funciona como un transportador del catión $H^{+}$. El complejo consiste de una proteína de membrana interna (*Czc*A), una proteína de membrana externa (*Czc*C) y una proteína asociada con ambas membranas (*Czc*B) (Figura 20) (Silver, 1996).

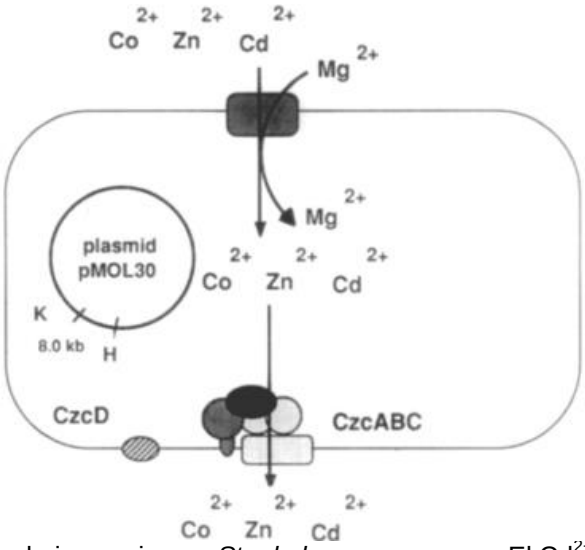



**Figura 20.** Resistencia a cadmio y a zinc en *Staphylococcus aureus*. El $Cd^{2+}$ es transportado dentro de la célula por un sistema de asimilación de manganeso y el $Zn^{2+}$ es transportado por un sistema desconocido (posiblemente un sistema de asimilación de magnesio). En presencia del plásmido ~125 8 (que contiene el determinante *cadA* localizado en un fragmento 3.5kb Bg/II-XbaI), el complejo de proteína *Cad*CA es inducido y activamente transporta ambos cationes fuera de la célula (Nies, 1992).

### 1.8.3.3 Cobalto.

El cobalto es requerido como elemento traza en procariontes para llevar a cabo una variedad de funciones metabólicas, pero altas concentraciones intracelulares de este metal de transición son tóxicas. Una de las estrategias de las bacterias para prevenir

el daño es exportar el metal en exceso por medio de sistemas de eflujo (Rodrigue *et al.*, 2005). El sistema *czc* es un sistema de eflujo que remueve Co(II), Zn(II), y Cd(II) que entra a las bacterias en una importación con Mg(II). El operón contiene varios genes: *Czc*A que es esencial para el transporte de catión y es la única proteína de las cuatro capaz de formar un túnel de membrana, además es capaz de mediar bajos niveles de resistencia a Co(II), Zn(II), y Cd(II) y es la más grande de las proteínas codificadas por el operón *czc. Czc*B transborda los metales de la membrana interna a la externa , previniendo el contacto con el periplasma; *Czc*C se soporta en *Czc*B para funcionar y puede actuar como un sustrato [Cd(II), Zn(II), y un interruptor Co(II)] para la bomba de eflujo y *Czc*D que se encarga de regular los genes de *czc* (Bruins *et al.*, 2000; Nies, 1992).

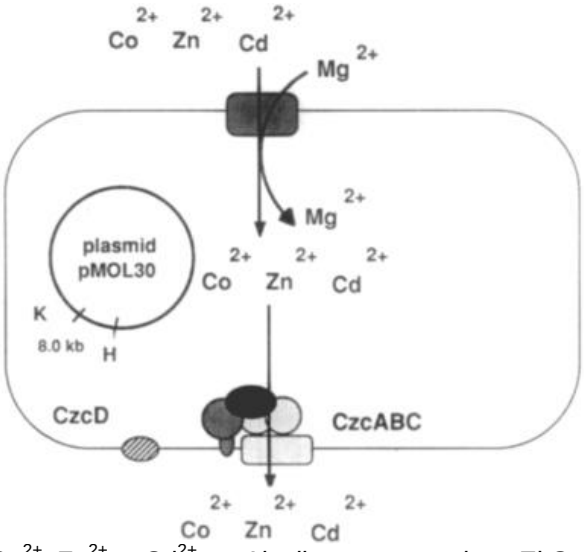



**Figura 21.** Resistencia a $Co^{2+}$, $Zn^{2+}$, y $Cd^{2+}$ en *Alcaligenes eutrophus*. El $Co^{2+}$, el $Zn^{2+}$, y el $Cd^{2}$ son transportados dentro de la célula por un sistema de asimilación de magnesio. En presencia del plásmido pMOL30 (que contiene el determinante *czc* localizado en un fragmento 8.0-kb KpnI-BarnHI), el complejo de proteína *Czc*CBA es inducido y activamente transporta ambos cationes fuera de la célula, *Czc*D está involucrado en la regulación de *czc* (Nies, 1992).

### 1.8.3.4 Cobre.

El cobre está ampliamente distribuido en la naturaleza y es frecuentemente encontrado en la corteza terrestre. Es un elemento traza esencial para los

organismos, jugando un papel importante en un gran número de procesos biológicos (Altimira *et al.*, 2012).

Los genes que confieren resistencia al cobre en las bacterias están frecuentemente presentes en plásmidos y organizados en un operón. La resistencia a cobre está codificada por los genes *cop* (*cop*ABCD) en varios Gram-negativos (Figura 21) y por los genes *pco* (*pco*ABCD) en *Escherichia coli* cepa RJ92. El gen *cop*A codifica a una oxidasa multi-cobre (*pco*A gen en *E. coli*), es uno de los principales genes para la resistencia al cobre en bacterias Gram-negativas, codifica la oxidasa multi-cobre que oxida Cu(I) a la forma química menos tóxica de Cu(II); *Cop*B, es una proteína de membrana exterior involucrada en el transporte de cobre a través de la membrana; *Cop*C es una proteína periplásmica capaz de adherir cobre y *Cop*D está asociada con la membrana interior donde participa en el transporte de cobre (Altimira *et al.*, 2012; Cervantes & Gutiérrez-Corona, 1994; Nies, 1992).

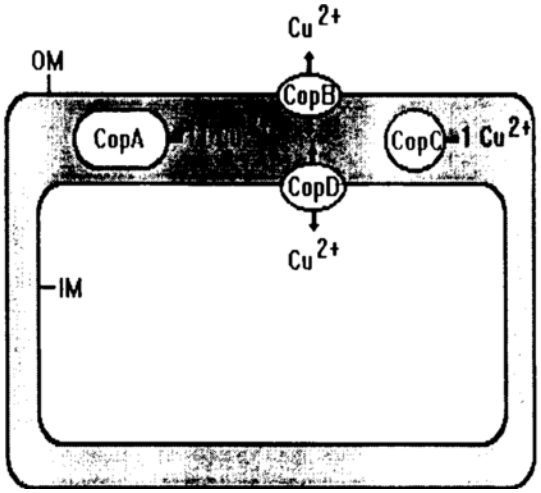



**Figura 22.** Locaciones celulares de los polipéptidos codificados por el determinante de resistencia al cobre del plásmido de *Pseudomonas wringae*. Las proteínas de la membrana externa (OM) y de la membrana interna (IM), *Cop*B y *Cop*D, respectivamente, participan en el transporte de cobre y/o en el eflujo. *Cop*A y *Cop*C son proteínas periplásmicas capaces de adherir iones cobre (Cervantes & Gutiérrez-Corona, 1994).

Otros sistemas de asimilación de cobre incluyen *Cut*A y *Cut*B. *Cut*A está además involucrado en la asimilación de zinc tanto como lo está en el transporte de cobre.

Además incluyen dos proteínas intracelulares de acarreo/almacenamiento de cobre (*Cut*E y *Cut*F), que pueden ser responsables de proteger a la célula bacteriana de la toxicidad del cobre y entregarlo a los sitios de síntesis de cuproproteínas. A los genes estructurales *cut*C y *cut*D les han sido asignados un papel en el eflujo de cobre (Figura 22). Las proteínas *Cut*C y *Cut*D son probablemente ATPasas de eflujo de cobre. El patrón de adhesión de cobre de la proteína *Cut*E está relacionada con una región ligadora de cobre. Un gene adicional, *cutR*, es responsable de regular el operón *cut* (Cervantes & Gutiérrez-Corona, 1994).

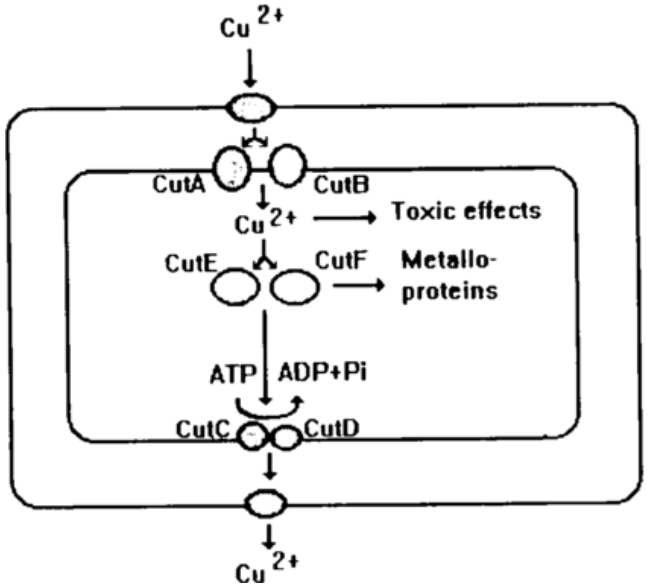



**Figura 23.** Modelo para el funcionamiento de los productos del gene cromosomal (*Cut*) involucrado en el metabolismo de cobre en *E. coli.* Las flechas indican el destino del cobre hasta que el exceso es extruido por la ATPasa de membrana (Cervantes & Gutiérrez-Corona, 1994).

### 1.8.3.5 Zinc

El zinc es un componente de muchas ADN y ARN polimerasas, además muchas proteínas regulatorias de genes en eucariontes (y algunos procariontes) contienen un patrón llamado “dedo de zinc” para el reconocimiento de ácidos nucleicos. Las fosfatasas alcalinas contienen $Zn^{2+}$ esencial, que juega un papel estructural, más que catalítico; el catión sirve para coordinar el grupo fosfato en el complejo enzima-

fosfato no-covalente. (Nies, 1992), además, el zinc es componente de más de 200 enzimas aisladas de especies diferentes donde es indispensable para su función catalítica y su estabilidad estructural. (Choudhury & Srivastava, 2001). Sin embargo, aunque es un ion metálico esencial, es tóxico a altas concentraciones, por ejemplo, es un potente inhibidor de los sistemas de transporte de electrones respiratorios de bacterias y mitocondrias. La toxicidad del zinc se ha hallado que es bastante baja comparada con otros metales como el Hg, Cd, Cu, Ni, Co y Pb. El límite de seguridad de concentración de zinc en agua potable es de 5 µg/ml. La resistencia a niveles tóxicos de zinc puede deberse a acumulación extracelular, al secuestro por metalotioneinas (MT)34–36, al secuestro físico intracelular o basada en eflujo (Choudhury & Srivastava, 2001).

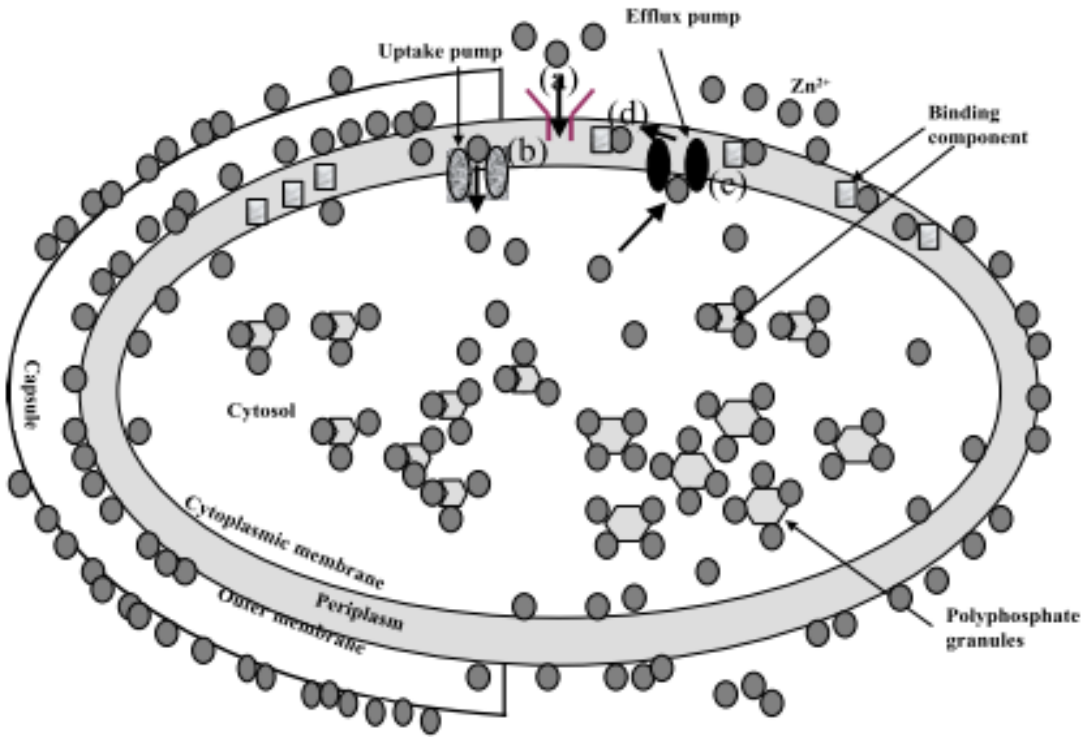



**Figure 24.** Mecanismos de resistencia al zinc en bacterias. Adhesión extracelular, adhesión sobre la membrana extracelular, exportación activa vía una bomba de eflujo (ATPasa tipo P o antiporteador de protón): después de que los iones zinc han sido tomados por rutas específicas o no específicas; secuestro por proteínas periplásmicas y/o citoplásmicas u otros ligandos, como gránulos de polifosfato (Choudhury & Srivastava, 2001).

## 2. Justificación

En el suelo del ex Lago de Texcoco existe una gran diversidad de microorganismos, que en general incluyen, arqueas, bacterias y hongos, los cuales tienen propiedades que no han sido analizadas. El análisis del genoma de microorganismos aislados de ambientes extremos puede contribuir a un conocimiento más completo sobre las características de interés de esos microorganismos, como es la resistencia a compuestos tóxicos, por ejemplo a metales pesados.

Con la anotación del genoma se pueden encontrar genes con funciones conocidas. Estos genes pueden estar expresados o no, pero con la anotación es posible hacer una predicción de potenciales características fisiológicas, de esta manera, la anotación podría facilitar la búsqueda de propiedades en microorganismos aislados antes de hacer las pruebas fisiológicas o de resistencia para la búsqueda de ciertas características, puesto que esta búsqueda experimental implica gasto de materiales y esfuerzos, y no garantiza que se encuentre la propiedad que se esté buscando.

## 3. Hipótesis

El ensamblaje y la anotación de la secuencia de *Texcoconibacillus texcoconensis* gen. nov., sp. nov., cepa 13CC$^{T}$ permitirá tener un conocimiento sobre los genes que le pueden conferir resistencia a diferentes compuestos, como es el caso del arsénico, cadmio, cobalto, cobre y zinc.

## 4. Objetivos

### 4.1 Generales

- Realizar el ensamblaje y anotación genómica de *T. texcoconensis*.
- Comprobar los resultados de la anotación genómica verificando la resistencia a arsénico, cadmio cobalto, cobre y zinc, predicha por la anotación.

### 4.2 Objetivos específicos

- Ensamblar el genoma de *T. texcoconensis* con el programa Velvet.
- Hacer la anotación del genoma con la herramienta en línea RAST.
- Construir arboles filogenéticos usando los genes *rpo*B y *rec*A para encontrar microorganismos similares a *T. texcoconensis* y usar la herramienta BLAST para hacer la comparación entre los genomas.
- Analizar el genoma de *T. texcoconensis* para la identificación de genes, entre ellos los de resistencia a arsénico, cadmio, cobalto, cobre y zinc.
- Realizar pruebas de resistencia a arsénico, cadmio, cobalto, cobre y zinc y determinar las Concentraciones Mínimas Inhibitorias (MICs) para comprobar la presencia de la función asociada a los genes predichos con RAST.

## 5. Metodología

### 5.1 Ensamble del genoma con VELVET

Previo al presente trabajo, Ruiz-Romero purificó y secuenció con Illumina Casava pipeline versión 1.8.2 el genoma de *Texcoconibacillus texcoconensis*. La información obtenida de la secuenciación fue usada en el presente trabajo.

La secuenciación se realizó con el Illumina Casava pipeline versión 1.8.2. Las lecturas recibidas “crudas” (raw) venían filtradas con el sistema de filtrado Illumina Chastity filtering y las lecturas que contenían adaptadores y/o señales de control PhiX fueron removidas usando un protocolo de filtrado propio del secuenciador, esto para garantizar la calidad de las mismas y fueron recibidas en un formato fastq. Con estos archivos de las secuencias forward (R1) y reverse (R2) se alimentó el programa FastQC, para corroborar la calidad de las mismas.

