Título

Estudio de la Estructura Poblacional de los géneros *Escherichia* y *Shigella*

Nombre Estudiante**:**
Lucía Chacón Vargas

Máster universitario en Bioinformática y bioestadística UOC-UB
Área 2

Consultor/a**: Carles Ventura Royo**
Profesor/a responsable de la asignatura: **José Luis Villanueva-Cañas**
Tutor externo**: Val Fernández Lanza**

Fecha entrega**: 5/Junio/2019**

A) Creative Commons:

B) GNU Free Documentation License (GNU FDL)

C) Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	Estudio de la Estructura Poblacional de los géneros *Escherichia* y *Shigella*
Nombre del autor:	*Lucía Chacón Vargas*
Nombre del consultor/a:	*José Luis Villanueva-Cañas*
Nombre del PRA:	*Carles Ventura Royo*
Nombre del tutor externo	*Val Fernández Lanza*
Fecha de entrega (mm/aaaa):	06/2019
Titulación::	*Máster universitario en Bioinformática y bioestadística UOC-UB*
Área del Trabajo Final:	*Genómica comparativa y evolución*
Idioma del trabajo:	*Español*
Palabras clave	*Escherichia, Shigella, Population Structure*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

Escherichia y *Shigella* se han diferenciado como dos géneros distintos. Sin embargo, con los avances en microbiología se ha visto que están fuertemente relacionadas. En este trabajo se ha realizado un estudio de las relaciones filogenéticas y pangenómicas de14.078 secuencias de Escherichia y 1.781 de Shigella.

Métodos: Se utilizaron secuencias genómicas y de aminoácidos de ambos géneros de la base de datos *RefSeq.*

Se eliminaron las secuencias genómicas redundantes y se generó la matriz de distancias con *MASH*.

Las cepas de Shigella y de Escherichia se clasificaron por filogrupos según el algoritmo del laboratorio de *Clermont*. Se crearon clústeres por género y por grupo filogenético con *UMAP* (en R) y con el software *Gephi*.

Con *Mmseqs2* se buscaron clústeres de las secuencias de aminoácidos. Se realizó el cálculo del pangenoma, coregenoma y genoma accesorio y se obtuvieron las gráficas correspondientes en R.

Resultados y conclusiones: Se observaron agrupaciones en *UMAP* muy relacionados entre sí, lo que indica la proximidad genómica de ambos géneros. Las cepas de Shigella clasificadas como filogrupo B1 (más del 90 % del total) se sitúan en agrupaciones muy cercanas al filogrupo B1 de *Escherichia.*

El pangenoma de ambos géneros juntos y de cada uno por separado sigue una distribución abierta. El tamaño del coregenoma va disminuyendo según aumenta el número de genomas. Bajando el umbral a 85% o 95% de genes compartidos, el tamaño del coregenoma se mantiene prácticamente constante.

Abstract (in English, 250 words or less):

Escherichia and Shigella have been deemed as two distinct bacterial genus. However, with the advances in microbiology it has been seen that they are strongly related.
A study of the phylogenetic and pangenomic relationships of 14,078 sequences of *Escherichia* and 1,781 of *Shigella* has been carried out.
Methods: Genomic and amino acid sequences from both genus were downloaded of the RefSeq database. Redundant genomic sequences were eliminated and the distance matrix was calculated with MASH.
Shigella and Escherichia strains were classified according to the Clermont laboratory algorithm. The corresponding clusters were obtained by genus and by phylogenetic group with UMAP (in R) and with Gephi.
Mmseqs2 was used for clustering protein sequences. The pangenome, coregenome and accessory genome were calculated and the corresponding graphs obtained in R.
Results and conclusions: It could be observed very related clusters, suggesting the genomic proximity of both genus.
Shigella strains classified as B1 phylogroup (more than 90 % of the total) were located in clusters very close to *Escherichia* phylogroup B1.
The pangenome of both genus together and of each genus separately follows an open distribution. The size of the coregenome decreases as the number of genomes increases. By lowering the threshold to 85% or 95% of shared genes, the size of the coregenome remains virtually constant

Índice

Lista de figuras

Figura 1. Descarga de las secuencias de la base de datos NCBI Assembly.

Figura 2. Estructura poblacional con UMAP (R). En rojo, Escherichia; En negro, Shigella. n_neighborgs=15.

Figura 3. Estructura poblacional con UMAP (R). En rojo, Escherichia; En negro, Shigella. n_neighborgs=5.

Figura 4. Número ideal de clústeres, calculados con la función fviz_nbclust del paquete factoextra de R.

Figura 5. Gráfico con los 2 clústeres resultantes de aplicar la función fviz_cluster de R.

Figura 6. Filogrupos obtenidos con el método de Clermont [27]

Figura 7. Interfaz de Gephi antes de aplicar el algoritmo de distribución (arriba) y buscando la estructura final (abajo). Filogrupos como atributo.

Figura 8. Clústeres resultantes de la aplicación del algoritmo ForceAtlas 2 en Gephi al archivo de distancias con los 300 pesos más bajos. En rojo, Escherichia; en negro, Shigella. Los puntos son los nodos del gráfico (las cepas).

Figura 9. Clústeres resultantes de la aplicación del algoritmo ForceAtlas 2 en Gephi al archivo de distancias con los 300 pesos más bajos. Cada color representa un grupo filogenético: (B1,Shigella B1), E, C, B2, F, (D,D1, D2), A1, (A0,A0a).

Figura 10. Estructuras finales, creadas en Gephi, por filogrupo con los 10 (a), 30 (b), 60 (c) o 100 (d) pesos más bajos.

Figure 11.Phylogenetic relationship among genome-sequenced *E. coli* and *Shigella* strains. The phylogenetic tree was generated by PhyML with amino-acid sequences of 1,273 core genes from completely sequenced *E. coli* and *Shigella* strains. Each color indicates the phylogenetic group of *E. coli* (red, A; yellow, B1; black, *Shigella*; blue, E; purple, D; green, B2). Bootstrap values (percentages of 1,000 replications) greater than 50% are shown at each node. *Escherichia fergusonii* ATCC 35469 were used for the out-group. The scale bar represents 0.001 nucleotide substitutions per site. De Kwak et al., [36].

Figura 12. Agrupaciones con UMAP por filogrupo

Figura 13. Agrupaciones con UMAP por género.

Figura 14. Pangenoma abierto (en azul) y cerrado (rojo). De [42]

Figura 15. Parte del archivo “TablaAnalysysPangenomas.tsv”. La primera columna son los clústeres, la segunda, las secuencias pertenecientes a cada clúster, y la tercera, el número de acceso de la proteína en RefSeq.

Figura 16. Genes nuevos, en función del número de cepas, en el genoma conjunto de Escherichia y Shigella. En el eje y, log de la media ± sd. Eje x: número de cepas. 10 repeticiones por punto.

Figura 17. Genes nuevos, en función del número de cepas, en el genoma de Escherichia. En el eje y, log de la media ± sd. Eje x: número de cepas. 10 repeticiones por punto.

