“Mycoplasma pneumoniae – PROYECTO DE BIOINFORMÁTICA 2”.

C. González-Quevedo[1]

[1]Escuela Nacional de Estudios Superior Unidad León
Universidad Nacional Autónoma de México
*cass_gnzl@comunidad.unam.mx

Resumen

Introducción: *M. pneumoniae* es una bacteria patógena exclusivamente humana y de distribución universal. Penetra por vía aérea y se adhiere a las células epiteliales respiratorias a través de las proteínas de su organela terminal.

Especie en estudio: *Mycoplasma pneumoniae*

Métodos: Se obtuvieron los datos de SRA publicados por el centro de regulación genómica, posteriormente se utilizaron programas bioinformáticos como FastQC, FastP, Bowtie2 y Spades. Además de plataformas como Augustus y BlastX del NCBI.

Resultados: Se pasó por un proceso de control y filtrado de calidad, mapeo, ensamble y anotación del genoma que se había seleccionado y se pudo encontrar que algunos de los genes para los que codificaba eran de una proteína de membrana, una GTPasa y una Thymidine phosphorylase de *Mycoplasma pneumonia*

Abstract

Background: M. pneumoniae is an exclusively human pathogenic bacterium with a universal distribution. It enters by air and attaches itself to respiratory epithelial cells through the proteins of its terminal organelle.

Species studied: *Mycoplasma pneumoniae*

Methods: The SRA data published by the genomic regulation center were obtained, later bioinformatics programs such as FastQC, FastP, Bowtie2 and Spades were used. In addition to platforms such as Augustus and BlastX from NCBI

Results: A process of quality control and filtering, mapping, assembly and annotation of the genome that had been selected was passed through and it was found that some of the genes for which it encoded were of a membrane protein, a GTPase and a Thymidine phosphorylase of Mycoplasma pneumonia

Los micoplasmas pertenecen a la familia Mycoplasmataceae de la clase Mollicutes (molli cutis: piel blanda). Son bacterias de pequeño tamaño, de hecho, son las bacterias más pequeñas con capacidad de división autónoma y vida libre[1]. La falta de pared celular condiciona muchas de las características del microorganismo, como su polimorfismo, que no se tiñan por la tinción de Gram, su resistencia a los antibióticos betalactámicos y su elevada sensibilidad a las variaciones de pH, temperatura, la tensión osmótica y los detergentes.[4]

La única protección externa es su membrana citoplásmica, la cual posee un gran número de lipoproteínas denominadas lipid-associated membrane proteins, muy antigénicas, y que son reconocidas por las células inmunitarias con unos receptores tipo Toll (toll-like receptors). Estas lipoproteínas pueden modular el sistema inmunitario e inducir la apoptosis celular o la muerte de las células inmunitarias. Además, poseen un citoesqueleto que forma un recubrimiento de elementos proteicos organizado helicoidalmente en forma de una red regular que envuelve por completo el citoplasma. Poseen una organela polar terminal multifuncional, asociada al citoesqueleto, compuesta por varias proteínas, que es esencial para la adherencia a las células del huésped, pero que participa, además, en el movimiento deslizante de estas bacterias y en su división celular[1,2].

La mayoría de las especies de micoplasmas no son patógenas y son habitantes comunes de las mucosas respiratorias o genitales. La especie de mayor importancia es *M. pneumoniae*, causante de infecciones respiratorias. [5]

M. pneumoniae es una bacteria patógena exclusivamente humana y de distribución universal. Penetra por vía aérea y se adhiere a las células epiteliales respiratorias a través de las proteínas de su organela terminal: la proteína P1 es una adhesina (citadhesina) de especial importancia en la patogenia y también es la diana de los principales anticuerpos que produce la respuesta inmunitaria del huésped[3]. También produce peróxidos que alteran el movimiento ciliar y daña a las células. [5]

El trabajo de secuenciación se realizó con los datos proporcionados por el centro de regulación genómica. Posteriormente se llevo a cabo un proceso de control de calidad, filtración, mapeo, ensamble y anotación de una pequeña porción del genoma. Todo esto con ayuda de algunas herramientas bioinformáticas como Fastqc, Fastp, Bowtie2, Spades, Augustus y Blast2Go.