Una vez realizado el análisis de la calidad de las lecturas se procedió a cargar estas en el programa Velvet versión 1.1.03, el cual está instalado en la supercomputadora Mazorka del Langebio, la cual cuenta con una capacidad de almacenamiento total de 67TB y 66 nodos, de los cuales 42 están prendidos, estos a su vez cuentan con un total de 532 núcleos, funcionando a 5.633 TFLOPS / Queue, teniendo 6 Queue: ensam, biofis, mem64, quad, twin y xeon; de estos para el Queue ensam, están asignados 2 nodos con 4 procesadores por nodo, 8 núcleos por procesador y 32 núcleos por nodo, en esta misma sección o queue cuenta con una memoria RAM de 260 GB por nodo y tiene un procesador E7-4820@ 2.00GHz que funciona a 0.512 TFLOPS.

Para poder iniciar el ensamblaje, se eligieron diferentes K-mers, partiendo de 50 pb, por ser el tamaño de inserto usado en la secuenciación, y este se subdividió, de esta manera se evaluaron 11 Kmer diferentes: 13, 15, 19, 23, 25, 29, 33, 35, 39, 43 y 45. Después de definir los K-mers a probar, se cargaron las secuencias “raw” en Velvet y se iniciaron las pruebas de ensamblaje con cada Kmer, dando como datos de salida: la cobertura (coverage), número de contigs, número de N´s y tamaño de ensamblaje (assembly length). Con base en estos parámetros, se hizo una comparación de los parámetros de salida del ensamblador y se eligió el mejor ensamblaje según esta comparación. El resultado del ensamblaje se guardó en un archivo con extensión .fasta. Con este archivo se hizo la anotación del genoma.

### 5.2 Anotación del genoma con RAST (Rapid Annotation using Subsystem Technology).

Se decidió usar la herramienta en línea RAST para hacer la anotación del genoma. Para hacer uso de esta, se ingresa al portal de RAST: http://rast.nmpdr.org/. Dentro

de esta página se debe generar un nuevo registro de usuario para después poder cargar los genomas que se desean anotar. Una vez cargado el genoma, el sistema comienza a procesar la información y por vía correo electrónico informa cuando el genoma está anotado. Es en este momento que se puede ingresar a RAST nuevamente para visualizar resultados, los cuales son presentados en una tabla donde viene el nombre del archivo que se cargó y el estatus del mismo (Figura 25). El usuario puede cargar tantos genomas como necesite y compararlos entre ellos dentro de RAST.

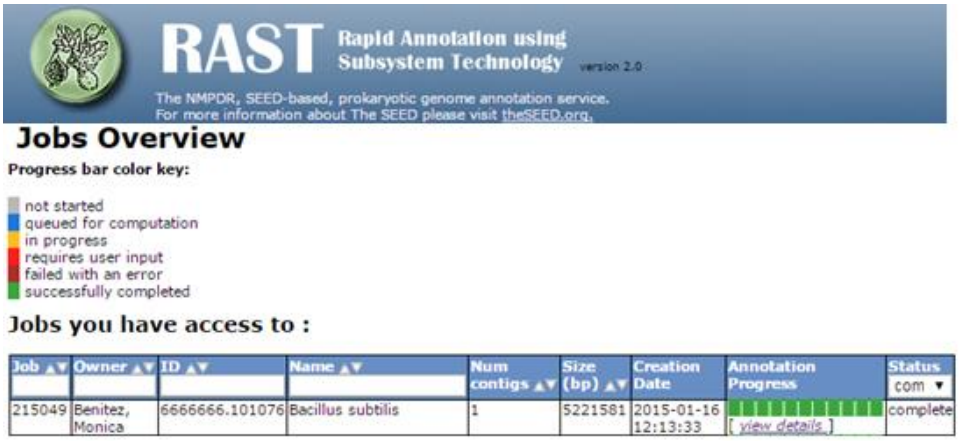



**Figura 25.** Sección dentro de RAST para cargar los genomas que se quieran anotar. En la tabla se muestran los detalles del genoma cargado y el estatus del mismo. Aquí se puede visualizar cuando se ha concluido la anotación del genoma.

Al seleccionar ver más detalles en la opción de progreso de la anotación, se cargan diferentes opciones de visualización con el ambiente SEED de los resultados de la anotación. Eligiendo la opción: Browse annotated genome in SEED Viewer (Figura 26) se despliega el resumen de la anotación. RAST ofrece dos opciones en esta sección: Subsystem Statistics y Features in Subsystems.


**Figura 26.** Visualización de los detalles del genoma cargado al sistema.

En la opción de Subsystem Statistics la primera aparecen los genes encontrados dentro del genoma en forma de gráfica, mostrado en base a una clasificación estos genes (Figura 27).

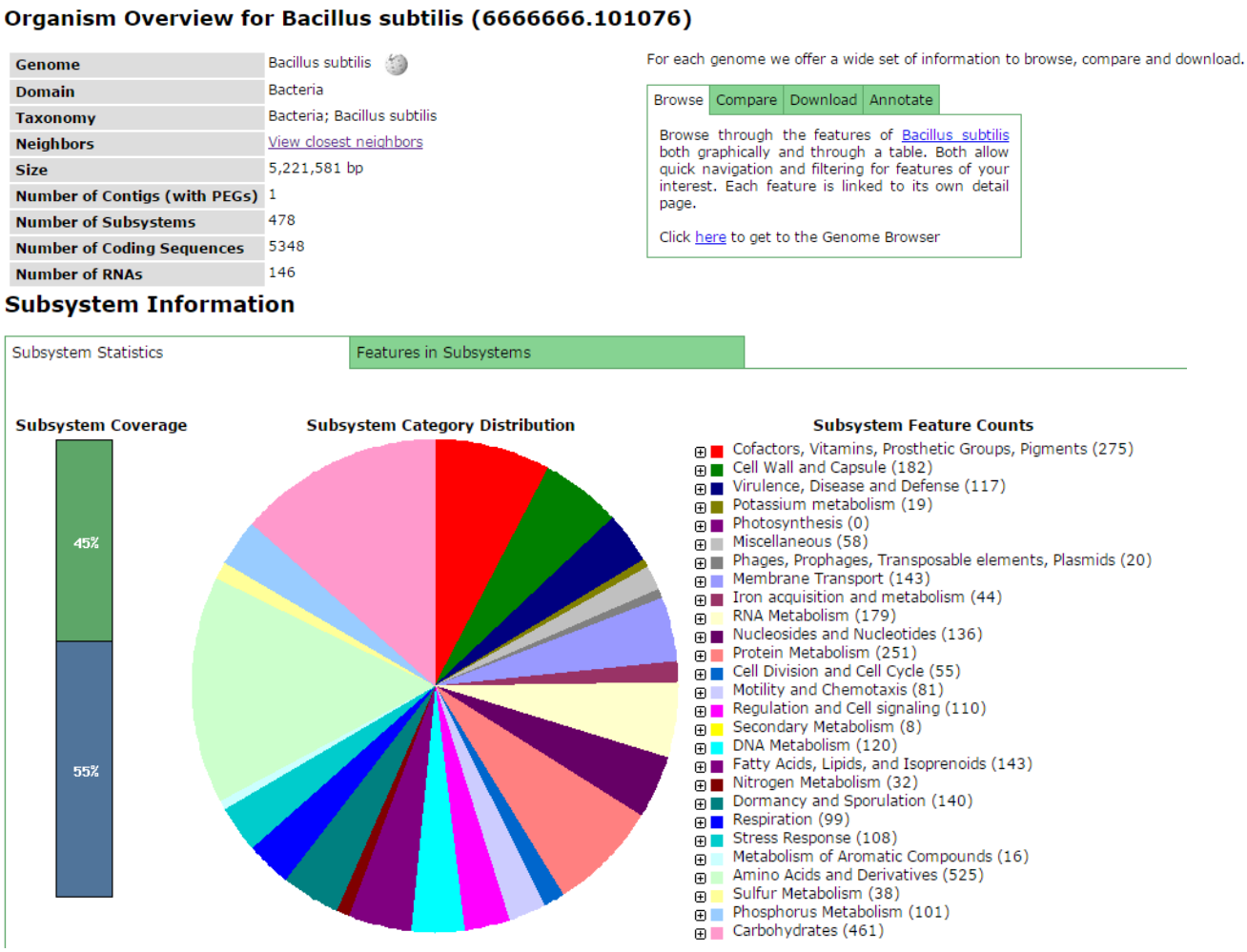
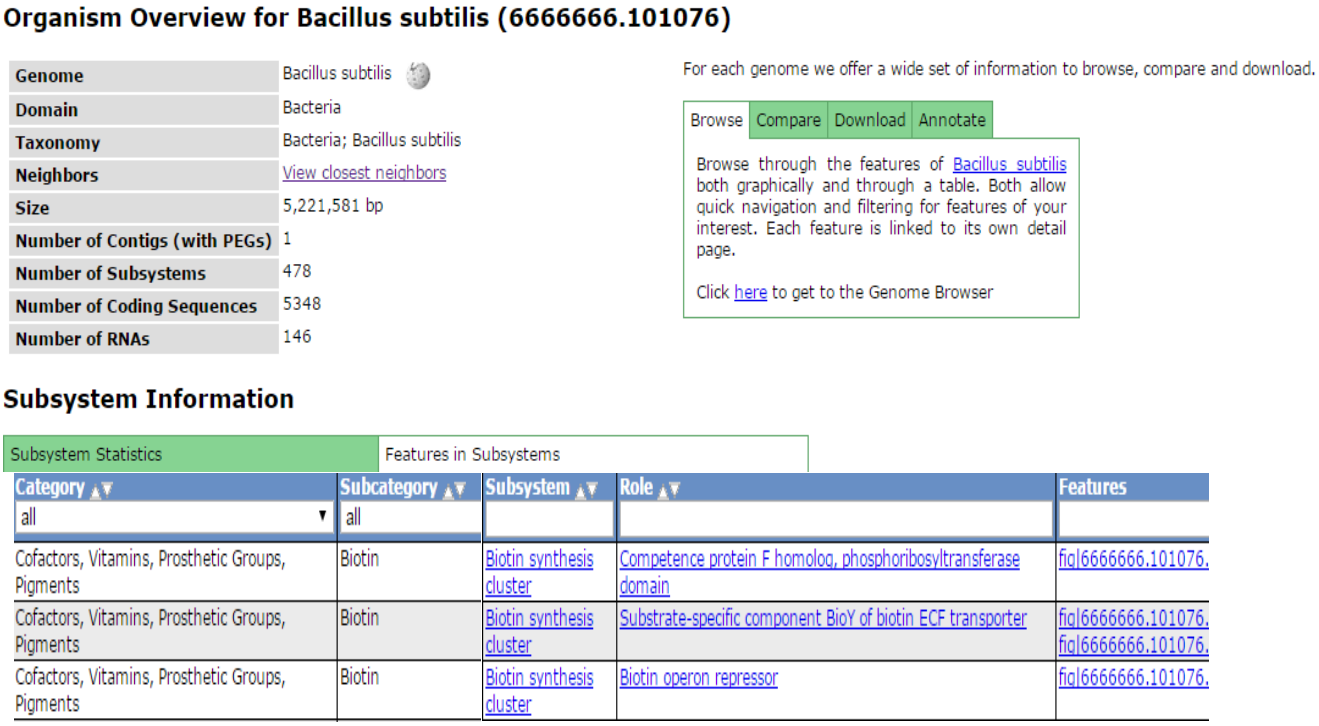



**Figura 27.** Visualización de Subsystem Statistics, donde está el resumen de los posibles genes identificados y la información sobre estos se presenta en una gráfica. Del lado derecho se pueden elegir alguna de las funciones y se despliegan más opciones.

En la opción de Features in Subsystems, que es la presentación de los genes en forma de tabla, aparecen el nombre de los genes y su posible función, además de otras opciones (Figura 28).



**Organism Overview for Bacillus subtilis (6666666.101076)**

| | |
|---|---|
| **Genome** | Bacillus subtilis |
| **Domain** | Bacteria |
| **Taxonomy** | Bacteria; Bacillus subtilis |
| **Neighbors** | View closest neighbors |
| **Size** | 5,221,581 bp |
| **Number of Contigs (with PEGs)** | 1 |
| **Number of Subsystems** | 478 |
| **Number of Coding Sequences** | 5348 |
| **Number of RNAs** | 146 |

For each genome we offer a wide set of information to browse, compare and download.

Browse | Compare | Download | Annotate

Browse through the features of Bacillus subtilis both graphically and through a table. Both allow quick navigation and filtering for features of your interest. Each feature is linked to its own detail page.

Click here to get to the Genome Browser

**Subsystem Information**

Subsystem Statistics | Features in Subsystems

| Category | Subcategory | Subsystem | Role | Features |
|---|---|---|---|---|
| all | all | | | |
| Cofactors, Vitamins, Prosthetic Groups, Pigments | Biotin | Biotin synthesis cluster | Competence protein F homolog, phosphoribosyltransferase domain | fig\|6666666.101076. |
| Cofactors, Vitamins, Prosthetic Groups, Pigments | Biotin | Biotin synthesis cluster | Substrate-specific component BioY of biotin ECF transporter | fig\|6666666.101076. fig\|6666666.101076. |
| Cofactors, Vitamins, Prosthetic Groups, Pigments | Biotin | Biotin synthesis cluster | Biotin operon repressor | fig\|6666666.101076. |

**Figura 28.** Visualización de Features in Subsystems en forma de tabla, donde se presenta la clasificación de los genes y sus posibles funciones.

Ambas opciones tienen enlaces en color azul, que despliegan más información sobre el gen seleccionado, por ejemplo ver la secuencia de la proteína, ver la homología de los genes con otros organismos, comparar ese gen con los de otros organismos, etc. (Figura 29).

**Compare Metabolic Reconstruction of Bacillus subtilis (A) and Bacillus cellulosilyticus DSM 2522 (B)**

The comparison of metabolic reconstruction will allow you to compare the **functioning parts** of two organisms. The notion of functioning is defined by having genes for all the functional roles that compose a variant of a subsystem.

The table below will list all genes which were associated with a subsytem in the respective organism. The first column will allow you to filter those that are unique to one organism, to the other, or common to both. The column 'SS active' will show you whether the subsystem this gene has been classified into was found to have an active variant in this organism.

If the gene cannot be found, you can click the find button in that cell to search for it.

display 15 items per page

displaying 1 - 15 of 2611 next» last»

| Presence all | Category all | Subcategory all | Subsystem | Role | Organism A | SS active A all | Organism B | SS active B all |
|---|---|---|---|---|---|---|---|---|
| A and B | Amino Acids and Derivatives | Alanine, serine, and glycine | Alanine biosynthesis | Alanine racemase (EC 5.1.1.1) | fig\|6666666.101076.peg.235, fig\|6666666.101076.peg.1958 | yes | fig\|6666666.68933.peg.352 | yes |
| A and B | Amino Acids and Derivatives | Alanine, serine, and glycine | Alanine biosynthesis | Branched-chain amino acid aminotransferase (EC 2.6.1.42) | fig\|6666666.101076.peg.1354, fig\|6666666.101076.peg.1748 | yes | fig\|6666666.68933.peg.1462, fig\|6666666.68933.peg.2089 | yes |

**Figura 29.** Comparación entre dos microorganismos haciendo una comparación de su metabolismo. Con esta opción RAST permite encontrar genes que comparten o genes que están presentes en un microorganismo y en el otro no.

## 5.3 Construcción de árboles filogenéticos usando como marcadores los genes de *rpo*B y *rec*A

Debido a que se trabajó con un microorganismo que es un género y especie nuevos, se decidió hacer un par de árboles filogenéticos usando las secuencias de los genes *rpo*B y *rec*A para comparar a *T. texcoconensis* con otros *Bacillus*, y de esta manera ver la posibilidad de usar el genoma conocido de alguno de los microorganismos que sean más cercanos a esta bacteria, como molde para completar el genoma de *T. texcoconensis* y también para poder comparar los genes presentes y que sean comunes entre los microorganismos a usar y *T. texcoconensis*.

Las secuencias de *rpo*B y *rec*A de *T. texcoconensis* se descargaron de los resultados de la anotación generada por el programa RAST. Estas dos secuencias se llevaron a la herramienta en línea BLASTn, para obtener a los microorganismos con secuencias similares a *T. texcoconensis*. Con los resultados del BLAST se buscaron las secuencias de *rpo*B y *rec*A en la base de datos de NCBI y se descargaron sus secuencias en archivos .fasta para construir una base de datos, en la cual se

guardaron un total de 68 secuencias de diferentes microorganismos, esto es 34 diferentes secuencias del *rpo*B y 34 de *rec*A.

Con la información obtenida, se compilaron los datos de las secuencias para hacer un solo archivo en formato con.tree, que es el tipo de extensión que lee el programa FigTree, el cual, se instala en una computadora con sistema operativo Linux. Al término del procesamiento de la información, nos genera una imagen de los datos alimentados ordenados en un árbol filogenético.

### 5.4 Análisis del genoma para identificación de genes de resistencia a metales pesados en *Texcoconibacillus texcoconensis*.

Con la anotación realizada en RAST, se pudo ver un listado de los posibles genes presentes en la cepa 13CC$^{T}$, así como la posible función. Dentro de estos genes, se eligieron los genes de resistencia a metales pesados para corroborar la predicción de esta función.

### 5.5 Pruebas de resistencia a metales pesados usando diferentes concentraciones de cada compuesto y Concentraciones Mínimas Inhibitorias (MIC´s).