Lista de tablas

1. Introducción

.1 Contexto y justificación del Trabajo

Se propone realizar un estudio de la estructura poblacional de los géneros *Escherichia* y *Shigella.*

Ambos géneros pertenecen a la Familia *Enterobacteriaceae*, Orden *Enterobacteriales*, Clase *Gammaproteobacteria*, Filo *Proteobacteria*, Dominio *bacteria*.

Escherichia coli es uno de los organismos modelo más importantes tanto en biología como en medicina. Han surgido muchos hallazgos del estudio de *E. coli*, incluida la conjugación bacteriana, la recombinación y la regulación genética. Más importante aún, *E. coli* desempeña funciones claves en el tracto intestinal de los humanos y otros vertebrados, especialmente en la sección inferior. Hay más de mil millones de células de E. coli en los intestinos de un ser humano sano [1].
Algunas cepas de *E. coli* pueden causar enfermedades intestinales o extra intestinales, como diarrea, infección urinaria, septicemia, neumonía y meningitis, en seres humanos y animales [2].

Shigella es un agente causal, específico para el ser humano, de la disentería bacilar, una enfermedad grave caracterizada por calambres abdominales, náuseas, fiebre y diarrea sanguinolenta y mucosa. Constituye una de las principales causas de disentería bacteriana y es una de las principales causas de mortalidad y morbilidad, especialmente en los países en vías de desarrollo [3]. Se estima que unos 164,7 millones de personas al año se infectan con *Shigella*. Shiga describió por primera vez a *Shigella* como *Bacillus dysenteriae* en 1898. Lo llamó Bacillus porque parecía estar relacionado con *Bacillus coli*, que ahora se conoce como *Escherichia coli* [4,5]. En la década de 1940, Ewing propuso clasificar cuatro especies en el nuevo género *Shigella*: *S. dysenteriae, S. flexneri, S. boydii* y *S. sonnei*, según las características antigénicas de esas especies [6].

Escherichia y *Shigella* se han diferenciado como dos géneros bacterianos distintos. Sin embargo, con los avances en microbiología se ha visto que están fuertemente relacionadas. Brenner et al. [7], ya en 1972, determinaron que la similitud nucleotídica entre *E. coli* y *Shigella* era de un 80 a un 90 %. Ambas pertenecen a la familia *Enterobacteriaceae*. Fenotípicamente comparten muchas características a pesar de que tienen entidades separadas tanto en epidemiología como en clínica.

Ambos géneros comparten una familia de toxinas, las *Shiga*, cuyos genes se consideran parte del genoma de los profagos lambdoide [8]. El origen más común de toxinas Shiga son las bacterias *S. dysenteriae* y el *grupo* Shigatoxigénico de *Escherichia coli* (STEC), el cual incluye el serotipo O157:H7 y otras *E. Coli* enterohemorrágicas [9].

Normalmente, se utilizan las características bioquímicas y serotípicas para identificar las especies, pero muchos aislados bacterianos no pueden ser diferenciados entre *Shigella* o *Escherichia* incluso usando métodos moleculares

como secuenciación genómica de rARN 16S o caracterización de proteínas mediante MALDI-TOF MS.
Todo esto representa un reto diagnóstico. Además, los cambios en el modelo de resistencia antimicrobiana con cambios en serotipos y serogrupos pone de manifiesto la necesidad de una correcta identificación [10,11].

.2 Objetivos del Trabajo

Objetivos generales

Se analizarán las relaciones filogenéticas y pan-genómicas entre *Escherichia* y *Shigella*, los tamaños de core-genoma y pan-genoma de los géneros y la estructura poblacional común que puedan tener.

Objetivos específicos

- √ Definir la estructura poblacional de *Escherichia – Shigella* .
- √ Definir la distribución de tamaños del genoma core, genoma accesorio y pangenoma.

1.3 Enfoque y método seguido

Este trabajo pretende analizar las agrupaciones, o clústeres entre ambos géneros y también sus relaciones filogenéticas, intra e inter especie.

También se emprendió un estudio del pan genoma y core genoma de ambos géneros, para la definición de su estructura poblacional.
El **pan genoma** describe el conjunto de todos los genes de una especie. En bacterias se presenta una gran variación genética entre individuos de la misma especie. Por lo tanto el pan genoma puede ser muy extenso.
El **core genoma** representa los genes comunes a todas las cepas de una especie.

Se trabajó con secuencias de genoma y proteínas descargadas de la base de datos RefSeq, tanto de *Escherichia* como de *Shigella*. Se eligió RefSeq porque, a diferencia de GenBank, proporciona un sólo registro para cada ADN, ARN o proteína.

Se utilizó software de libre acceso y ejecutado en el ordenador desde la consola de Unix o desde el sistema operativo de Mac o Windows (se utilizaron ambos sistemas, eligiendo uno u otro por motivos prácticos) o funciones de R, dependiendo del proceso.

- Con *Mash* [12], desde la consola de Unix, redujimos las secuencias a esquemas pequeños y representativos, con los que se estimaron las distancias de mutación global y se eliminaron las secuencias redundantes. Con solo los esquemas (obtenidos mediante la función *sketch*), que pueden ser miles de veces más pequeños, se puede calcular la distancias de grandes conjuntos de datos (mediante la función *dist*). Este algoritmo es apropiado para hallar clústeres, construir árboles, etc.

- Con el paquete *igraph*, de R [13], de análisis de redes y visualización, se crearon clústeres y se guardó una única secuencia de cada grupo.
- Con *UMAP* (Uniform Manifold Approximation and Projection) desde R, que es una técnica de aprendizaje múltiple para la reducción de dimensiones basada en geometría de Riemann y topología algebraica [14], se definieron clústeres.
- Se usó *Mmseqs2* (Many-against-Many searching), desde la consola de Unix, para buscar y agrupar grandes conjuntos de secuencias [15] de proteínas. Se calculó el tamaño del coregenoma y del pangenoma.
- Con el software Gephi [16], desde el sistema operativo de Windows o Mac para la exploración de gráficos en el estudio de clústeres.

1.4 Planificación del Trabajo

Los recursos y las tareas que se llevaron a cabo son las siguientes:

- Recursos
  - Conexión mediante VPN al un ordenador de 64 G de RAM de la Unidad de Bioinformática del Hospital Ramón y Cajal, Madrid.
  - Ordenador de sobremesa con sistema operativo Windows de 8 G de RAM, procesador intel Core 2.
  - Ordenador portátil con sistema operativo macOS y sistema operativo Ubuntu,16 G de RAM, procesador intel Core i7.

