Universidad de Murcia

Facultad de Biología

Estudio del pangenoma de *Clostridium perfringens*

Análisis de datos ómicos

Autores:
Manuel Piñero Hernández
Guillermo Sánchez-Cid Bueno

Abril 2022

Índice

1. Introducción

Clostridium perfringens es una bacteria (filo *Firmicutes*) Gram-positiva anaeróbica y formadora de esporas la cual ha logrado tener un gran éxito evolutivo a la hora de adaptarse a una amplia variedad de ambientes y hospedadores. Prueba de ello es la numerosa cantidad de casos en los que se ha logrado aislar en circunstancias tan dispares como el intestino de perro, humanos, comida o en aguas contaminadas. Si bien esta especie es conocida por sus toxinas y su alta patogenicidad, existen otras cepas que utilizan otras vías para infectar hospedadores, son capaces de vivir gran parte de su ciclo de vida en fómites o incluso no son patogénicas. El conjunto de genes de todas las cepas de esta especie, también conocido como pangenoma, debe ser un conjunto muy variado de adaptaciones evolutivas que han permitido a *Clostridium perfringens* desenvolverse con éxito en muchas circunstancias.

Recientemente, el análisis de la secuenciación del genoma completo (Whole-Genome Sequence - WGS) de 56 cepas de *C. perfringens* reveló un genoma altamente divergente y con una significativa cantidad de transferencia horizontal de genes [4]. Esta variabilidad genómica se ha observado en aspectos como el contenido episómico, el tamaño del cromosoma y los elementos móviles. Sin embargo, también se ha observado un alto grado de preservación en regiones conservadas con alta homología en el cromosoma de las cepas de esta especie [1].

Es, por tanto, una bacteria de gran interés para la salud pública, por lo que ha sido objeto de muchas investigaciones. En este trabajo tenemos como propósito explorar el pangenoma de *Clostridium perfringens* con la ayuda de diversas herramientas implementadas por la comunidad científica.

2. Resultados y discusión

2.1. Análisis del pangenoma

El análisis del pangenoma con el software Roary ha separado los genes de las 10 cepas en las siguientes clases:

- **Core**: se han identificado 1680 genes (18, 7 % del pangenoma). En la figura 2B se divide el core en dos. El core propiamente dicho, corresponde a clusters de genes presentes en las 10 cepas (en entre el 99 % y 100 % de las cepas). El soft-core es el formado por los clusters de genes encontrados en entre el 95 % y 99 % de las cepas (9 cepas). En nuestro caso el valor para el soft-core es 0.
- **Accesorio**: el genoma accesorio de las 10 cepas consideradas asciende a 7317 genes (81, 3 % del pangenoma). Podemos dividirlo en los genes del «caparazón» (*shell genes*) y en los de la «nube» (*cloud genes*). Los genes del caparazón son los compartidos por entre el 15 % y 95 % de las cepas (entre 2 y 9 cepas en nuestro estudio). Por otro lado, los genes de la nube son los presentes en menos del 15 % de las cepas. En nuestro estudio los genes de la nube representan genes exclusivos a una cepa. Los genes del caparazón son 5074, y los de la nube 2243.

El total de genes de estas 10 cepas de *Clostridium perfringens* es 8997.

La figura 1 muestra una vista linealizada del pangenoma de las 10 cepas. Para ello, se construyó un árbol filogenético en base a los genes del core con el software FastTree (figura 1A). Este programa produce árboles filogenéticos con mayor rapidez que otros programas de amplio uso en el campo de la filogenómica, como RAxML 7, y utiliza el método bayesiano Approximate Maximum Likelihood Estimation (AMLE).

La figura 1B representa la matriz del pangenoma. Las líneas azules identifican regiones de sintenia, alineadas junto al árbol filogenético comentado más arriba. La línea amarilla vertical separa el genoma core, constituido por regiones de sintenia en todas las cepas, del genoma accesorio. Dentro del genoma accesorio se observan regiones que están presentes en solo una cepa o en varias.