Para la preparación de esta prueba, se preparó un medio de cultivo sólido que contenía (g/l): casaminoácidos, 7.5; extracto de levadura, 10; citrato trisódico, 3.0; agar, 20; y (esterilizado por separado) $NH_4Cl$, 1; LiCl, 0.1; $CaSO_4$ $2H_2O$, 0.17; $MgSO_4$. $7H_2O$, 0.24; KCl, 10; $KH_2PO_4$, 1; $Na_2CO_3$, 10 y NaCl, 200, ajustando el pH a 8.9, como describe Ruiz–Romero *et al.*, (2013).

Al medio anteriormente mencionado, en la parte inorgánica, se agregaron las diferentes concentraciones de cada uno de los metales a probar, tal y como se indica

en la Tabla 1, y se inocularon las cajas con *T. texcoconensis*. Después las cajas se incubaron a 37°C durante 72 horas, ya que primero se confirmó que la cepa podía crecer en presencia de los diferentes metales.

**Tabla 1.** Diferentes concentraciones de metales que se agregaron al medio de cultivo para la prueba de confirmación del crecimiento.

| Metal | Compuesto usado | Concentración | Referencia |
|---|---|---|---|
| **Arsénico (III)** | $As_2O_3$ | 1-3 mM | Liao *et al.*, 2011 |
| **Arsénico (V)** | $Na_2HAsO_4$ | 5, 9 y 20 mM | Liao *et al.*, 2011 |
| **Cadmio** | $3CdSO_4 \cdot 8H_2O$ | 500, 100, 50 y 5 µg de $Cd^{2+}$ $ml^{-1}$ | Lee *et al.*, 2001; Sochor *et al.*, 2011 |
| **Cobalto** | $CoCl_2$ | 1, 3 y 8 mM. | Xiong and Jayaswal, 1998. |
| **Cobre (II)** | $CuSO_4$ | 2 , 5,10 mM | Elguindi *et al.*, 2011 |
| **Zinc** | $ZnSO_4$ | 0.1, 0.2 y 1 mM. | Xiong and Jayaswal, 1998 |

Una vez confirmado que la cepa podía crecer a diferentes concentraciones de metales se aumentó la concentración de estos hasta llegar a las Concentraciones Mínimas Inhibitorias (MIC´s) para cada uno de ellos. La MIC se define como la concentración del compuesto tóxico, en este caso del metal, que inhibe completamente el crecimiento de las bacterias en medio solido después de 72 horas de incubación. (Liao *et al.*, 2011).

## 6. Resultados y discusión

### 6.1 Ensamblaje

Al hacer la evaluación de la calidad de las lecturas recibidas del proceso de secuenciación con el programa FastQC obtuvimos un reporte de calidad, el cual permitió corroborar que las lecturas a usar tenían una buena calidad. Las lecturas

evaluadas se denominaron R1 para forward y R2 para reverse, de esta manera se presentan los resultados de la evaluación de la calidad de ambas lecturas.

- **Informe de calidad del archivo R1**

Como se puede observar en la Figura 30, se presenta el reporte de calidad generado por FastQC, donde en la parte izquierda se muestran los parámetros evaluados para la calidad, llamados módulos. Cada uno de los módulos se presenta con un símbolo y un color. El símbolo de aprobación (✓) aparece en color verde e indicaría que en ese módulo los datos son normales; se tiene un símbolo de advertencia (!) en color naranja, el cual indicaría que los datos del módulo son anormales y un último símbolo, que es de desaprobación (X) en color rojo, el cual indicaría que los datos del módulo son muy inusuales. Todo lo anterior se basa en rangos establecidos como aceptables por el programa FastQC para las secuencias de ADN.

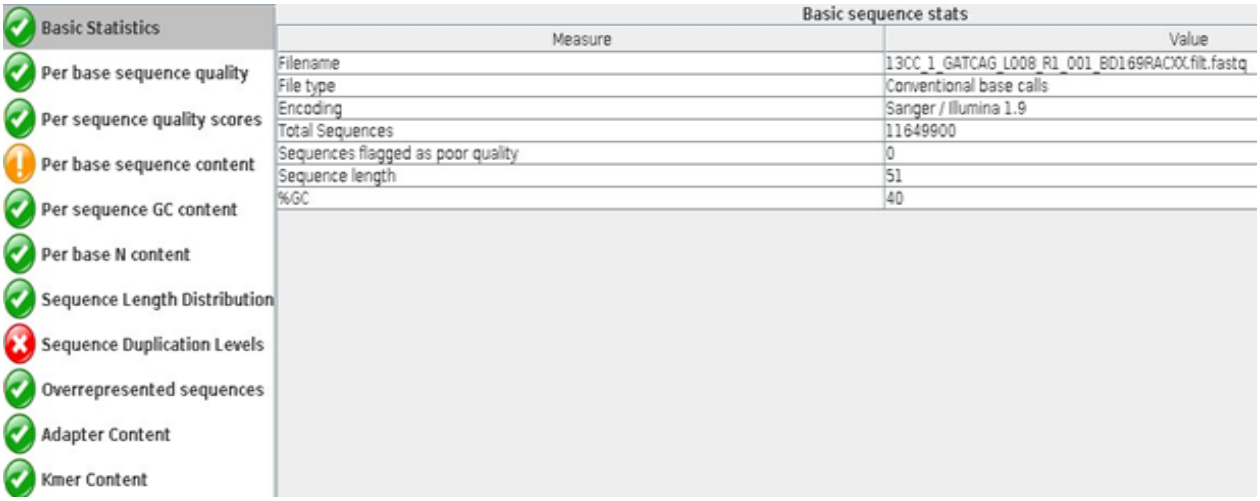



**Figura 30.** Reporte de calidad de FastQC a análisis de calidad de lecturas del primer archivo de lecturas, denominado R1.

En el caso del contenido de bases de cada secuencia, donde se tiene una señal de advertencia, indica que tenemos una diferencia entre el contenido de A y T, o G y C, pues según el programa el contenido de bases no debería tener una diferencia

mayor a 10%; sin embargo, en el archivo cargado (R1), la diferencia de la cantidad de bases de cada secuencia es mayor al 10% (Figura 31).

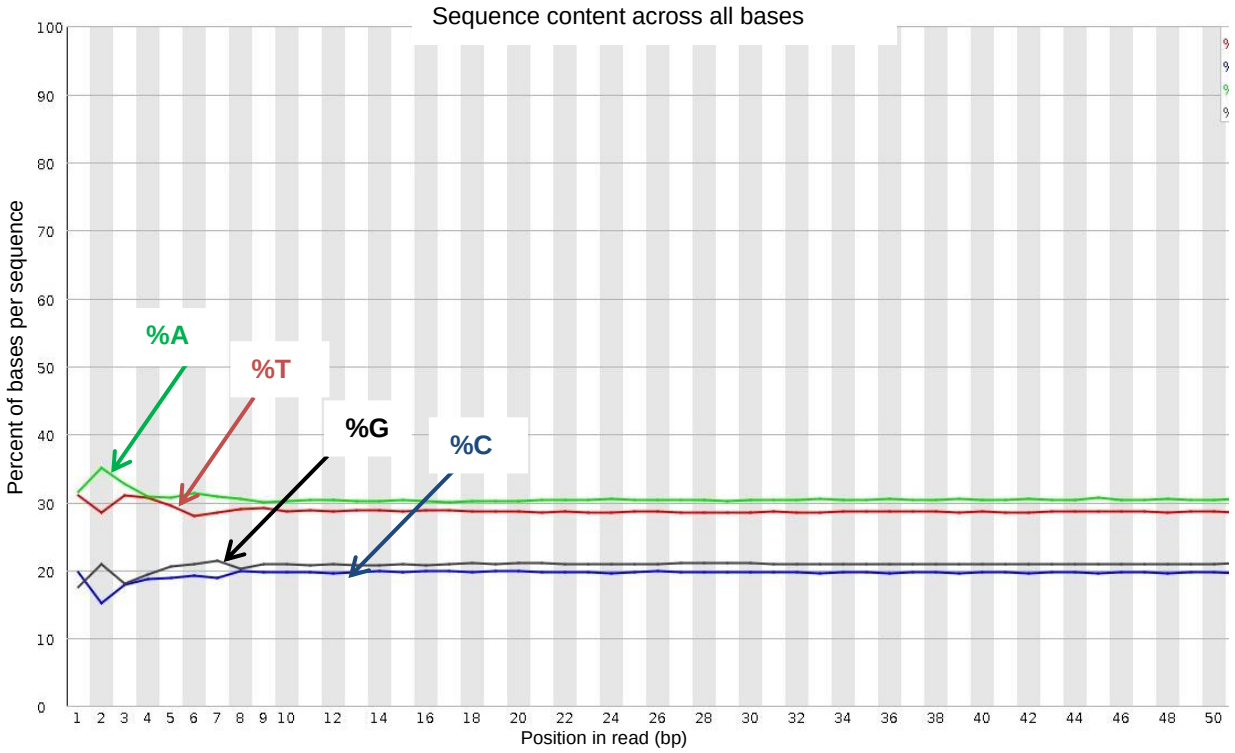



**Figura 31.** Contenido de bases de cada secuencia, nos indica que tenemos una diferencia entre el contenido de A y T, o G y C, en este caso es mayor al 10%.

La señal en rojo que aparece en los resultados con la marca de desaprobación (X), está indicando el nivel de duplicación de secuencia, este parámetro se refiere a que un bajo nivel de duplicación puede indicar una alta cobertura en la secuencia de interés, pero un alto nivel de duplicación puede indicar algún tipo de sesgo en la secuencia. Por tanto, en este caso, se emite una señal de error porque las secuencias no únicas constituyen el 47.9% del total (Figura 32).

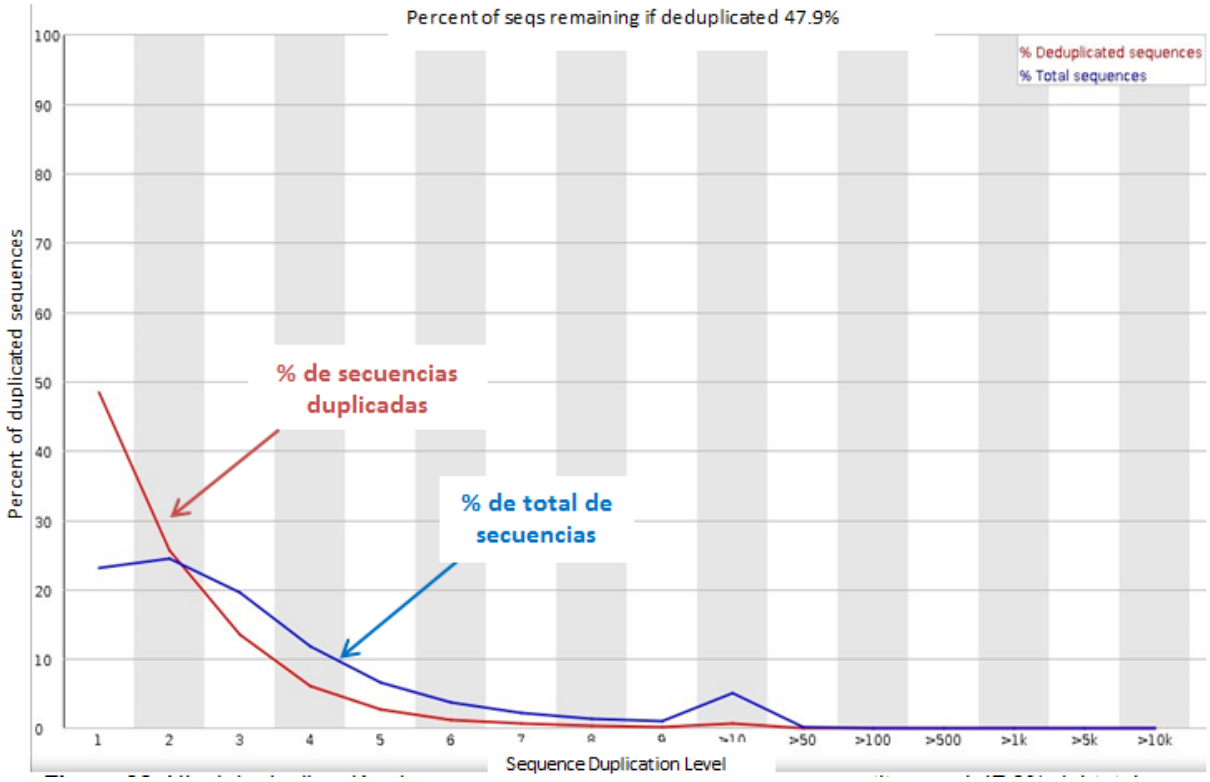



**Figura 32.** Nivel de duplicación de secuencia las secuencias no únicas constituyen el 47.9% del total.

En cuanto a los otros parámetros, al tener una señal en verde, nos indica que la evaluación fue satisfactoria y por tanto son parámetros que están dentro de un rango aceptable. Como es el caso de la calidad de las lecturas, en el caso del archivo R1, se puede ver que las lecturas caen dentro del rango de color verde, lo que significa un buen porcentaje de calidad en cada posición de la secuencia (Figura 33a). Y para el contenido de GC por secuencia, para el que se espera que se tenga una distribución normal, cosa que cumplen las secuencias del archivo R1 (Figura 34a).

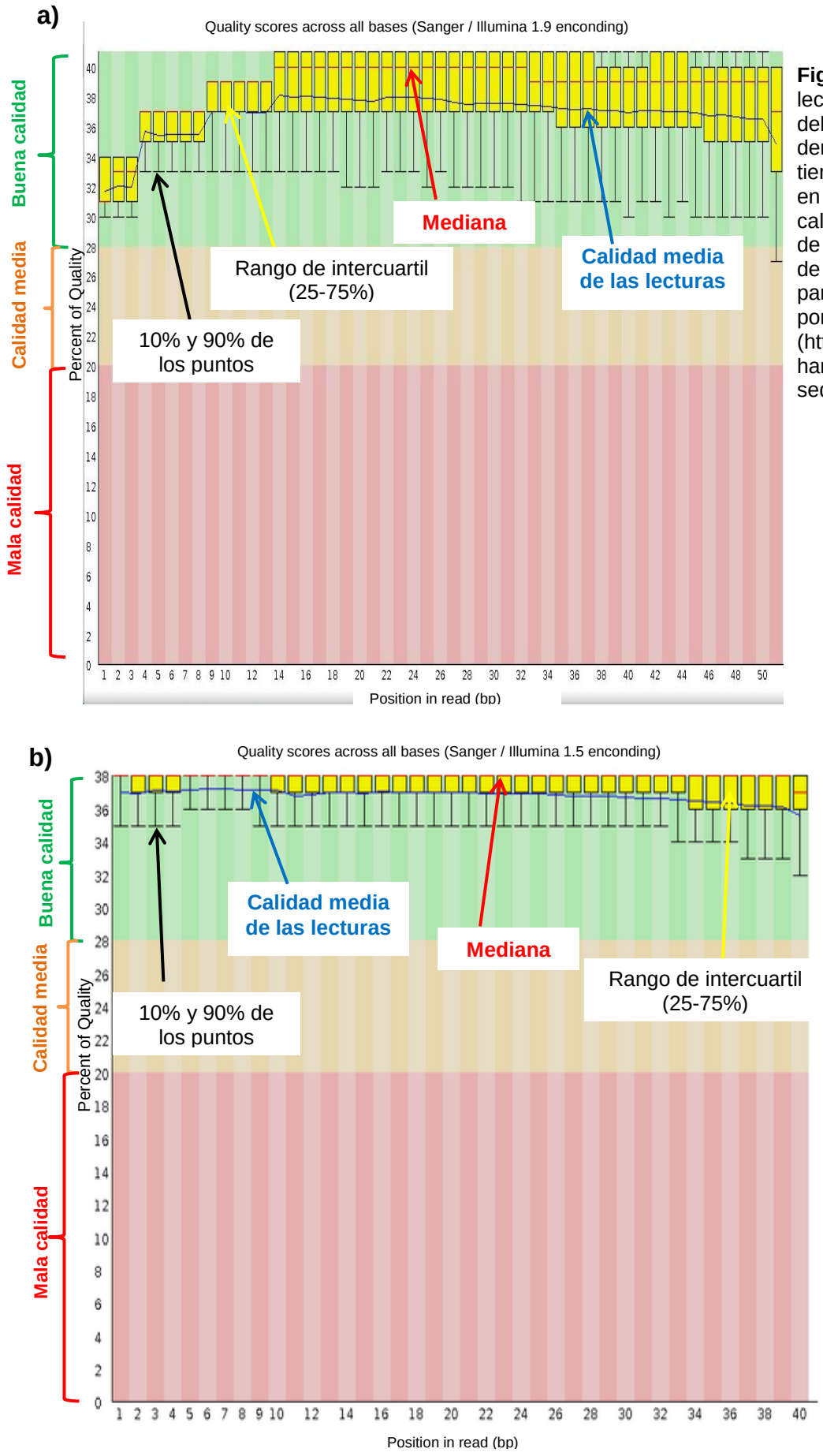



**Figura 33.** Calidad de las lecturas por base secuenciada del primer archivo de lecturas, denominado R1. . En el eje x se tiene la posición del read (pb) y en el eje y se presenta la calidad. a) Reporte de calidad de la cepa 13CC y b) Ejemplo de reporte de calidad de FastQC para comparar datos generados por Illumina de buena calidad (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html).