- Tareas realizadas
  - Búsqueda bibliográfica
  - Descarga de las secuencias genómicas y de aminoácidos de *Escherichia* y *Shigella* de la base de datos *RefSeq.*
  - Instalación del software necesario en Unix, Windows y/o macOS.
  - Reducción de las secuencias genómicas a sketches y matriz de distancias con *mash*.
  - Eliminación de genomas redundantes con *igraph* en R.
  - Matriz de distancias de secuencias no redundantes con *mash*
  - Representación de la estructura poblacional del conjunto de ambos géneros con la función *umap* en R.
  - Clustering de las secuencias de nucleótidos con *mclust* en R.
  - Clustering de las secuencias de nucleótidos con *k-means* en R.
  - Obtención de grupos filogenéticos, aplicando el script del laboratorio de Clermont a las secuencias genómicas.
  - Creación de los archivos necesarios para la obtención de clústeres.
  - Obtención y visualización de clústeres mediante el software *Gephi*.
  - Obtención y visualización de clústeres con la función *umap* en R.
  - Clustering de las secuencias de aminoácidos con el software *MMseqs2.*
  - Cálculo, representación visual y creación de tablas del pangenoma y coregenoma de *Escherichia* y *Shigella.*
  - Análisis de los resultados anteriores.
  - Redacción de la Memoria

- Planificación temporal

	Febrero	Marzo	Abril	Mayo	Junio
Definición del Proyecto Lectura de Bibliografía Entrega de PEC 0	20	4			
Descarga de secuencias Instalación de software Entrega de PEC1		5 18			
Reducción y cálculo de distancias de secuencias genómicas		18 25			
Estudio de la estructura poblacional con *igraph, UMAP* y *k-means*. Obtención de clústeres con *Mmseqs2* Entrega de PEC2		25	24. 24		
Clasificación de grupos filogenéticos según *Clermont*. Generación de los archivos para estudio de clústeres.			25	8	
Creación y análisis de clústeres en Umap y Gephi.				9 13	
Cálculo, representación visual, creación y análisis de tablas del pangenoma, coregenoma y accesorio Entrega de PEC3				14. 20	
Redacción y entrega de la memoria				21	5

1.5 Breve sumario de productos obtenidos

- Matriz de distancias de secuencias genómicas no redundantes de *Escherichia* y *Shigella*
- Definición de la estructura poblacional y gráficos correspondientes con *UMAP* y *Gephi*.
- Obtención de la estructura filogenética mediante el algoritmo de *Clermont* y figuras correspondientes con *UMAP* y *Gephi*.
- Cálculo de los tamaños del pangenoma, coregenoma y genoma accesorio y gráficos correspondientes.

1.6 Breve descripción de los otros capítulos de la memoria

- En el capítulo 2 se desarrollará el estudio de la estructura poblacional con los siguientes puntos:

Introducción
Descarga de secuencias.
Obtención de la matriz de distancias.
Creación de gráficos con k-means, UMAP y Gephi.
Obtención de la estructura filogenética.
Creación de gráficos con Gephi y UMAP.

- En el capítulo 3, el estudio del Pangenoma y Coregenoma:
  Introducción
  Clustering de las secuencias de aminoácidos
  Cálculo del pangenoma y coregenoma con sus gráficos correspondientes.

- El capítulo 4 son las conclusiones

2. Estudio de la Estructura Poblacional.

.1. Descarga de las secuencias

La base de datos *GenBank* [17] es parte del *International Nucleotide Sequence Database Collaboration (INSDC),* junto con el *European Nucleotide Archive* y el *DNA Data Bank of Japan (DDBJ)*. Los datos se intercambian diariamente entre los tres colaboradores para lograr una cobertura mundial completa.
Sin embargo, GenBank puede ser muy redundante para algunos loci.
Por lo tanto, se decidió obtener las secuencias de la base de datos de *RefSeq* [18], ya que esta base de datos pasa por un proceso de curación más exhaustiva. Sus genomas son copias de los de GenBank.

Se descargaron las secuencias *Genomic FASTA* y *Protein FASTA* de Assembly-NCBI [19], que contiene los genomas ensamblados de la base de datos RefSeq (Fig.1), tanto del género *Escherichia* como del género *Shigella*. A 20 de Febrero de 2019 había 14.078 secuencias de *Escherichia* y 1.781 de *Shigella*.

Figura 1. Descarga de las secuencias de la base de datos NCBI Assembly.

.2. Reducción dimensional de las secuencias genómicas mediante *mash*.

En 1990, cuando se publicó *BLAST* [20] había menos de 50 millones de secuencias nucleotídicas en las bases públicas. Hoy día cualquier secuenciación puede producir 1 trillón de bases. Por lo tanto, ha surgido la necesidad de poner en marcha métodos nuevos que permitan gestionar tal ingente cantidad de datos.
El grupo de Phillippy en 2016 [12] creó la herramienta *Mash*, basada en la técnica *MinHash*,que reduce grandes secuencias o conjuntos de secuencias a representaciones esquemáticas o sketches. Esta técnica empezó a utilizarse para la búsqueda de páginas web o imágenes muy similares [21,22]. En 2015, ya se empezó a usar en problemas de montaje genómico [23], clustering de genes 16S ADNr [24] y clustering de secuencias [25]. Requiere poco gasto computacional y, por lo tanto es muy apropiada para trabajar con grandes datos en genómica.

Mash crea, manipula y compara sketches de datos genómicos y también puede calcular la tasa de mutación entre secuencias desde sus sketches con la función *dist*, que devuelve un cálculo del índice de Jaccard (fracción de k-mers compartidos), un valor P y la distancia, que calcula la tasa de mutación basándose en un modelo evolutivo simple [26].

En este trabajo, se utilizó *mash* para la reducción dimensional y el cálculo la matriz de distancias de las secuencias genómicas.

Esquema de trabajo:
Antes de proceder a la obtención de sketches, se añade el prefijo *shig* o *esche* a cada secuencia de Shigella o Escherichia, respectivamente, para facilitar operaciones posteriores, mediante el comando:

rename 's/^/esche_/' *

Después, se reparten las secuencias en carpetas, para que el gasto computacional sea menor, mediante la orden:

ls|parallel -n(*nº de items en cada carpeta*) mkdir {#}\;mv {} {#}

Una vez hecho ésto, se generan los sketches de cada carpeta con la orden de mash:

mash sketch -o ../skgenesche1 *.fna

donde -o es la opción para que genere un archivo terminado en msh, *skgenesche1*, es el nombre del archivo resultante y *./fna, para que se aplique a cada archivo de la carpeta terminado en fna.
Se hace para cada carpeta. Luego, se unen todos los archivos de sketches en un único archivo, mediante la orden:

mash paste genall.msh sketch1 sketch2....etc

donde *genall.msh* es el archivo, que se va a generar, con todos los sketches de ambos géneros y *sketch1*... etc, son cada uno de los archivos de sketches creados anteriormente.

.3. Eliminación de secuencias genómicas redundantes y obtención de la matriz de distancias correspondiente.

Es frecuente encontrar en la base de datos varias entradas para una misma secuencia o muy similar. Esto puede distorsionar los resultados, ya que se carga el peso sobre un género o un grupo, en detrimento de otros.
Para evitar esta situación, se trató de eliminar secuencias con distancias menores de 0,0001. Para ello, se siguieron los siguientes pasos:

√ Se calcularon las distancias de todas contra todas con *mash*, obteniendo aquéllas cuya distancia fuera menor de 0,0001 con:

mash.dist -d 0.0001 genall.msh genall.msh > distred.tsv

donde 0,0001 es la distancia máxima que se registrará y *genallmsh.tsv* es el archivo donde se vuelca la matriz de distancias.

√ Mediante el paquete *igraph*, de R (Anexo 1), se agrupan las secuencias por semejanza y se obtiene el archivo correspondiente (*components.txt*)

√ Se conserva una secuencia representativa de cada grupo, eliminando el resto, mediante un script en Perl. Nuestro archivo final contenía 13.483 secuencias, 1.268 de *shigella* y 12.215 de *escherichia*.