Además, se puede observar que cepas más cercanas filogenéticamente tienen patrones de sintenia más similares. Obsérvese, por ejemplo, las regiones alineadas entre ATCC3626 y JFP981, que forman un clado propio.

En general, el árbol filogenético producido concuerda con los resultados expuestos por Kiu et. al 2017 , en el que se analizó el pangenoma de 56 cepas de *Clostridium perfringens*, incluidas las 10 utilizadas

para este trabajo [3]. Por ejemplo, JGS1495 y JGS1987 forman parte del Clado 3 en el artículo original, y son las únicas cepas de este clado utilizadas en este trabajo. El árbol producido aquí las agrupa en un clado propio, lo que es indicativo de un buen análisis. Podemos decir algo similar sobre el clado formado por ATCC3626, JFP981, 1207 CPER y JJC, las cuales forman parte del Clado 2 en el artículo original. Por otro lado, JP838 y MJR7757A forman parte del Clado 4, y son linajes hermanos en el árbol de este trabajo. Para terminar, SM101 (Clado 1) tiene como parientes más cercanos a las cepas del Clado 4. JGS1721 (Clado 2) aprece como un outgroup.

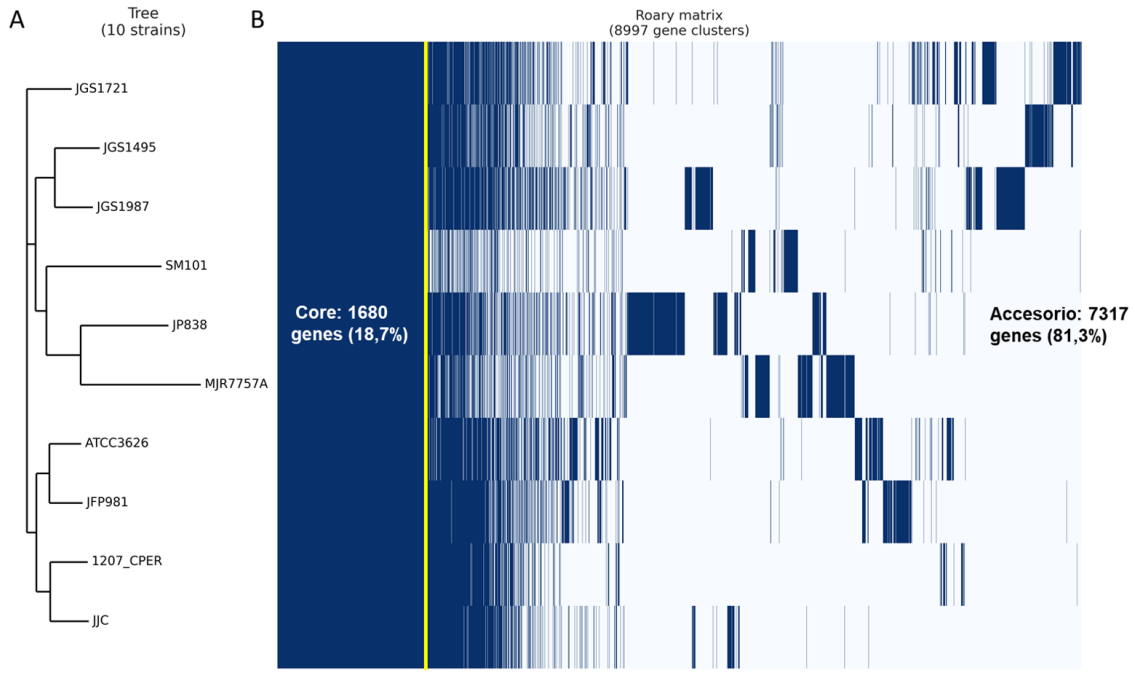

Figura 1: **A)** Árbol filogenético de las 10 cepas estudiadas, basado en el alineamiento del genoma core. **B)** Matriz lineal del pangenoma, alineada junto al árbol filogenético. La línea amarilla vertical separa el genoma core del genoma accesorio.