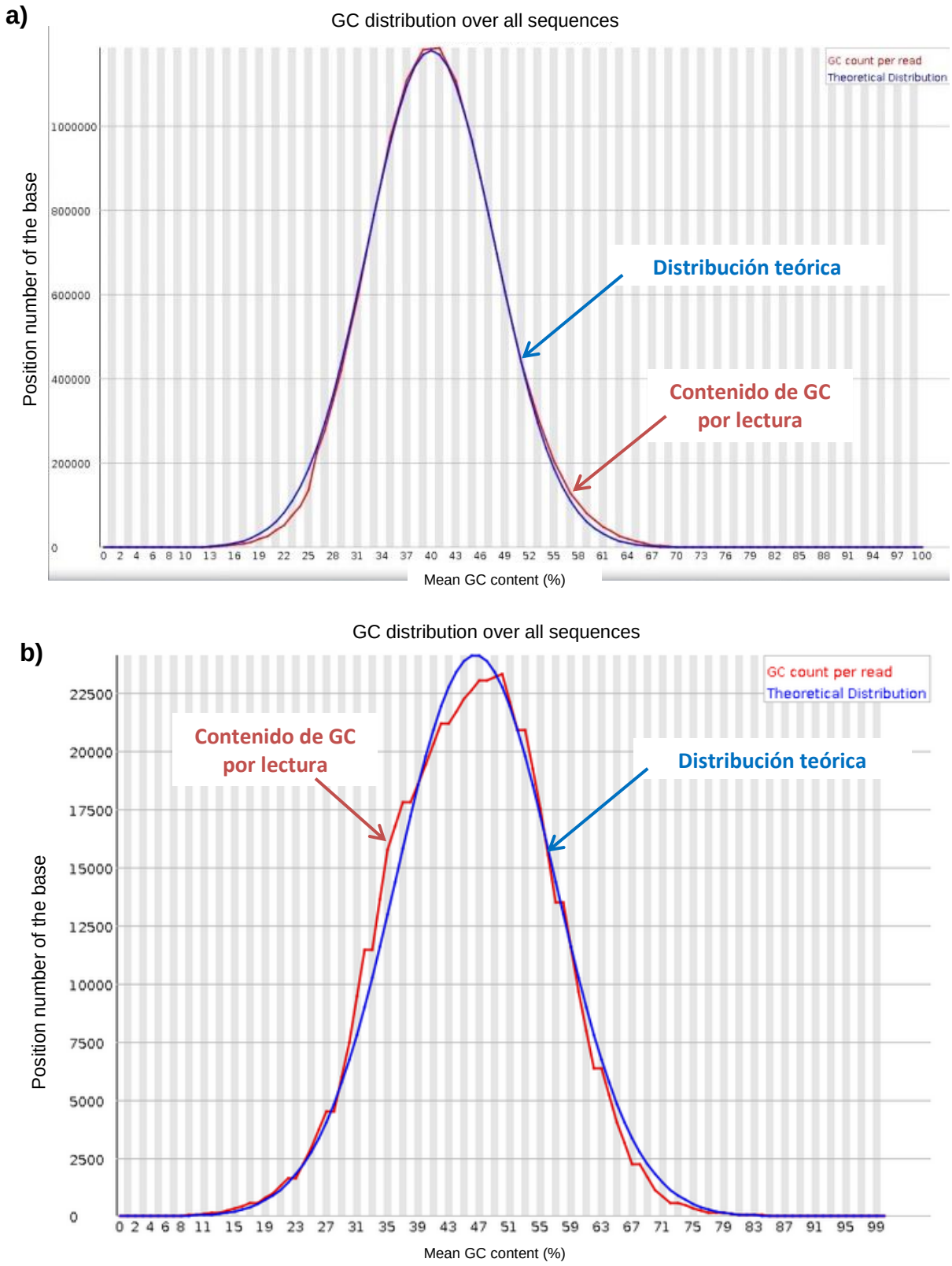



**Figura 34.** Calidad de las lecturas por contenido de GC por secuencia del primer archivo de lecturas, denominado R1. En el eje X se tiene el contenido de GC promedio (%) y en el eje Y se presenta la numero de posición de las bases. a) Reporte de calidad de la cepa 13CC y b) Ejemplo de reporte de calidad de FastQC para comparar datos generados por Illumina de mala calidad con los de buena calidad (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html).

- **Informe de calidad del archivo R2**

En el resumen de los resultados del reporte de calidad generados por FastQC, hay dos módulos en naranja, que son señales de advertencia y el resto están en color verde lo que significa que está bien la secuencia en estos módulos (Figura 35).

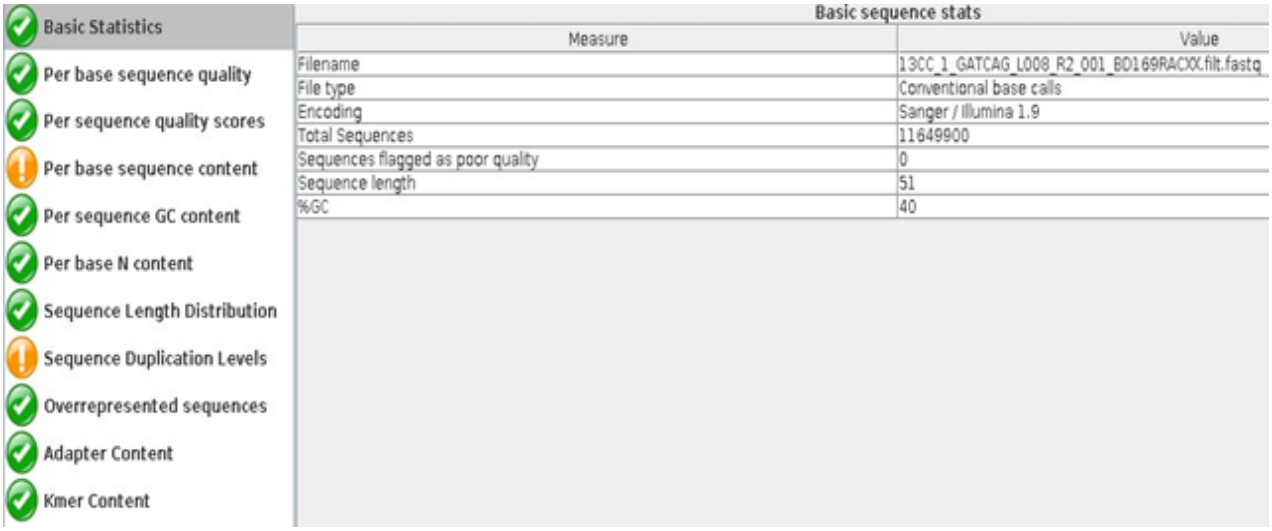



**Figura 35.** Reporte de calidad de FastQC a análisis de calidad de lecturas del primer archivo de lecturas, denominado R2.

Para el contenido de bases de cada secuencia, este tiene un resultado semejante al del archivo R1, es decir, aparece con una señal de advertencia en color naranja, lo que indica que tenemos una diferencia entre el contenido de A y T, o G y C, pues según el programa el contenido de bases no debería tener una diferencia entre estas bases mayor a 10%; sin embargo, en las secuencias del archivo R2, la diferencia de la cantidad de bases es mayor al 10%, tal y como se puede en la Figura 36.

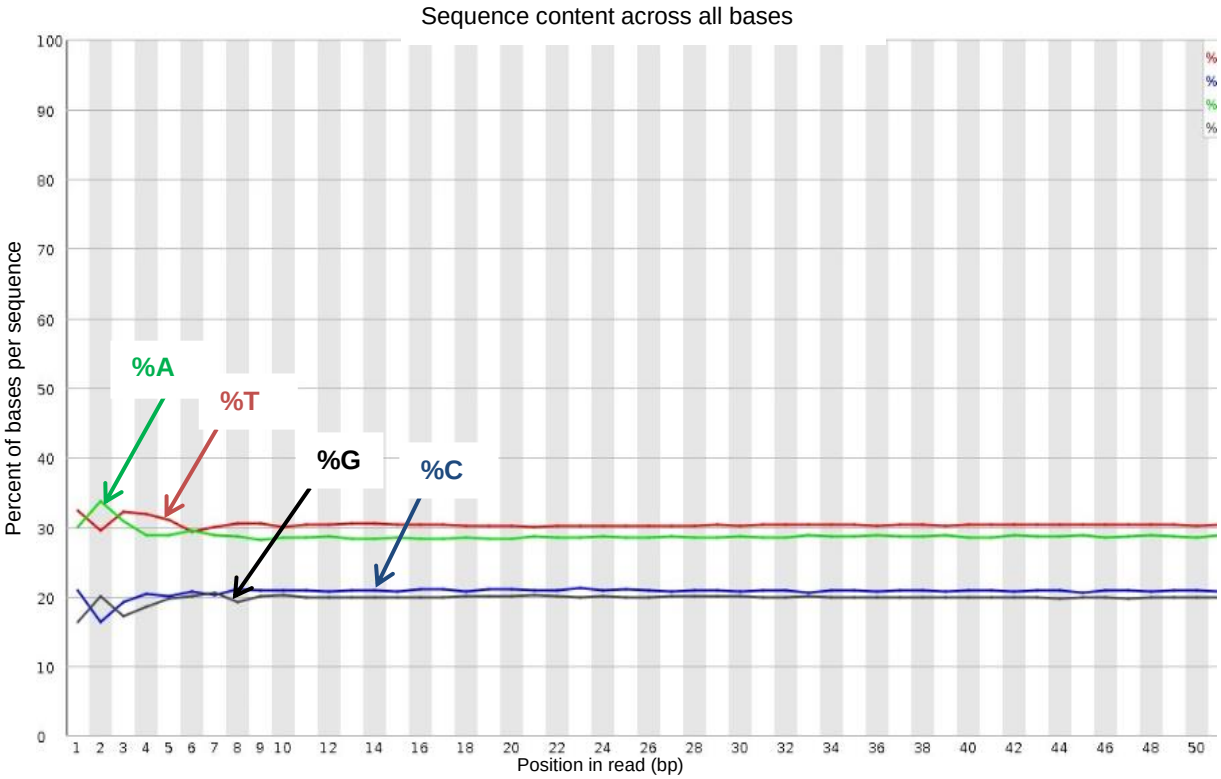



**Figura 36.** Contenido de la secuencia por base, nos indica que tenemos una diferencia entre el contenido de A y T, o G y C, en este caso es mayor al 10%.

Para el módulo del nivel de duplicación de secuencia, en esta secuencia apareció en color naranja a diferencia de la secuencia R1, donde se tuvo una señal en rojo. Lo que indica entonces este módulo, es que las secuencias no únicas constituyen más del 20% de las secuencias no únicas del total, pues se obtuvo que estas abarcan el 51.04 % del total de las secuencias, es decir, que el nivel de duplicación es relativamente alto (Figura 37).

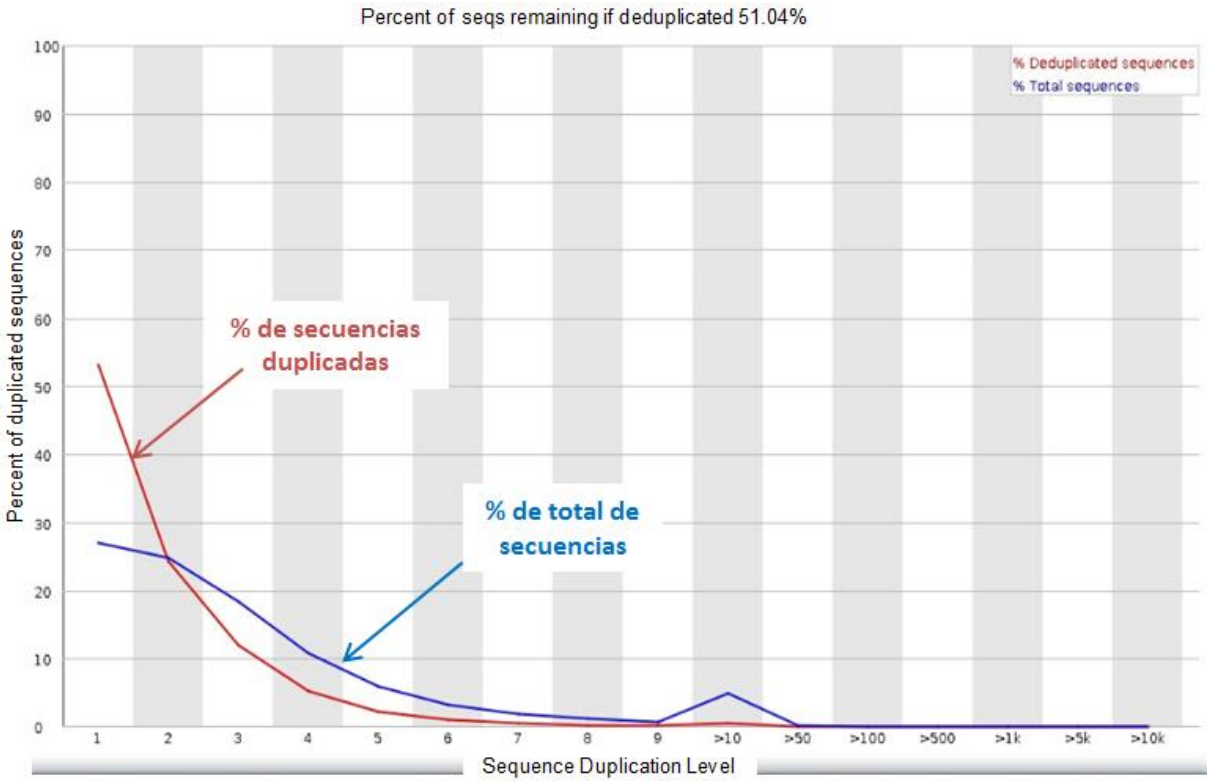
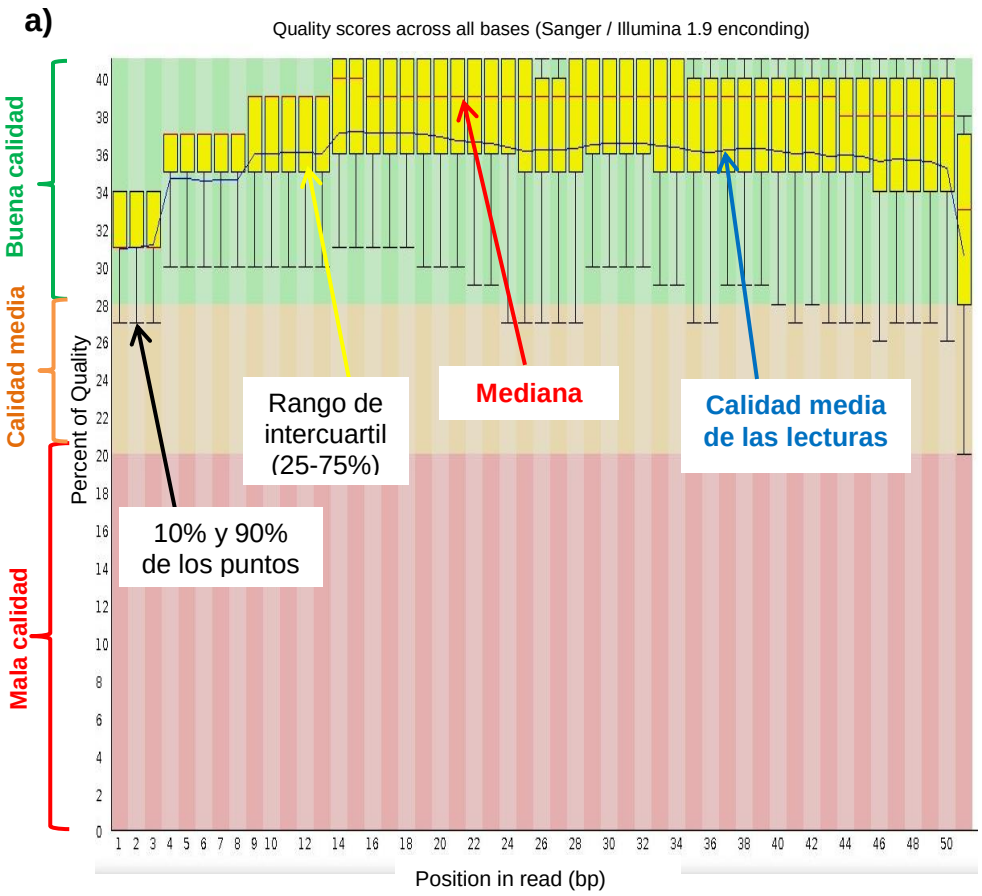



**Figura 37.** Nivel de duplicación de secuencia las secuencias no únicas constituyen el 51.04% del total.

En cuanto a los otros parámetros, también presentan una señal en verde, por lo que al igual que la secuencia R1, nos están indicando que la evaluación fue satisfactoria y por tanto son parámetros que están dentro de un rango aceptable. En la Figura 38a, se puede ver que las lecturas caen dentro del rango de color verde, aunque presentan más variación que el caso de las lecturas de la secuencia R1. Y en cuanto al contenido de GC por secuencia, la distribución de la gráfica es la misma que para la secuencia R1 (Figura 39a).