Descarga y control de calidad

Se seleccionó la secuencia ERR5948382 de P. pneumoniae desde el SRA que fue proporcionada por el centro de regulación genómica en su trabajo de investigación “Rational engineering of Mycoplasma pneumoniae as attenuated chassis”, en donde además de otra información se proporcionan los datos básicos de la librería de dicha secuencia.

Organism: Mycoplasma pneumoniae M129
Instrument: Illumina MiSeq
Strategy: WGS
Source: Genomic
Selection: Size fractionation
Layout: Paired

Posterior a su descarga, se realizó un control de calidad con la ayuda de FastQC la cual es una herramienta que presenta un output HTML con estadísticas básicas, calidad por secuencia de base, contenido de GC, niveles de duplicación y el contenido de adaptadores. Todo de una manera muy gráfica por lo que es más sencillo su interpretación.

Filtrado de secuencias

Se hizo uso de FastQ en el que además de los parámetros básicos se le agregó un parámetro un poco mas especifico para tener filtrados de mejor calidad y al final se realizó un conteo total de la longitud de cada archivo comparándolos con la longitud de los originales. Esto con la finalidad de ver las diferencias después del filtrado aplicado.

Mapeo

Se realizó una descarga de un genoma de referencia que en este caso fue el de *Mycoplasma genitalium* G37 y con este se realizó un índice para poder realizar dicho mapeo con la ayuda de Bowtie2 y posteriormente se filtraron las secuencias que no fueron mapeadas, con la finalidad de obtener mejor calidad en nuestro archivo.

Ensamble de Genoma

Tomando en cuenta que en la secuencia trabajada se utilizó la tecnología de Illumina, fue necesario activar un environment en donde se tenia la herramienta "Spades" que precisamente esta diseñada para ensambles cortos para tecnologías como Illumina.

Posterior a su ensamble se comprobó la calidad con Quast, primero se verificó la calidad del ensamble en general y posteriormente se comprobó la calidad con el genoma de referencia. Esto nos arrojo los resultados en un output HTML.

Anotación

Se realizó un conteo general de scaffolds para ver la cantidad que se tenían en el archivo generado del ensamble. Posteriormente se hizo uso de la plataforma FAS Center for Systems Biology para convertir el archivo en scaffolds en un formato tabular para posteriormente poder cortar la columna con la longitud de cada uno de los scaffolds. Teniendo esto se seleccionó el primer scaffold que presentaba una longitud de 316434 con el nombre NODE_1_length_316434_cov_797.810397 y se separo del resto de los scaffolds.

Posteriormente se usó la plataforma Augustus que nos ayudó a determinar los genes que se tenían en el scaffold seleccionado. Esto en comparación con el algoritmo de *Escherichia coli* por ser el más cercano a nuestro organismo de estudio *M. pneumoniae.*

Para la anotación funcional de los genes se usó la plataforma del NCBI especialmente BLASTX en el cual se le dejaron los parámetros por default a excepción del parámetro "Max target sequences" que se cambio a 10 y el parámetro "Expect treshold" que se cambio a 0.005.

Resultados

Al realizar el control de calidad se obtuvieron dos resultados debido a que era una secuencia pareada. En las estadísticas básicas se proporciona la información general sobre dichas secuencias.

Measure	Value
Filename	ERR5948382_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1301991
Sequences flagged as poor quality	0
Sequence length	301
%GC	41

Figura 1 Estadísticas básicas del reporte de FastQC para la secuencia ERR5948382_1.fastq

De manera general se observa que en ambos archivos HTML la calidad de las secuencias es relativamente buena ya que los porcentajes de calidad se encuentran la mayoría por arriba del 30 Phred score, el porcentaje de GC se presenta de manera estable por lo que no se percibe contaminación. El porcentaje de N es cero, lo cual quiere decir que todas las bases fueron llamadas y no hay ninguna que no haya sido reconocida por el programa. Los niveles de duplicación de la mayor proporción de la secuencia, al igual que las secuencias sobre representadas y los adaptadores se encuentran en cero, lo que nos indica buena calidad.