La figura 2A muestra un gráfico de frecuencias que identifica el número de genes idénticos según el número de genomas estudiados. Podemos observar que existen muchos genes únicos de una cepa (5074, como se indica en la figura 2B). En el lado opuesto de la gráfica aparecen los genes que son compartidos por las 10 cepas, o lo que es lo mismo, el genoma core, 1680 genes (ver figura 2A).

Destacar que, en la figura 2B, se observa de una forma más simplificada la distribución de genes en función del número de cepas que tienen en común dichos genes mediante un diagrama de sectores o de tarta. Como se ha mencionado arriba, se establecen cuatro grupos: core, soft-core, shell y cloud. Los dos diagramas son coherentes entre sí y nos muestran una visión del pangenoma a distintos niveles de detalle.

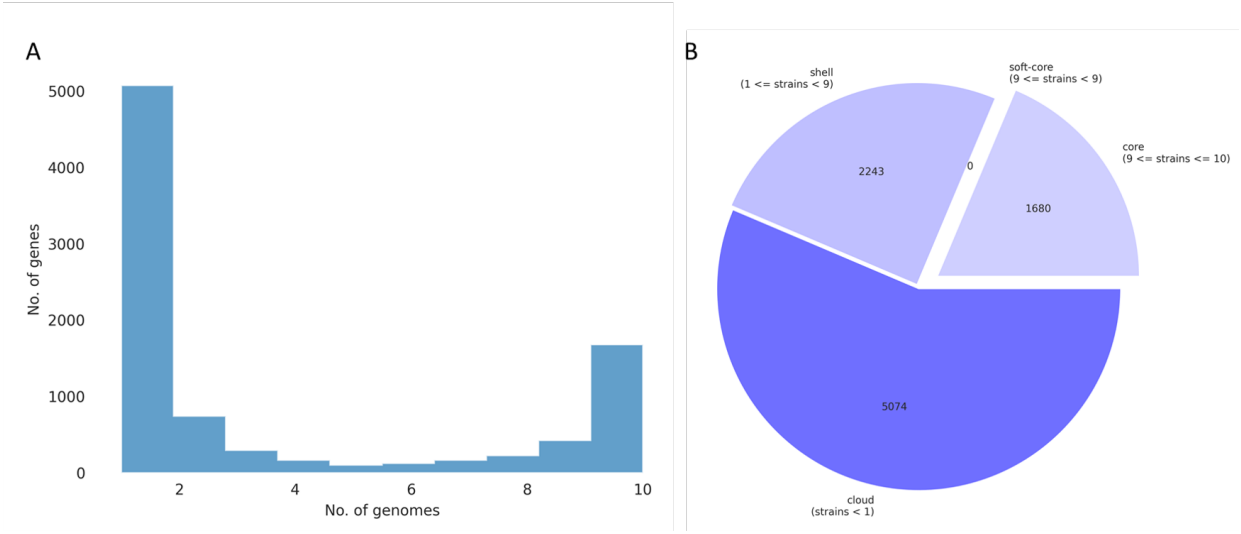

Figura 2: **A)** Gráfico de frecuencias de la distribución de los genes (eje de ordenadas: "No. of genes") en función de el número de cepas que los presentan (eje de abcisas: "No. of genomes"). **B)** Diagrama de sectores de la clasificación de los genes en función de 4 grupos establecidos según el siguiente criterio: core (comunes a 10 cepas), soft-core (comunes a 9 cepas), shell (comunes a 2-9 cepas) y cloud (exclusivos de 1 cepa).

2.2. Alineamiento de los genomas

En base al cladograma de la figura 1A, se ha creado un diagrama en forma de anillos para obtener una perspectiva más general del pangenoma de *Clostridium perfringens* (figura 3). Para ello, hemos seleccionado 5 candidatos en función de los clados generados en nuestro árbol filogenético. Las cepas seleccionadas son: JP838, JGS1495, ATCC3626, JGS1721 y SM101. Para realizar un gráfico como este, es necesario realizar un alineamiento de los genomas escogidos; y a su vez, para alinear los genomas, necesitamos un genoma de referencia contra el que alinear el resto de candidatos. La herramienta utilizada para el alineamiento es BLASTN y el genoma de referencia es la cepa JP838.