√ Con ese archivo final, se vuelven a hacer sketches, y se calcula la matriz de distancias con mash. Esta vez:

```
mash dist -t genall2.msh genall2.msh > gensinred.tsv
```

donde *gensinred.tsv* es el archivo con la matriz de distancias de las secuencias no redundantes, que se utilizará posteriormente en diferentes apartados.

.4. Estudio de la estructura poblacional con *UMAP*

Intentaremos definir la estructura poblacional conjunta de ambos géneros, utilizando métodos de agrupamiento.

UMAP [14] es un algoritmo de reducción de dimensiones, utilizado para buscar estructuras ocultas en los datos. Está basado en grafos k-neighbour. Permite la posibilidad de visualizar los datos y, revelar así, patrones y grupos de datos más o menos similares. Tiene la ventaja sobre otros métodos, como t-SNE, que es más rápido.

“A diferencia de la PCA (análisis de componentes principales) pero similar a otros enfoques como t-SNE, se centra en la agrupación local, lo que significa que, aunque las observaciones "similares" deben agruparse, no intenta preservar la estructura global exacta entre todas las observaciones... Esta propiedad puede representar los datos de una manera más intuitiva (y visualmente interesante)” [28].

Este algoritmo también presenta sus inconvenientes: “Al igual que t-SNE, no conserva completamente la densidad y puede provocar escisiones, no reales, en agrupaciones” [29]

El parámetro n_neighbors controla la manera en que UMAP equilibra la estructura local frente a la global. Valores bajos obligarán a UMAP a concentrarse en una estructura muy local, en detrimento del panorama global, mientras que valores más altos tendrán en cuenta vecindarios más grandes de cada punto, perdiendo la estructura de los detalles finos.

Se trabajó con el archivo de secuencias genómicas, no redundantes, para ambos géneros, *gensinred.tsv.*

Se realizaron 2 gráficos, cambiando la opción *n_neighborgs*, de 15 (por defecto) a 5.

Observamos en la Figura 2 (n_neighborgs=15) una estructura en que *shigella* y *escherichia* se presentan en grupos más o menos diferenciados, pero ocupando agrupamientos indistinguibles entre ambos géneros.

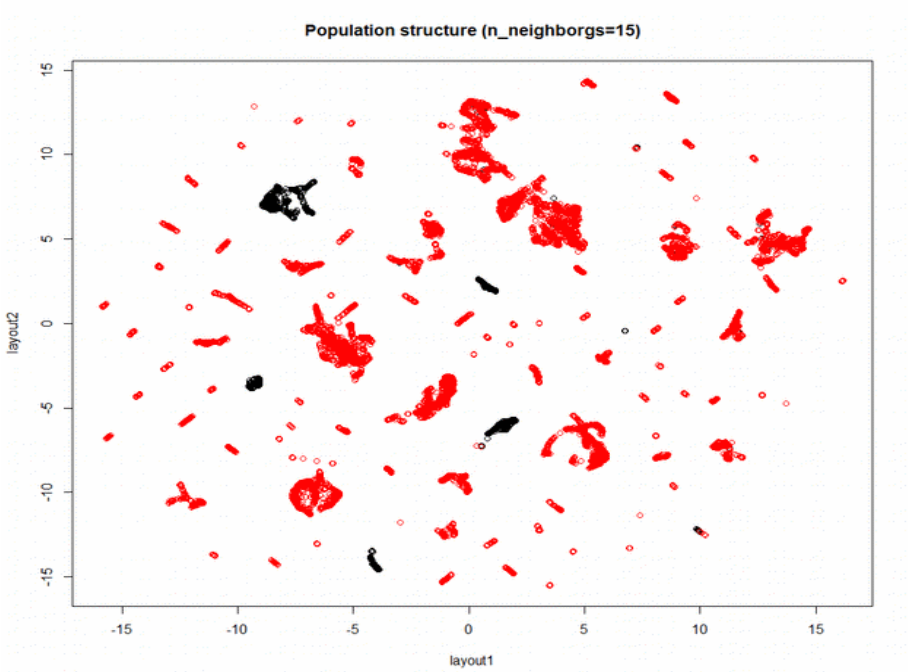

Figura 2. Estructura poblacional con UMAP (R). En rojo, Escherichia; En negro, Shigella. n_neighborgs=15.

En la Figura 3 (n_neighborgs = 5) se aprecia una menor diferenciación inter e intra especie, aunque se siguen distinguiendo grupos diferenciados.

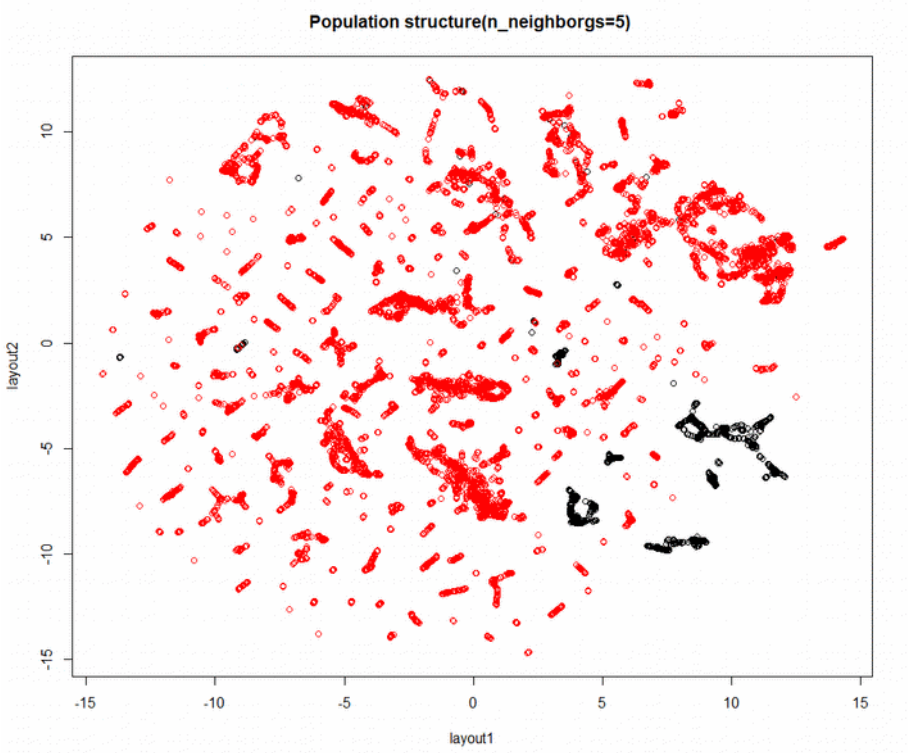

Figura 3. Estructura poblacional con UMAP (R). En rojo, Escherichia; En negro, Shigella. n_neighborgs=5.

Se observan, claramente grupos de *shigella* diferenciados de *escherichia*. Sin embargo, algunas *shigellas* están muy estrechamente relacionadas con *escherichia*.