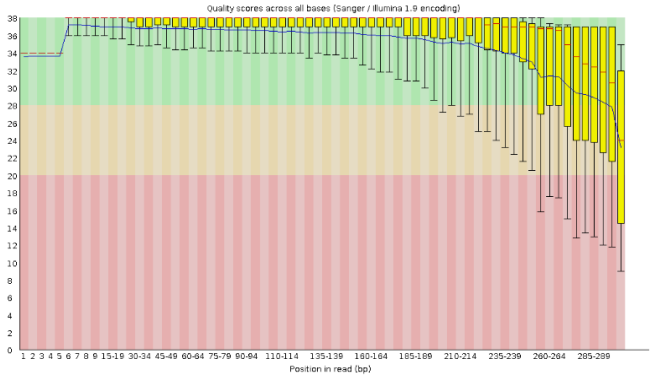

Figura 2 Calidad por base de la secuencia ERR5948382_1.fastq

Posteriormente al filtrar las secuencias se pudo visualizar que, si hubo una diferencia al menos en la eliminación de los adaptadores, con lo que al final en la comparación nos quedaron los siguientes tamaños:

```
5207964 ERR5948382_1.fastq
5207964 ERR5948382_2.fastq

5142612 Filtrado_1.fastq
5142612 Filtrado_2.fastq
```

Con una cobertura total de 391,899,291 nucleótidos.

En el ensamble del genoma uno de los resultados fue un gráfico con los contigs obtenidos por tamaño de referencia. (Fig. 3)

Finalmente, en la anotación con Augustus se obtuvó el grafico con la representación de algunos genes por región. (Fig. 4). Además de obtener resultados en texto plano con la misma predicción de genes, la secuencia total del scaffold seleccionado, la predicción de secuencias de aminoácidos y de secuencias codificantes.

Se seleccionaron solamente 3 genes para hacer la anotación funcional, los cuales fueron los genes “g188.t1”, “g189.t1” y “g190.t1”.

La mejor coincidencia del BlastX para el gen “g188.t1” fue una proteína de membrana especialmente de *Mycoplasma pneumon*iae. Por otro lado, la mejor coincidencia para el gen “g189.t1” fue una GTPasa ObgE de *Mycoplasma pneumoniae* y finalmente para el gen “g190.t1” la mejor coincidencia fue una Thymidine phosphorylase de *Mycoplasma pneumonia*