El diagrama de anillos parece ser coherente con el resto de gráficas discutidas en apartados anteriores. En general, se observa una alta identidad de secuencia ($90 - 100\,\%$), representado por los colores más intensos; aunque, por otro lado, también es evidente que existe a lo largo del alineamiento regiones puntuales de menor identidad (menos del $90\,\%$).

De esta forma, se puede corroborar que, a pesar de ser cepas de una misma especie, el pangenoma de *Clostridium perfringens* es muy divergente en términos evolutivos y muestra una gran riqueza de genes exclusivos a sus respectivas cepas. Este pangenoma parece responder a una radiación evolutiva fruto de la necesidad de adaptación de la especie a diversos ambientes y su notable éxito a la hora de poblar hábitats tan dispares.

Por otro lado, también se aprecian dos grandes huecos en la parte superior-izquierda del diagrama. Esto puede deberse a errores instrumentales a la hora de secuenciar el genoma de referencia. Es decir, el genoma de la cepa JP838 no ha sido secuenciado en su totalidad. Por lo tanto, no se ha podido tener en cuenta la totalidad del genoma de las cepas en el proceso de alineamiento.

Regiones de baja identidad de secuencia, como la encontrada en la parte superior izquierda (entorno a las 4000 kbp), probablemente correspondan a regiones de genes del genoma accesorio.

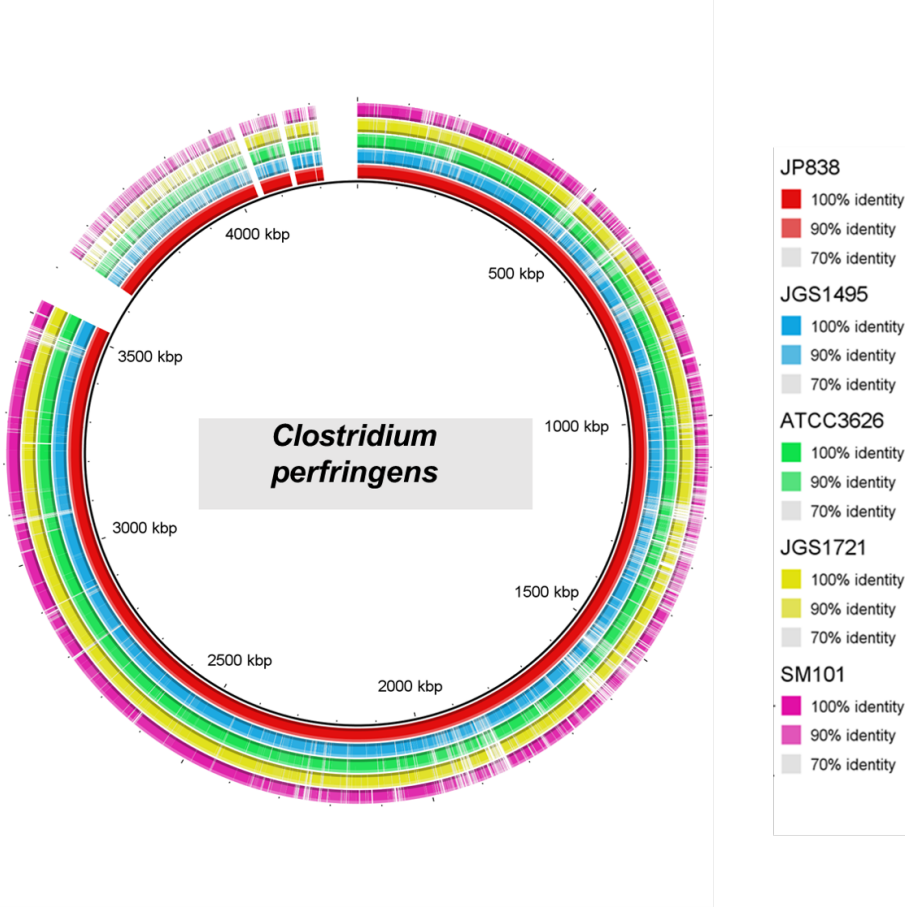

Figura 3: Diagrama circular del pangenoma de *Clostridium perfringens*. Los genomas representados se distribuyen de la siguiente forma: JP838, genoma de referencia (rojo); JGS1495 (azul); ATCC3626 (verde); JGS1721 (amarillo) y SM101 (violeta). El porcentaje de identidad de los genomas con respecto al genoma de referencia está representado por tres niveles o intensidades de color: 100 % de identidad, color intenso; 90 % de identidad, color más claro; 70 % de identidad, gris; 0 % de identidad, blanco.