.5. Agrupaciones (clustering) de las secuencias con *k-means*

K-means clustering [30] es el algoritmo de aprendizaje automático, no supervisado, más utilizado para dividir un conjunto de datos dado en un conjunto de k grupos (es decir, k clústeres), donde k representa el número de grupos pre-especificados por el investigador. Clasifica los objetos en múltiples grupos, de modo que los objetos dentro del mismo grupo son lo más similares posible (es decir, alta similitud intraclase), mientras que los objetos de diferentes grupos son tan diferentes como sea posible (es decir, baja similitud interclase). En *k-means*, cada grupo está representado por su centro (es decir, el centroide) que corresponde a la media de los puntos asignados al grupo.

Se realizó un estudio con *k-means* en R (véase script *k-means* en el apartado Anexos).
En primer lugar con la función *fviz_nbclust* del paquete *factoextra* se obtendrá el número ideal de clústeres para nuestros datos, de modo que la variación total intra-clúster sea minimizada [23]:

$$\text{minimize}\left(\sum_{k=1}^{k} W(Ck)\right)$$

donde C_k es el cluster k^{th} y $W(C_k)$ es la variación intra-cluster.

En este caso, el número ideal de clústeres, según *fviz_nbclust* fue de 2 (Figura 4)

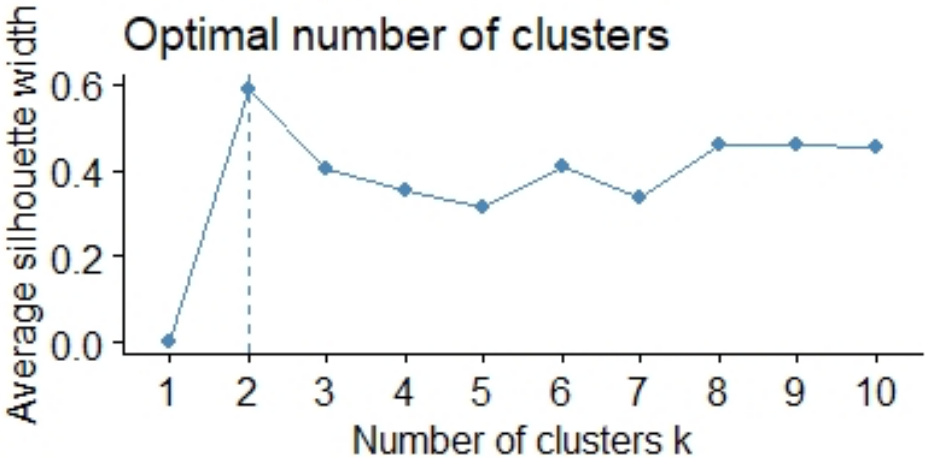

Figura 4. Número ideal de clústeres, calculados con la función fviz_nbclust del paquete factoextra de R.

El siguiente paso fue crear el gráfico correspondiente (figura 5) con la función *fviz_cluster.*

Figura 5. Gráfico con los 2 clústeres resultantes de aplicar la función fviz_cluster de R.

Observamos en la figura 5, que esa distribución podría no ser la adecuada para nuestros datos. Se observa un clúster con algunos elementos y otro con la mayoría de los datos, sin ninguna conexión entre ambos.
Se realizaron gráficos con 4 y 5 clústeres (datos no mostrados) que tampoco presentaron una distribución coherente.
Se concluyó que *k-means* no era el algoritmo adecuado para agrupar los datos de secuencias genómicas. “Este algoritmo es ampliamente usado por su sencillez y rapidez comparado con otros. También puede manejar grandes conjuntos de datos. Sin embargo, es sensible a los valores atípicos y pueden producirse resultados diferentes si se cambia el orden de sus datos” [31].
En la figura 5 se observa cómo algunos valores atípicos han podido distorsionar la forma de los clústeres.

.6. Obtención de la estructura filogenética

El género Escherichia está compuesto por *E. albertii, E. fergusonii, E. coli* y cinco clades crípticos de *Escherichia*. Además, las especies de *E. coli* se pueden dividir en siete filogrupos principales denominados A, B1, B2, C, D, E y F. Ya que los estilos de vida y/o los huéspedes son distintivos de cada filogrupo, su identificación es clave para estudios epidemiológicos.
Las pruebas fenotípicas clásicas no logran identificar los filogrupos [32].
El laboratorio de Clermont [33] ha desarrollado ensayos de PCR que permiten la identificación de la mayoría de estas especies/filogrupos y han desarrollado el método *ClermonTyping* y su web, *ClermonTyper*, que permite asignar una secuencia de cepa a *E. albertii, E. fergusonii, Escherichia clades I – V, E .coli,* así como a los siete principales filogrupos de *E. coli*. Este método *in silico* muestra una concordancia del 99,4% con los ensayos de PCR in vitro. *ClermonTyper* es un recurso libre para la comunidad científica [32,34].

Aunque desarrollado para *Escherichia*, se decidió utilizar este algoritmo para el estudio de la estructura filogenética de ambos géneros, ya que esto nos permitía observar las relaciones de cada grupo filogenético de *Escherichia* con

los grupos de *Shigella*. Una vez aplicado el algoritmo de Clermont, se obtuvieron los filogrupos para ambos géneros (Figura 6). En resumen, los 10 grupos con mayor representación fueron: *B1, A1, E, B2-sgI, A0a, D1, F, B2-sgII, C* y *A0*.

Figura 6. Filogrupos obtenidos con el método de Clermont [27]

.6.1. Generación de los archivos necesarios.

Para generar los archivos que se usaron en este apartado, se creó un script en R (véase Script *Clermont.R* en Anexos). En resumen:

1. Lectura de la matriz de distancias genómicas, creada con *mash* (archivo *"gensinred"*).

2. Transformación de la matriz unimodal en un objeto igraph, en el que se conserva la mitad de la imagen especular que forma la matriz diagonalmente, y se crea un dataframe con los datos.

3. Se genera un archivo, *"Seminetxxx.tsv",* con los 10, 30, 60,100 o 300 pesos (siendo peso la distancia genómica entre dos secuencias) más bajos de la tabla. El archivo contiene las siguientes variables: *Source* y *Target,* que son las secuencias cuya distancia queda registrada en la variable *Weight*, y el tipo de gráfico que se generará (dirigido o no dirigido), *Type*.
4. Lectura del archivo *"Clermont.tsv"*, que después de darles nombre, contiene las siguientes variables: *ID*, es la identidad de la secuencia según nuestro archivo primitivo, con los prefijos *esche* o *shig*, adjudicados en una fase previa del estudio, *group*, es el grupo filogenético de cada secuencia, según el algoritmo de Clermont. *Dist*, es la variable *distancia*, *pvalue* y *sketch* (otra magnitud de *distancia*, generada con *mash*).

5. Después de dar forma al archivo, modificando nombres de variables, se genera el archivo "*NodosClermon100.tsv*" con las

   siguientes variables: *ID, Specie, Accession, phylogroup* (group en el archivo anterior)*, Dist, pvalue y sketch*.

6. Para aplicar *UMAP* (Uniform Manifold Approximation and Projection) [14], se reestructura la matriz para convertirla en un data.frame y se genera el gráfico correspondiente.