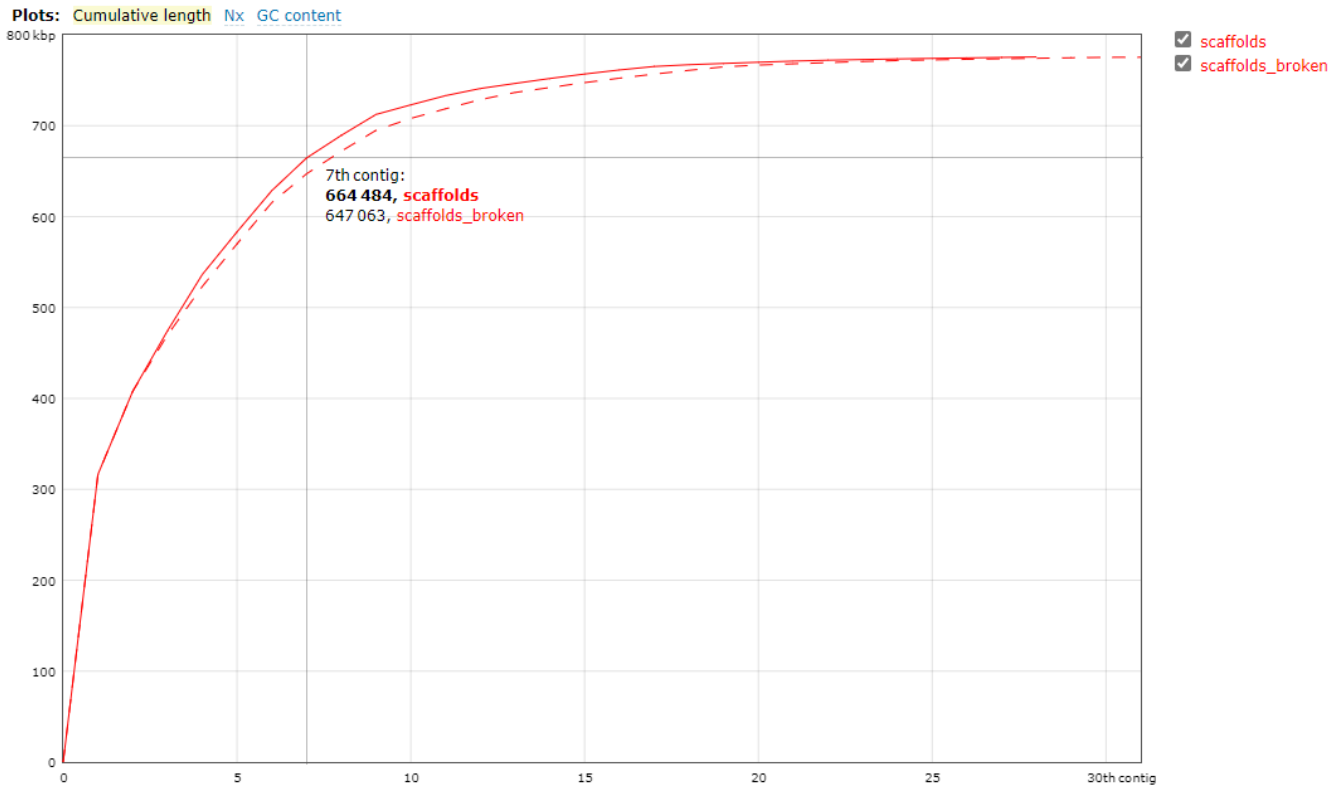

Figura 3Contigs ordenados del más largo (#1) al más corto

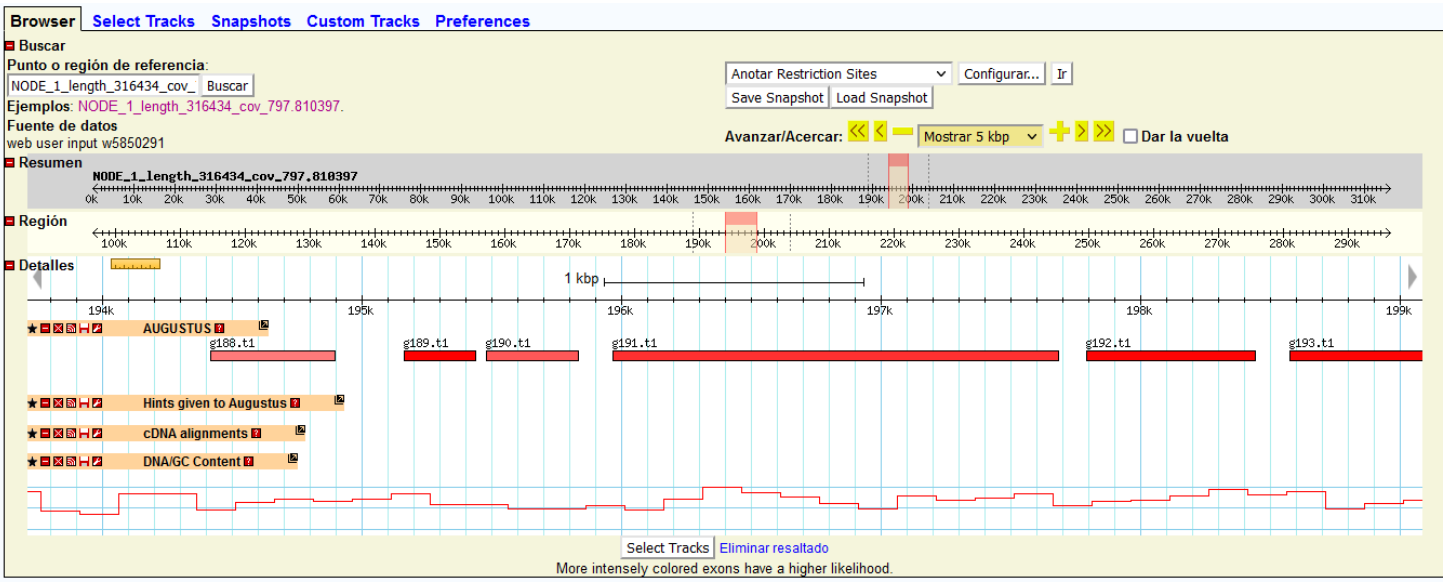

Figura 4 Predicción de genes por Augustus determinada por seccion de 5kbp

PROCEDIMIENTO

Todo se realizó desde GAIA. Se seleccionó el organismo *Mycoplasma pneumoniae* y se descargó una secuencia de SRR, en la cual se aplicó el comando

```
fastq-dump --split-3 ERR5948382
```

Esto debido a que eran secuencias pareadas y se necesitaba que estuvieran en archivos separados

1) Control de calidad

A dichos archivos descargados se les aplicó un control de calidad con el comando

```
fastqc ERR5948382_1.fastq
```

```
fastqc ERR5948382_2.fastq
```

Al terminar de aplicar el control de calidad se crearon unos archivos HTML los cuales se descargaron con el comando

```
scp -P 10022 cassgonzalez@132.247.172.26:/projects/cassgonzalez/Intento2/ERR5948382_1_fastqc.html .
```

```
scp -P 10022 cassgonzalez@132.247.172.26:/projects/cassgonzalez/Intento2/ERR5948382_2_fastqc.html .
```

En estos archivos se pudo analizar la calidad que tenían.

2) Filtrado de secuencias

Para realizar el filtrado además de los parámetros básicos se le agregó uno un poco más específico para tener un mejor filtrado. Se uso el comando

```