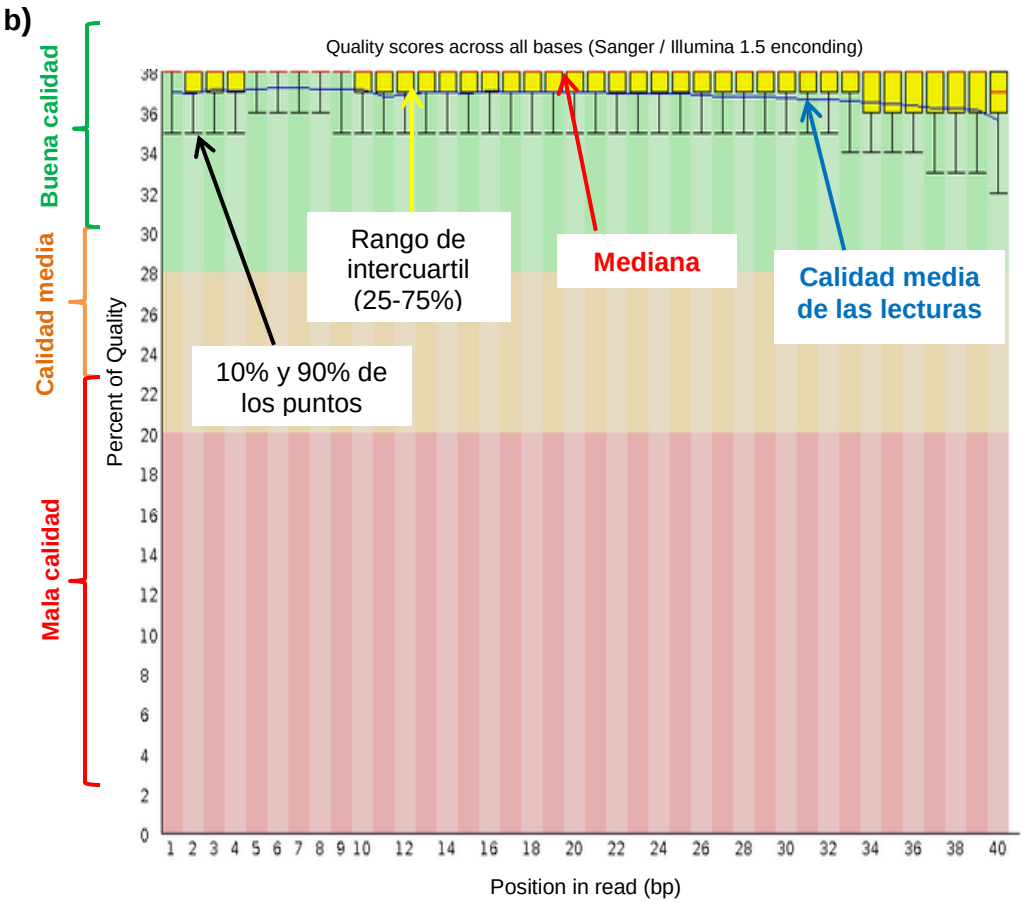
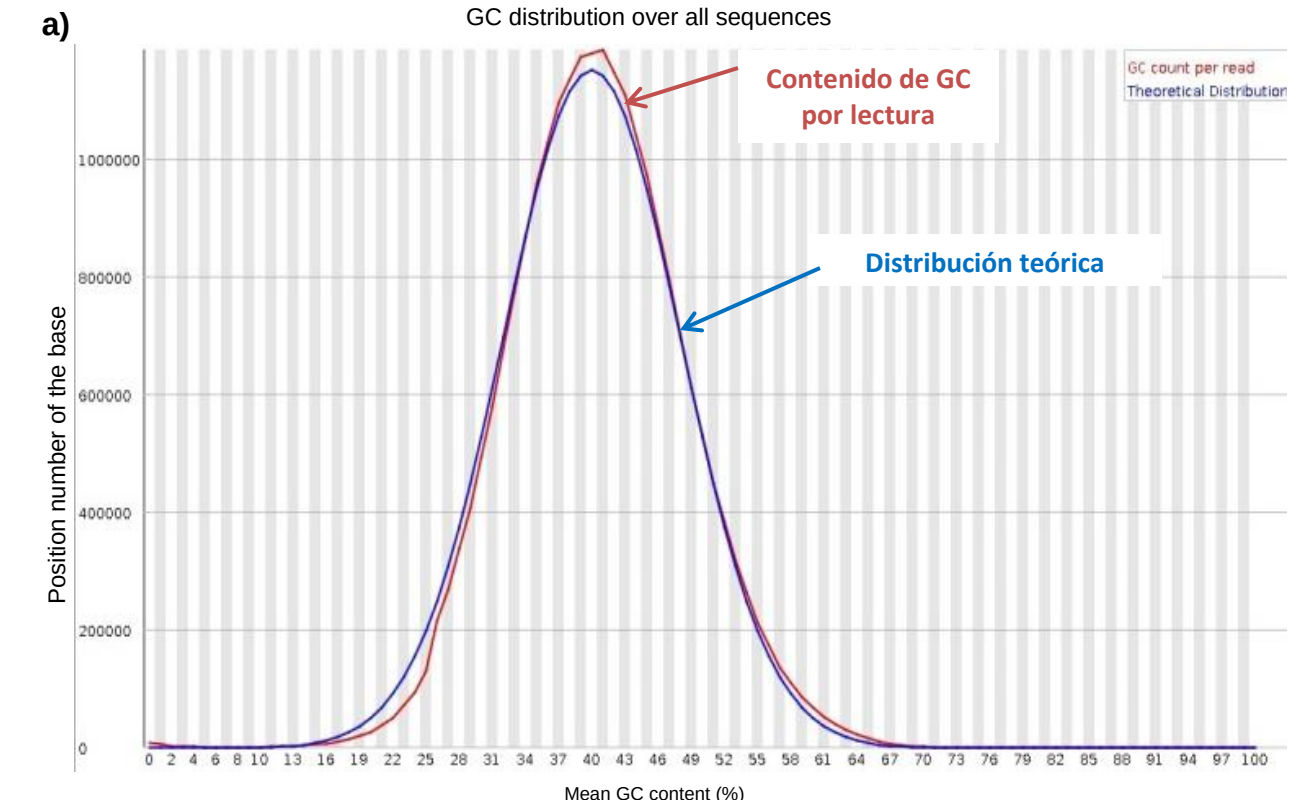




**Figura 38.** Calidad de las lecturas por base secuenciada del primer archivo de lecturas, denominado R2. En el eje x se tiene la posición del read (pb) y en el eje y se presenta la calidad. a) Reporte de calidad de la cepa 13CC y b) Ejemplo de reporte de calidad de FastQC para comparar datos generados por Illumina de buena calidad.



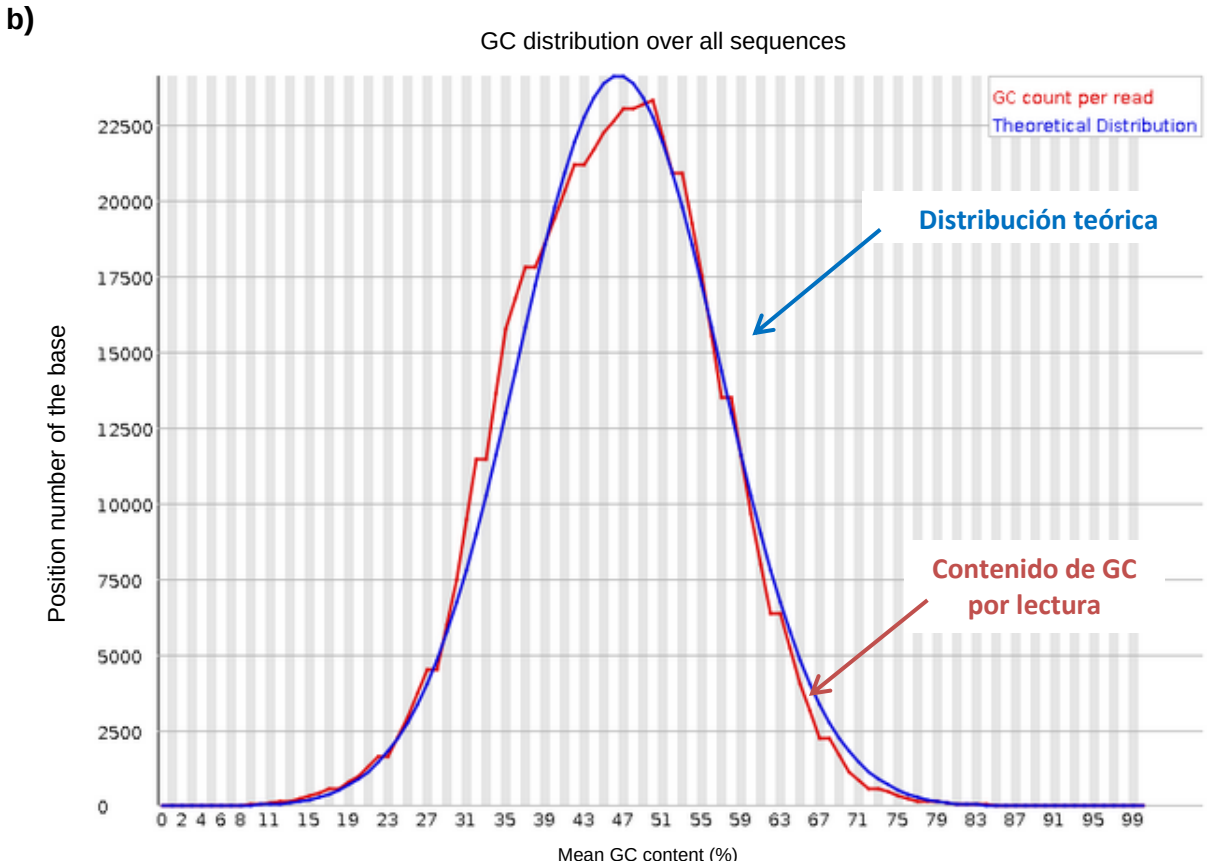




**Figura 39.** Calidad de las lecturas por contenido de GC por secuencia del primer archivo de lecturas, denominado R2. a) Reporte de calidad de la cepa 13CC y b) Ejemplo de reporte de calidad de FastQC para comparar datos generados por Illumina de mala calidad con los de buena calidad (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html).

Una vez confirmada la calidad de las lecturas, se cargaron en Velvet y se probaron los diferentes valores de K-mers mencionados en la metodología. Se fueron obteniendo diferentes resultados según los parámetros de salida que se pidieron, tal como se muestran en la Tabla 2.

**Tabla 2.** Resultados de diferentes ensamblajes generados por Velvet por las variaciones de K-mer.

| Kmer | Cobertura | Número de contigs | Número de N´s | Tamaño de ensamblaje (pb) |
|---|---|---|---|---|
| **13** | 0 | 0 | 0 | 0 |
| **15** | 110.87 | 1333 | 1772628 | 1772628 |
| **19** | 327.48 | 97 | 18731 | 3370962 |
| **23** | 338 | 83 | 6700 | 3382668 |
| **25** | 296 | 69 | 5622 | 3387585 |
| **29** | 305.78 | 60 | 5917 | 3395935 |
| **33** | 254.96 | 56 | 4188 | 3400635 |
| **35** | 231.92 | 54 | 4863 | 3404529 |
| **39** | 199.68 | 45 | 3486 | 3407401 |
| **43** | 139.84 | 44 | 6824 | 3221851 |
| **45** | 251.53 | 13 | 467 | 26629 |

De los ensamblajes obtenidos en Velvet se eligió el del K-mer=29, ya que este presentó un valor alto de cobertura, pocos contigs, así como numero de N´s y un tamaño de ensamblaje que nos permite considerar que se tiene representado gran parte del genoma.

Previo a esto se hizo un cálculo de la cobertura esperada, dividiendo el número del total de bases entre el tamaño del genoma esperado, dato que se agrega en el reporte entregado por la empresa que secuenció el genoma. El valor obtenido fue de 342X. Por lo que vemos, que si hubiésemos elegido el valor más cercano a esta cobertura como base para el ensamblaje, hubiésemos optado por un genoma con más contigs y más N´s.

Es así como el genoma resultante quedó de una longitud de 3, 395, 935 pb, organizado en 60 contigs y con un estimado de N´s de 5917, además de una cobertura alta, de 306X.

Para la visualización del genoma ensamblado se usó el programa CGview (Figura 40). En el cual se observan marcados los marcos de lectura abierta (ORF), los sitios de inicio y paro de los genes y el contenido de GC del genoma.

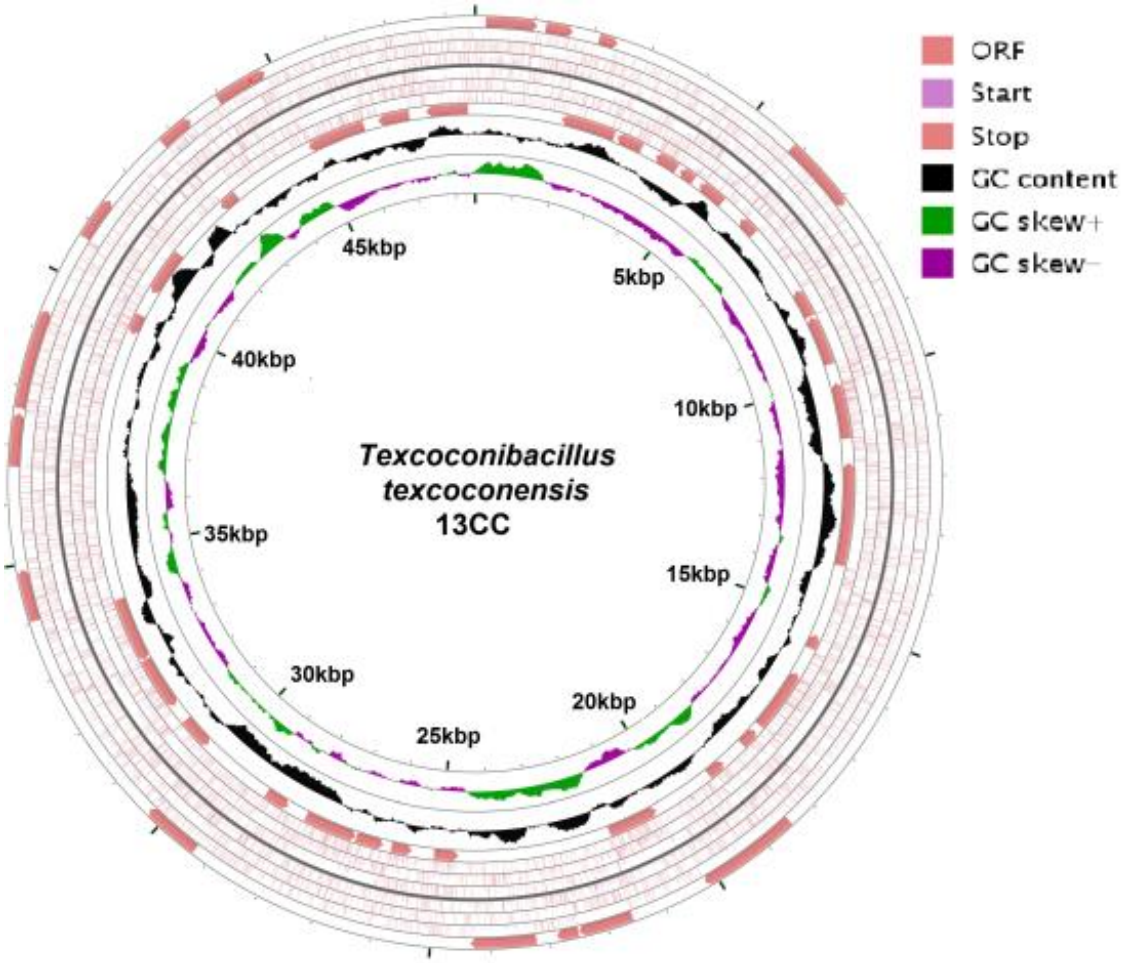



**Figura 40.** Visualización del genoma de *T. texcoconensis* generada por el programa CGview.

### 6.2 Arboles filogenéticos

Con la comparación mediante BLAST de los genes *rpo*B y *rec*A con los de otros microorganismos se construyeron dos árboles filogenéticos. En el árbol de *rpo*B, *Bacillus selenitireducens* y *Bacillus cellulosilitycus* están muy cercanos a *Texcoconibacillus texcoconensi*s (Figura 41), y como se puede observar en los valores presentados en cada una de las ramas, estas tienen un valor superior al 0.8 (80% de cercanía), lo que nos indica que está suficientemente sustentado que estos dos microorganismos sean los más cercanos a *T. texcoconensis*, sin embargo, el valor de la rama en donde están *B. cellulosilitycus* y *T. texcoconensis* no supera el 0.86, lo que nos indica que aunque están cercanos y la rama está sustentada con los datos del gen *rpo*B, estos microorganismos presentan aun cuanto menos un 14% de distanciamiento. En cuanto al árbol de *rec*A, el microorganismo más cercano a *T. texcoconensis* es *Bacillus selenitireducens*, pero tal como se muestra en este árbol, la rama que los une no está muy sustentada, pues solo tiene un poco más del 0.6, lo que significa que tienen una cercanía filogenética poco mayor del 60%. Y si se observa las ramas de donde inicia, se puede observar que esta *B. cellulosilitycus,* pero la rama que emerge de ahí sólo tiene un sustento del 0.7, lo que nos dice que hay suficientes diferencias en este gen como para que no se puedan establecer relaciones cercanas entre las filogenias de los microorganismos usados (Figura 42). Otros microorganismos podrían ser más cercanos, pero sus genomas no están reportados en las bases consultadas (NCBI).