.7. Agrupaciones (clustering) de las secuencias con *Gephi*

Gephi [16] es una herramienta para análisis de datos y exploración de gráficos. Permite la manipulación sencilla de datos, para mayor comprensión de los gráficos.
Gephi utiliza algoritmos basados en la fuerza. Los gráficos pueden explorarse mediante herramientas muy sencillas como la extracción de datos, en forma de tabla, de una zona ampliada concreta para su manipulación posterior, colores diferentes según características que se pueden asignar por el usuario, etc.
Los pasos para obtener gráficos son los siguientes:

1.- Se carga el archivo *Seminetxxx.tsv*. Escogemos las opciones por defecto del programa, excepto para *Tipo de grafo*, que en nuestro caso es No *Dirigido.*

2.- Se carga el archivo *NodosClermont100.tsv*. Se escogen las opciones por defecto del programa, excepto para *Tipo de grafo*, que es No dirigido, y añadimos el archivo a nuestro espacio de trabajo.

3.- Con ambos archivos cargados en el mismo espacio de trabajo, se elige la distribución *ForceAtlas 2* [35]: Sigue un modelo de atracción lineal y repulsión lineal con unas pocas aproximaciones (simulación BarnesHut). Mucho más rápida que la distribución *ForceAtlas* y de una calidad similar.

4.- Los gráficos se pueden dividir en diferentes colores por atributos de nodos o aristas. En nuestro caso, por: *accession, dist, phylogroup, pvalue, sketch* o *specie.*

5.- Cuando el gráfico ha alcanzado la estructura definitiva, se puede ampliar una zona de interés y exportar esos datos como tabla.

En la figura 7 se muestra la interfaz de Gephi, una vez cargados los archivos, y filogrupos escogido como atributo, antes de aplicarle el algoritmo de fuerza (arriba) y en busca de la estructura final (abajo).

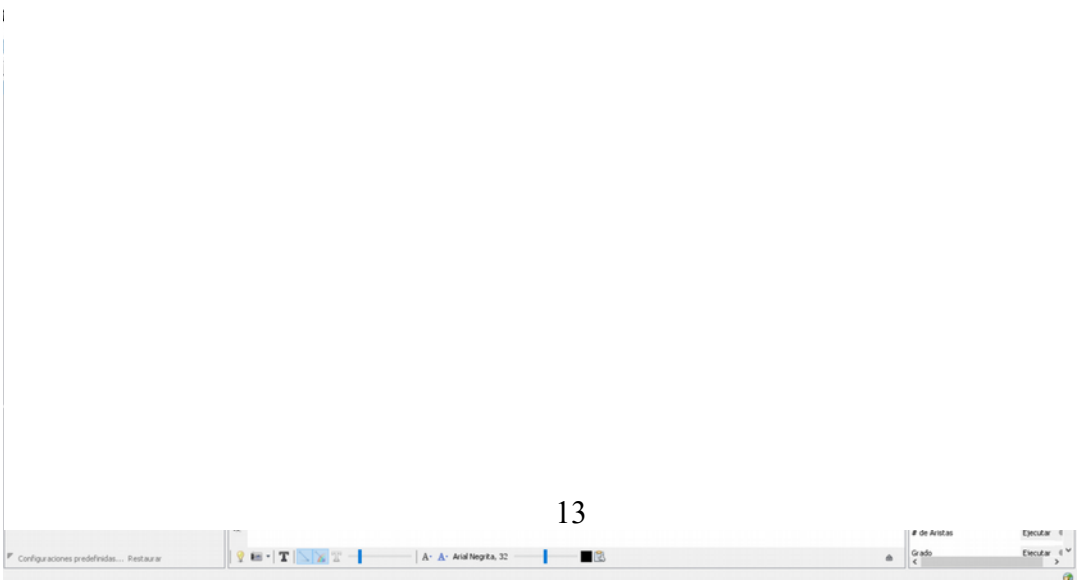

Figura 7. Interfaz de Gephi antes de aplicar el algoritmo de distribución (arriba) y buscando la estructura final (abajo). Filogrupos como atributo.

Se realizaron dos tipos de análisis de distribución: Por género y por grupos filogenéticos.

2.7.1. Por género.

En la figura 8, podemos observar la estructura resultante de la aplicación del algoritmo *ForceAtlas 2* al conjunto de datos con los 300 pesos más bajos de la matriz de distancias genómicas de ambos géneros. En rojo, *Escherichia*, en negro, *Shigella.*

Figura 8. Clústeres resultantes de la aplicación del algoritmo ForceAtlas 2 en Gephi al archivo de distancias con los 300 pesos más bajos. En rojo, Escherichia; en negro, Shigella. Los puntos son los nodos del gráfico (las cepas).

Se aprecian unos 8 clústeres bien definidos, entre ellos el correspondiente al género *Shigella* (en negro, en la figura 8). Hay que tener en cuenta que estas distancias son relativas; sin embargo, y aún por realizar un estudio más exhaustivo, se observa que *Shigella* es más próxima a algunos agrupamientos de *Escherichia* que algunas *Escherichias* entre sí. También se pueden apreciar, a simple vista algunas cepas de *Shigella* en otros clústeres.

2.7.2. Por filogrupo

Los archivos utilizados para este análisis son los mismos que para el estudio por género. En este caso, definimos "*phylogroup"* como atributo para su visualización en *Gephi.*
En la figura 9 se muestran los agrupamientos con anotaciones por filogrupo. El archivo utilizado es el mismo que el de la figura 8 con los 300 pesos más bajos.

Figura 9. Clústeres resultantes de la aplicación del algoritmo ForceAtlas 2 en Gephi al archivo de distancias con los 300 pesos más bajos. Cada color representa un grupo filogenético: (B1,Shigell-B1), E, C, B2, F, (D,D1, D2), A1, (A0,A0a).

Se intentaron cargar en Gephi archivos mayores al de 300, pero el tiempo que consumía el programa en buscar la estructura, y la dificultad en manipular la imagen correspondiente hizo que se decidiera trabajar con archivos de menor peso. Se eligió el de 300 en un compromiso entre calidad y manejabilidad. Quisimos asegurarnos de que la estructura poblacional no era muy diferente dependiendo del número de cepas empleadas. Para ello, también se trabajó con archivos con los 10, 30, 60 o 100 pesos más bajos de la matriz de distancias (figura 10).

Figura 10. Estructuras finales, creadas en Gephi, por filogrupo con los 10 (a), 30 (b), 60 (c) o 100 (d) pesos más bajos.

Se observa que los mismos clústeres se forman ya con el archivo de 10 y, según aumenta la población, van adquiriendo densidad.

Podemos apreciar 8 clústeres:

- Grupos A0, A0a y A1
- Grupos D, D1 y D2
- Grupo F
- Grupo B2
- Grupo E
- Grupo C
- Grupo B1
- Grupo Shigella_B1

No se han tenido en cuenta otros grupos minoritarios, cuyas secuencias se encuentran dispersas sin formar clústeres.
Se comprueba, según estos resultados, que la clasificación de Clermont aplicada a *Shigella* es correcta, ya que las *shigellas* se agrupan en los clústeres definidos por el algoritmo, no sólo el grupo B1, sino también el resto de los grupos (datos no mostrados).
Según esta clasificación gran parte de las secuencias de *Shigella* (un 94,3%) se clasificaron como grupo B1. Se puede observar en las figuras 8 y 9 cómo *Shigella* se posiciona a muy corta distancia del grupo B1 de *Escherichia*.