fastp -i ERR5948382_1.fastq -I ERR5948382_2.fastq -o Filtrado_1.fastq -O Filtrado_2.fastq >&  Filtrado.log
```

Cuando terminó de correr el código anterior se aplico

```
Wc -l *.fastq
```

Esto se hizo con la finalidad de hacer una comparación entre los archivos originales y ver que tanto mejoraron con el filtrado que se les aplicó

3) Cálculo de cobertura

Después de haber realizado el control de calidad nos podemos dar cuenta en los reportes generados en HTML que el total de las secuencias es de 1301991 y la longitud de cada una es de 301 nucleótidos. Con eso podemos saber que la cobertura total es de

```
1301991*301 = 391,899,291
```

4) Mapeo (si existe referencia)

Para poder iniciar el mapeo fue necesario empezar con descargar un genoma de referencia desde el NCBI. En este caso se escogió "*Mycoplasma genitalium G37*".

Se subió a GAIA con el comando

```
scp -P 10022 /root/sandbox/Ref.fa cassgonzalez@132.247.172.26:/projects/cassgonzalez
```

Posteriormente se debe realizar un índice, el cual se hizo con

```
bowtie2-build Ref.fa RefParaERDB
```

Ya teniendo nuestro índice se corrió el mapeo con

```
bowtie2 --maxins 1000 -x RefParaERDB -1 /projects/cassgonzalez/Intento2/Filtrado_1.fastq -2
/projects/cassgonzalez/Intento2/Filtrado_2.fastq -S Mapeo3Proyecto.sam
```