2.3. Introducción a la anotación funcional del pangenoma

En las siguientes líneas se presentarán dos opciones para la anotación funcional de genes del pangenoma: clusterProfiler y Clusters of Orthologous Groups (COG).

Aunque son herramientas de gran interés, creemos que su utilización excede los objetivos de este trabajo. Sin embargo, se ha realizado un estudio preliminar de ambas herramientas para saber utilizarlas en un futuro si fuera necesario.

2.3.1. Enriquecimiento con clusterProfiler

El fichero `gene_presence_absence.Rtab` es uno de los de mayor utilidad que produce Roary. Contiene una matriz binaria con la ausencia o presencia de cada gen (indicado por su ID de Uniprot) en cada cepa (0 si el gen no está presente, y 1 si lo está). Este fichero puede cargarse en R o Excel para manipularlo de tal que manera que se obtengan los genes propios del genoma core y del genoma accesorio. Se han separado los genes del core, el caparazón y la nube en sendos archivos `.xlsx` (`core.xlsx`, `shell.xlsx` y `cloud.xlsx`, respectivamente). Los datos se encuentran en el enlace mostrado en la sección 4.6.

El software de R clusterProfiler 4.0 permite la anotación funcional de genes a partir de un fichero

que contenga el ID de cada gen (puede estar en formato Refseq o Uniprot, entre otros) [7]. La anotación funcional consiste en el enriquecimiento en base a términos de ontologías como la Gene Ontology (GO) y la Kyoto Encyclopedia of Genes and Genomes (KEGG). Tras el enriquecimiento se pueden generar gráficos descriptivos de los resultados, con paquetes como ggplot2.

El paquete clusterProfiler necesita una base de datos de anotación para los genes del organismo que se está estudiando. Para ello, existen paquetes de anotación de diferentes organismos. Por ejemplo, uno de los paquetes de anotación de *E. coli* es org.EcK12.eg.db. Se pueden generar paquetes de anotación para organismos concretos si existe una anotación funcional de su genoma. La librería de R adecuada para ello es AnnotationForge.

No se ha encontrado un paquete de anotación para los genomas de las cepas de *C. perfringens* estudiadas. Por tanto, un siguiente paso en nuestro análisis habría sido generar el paquete de anotación, que más tarde sería usado en clusterProfiler.

2.3.2. Enriquecimiento con Clusters of Orthologous Groups (COG)

Otra forma posible de realizar la anotación funcional del pangenoma es la utilización de la base de datos COG, y, en concreto, de la herramienta en línea eggNOG-mapper v2. Para ello, el archivo de entrada se corresponde con las secuencias de los genes que se quieren anotar. El fichero `pan_genome_reference.fa` de Roary contiene todos los genes del pangenoma de las 10 cepas estudiadas. Por tanto, este fichero puede manipularse para estudiar funcionalmente el pangenoma.

3. Conclusión

El análisis del pangenoma de 10 cepas *Clostridium perfringens* provenientes de una gran variedad de hospedadores muestra un pangenoma amplio y diverso. De los 8997 genes identificados, solo el $18,7\,\%$ (1680 genes) pertenecen al genoma core, y el $81,3\,\%$ (7317 genes) al genoma accesorio. Estos datos van en consonancia con los obtenidos por Kiu et. al 2017: tras estudiar 56 cepas encontraron 11667 genes, donde solo el 12,6 % pertenecían al genoma accesorio [3]. Los autores de este estudio concluyen que *C. perfringens* posee el pangenoma más diverso de entre todas las Gram-positivas estudiadas hasta el momento. El análisis filogenómico de las cepas basado en el genoma core también concuerda con estudios previos.