Este resultado es similar a clasificaciones filogenéticas realizadas por otros autores (Véase Figura 11, obtenida de Kwak et al. [36]).

Con respecto al orden de divergencia, Wang y colaboradores, basándose en árboles filogenéticos de los genes de mantenimiento, indicaron que el grupo D divergió primero y que A y B1 son grupos hermanos que se separaron más tarde [37]. Un análisis posterior indicó que quizás B2 en lugar de D es el ancestro [38]. El equipo de Touchon et al. [39] afirma que el grupo D (seguido del grupo B2) diverge primero.
Sims et al. [40] postulan un orden de divergencia, relativo al grupo externo *Escherichia fergusonii* de la siguiente manera: B2, D, *S.dysenteriae*, E, *Shigella*, B1 y A.

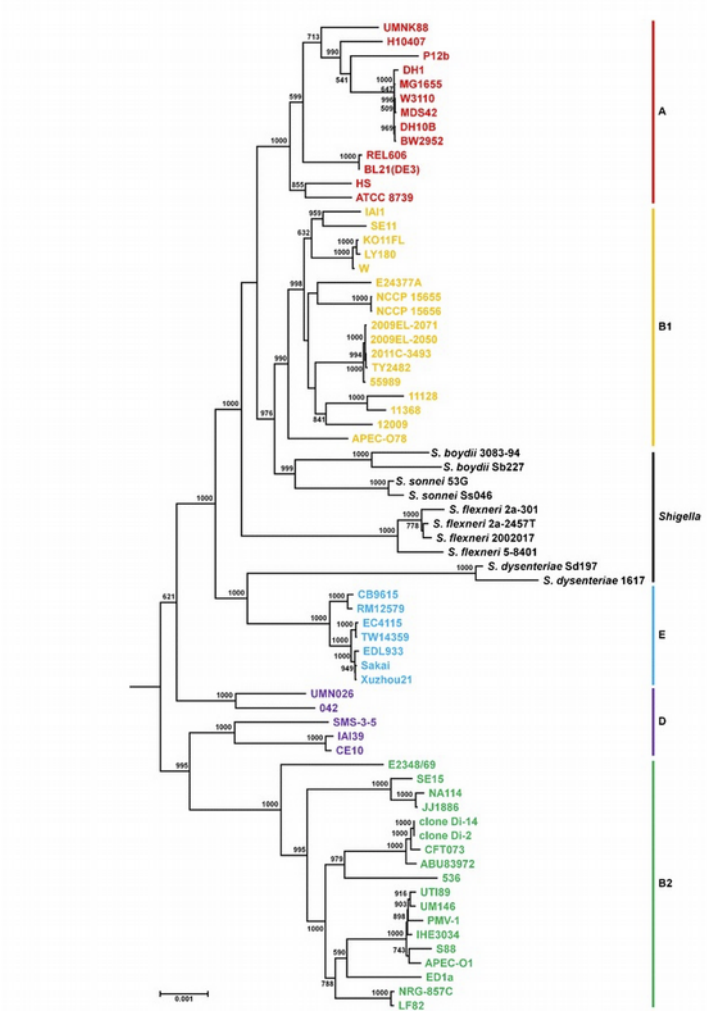

Figure 11.Phylogenetic relationship among genome-sequenced E. coli *and* Shigella *strains. The phylogenetic tree was generated by PhyML with amino-acid sequences of 1,273 core genes from completely sequenced* E. coli *and* Shigella *strains. Each color indicates the phylogenetic group of* E. coli *(red, A; yellow, B1; black,* Shigella*; blue, E; purple, D; green, B2). Bootstrap values (percentages of 1,000 replications) greater than 50% are shown at each node.* Escherichia fergusonii *ATCC 35469 were used for the out-group. The scale bar represents 0.001 nucleotide substitutions per site. De Kwak et al., [36].*

2.8 Agrupaciones de las secuencias con *UMAP*

Ya se hizo referencia a UMAP en la sección 2.4. de este trabajo.
En este caso, trabajaremos con grupos filogenéticos
Se generó un script en R que se puede leer en *Script Clermont en R - Umap-grupos filogenéticos* en la sección *Anexos*.
Se utilizó el archivo con la clasificación filogenética de Clermont y las distancias entre cepas. En las figuras 12 y 13 se muestra el resultado del agrupamiento, en función del filogrupo o del género, respectivamente.

La representación de las agrupaciones en función del filogrupo no es fácil de interpretar, ya que hay muchos colores indistinguibles entre sí.
Su análisis requeriría un estudio estadístico del archivo del que se generó el gráfico.

Figura 12. Agrupaciones con UMAP por filogrupo

Figura 13. Agrupaciones con UMAP por género.

. Estudio del Pangenoma y Coregenoma

3.1. Introducción

En el año 2005 ya se conocían las secuencias completas de 250 especies bacterianas, definiendo “especie” como un grupo de organismos que comparten más del 97 % de identidad de secuencia en los genes de bajas tasas de evolución para RNAr 16S.

Según se iban secuenciando más genomas, se iban encontrando nuevos genes para la misma especie. Tettelin y colegas [41] comprendieron que estaban muy lejos de tener los suficientes genomas para caracterizar todos los genes de una especie dada y, buscando una manera de representar esa diversidad, propusieron el concepto de *pangenoma,* definido como el conjunto de todos los genes de una especie. Este conjunto de genes se divide en genes *core*, presentes en todas las cepas estudiadas, genes *accesorios*, encontrados solo en algunas cepas y genes *únicos*, restringidos a una sola cepa [41].
Si el número de nuevos genes encontrados con cada nuevo genoma no alcanza una meseta, ese pangenoma es, teóricamente, infinito, y se dice que es *abierto* (Figura 14).

El concepto de pangenoma ha sido ampliamente adoptado por grupos de investigación que estudian la diversidad intra e inter taxones bacterianos. Además de las altas tasas de recombinación y los elementos genéticos móviles, que desde hace tiempo se sabe que son los impulsores de la diversidad procariota, la transferencia horizontal de genes (intercambio directo o indirecto de material genético entre organismos no relacionados) también contribuye a la diversidad individual entre bacterias [42].

Figura 14. Pangenoma abierto (en azul) y cerrado (rojo). De [42]

Es indiscutible la importancia que el análisis del pan genoma y core genoma tiene para definir la estructura de una especie. Los estudios del pan genoma son útiles en diferentes análisis [43]:

- √ "Caracterizar las cepas por su conjunto de genes individuales (por ejemplo, detectar factores de virulencia presentes solo en una cepa particular de una especie)
- √ Desarrollar vacunas contra cepas patógenas.
- √ Detección, identificación y seguimiento de nuevas cepas en muestras de metagenómica basadas en su subconjunto de genes individuales del pan genoma de las especie.
- √ Estudiar el impacto evolutivo de la transferencia horizontal de genes.
- √ Explorar la diversidad de cepas en estudios genómicos de poblaciones ambientales."

Para este estudio, se siguieron los siguientes pasos:

.2. Clustering de las secuencias de aminoácidos con *MMseqs2*

Se utilizaron las secuencias no redundantes de proteínas descargadas de Assembly-NCBI, de la base de datos RefSeq.