Ahora para poder descargar el archivo a nuestra computadora no era tan sencillo ya que era muy pesado por lo que fue necesario eliminar las secuencias que no se mapearon y para esto se uso

```
awk '$3!="*"' Mapeo3Proyecto.sam >MapeoListo.sam
```

Ahora si ya es mas sencillo descargar el archivo con

```
scp -P 10022 cassgonzalez@132.247.172.26:/projects/cassgonzalez/MapeoListo.sam .
```

5) Ensamble de genoma

Para realizar el ensamble del genoma primero es necesario activar el environment con

```
Conda actívate spades
```

Se utilizó spades debido a que la tecnología que se utilizó para mi secuencia fue Illumina

Para correr el ensamble se utilizó

```
spades.py -k 33,37,41 -t 1 -m 7 --pe1-1 /projects/cassgonzalez/Intento2/Filtrado_1.fastq --pe1-2
/projects/cassgonzalez/Intento2/Filtrado_2.fastq -o spades2_ER
```

Después de que se corrió el ensamble, se comprueba la calidad con

```
quast --split-scaffolds -t 1 /projects/cassgonzalez/spades2_ER/scaffolds.fasta
```

Y también se comprobó la calidad con la referencia, pero para esto fue necesario moverse a la carpeta que se generó con el código anterior "quast results". Dentro de la carpeta se aplicó el siguiente código

```
quast.py --split-scaffolds -t 1 -r ../Ref.fa /projects/cassgonzalez/spades2_ER/scaffolds.fasta
```

Este último código nos generó una carpeta que contiene dos archivos HTML que son los que más no interesa visualizar. Para eso vamos a descargarlos con

```
scp -P 10022
cassgonzalez@132.247.172.26:/projects/cassgonzalez/quast_results/results_2021_12_05_20_09_50/report.html .
```

```
scp -P 10022
cassgonzalez@132.247.172.26:/projects/cassgonzalez/quast_results/results_2021_12_05_20_09_50/icarus.html .
```

6) Anotación (de una porción)

Ya que se tiene el archivo fasta que se generó del ensamble entonces primero se necesita ver cuántos scaffolds se tienen y eso se hace con

Grep '>' -c scaffolds.fasta

Con eso nos damos cuenta que se tienen 469

Ahora para poder seleccionar solo una porción o bien solo un scaffold es necesario primero convertir nuestro archivo scaffolds.fasta a un formato tabular, eso lo hacemos con la ayuda de la siguiente pagina

http://archive.sysbio.harvard.edu/CSB/resources/computational/scriptome/UNIX/Tools/Change.html

Para convertir en tabular se usó el código:

```
perl -e ' $count=0; $len=0; while(<>) { s/\r?\n//; s/\t/ /g; if (s/^>//) { if ($. != 1)
{ print "\n" } s/ |$/\t/; $count++; $_ .= "\t"; } else { s/ //g; $len += length($_) }
print $_; } print "\n"; warn "\nConverted $count FASTA records in $. lines to tabular
format\nTotal sequence length: $len\n\n"; ' scaffolds.fasta > scaffolds.tab
```

Posteriormente se debe saber cuál es la longitud de los scaffolds y esa esta dada por le última columna del archivo que se generó del código pasado, por lo que se usó el siguiente código para saberlo

```
perl -e ' $col=-1; while (<>) { s/\r?\n//; @F = split /\t/, $_; $len = length($F[$col]); print "$_\t$len\n" } warn "\nAdded column with length of column $col for $. lines.\n\n"; ' scaffolds.tab > seqs_length.tab
```

Se visualizó el archivo con

```
less seqs_length.tab
```

Como el archivo es muy grande no se puede visualizar muy bien por lo fue necesario cortar la última columna con

```
Cut -f 4 seqs_length.tab > OnlyLength.list
```

Visualizamos la columna con

```
less OnlyLength.list
```

Y en esta columna solo se presenta el tamaño de los diferentes scaffolds que se generaron, de esta manera se seleccionó el primero que tenía una longitud de 316434