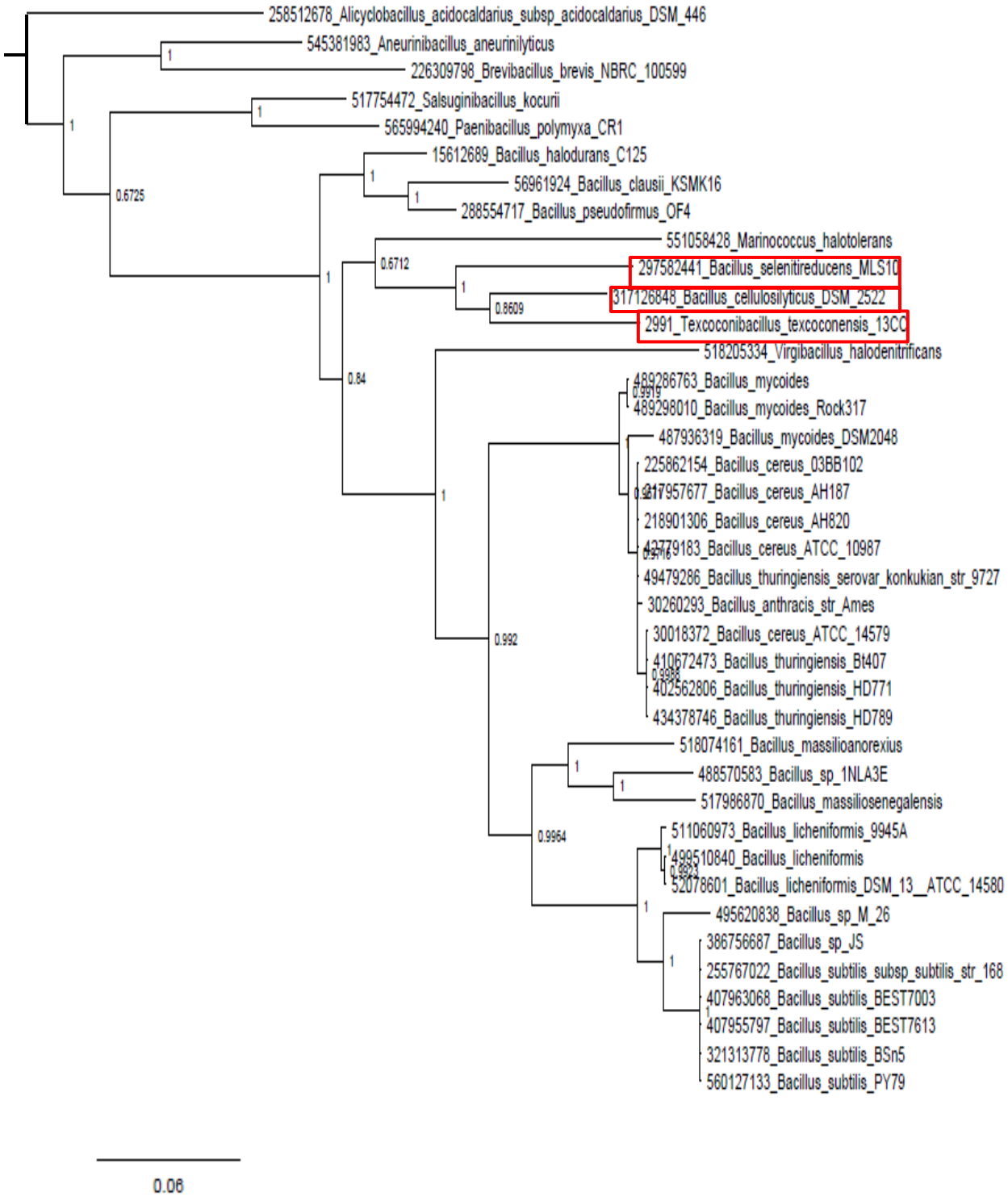



**Figura 41.** Árbol filogenético usando *rpo*B como marcador.

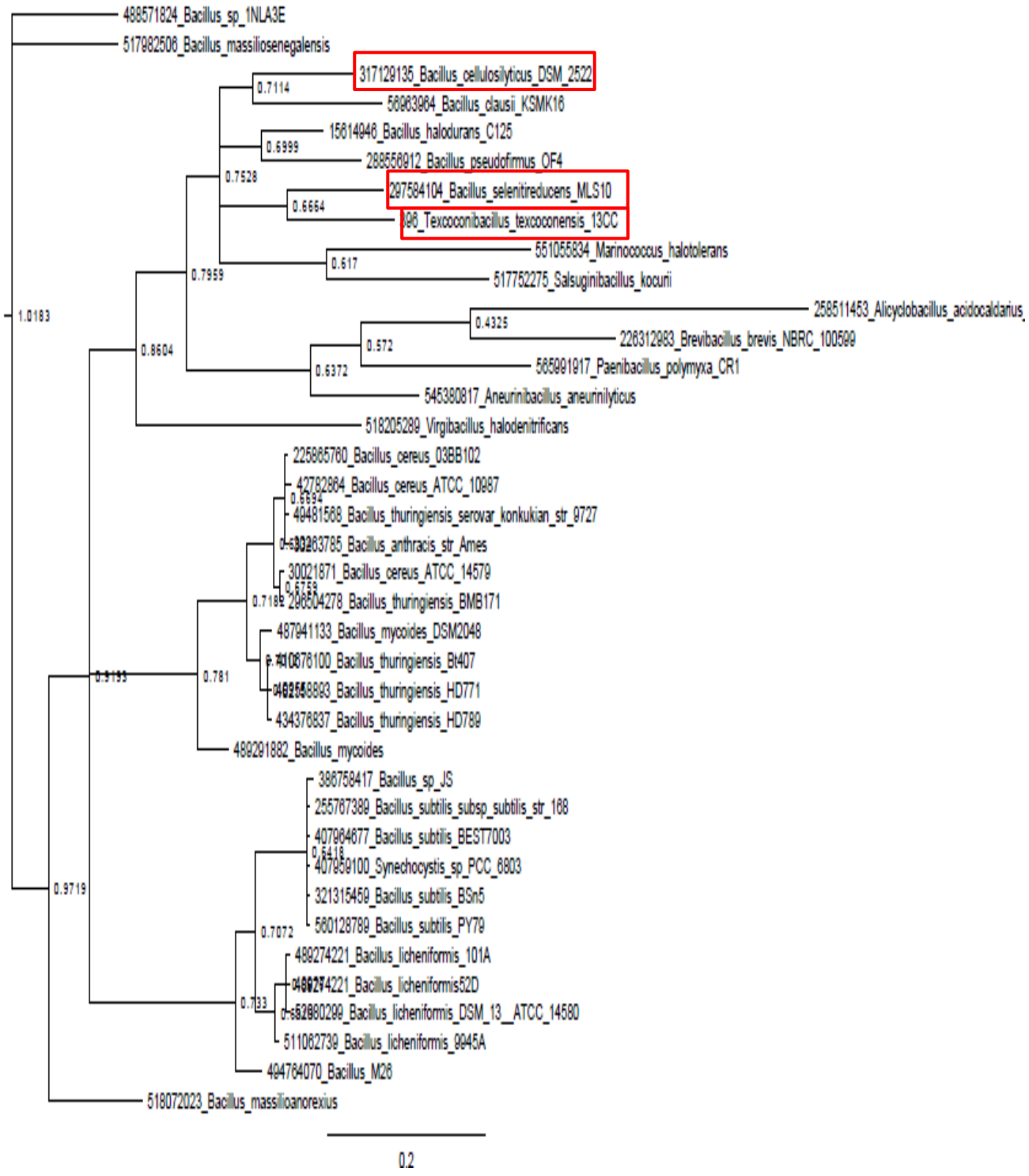



**Figura 42.** Árbol filogenético usando *rec*A como marcador.

### 6.3 Anotación mediante RAST y análisis del genoma para identificación de genes de resistencia a metales pesados en *Texcoconibacillus texcoconensis.*

El genoma ensamblado se cargó en RAST para generar la Anotación. En la primer visualización aparecen los detalles del genoma que se carga, en este caso que está organizado en 60 contigs y que tiene un tamaño de ensamble de 3, 395, 935 bp (Figura 43).

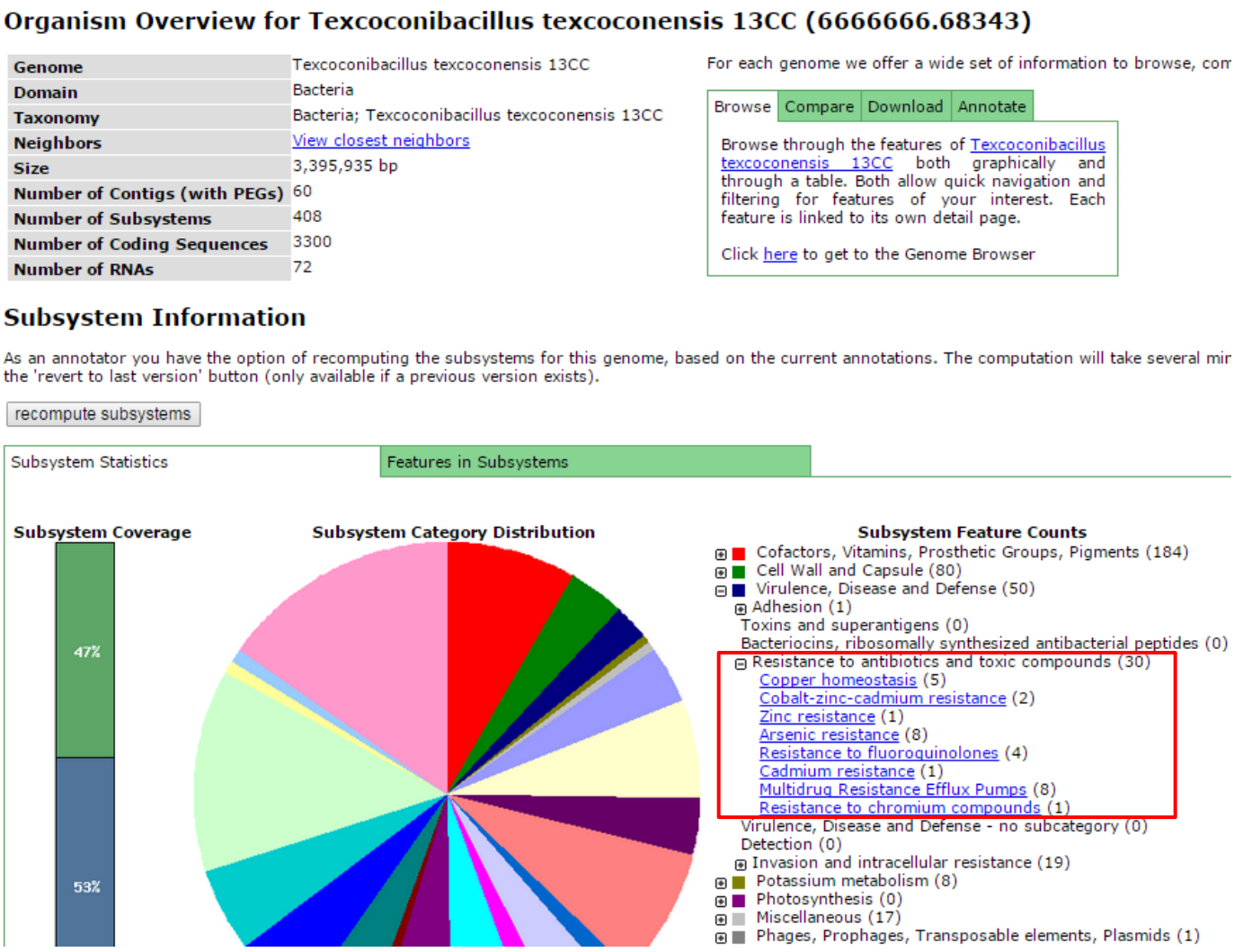



**Figura 43.** Visualización de generalidades en el genoma de la cepa 13CC una vez anotado por RAST.

Además de esto se pudieron visualizar los sistemas con los que cuenta el microorganismo en su genoma, en el caso de *T. texcoconensis*, observamos que

tiene un total de 1925 posibles genes y de estos 11 genes sugieren resistencia a arsénico, cadmio, cobalto, cobre y zinc. Estos los localizamos en la sección de "Virulencia, enfermedad y defensa" en la subclasificación de "resistencia a antibióticos y compuestos tóxicos" (Tabla 3).

**Tabla 3.** Genes de resistencia a metales pesados presentes en *Texcoconibacillus texcoconensis*, cepa 13CC$^{T}$

| Category | Subcategory | Subsystem | Role |
|---|---|---|---|
| Virulence, Disease and Defens | Resistance to antibiotics and toxic compound | | |
| Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Copper homeostasis | Cytochrome c heme lyase subunit CcmF |
| Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Copper homeostasis | Copper-translocating P-type ATPase (EC 3.6.3.4) |
| Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Copper homeostasis | Copper resistance protein CopC |
| Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Cobalt-zinc-cadmium resistance | Cobalt-zinc-cadmium resistance protein |
| Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Cobalt-zinc-cadmium resistance | Transcriptional regulator, MerR family |
| Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Zinc resistance | Response regulator of zinc sigma-54-dependent two-cor system |
| Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Arsenic resistance | Arsenical-resistance protein ACR3 |
| Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Arsenic resistance | Arsenical resistance operon trans-acting repressor ArsD |
| Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Arsenic resistance | Arsenical resistance operon repressor |
| Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Arsenic resistance | Arsenical pump-driving ATPase (EC 3.6.3.16) |
| Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Arsenic resistance | Arsenate reductase (EC 1.20.4.1) |
| Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Resistance to fluoroquinolones | DNA gyrase subunit B (EC 5.99.1.3) |
| Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Resistance to fluoroquinolones | DNA gyrase subunit A (EC 5.99.1.3) |
| Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Resistance to fluoroquinolones | Topoisomerase IV subunit B (EC 5.99.1.-) |
| Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Resistance to fluoroquinolones | Topoisomerase IV subunit A (EC 5.99.1.-) |

Con esta predicción se decidió hacer la comprobación de la resistencia a arsénico, cadmio, cobalto, cobre y zinc y así poder corroborar los genes anotados.

### 6.4 Comparación de *B. selenitireducens* y *B. cellulosilitycus* con *T. texcoconensis*

Con la información generada con los dos árboles filogenéticos, se hizo una comparación con BLAST en NCBI de los genomas, con la finalidad de obtener un genoma de referencia para así poder completar el ensamblaje previamente realizado;

sin embargo no se obtuvo un genoma con similitud al de *T. texcoconensis*; así que se optó por descargar los genomas de *B. selenitireducens* y *B. cellulosilitycus* y se cargaron en el servidor RAST. Después se hizo una comparación por función con *T. texcoconensis* para poder ver su similitud con éste a nivel de metabolismo. Se obtuvo poca similitud entre los microorganismos, pues en los resultados de la comparación por función, se encontró que *B. selenireducens* tiene 1670 genes en común con *T. texcoconensis*, de los cuales 11 son genes de resistencia a metales pesados (Tabla 4). Mientras que en el caso de *Bacillus cellulosilitycus*, obtuvimos que comparten 2212 genes, de los cuales 7 son genes de resistencia a metales pesados (Tabla 5).