Un estudio funcional del pangenoma habría revelado, entre otros aspectos, una gran proporción de genes de transposasas e integrasas. La fuente principal de adición de genes al pangenoma de esta especie es la transferencia horizontal de genes, primariamente a través de la integración de fagos en el genoma bacteriano.

Un aspecto curioso es la ausencia de sistemas CRISPR-Cas en esta especie. Se ha postulado que la infección por fagos constituye una fuente de diversidad genómica que permite la adaptación a ambientes dispares en *C. perfringens*. Por tanto, los sistemas de defensa frente a fagos, como CRISPR-Cas, no proporcionarían una ventaja adaptativa.

4. Metodología

La figura 4 representa el flujo de trabajo seguido durante la realización de este trabajo. En la presente sección se explican en más detalle cada uno de los pasos.

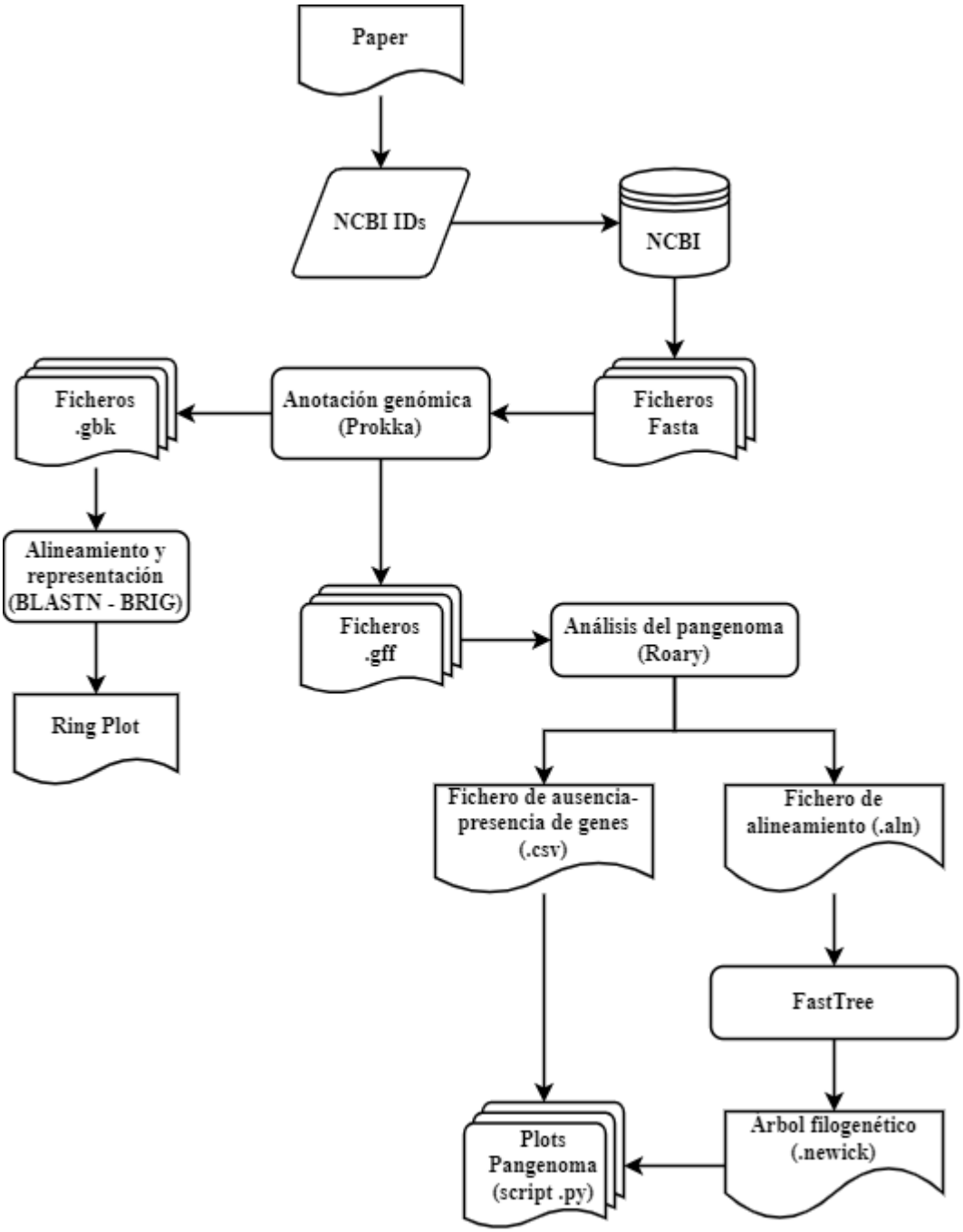

Figura 4: Flujograma que expone los pasos seguimos durante la realización de este trabajo. Explicación en el texto.

4.1. Selección de genomas

Se han escogido genomas de 10 cepas de *Clostridium perfringens*, atendiendo a su diversidad ecológica y fenotípica, con especial interés en la selección de cepas procedentes de distintos hospedadores y que expresen diferentes toxinas. Se ha utilizado como referencia el artículo de Kiu et al. 2017 para obtener las cepas [3]. Estas cepas se han secuenciado con técnicas como PacBio SMRT DNA Sequencing. Las secuencias ya ensambladas se han descargado de NCBI (sección Assembly) en formato FASTA, y utilizando Refseq como base de datos de origen.

Nombre	**Toxinotipo**	**Origen**	**Enfermedad del hospedador**	**NCBI ID**
SM101	A	Carne de consumo	Intoxicación alimentaria	SAMN02604026
JGS1721	D	Oveja	Enterotoxemia	SAMN02436277
JFP981	A	Aves de corral	Enteritis necrotizante	SAMN05323904