Para poder seleccionar solo ese scaffold de todo el documento primero fue necesario saber cuál era el nombre del scaffold y eso se realizó con

```
head -n 1 seqs_length.tab | cut -f 1
```

De esta manera el nombre fue NODE_1_length_316434_cov_797.810397

Antes de cortar el scaffold seleccionado se creó un archivo para mandar ahí la información del scaffold. Esta se creó con

```
Cat > CassG_1.list
```

Y en la primera línea se escribió el nombre del scaffold “NODE_1_length_316434_cov_797.810397”

Para separar el scaffold seleccionado se utilizó

```
perl -e ' ($id,$fasta)=@ARGV; open(ID,$id); while (<ID>) { s/\r?\n//; /^>?(\S+)/; $ids{$1}++; } $num_ids = keys %ids; open(F, $fasta); $s_read = $s_wrote = $print_it = 0; while (<F>) { if (/^>(\S+)/) { $s_read++; if ($ids{$1}) { $s_wrote++; $print_it = 1; delete $ids{$1} } else { $print_it = 0 } }; if ($print_it) { print $_ } }; END { warn "Searched $s_read FASTA records.\nFound $s_wrote IDs out of $num_ids in the ID list.\n" } ' CassG_1.list scaffolds.fasta > Scaffold1.fna
```

Finalmente descargamos ese archivo que se generó con el siguiente código

```
scp -P 10022 cassgonzalez@132.247.172.26:/projects/cassgonzalez/spades2_ER/Scaffold1.fna .
```

Para poder hacer la anotación se hizo uso de la herramienta de Augustus

http://bioinf.uni-greifswald.de/augustus/submission.php

En esta página además de subir nuestra secuencia se seleccionó el organismo de *Escherichia coli* ya que era el organismo más cercano (por ser una bacteria) para el que hay un algoritmo en esta plataforma y los demás parámetros se quedaron por default. Al finalizar nos generó varios archivos con resultados en texto plano y también unos gráficos sobre los genes que se encontraron.

CONCLUSIONES

La bacteria *Mycoplasma pneumoniae* causa infecciones que afectan al tracto respiratorio superior e inferior en personas de todas las edades[4]. Además de que tiene una fácil transmisión y forma de contagio sin distinciones de genero y edad. Se puede presentar en cualquier estación del año con una ligera tendencia a elevarse en verano. Por esta razón, entre muchas otras este organismo presenta especial interés en la investigación. En este trabajo se llevó a cabo una secuenciación y ensamble de genoma con un genoma de referencia como base, además de que se llevó a cabo una pequeña parte de anotación funcional, sim embargo no se hizo uso de todas las herramientas bioinformáticas que están disponibles hoy en día. Por lo que se recomienda aplicar las herramientas más especializadas para un mejor resultado ya que es solo una pequeña parte de lo mucho que hay que estudiar y descubrir de esta bacteria

REFERENCIAS

1. Waites KB, Talkington DF. Mycoplasma pneumo niae and its role as human pathogen. Clin Microbiol Rev. 2004;17:697-728.
2. Baum SG. Mycoplasma pneumoniae and atypical pneumonia. En: Principles and practice of infectious diseases. Mandel GI, Bennett JE, Dolin R, editores. 7.ª ed. New York: Chur chill Livingstone; 2010. p. 2481-8.
3. Ciesielski, C. A., Blaser, M. J., & Wang, W. L. (1986). Serogroup specificity of Legionella pneumophila is related to lipopolysaccharide characteristics. Infection and immunity, 51(2), 397-404
4. Atkinson TP, Balish MF, Waites KB. Epidemiology, clinical manifestations, pathogenesis and laboratory detection of M. pneumoniae. FEMS Microbial Rev. 2008;32:956-73.
5. Gonzalo De Liria, C. R., & Hernández, M. M. (2013). Infecciones causadas por Mycoplasma pneumoniae. Anales de Pediatría Continuada, 11(1), 23–29. https://doi.org/10.1016/s1696-2818(13)70114-8