**Tabla 4.** Comparación de la reconstrucción del metabolismo de *B. selenireducens* (A) y *T. texcoconensis* (B).

| Presence | Category | Subcategory | Subsystem | Role | Organism A | SS active A | Organism B | SS active B |
|---|---|---|---|---|---|---|---|---|
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Arsenic resistance | Arsenate reductase (EC 1.20.4.1) | fig\|6666666.70227.peg.635, fig\|6666666.70227.peg.1230 | yes | fig\|6666666.86893.peg.790, fig\|6666666.86893.peg.2726 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Arsenic resistance | Arsenical pump-driving ATPase (EC 3.6.3.16) | fig\|6666666.70227.peg.815 | yes | fig\|6666666.86893.peg.2728 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Arsenic resistance | Arsenical resistance operon repressor | fig\|6666666.70227.peg.636 | yes | fig\|6666666.86893.peg.1106 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Arsenic resistance | Arsenical resistance operon trans-acting repressor ArsD | fig\|6666666.70227.peg.814 | yes | fig\|6666666.86893.peg.2727 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Arsenic resistance | Arsenical-resistance protein ACR3 | fig\|6666666.70227.peg.2772 | yes | fig\|6666666.86893.peg.2116, fig\|6666666.86893.peg.2356, fig\|6666666.86893.peg.2725 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Beta-lactamase | Beta-lactamase class C and other penicillin binding proteins | fig\|6666666.70227.peg.2073 | yes | fig\|6666666.86893.peg.2565 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Cadmium resistance | Cadmium efflux system accessory protein | fig\|6666666.70227.peg.362 | yes | fig\|6666666.86893.peg.3115 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Cobalt-zinc-cadmium resistance | Cobalt-zinc-cadmium resistance protein | fig\|6666666.70227.peg.2562, fig\|6666666.70227.peg.3054 | yes | fig\|6666666.86893.peg.2469 | yes |

**Tabla 5.** Comparación de la reconstrucción del metabolismo de *B. cellulosilitycus* (A) y *T. texcoconensis* (B)

| Presence | Category | Subcategory | Subsystem | Role | Organism A | SS active A | Organism B | SS active B |
|---|---|---|---|---|---|---|---|---|
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Arsenic resistance | Arsenate reductase (EC 1.20.4.1) | fig\|6666666.68933.peg.284, fig\|6666666.68933.peg.3582 | yes | fig\|6666666.68343.peg.904, fig\|6666666.68343.peg.1903 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Arsenic resistance | Arsenical resistance operon repressor | fig\|6666666.68933.peg.282, fig\|6666666.68933.peg.2985 | yes | fig\|6666666.68343.peg.1394 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Arsenic resistance | Arsenical-resistance protein ACR3 | fig\|6666666.68933.peg.196, fig\|6666666.68933.peg.283, fig\|6666666.68933.peg.286 | yes | fig\|6666666.68343.peg.1550, fig\|6666666.68343.peg.1904, fig\|6666666.68343.peg.2789 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Cadmium resistance | Cadmium efflux system accessory protein | fig\|6666666.68933.peg.199, fig\|6666666.68933.peg.329, fig\|6666666.68933.peg.3209 | yes | fig\|6666666.68343.peg.2607 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Cobalt-zinc-cadmium resistance | Cobalt-zinc-cadmium resistance protein | fig\|6666666.68933.peg.1570, fig\|6666666.68933.peg.2157 | yes | fig\|6666666.68343.peg.1140 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Cobalt-zinc-cadmium resistance | Transcriptional regulator, MerR family | fig\|6666666.68933.peg.1103, fig\|6666666.68933.peg.1915 | yes | fig\|6666666.68343.peg.1464 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Copper homeostasis | Copper-translocating P-type ATPase (EC 3.6.3.4) | fig\|6666666.68933.peg.330, fig\|6666666.68933.peg.2483, fig\|6666666.68933.peg.3173, fig\|6666666.68933.peg.3625, fig\|6666666.68933.peg.4289 | yes | fig\|6666666.68343.peg.490, fig\|6666666.68343.peg.786, fig\|6666666.68343.peg.2605 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Cobalt-zinc-cadmium resistance | Transcriptional regulator, MerR family | fig\|6666666.70227.peg.2292, fig\|6666666.70227.peg.3395 | yes | fig\|6666666.86893.peg.2029 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Copper homeostasis | Copper-translocating P-type ATPase (EC 3.6.3.4) | fig\|6666666.70227.peg.361, fig\|6666666.70227.peg.1963, fig\|6666666.70227.peg.3083 | yes | fig\|6666666.86893.peg.225, fig\|6666666.86893.peg.1646, fig\|6666666.86893.peg.3020, fig\|6666666.86893.peg.3113 | yes |
| A and B | Virulence, Disease and Defense | Resistance to antibiotics and toxic compounds | Copper homeostasis | Cytochrome c heme lyase subunit CcmF | fig\|6666666.70227.peg.1408 | yes | fig\|6666666.86893.peg.1120 | yes |

## 6.5 Pruebas de resistencia a metales pesados usando diferentes concentraciones de cada compuesto y Concentraciones Mínimas Inhibitorias (MIC´s).

### 6.4.1 Pruebas de resistencia a metales pesados

*T. texcoconensis* creció en el medio de cultivo adicionado con todos los metales pesados probados, los cuales, como ya se mencionó, fueron: arsénico, cadmio,

cobalto, cobre y zinc. Las concentraciones a las cuales se registró crecimiento del microorganismo se presentan en la Tabla 6.

**Tabla 6.** Resultados del crecimiento de la cepa 13CC a diferentes concentraciones de metales.

| Metal | Concentración |
|---|---|
| **Arsénico (III)- $As_2O_3$** | 1 y 3 mM |
| **Arsénico (V)- $Na_2HAsO_4$** | 5, 9 y 20 mM |
| **Cadmio- $3CdSO_4 \cdot 8H_2O$** | 5µg |
| **Cobalto- $CoCl_2$** | 1, 3 y 5 mM |
| **Cobre (II)- $CuSO_4$** | 2, 5,10 mM |
| **Zinc- $ZnSO_4$** | 0.2 mM |

Debido a que hubo crecimiento de *T. texcoconensis* a las diferentes concentraciones de metales presentadas, se pudo confirmar que los genes identificados con la anotación están funcionando, puesto que las proteínas para los que estos codifican le permitieron al microorganismo crecer en diferentes concentraciones de metales, concentraciones a las cuales otros microorganismos aislados de lugares contaminados con arsénico pueden crecer, como es el caso de *Bacillus arsenicus* a/3$^T$ reportado por Shivaji en 2005 y *Bacillus indicus* reportado por Suresh en 2004, los cuales pueden crecer también a una concentración de 20 mM de As(V), pero solo el último crece a 3 mM de As(III). Por otro lado a la cual *B. arseniciselenatis* DSM 15340$^T$ (Shivaji, 2005), el cual a pesar de haber sido aislado de un cuerpo de agua contaminado con arsénico, no tuvo la capacidad de crecer a esta concentración.

Para el caso del cadmio, se probaron concentraciones de 500, 100, 50 y 5 µg de cadmio por mililitro, pero solo se registró crecimiento a la concentración de 5 $\mu g \cdot mL^{-1}$, este valor es diez veces menor que la concentración a la que puede crecer *S. aureus*

que es 50µg·mL$^{-1}$ (Sochor *et al.*, 2011). En cuanto al cobalto, se probaron tres concentraciones: 1, 3 y 8 mM. *T. texcoconensis* creció a 1 y 3mM, que es menor a lo reportado por Schmidt & Schlegel, 1989 para *Alcaligenes eutrophus* que puede crecer a una concentración de 7.5 mM de cobalto. Para cobre, *T. texcoconensis* creció a 2, 5 y 10 mM, valores similares a lo reportado por Elguindi *et al.*, 2011, para *E. coli*, la cual crece concentraciones superiores a 20 mM. Finalmente el zinc, se probaron concentraciones 0.1, 0.2 y 1 mM, de las cuales, *T. texcoconensis* solo creció a 0.1 y 0.2 mM, los cuales son valores bajos en comparación con lo reportado por Xiong & Jayaswal, 1998 para *S. aureus* RN-ZC que puede crecer a concentraciones mayores de 10 mM de zinc.

De esta manera se confirmó que *T. texcoconensis* tenía la capacidad de crecer a diferentes concentraciones de diferentes metales, tal y como se predijo en la anotación.

### 6.4.2 Concentraciones Mínimas Inhibitorias (MIC´s)

Por los resultados de las pruebas anteriores, es decir, que presentó crecimiento, se aumentaron las concentraciones de estos para así poder obtener el valor de las Concentraciones Mínimas Inhibitorias para cada uno de estos metales.

Al igual que con las pruebas de resistencia a metales pesados, la determinación de los MIC´s se hicieron con: $As_2O_3$, $Na_2HAsO_4$, $3CdSO_4{\cdot}8H_2O$, $CoCl_2$, $CuSO_4$ y $ZnSO_4$, incrementando la concentración de todos estos metales. De esta manera se determinaron las MIC´s, tal y como se muestra en la Tabla 7.

**Tabla 7.** Concentraciones mínimas inhibitorias (MICs) para cada metal evaluado, y la comparación con otros microorganismos

| Microorganismo | MIC[a] (mM) | | | | | |
|---|---|---|---|---|---|---|
| | $As^{+3}$ | $As^{+5}$ | $Cd^{+2}$ | $Co^{+2}$ | $Cu^{+2}$ | $Zn^{+2}$ |
| *T. texcoconensis* | 7 | 625 | $3.8x10^{-4}$ | 6 | 16 | 0.5 |
| *Bacillus sp. AR-9* [1] | 5 | 200 | N.D | N.D | N.D | N.D |
| *E. faecium* [3] | N.D | N.D | N.D | N.D | 24 | N.D |
| *S. aureus* RN-ZC [4] | N.D | N.D | N.D | 5 | N.D | >10 |
| *S. aureus* RN-MZ [4] | N.D | N.D | N.D | 0.5 | N.D | 0.5 |
| *E. coli* [3,5] | N.D | N.D | N.D | 0.05 | 20 | N.D |
| *Alcaligenes eutrophus* [2] | N.D | N.D | N.D | 20 | N.D | N.D |
| *P. putida 06909* [6] | N.D | N.D | 1.7 | 0.3 | 2 | 11.5 |

a. MIC: es la concentración más baja del compuesto que inhibe totalmente el crecimiento de los microorganismos en medio sólido después de 72h de incubación. N.D. No Data.
1. Liao *et al.*, 2011; 2. Schmidt & Schlegel, 1989; 3. Elguindi *et al.*, 2011; 4. Xiong & Jayaswal, 1998; 5. Rodrigue *et al.*, 2005; 6. Lee *et al.*, 2001.

Como se puede observar en la Tabla 7, la concentración de As (V) que resiste *T. texcoconensis* en comparación con otros microorganismos es poco más de tres veces a lo reportado por Liao *et al.*, 2011 al hacer la prueba de MIC con *Bacillus sp.* AR-9, el cual pudo crecer hasta 200 mM.

Por otra parte, es interesante que a una concentración de 0.5 mM de zinc la cepa 13CC ya no presente crecimiento, lo que llama la atención, pues el arsénico y el cobalto son más tóxicos que el zinc y si pudo crecer a concentraciones muchas más altas de estos compuestos.

### 6.4.2.1 Concentraciones Mínimas Inhibitorias (MIC´s) con arsénico

A una concentración de 625 mM, la cepa $13CC^T$ de arseniato no creció, mientas que a 620 mM aún se observó crecimiento (Figura 44 b y c). En cuanto al arsenito, a 7 mM de $As_2O_3$, la cepa ya no presenta crecimiento, mientras que a 6 mM aún se observó crecimiento de esta (Figura 44 d y e).

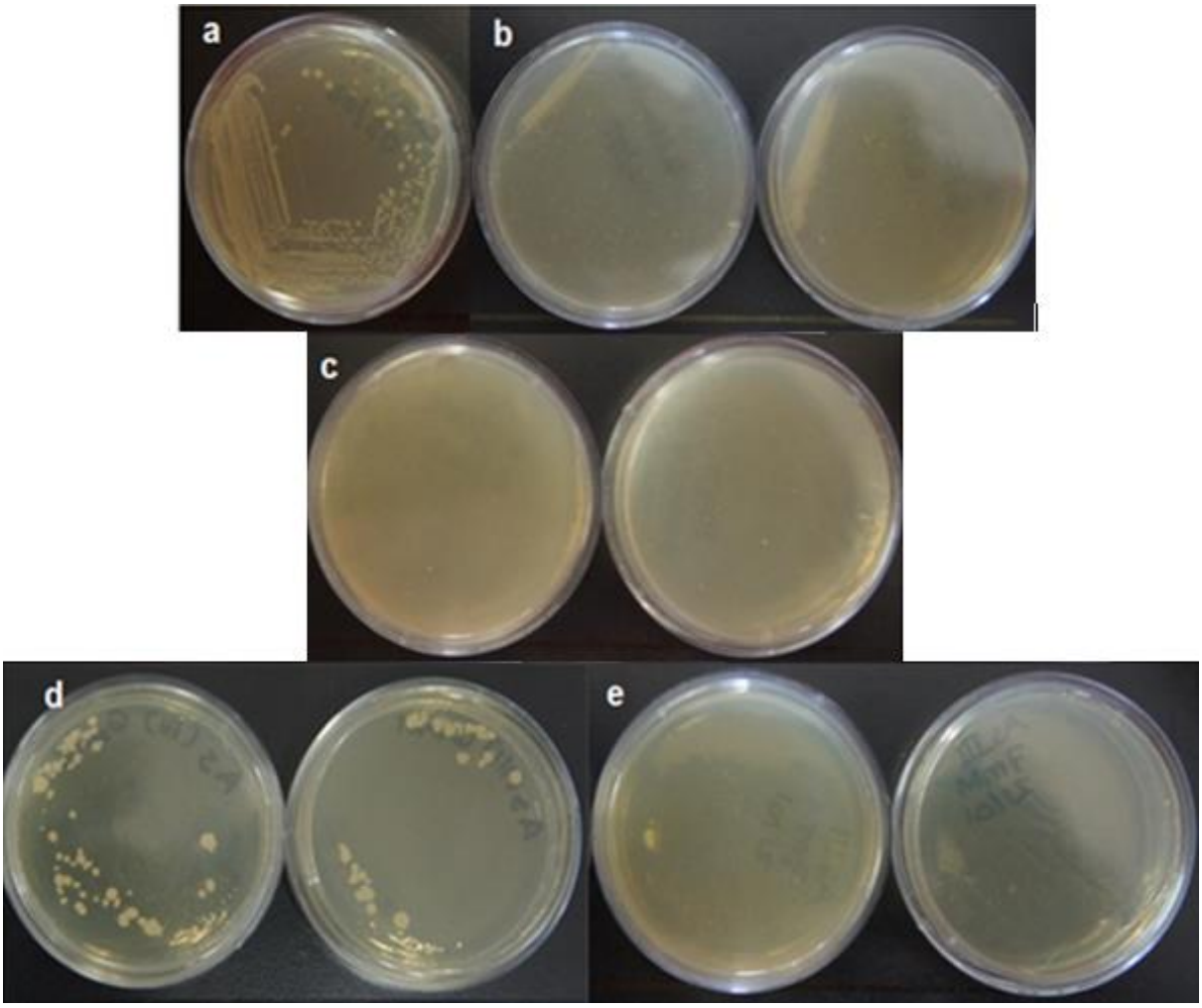



**Figura 44.** Resultados MIC para As (III) y As (V); a) Control, cepa sin metales; b) 620 mM de $As^{+5}$; c) 625 mM de $As^{+5}$; d) 6 mM de $As^{+3}$; e) 7 mM de $As^{+3}$.

### 6.4.2.2 Concentración Mínima Inhibitoria (MIC) con cadmio

Para el cadmio, la MIC fue de $3.8x10^{-4}$ mM, que son aproximadamente 9 µg de $Cd^{+2}$. En comparación con *P. putida 06909* (Lee *et al.*, 2001), en *T. texcoconensis* es mucho más baja la concentración mínima inhibitoria de este metal en el microorganismo. Por lado se registró crecimiento a una concentración de $3.38x10^{-4}$ mM, que corresponde a 8 µg de $Cd^{+2}$. En la Figura 45b se puede ver el crecimiento de la cepa $13CC^{T}$, aunque en comparación con el control es mucho menor, pero en contraste, en la Figura 45c ya no se visualiza ninguna colonia en el medio.

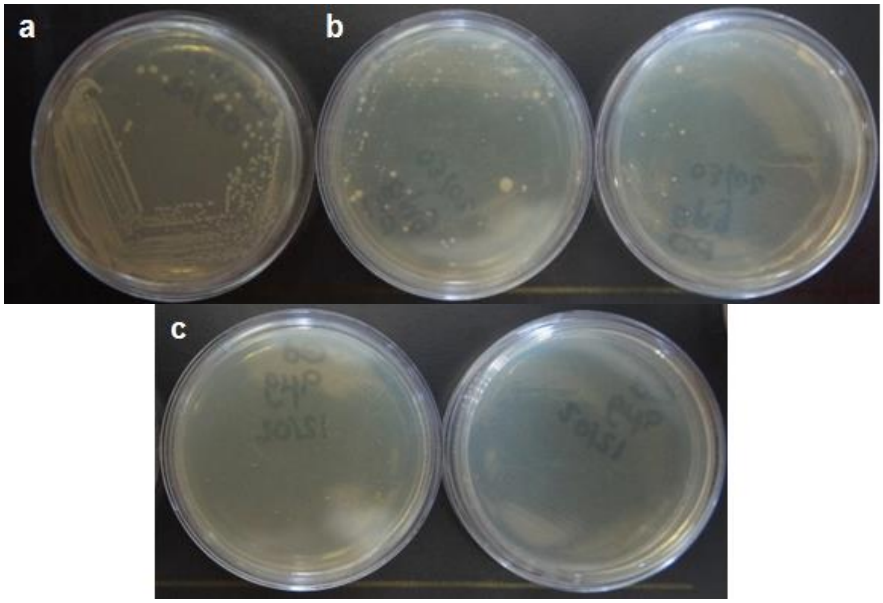



**Figura 45.** Resultados MIC para cadmio; a) Control, cepa sin metales; b) $3.8x10^{-4}$ mM de $Cd^{+2}$; c) $3.38x10^{-4}$ mM de $Cd^{+2}$.

### 6.4.2.3 Concentración Mínima Inhibitoria (MIC) con cobalto

La MIC fue de 6 mM, mientras que a 5 mM aún se tuvo crecimiento, esto se puede observar en la Figura 46b, donde el color rojo es característico del cobalto. En la Figura 46c, se observa que ya no hay crecimiento de la cepa, por lo que corresponde con la concentración mínima inhibitoria.