1207_CPER	A	Humano (sangre)	Paciente de UCI	SAMN03197169
JJC	A	Suelo	No se aplica	SAMN02317206
ATCC3626	B	Cordero	Disentería	SAMN02436295
JGS1495	C	Cerdo	Diarrea	SAMN02436294
JGS1987	E	Ternero	Enteritis hemorrágica	SAMN02436167
MJR7757A	A	Humano (vagina)	Proyecto Microbioma Humano	SAMN03842618
JP838	A	Perro	Gastroenteritis hemorrágica	SAMN03377063

Cuadro 1: Tabla que muestra algunos metadatos sobre los genomas escogidos para la realización de este trabajo.

4.2. Anotación de genomas

La herramienta de software utilizada para la anotación de los genomas bacterianos ha sido Prokka, versión 1.14.6 [6]. Este programa se encuentra disponible en el cluster Dayhoff del Máster.

A la hora de ejecutar el programa es importante indicar el reino al que pertenece el organismo de estudio. Otras opciones que se deben incluir son: número de cores a utilizar, prefijo de los ficheros generados, etiqueta que se le va a poner a cada locus que se identifique y el nombre del directorio de destino.

```
#Ejemplo de ejecución con una de las cepas
$ prokka --cpus 8 --kingdom Bacteria --prefix JP838 --locustag JP838
./JP838_genome.fasta
```

El output obtenido es una gran variedad de ficheros que servirán posteriormente para futuros análisis y visualizaciones del genoma.

4.3. Obtención y análisis del pangenoma

Se ha hecho uso del software Roary en su versión 3.11.2 para obtener el pangenoma de las 10 cepas de *Clostridium perfringens* estudiadas [4]. Los archivos de entrada que se utilizan son los `.gff` generados con Prokka en las 10 cepas.

```
##Ejecución de Roary

$ roary -f roaryresults -p 4 -e -n -v --maft gffs/*.gff

## -e create a multiFASTA alignment of core genes using PRANK
## -n fast core gene alignment with MAFFT, use with -e
## -v verbose output to STDOUT
```

El software FastTree permite la realización de árboles filogenéticos basados en el método de Approximate Maximum Likelihood Estimation (AMLE) [5]. El archivo necesario para ello es core_gene_alignment.aln, producido con Roary. Se crea un archivo `.newick`, un formato popular en la representación de árboles filogenéticos.

```
$ FastTree -nt -gtr roaryresults/core_gene_alignment.aln >
roaryresults/mytree.newick
```

Por último, se ha hecho uso de este repositorio de Github para crear diferentes gráficos, los cuales son de utilidad en el análisis del pangenoma. Cabe destacar los gráficas de matriz del pangenoma, y el diagrama de tarta del pangenoma. Se ha utilizado un script de Python que utiliza paquetes como Biopython, Seaborn, Matplotlab, Pandas y Numpy. Los ficheros de entrada necesarios para realizar las figuras son el archivo `.newick` y el `gene_presence_absecense.csv` generado con Roary.

```
## Guardar en .png.
$ python3 roary_plots.py --labels roaryresult/mytree.newick roaryresult/gene_presence_absence.csv

###Guardar en .svg para poder editarla.

$ python roary_plots.py --labels --format svg roaryresult/mytree.newick
roaryresult/gene_presence_absecense.csv
```

4.4. Ring-plot del pangenoma

Para obtener una representación gráfica y realizar la comparación circular de los genomas estudiados de una manera más visual, se ha utilizado BRIG v.0.95 (BLAST Ring Image Generator) [2]. Para realizar el alineamiento, se ha ejecutado un BLASTN (integrado en BRIG) de los genomas bacterianos, y el propio software genera una representación de los mismos en forma de anillos. Acepta una gran variedad de formatos; en este caso, se utilizó uno de los formatos de GeneBank, `.gbk`.

El genoma de refererencia utilizado para la representación es *Clostridium perfringens* JP838. El BLASTN ha tomado el genoma de referencia y ha alineado el resto de genomas de las otras cepas contra este.

4.5. Flujograma

El diagrama de flujo se ha realizado mediante la web-app diagrams.net.

4.6. Disponibilidad de los datos

Los ficheros y archivos generados durante la realización de este proyecto están disponibles en un repositorio de Google Drive.

Referencias

[1] Mostafa Y Abdel-Glil y col. "Comparative in silico genome analysis of Clostridium perfringens unravels stable phylogroups with different genome characteristics and pathogenic potential". En: *Scientific reports* 11.1 (2021), págs. 1-15.

[2] Nabil-Fareed Alikhan y col. "BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons". En: *BMC genomics* 12.1 (2011), págs. 1-10.

[3] Raymond Kiu y col. "Probing genomic aspects of the multi-host pathogen Clostridium perfringens reveals significant pangenome diversity, and a diverse array of virulence factors". En: *Frontiers in microbiology* 8 (2017), pág. 2485.

[4] Andrew J. Page y col. "Roary: rapid large-scale prokaryote pan genome analysis". En: *Bioinformatics* 31.22 (jul. de 2015), págs. 3691-3693. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv421. eprint: https://academic.oup.com/bioinformatics/article-pdf/31/22/3691/17122651/btv421.pdf. URL: https://doi.org/10.1093/bioinformatics/btv421.

[5] Morgan N. Price, Paramvir S. Dehal y Adam P. Arkin. "FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix". En: *Molecular Biology and Evolution* 26.7 (abr. de 2009), págs. 1641-1650. ISSN: 0737-4038. DOI: 10.1093/molbev/msp077. eprint: https://academic.oup.com/mbe/article-pdf/26/7/1641/13642970/msp077.pdf. URL: https://doi.org/10.1093/molbev/msp077.

[6] Torsten Seemann. "Prokka: rapid prokaryotic genome annotation". En: *Bioinformatics* 30.14 (mar. de 2014), págs. 2068-2069. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu153. eprint: https://academic.oup.com/bioinformatics/article-pdf/30/14/2068/7250406/btu153.pdf. URL: https://doi.org/10.1093/bioinformatics/btu153.

[7] Tianzhi Wu y col. "clusterProfiler 4.0: A universal enrichment tool for interpreting omics data". En: *The Innovation* 2.3 (2021), pág. 100141.