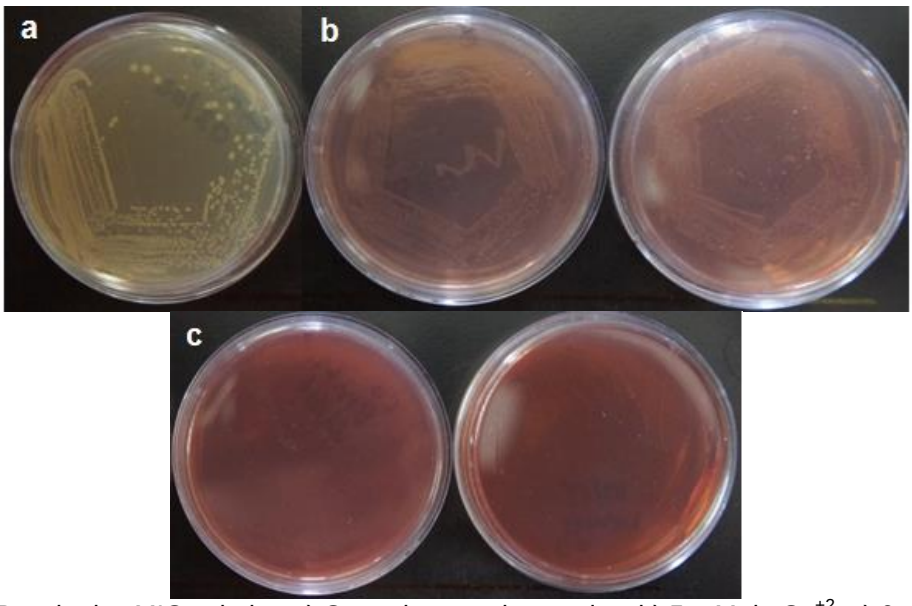



**Figura 46.** Resultados MIC cobalto; a) Control, cepa sin metales; b) 5 mM de $Co^{+2}$; c) 6 mM de $Co^{+2}$.

### 6.4.2.4 Concentración Mínima Inhibitoria (MIC) con cobre

En el caso del cobre, la MIC fue de 16 mM de $Cu^{+2}$ y a una concentración de 15 mM si se tuvo crecimiento. En la Figura 47 se presentan las cajas con la adición de cobre, donde se puede ver en la Figura 47b que aún hay crecimiento del microorganismo, mientras que en la Figura 46, la cepa ya no pudo crecer.

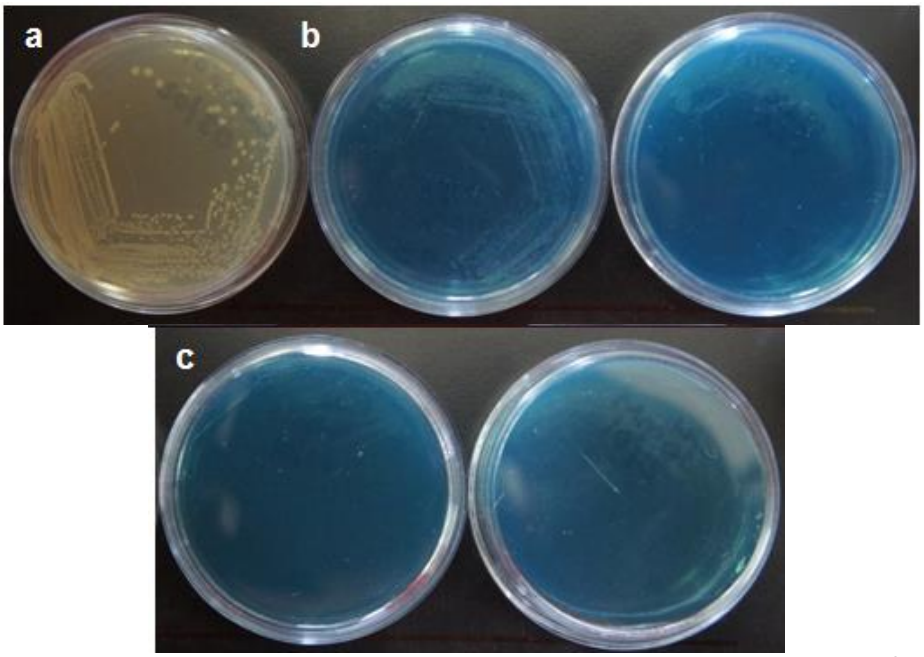



**Figura 47.** Resultados MIC para cobre; a) Control, cepa sin metales; b) 15 mM de $Cu^{+2}$; c) 16 mM de $Cu^{+2}$.

### 6.4.2.5 Concentración Mínima Inhibitoria (MIC) con zinc.

En cuanto al zinc, la cepa sólo pudo crecer hasta una concentración de 0.45 mM, aunque es muy poco el crecimiento que presentó el microorganismo, como se puede ver en la Figura 48b; y a una concentración de 0.5 mM de $ZnSO_4$ se llegó a la concentración mínima inhibitoria, por lo que ya no presenta crecimiento el microorganismo (Figura 48c).

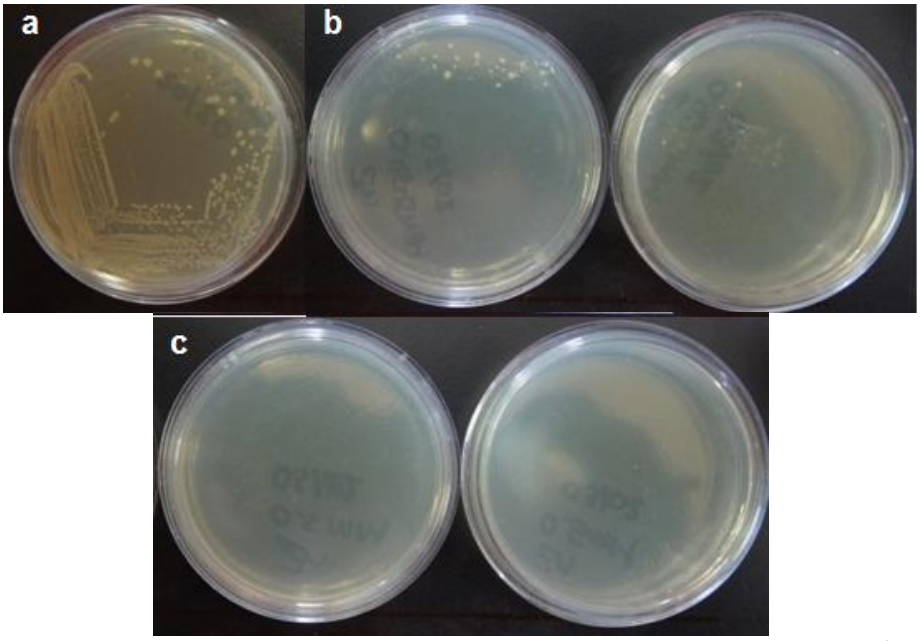



**Figura 48.** Resultados MIC para zinc; a) Control, cepa sin metales; b) 0.45 mM de $Zn^{+2}$; c) 0.5 mM de $Zn^{+2}$.

## 7. Conclusiones

Se logró obtener un ensamblaje del genoma de *T. texcoconensis* cepa 13CC$^{T}$, el cual está organizado en 60 contigs y tiene una longitud de 3, 395, 935 pb, así como un estimado de N´s de 5917.

Al hacer dos árboles filogenéticos basándonos en los genes *rpo*B y *rec*A, pudimos ver que aunque hay dos microorganismos que son filogenéticamente cercanos a *T. texcoconensis*, *B. selenitireducens* y *B. cellulosilitycus;* no lo son suficientemente para usarlos de molde para mejorar el ensamblaje, pues al hacer la comparación de funciones metabólicas con RAST, se encontró que estos solo tienen en común con *T. texcoconensis* algunos genes, 11 con *B. selenitireducens* y 7 con *B. cellulosilitycus.*

Con la anotación, se lograron identificar 1925 posibles genes, de los cuales 11 genes son de resistencia a metales pesados.

El ensamblaje y anotación del genoma de la cepa 13CC$^{T}$ permitieron identificar genes de resistencia a arsénico, cadmio, cobalto, cobre y zinc en este microorganismo, por lo que al hacer las pruebas para corroborar que estos genes se estuvieran expresando, pudimos ver crecimiento del microorganismo y además se determinó la concentración mínima inhibitoria para cada uno de los metales antes mencionados, con esto se corroboró que los genes encontrados en la anotación están activos.

Las concentraciones mínimas inhibitorias fueron de 625 mM para de arseniato, 7 mM para arsenito, 3.8 x$10^{-4}$ mM para cadmio, 6 mM para cobalto y 0.5 mM para zinc. De estas concentraciones, se ve que la resistencia a As(V) de este microorganismos es hasta tres veces mayor que lo reportado para *Bacillus sp.* AR-9.

## 8. Recomendaciones

Es una buena opción usar el ensamblaje y la notación de un genoma para poder tener una primera idea de las características fisiológicas que puede presentar un microorganismo, de esta manera se pueden hacer pruebas experimentales posteriores, según la información que se vaya obteniendo mediante la anotación. Es por ello que recomienda hacer este análisis antes de trabajar con algún microorganismo del cual no se tenga información sobre sus capacidades y características, de esta manera se puede ahorrar tiempo, dinero y esfuerzo.

El haber hecho crecer a la cepa 13CC$^{T}$ en diferentes concentraciones de diferentes metales nos permitió corroborar que los genes presentes se están expresando; sin embargo, se puede explorar en futuros trabajos, cuáles son los mecanismos de resistencia precisos en este microorganismo, puesto que no se sabe cómo es que están siendo activados los genes presentes en esta cepa.

Por otro lado, al conocer cómo es que se lleva a cabo la resistencia a metales en esta bacteria, se podrían proponer usos de este microorganismo para apoyar con la recuperación de algún sitio contaminado con metales como es el arsénico, cobre o cobalto; ya que son los metales a los que presentó mayor resistencia, según lo comparado con la bibliografía.

## 9. Bibliografía


Achour, A. R., Bauda, P., & Billard, P. (2007). Diversity of arsenite transporter genes from arsenic-resistant soil bacteria. *Research in Microbiology*, *158*(2), 128–137. http://doi.org/10.1016/j.resmic.2006.11.006

Altimira, F., Yáñez, C., Bravo, G., González, M., Rojas, L. a., & Seeger, M. (2012). Characterization of copper-resistant bacteria and bacterial communities from copper-polluted agricultural soils of central Chile. *BMC Microbiology*, *12*(1), 193. http://doi.org/10.1186/1471-2180-12-193

Aziz, R. K., Bartels, D., Best, A. a, DeJongh, M., Disz, T., Edwards, R. a, … Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, *9*, 75. http://doi.org/10.1186/1471-2164-9-75

Bruins, M. R., Kapil, S., & Oehme, F. W. (2000). Microbial resistance to metals in the environment. *Ecotoxicology and Environmental Safety*, *45*(3), 198–207. http://doi.org/10.1006/eesa.1999.1860

Cervantes, C., & Gutiérrez-Corona, F. (1994). Copper resistance mechanism in bacteria and fungi. *FEMS Microbiol. Rev.*, *14*, 121–138. http://doi.org/doi:10.1111/j.1574-6976.1994.tb00083.x

Choudhury, R., & Srivastava, S. (2001). Zinc resistance mechanisms in bacteria. *Current Science*, *81*(7), 768–775. http://doi.org/10.1172/JCI57235.M.

Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements, 1394–1403. http://doi.org/10.1101/gr.2289704

Elguindi, J., Moffitt, S., Hasman, H., Andrade, C., Raghavan, S., & Rensing, C. (2011). Medium Influence Survival of Copper-Ion Resistant Bacteria, *89*(6), 1963–1970. http://doi.org/10.1007/s00253-010-2980-x.Metallic

Hassen, a., Saidi, N., Cherif, M., & Boudabous, a. (1998). Resistance of environmental bacteria to heavy metals. *Bioresource Technology*, *64*(1), 7–15. http://doi.org/10.1016/S0960-8524(97)00161-2

Hernandez, D., François, P., Farinelli, L., Østerås, M., & Schrenzel, J. (2008). De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Research*, *18*(5), 802–809. http://doi.org/10.1101/gr.072033.107

Lee, S.-W., Glicmann, E., & Cooksey, D. A. (2001). Chromosomal Locus for Cadmium Resistance in. *Microbiology*, *67*(4), 1437–1444. http://doi.org/10.1128/AEM.67.4.1437

Liao, V. H. C., Chu, Y. J., Su, Y. C., Hsiao, S. Y., Wei, C. C., Liu, C. W., … Chang, F. J. (2011). Arsenite-oxidizing and arsenate-reducing bacteria associated with arsenic-rich groundwater in Taiwan. *Journal of Contaminant Hydrology*, *123*(1-2), 20–29. http://doi.org/10.1016/j.jconhyd.2010.12.003

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, *9*, 387–402. http://doi.org/10.1146/annurev.genom.9.081307.164359

Markowitz, V. M., Chen, I. M. a, Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., … Kyrpides, N. C. (2012). IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research*, *40*(D1), 115–122. http://doi.org/10.1093/nar/gkr1044

Mavromatis, K., Chu, K., Ivanova, N., Hooper, S. D., Markowitz, V. M., & Kyrpides, N. C. (2009). Gene context analysis in the Integrated Microbial Genomes (IMG) data management system. *PloS One*, *4*(11), e7979. http://doi.org/10.1371/journal.pone.0007979

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithm for Next-Ganeration Sequencing data. *Genomics*, *95*(6), 315–327. http://doi.org/10.1016/j.ygeno.2010.03.001.Assembly

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., & Marshall, D. (2009). Tablet-next generation sequence assembly visualization. *Bioinformatics*, *26*(3), 401–402. http://doi.org/10.1093/bioinformatics/btp666

Muller, D., Médigue, C., Koechler, S., Barbe, V., Barakat, M., Talla, E., … Bertin, P. N. (2007). A tale of two oxidation states: Bacterial colonization of arsenic-rich environments. *PLoS Genetics*, *3*(4), 0518–0530. http://doi.org/10.1371/journal.pgen.0030053

Nies, D. H. (1992). Resistance to cadmium, cobalt, zinc, and nickel in microbes. *Plasmid*, *27*(1), 17–28. http://doi.org/10.1016/0147-619X(92)90003-S

Oren, A. (2002a). Diversity of halophilic microorganisms: environments, phylogeny, physiology, and applications. *Journal of Industrial Microbiology & Biotechnology*, *28*(1), 56–63. http://doi.org/10.1038/sj/jim/7000176

Oren, A. (2002b). Molecular ecology of extremely halophilic Archaea and Bacteria. *FEMS Microbiology Ecology*, *39*, 1–7.

Overbeek, R., Disz, T., & Stevens, R. (2004). The SEED: A peer-to-peer environment for genome annotation. *Communications of the ACM*, *47*(11), 47–51. http://doi.org/10.1145/1029496.1029525

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., … Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using

Subsystems Technology (RAST). *Nucleic Acids Research*, *42*(D1), 206–214. http://doi.org/10.1093/nar/gkt1226

Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, *52*(4), 413–435. http://doi.org/10.1007/s13353-011-0057-x

Paszkiewicz, K., & Studholme, D. J. (2010). De novo assembly of short sequence reads. *Briefings in Bioinformatics*, *11*(5), 457–472. http://doi.org/10.1093/bib/bbq020

Ramírez D., N., Serrano R., J. A., & Sandoval T., H. (2006). Actinomicetos halófilos en México. *Revista Mexicana de Ciencias Farmacéuticas*, *37*, 56–71.

Rodrigue, A., Effantin, G., & Mandrand-Berthelot, M.-A. (2005). Identification of rcnA ( yohM ), a Nickel and Cobalt Resistance Gene in Escherichia coli Identification of rcnA ( yohM ), a Nickel and Cobalt Resistance Gene in Escherichia coli, *187*(8). http://doi.org/10.1128/JB.187.8.2912

Rodríguez, J. M. (2004). Secuenciación de genomas. *Arbor*, *698*(Febrero), 285–310.

Ruiz-Romero, E., Coutiño-Coutiño, M. D. L. A., Valenzuela-Encinas, C., López-Ramírez, M. P., Marsch, R., & Dendooven, L. (2013). Texcoconibacillus texcoconensis gen. nov., sp. nov., alkalophilic and halotolerant bacteria isolated from soil of the former lake Texcoco (Mexico). *International Journal of Systematic and Evolutionary Microbiology*, *63*(PART9), 3336–3341. http://doi.org/10.1099/ijs.0.048447-0

Silver, S. (1996). Bacterial resistances to toxic metal ions - A review. *Gene*, *179*(1), 9–19. http://doi.org/10.1016/S0378-1119(96)00323-X

Sochor, J., Zitka, O., Hynek, D., Jilkova, E., Krejcova, L., Trnkova, L., … Kizek, R. (2011). Bio-sensing of cadmium(II) ions using Staphylococcus aureus. *Sensors*, *11*(11), 10638–10663. http://doi.org/10.3390/s111110638

Tsai, S. L., Singh, S., & Chen, W. (2009). Arsenic metabolism by microbes in nature and the impact on arsenic remediation. *Current Opinion in Biotechnology*, *20*(6), 659–667. http://doi.org/10.1016/j.copbio.2009.09.013

Ventosa, a., & Nieto, J. J. (1995). Biotechnological applications and potentialities of halophilic microorganisms. *World Journal of Microbiology & Biotechnology*, *11*(1), 85–94. http://doi.org/10.1007/BF00339138

Xiong, a., & Jayaswal, R. K. (1998). Molecular characterization of a chromosomal determinant conferring resistance to zinc and cobalt ions in Staphylococcus aureus. *Journal of Bacteriology*, *180*(16), 4024–4029.

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. http://doi.org/10.1101/gr.074492.107

*Paginas de internet:*

Burrows-Wheeler Aligner, 28/02/2010, [citado 06/04/2015], Formato html, Disponible en Internet: http://bio-bwa.sourceforge.net/

IMG Data Management, U.S. Department of Energy, Office of science, IMG version 4.530 June 2015, [citado 06/04/2015], Formato html, Disponible en Internet: http://darlinglab.org/mauve/user-guide/viewer.html