MMseqs2 (Many-against-Many searching) es un paquete de software para buscar y agrupar conjuntos de grandes secuencias [15,44].
Según M.Steinegger y J. Söding [15]: "*MMseqs2 cierra la brecha de costo y rendimiento entre la secuenciación y el análisis computacional de las secuencias de proteínas. Sus importantes ganancias en velocidad y sensibilidad deberían facilitar el análisis de grandes conjuntos de datos e*

incluso todo el espacio de secuencia de proteínas genómicas y metagenómicas a la vez".
Los autores reclaman una sensibilidad igual a la de BLAST, siendo 35 veces más rápido.

Para buscar clústeres, se usó el algoritmo *easy linclust*, que puede trabajar con archivos FASTA y está recomendado para grandes conjuntos de datos [45].
El resultado fue un archivo al que se llamó "*TablaAnalysysPangenomas.tsv*" , que consta de tres columnas: la primera son las secuencias por clústeres, la segunda son las secuencias agrupadas en cada clúster y la tercera, el número de acceso de la proteína en RefSeq. Se puede ver una parte del archivo en la figura 15.

Figura 15. Parte del archivo "TablaAnalysysPangenomas.tsv". La primera columna son los clústeres, la segunda, las secuencias pertenecientes a cada clúster, y la tercera, el número de acceso de la proteína en RefSeq.

.3. Cálculo del pangenoma y coregenoma

Partiendo de ese archivo, se generó un script en R (Véase *Script Pangenoma en R*, en la sección Anexos) con el que iniciamos el estudio del pangenoma y coregenoma.

Se calculó 10 veces el número de genes para cada tamaño de genoma utilizados (5, 10, 25, 50, 100, 200, 300, 400, 800 0 1000). Después se obtuvo la media para cada punto, y la desviación estándar, tanto para el pangenoma como el coregenoma.
Se obtuvo el tamaño del genoma accesorio, restando el tamaño del pangenoma de el del coregenoma (Tabla 1).

Se generaron los gráficos correspondientes con el paquete *ggplot* de R.

Como se puede observar en las figuras 16, 17 y 18 el número de genes nuevos aumenta de manera exponencial según aumenta el número de cepas. Es decir, sigue la ley de Heaps [46], originalmente formulada en el campo del lenguaje: la tasa a la cual se encuentran nuevas palabras en un texto dado, decrece según se amplía el tamaño de dicho texto. Esto es válido en el caso del pangenoma de las cepas analizadas: según aumenta el número de genomas, decrece la cantidad de nuevos genes. Se observa un pangenoma, tanto para ambos géneros conjuntos, como para *escherichia* o *shigella*, que no alcanza una meseta. Es teóricamente infinito. Se dice que es un pangenoma abierto.

Si se observa la tabla 1, vemos que el tamaño del pangenoma para ambos géneros conjuntos, no obedece a la suma del tamaño del pangenoma para cada género por separado. Esto es lógico si pensamos que ambos géneros comparten genes. Por lo tanto, nuevos genes para uno de los géneros, pueden no serlo para la suma de ambos.

El tamaño del coregenoma (genes compartidos) decrece según aumenta el número de genomas (Figuras 16, 17 y 18 y Tabla 1). Es decir, al aumentar el tamaño del genoma, la población va adquiriendo nuevos genes que ya no son compartidos por todos los genomas.

Figura 16. Genes nuevos, en función del número de cepas, en el genoma conjunto de Escherichia y Shigella. En el eje y, log de la media ± sd. Eje x: número de cepas. 10 repeticiones por punto. En verde, pangenoma. En rojo, coregenoma.

Figura 17. Genes nuevos, en función del número de cepas, en el genoma de Escherichia. En el eje y, log de la media ± sd. Eje x: número de cepas. 10 repeticiones por punto. En verde, pangenoma. En rojo, coregenoma.

Figura 18. Genes nuevos, en función del número de cepas, en el genoma de Shigella. En el eje y, log de la media ± sd. Eje x: número de cepas. 10 repeticiones por punto. En verde, pangenoma. En rojo, coregenoma.

Nº de Genomas	Escherichia&Shigella			Escherichia		Shigella	
	Pan	Core	Accesorio	Pan	Core	Pan	Core
5	8380,6	2910,4	5470,2	7408,2	2796,2	1586,2	424,7
10	10998,0	2699,1	8298,9	10374,8	2484,4	2120,2	371,6
25	16735,8	2205,0	14530,8	15498,8	2047,0	3068,0	282,1
50	22700,8	1672,6	21028,2	21211,8	1650,5	4639,0	210,1
100	31615,9	1465,4	30150,5	30065,4	1330,4	7236,9	139,5
200	45113,7	882,8	44230,9	42527,5	876,0	11756,9	82,7
300	55041,9	641,7	54400,2	52939,2	686,4	14195,2	44,7
400	64086,6	402,9	63683,7	61235,1	487,0	18812,5	42,7
800	92803,8	202,9	92600,9	89263,9	234,4	27507,4	18,6
1000	106507,0	162,9	106344,1	102599,0	162,6	33983,8	14,9

Tabla 1. Media (n=10) del número de genes en pangenoma (Pan), coregenoma (Core) o Accesorio (en amarillo) para Escherichia&Shigella, Esherichia o Shigella (100 % de identidad) según número de genomas

El método empleado con más frecuencia para el estudio del coregenoma es el cálculo de los genes presentes en el 100% de las muestras. Esto conlleva algunas desventajas. Por ejemplo, puede haber variantes raras que no tengan algunos genes que, de otra manera podrían ser incluidos dentro del

coregenoma [47]. Kaas et al. en 2012 [48] propusieron el concepto *softcore*: todos los genes encontrados en, al menos, el 95% de todos los genomas.

En este trabajo, también se calculó el tamaño del coregenoma para cutoff de 85% y de 95%. Es decir, con una restricción del 85% o 95% de nuevos genes. En la tabla 2 y figura 19, se observa que el número de genes nuevos se mantiene prácticamente constante con estos umbrales. Sin embargo, como ya habíamos comentado anteriormente, sin aplicar ninguna restricción, el tamaño del coregenoma va decreciendo al aumentar el número de genomas.

Nº de Genomas	Escherichia Core 85%	Shigella Core 85%	Esch&Shig Core 85%	Esch&Shig Core 95%
5	2729,4	2622,9	3009,9	3060,6
10	2764,9	2898,5	3206,9	2527,9
25	2738,2	2875,2	3120,6	2783,1
50	2756,9	2926,5	3139,8	2636,3
100	2765,4	2946,8	3159,6	2696,9
200	2519,7	2882,3	3158,7	2671,8
300	2761,8	2891,4	3160,7	2652
400	2758,4	2863,3	3161,4	2639,7
800	2757,7	2868,2	3149,3	2648,8
1000	2757,1	2866,7	3147,4	2640,9

Tabla 2. Media (n=10) del número de genes en coregenoma para Escherichia&Shigella, Esherichia o Shigella (85 % o 95% de identidad) según número de genomas.

Figura 19. Representación de la media (n=10) del número de genes del core-genoma respecto al número de genomas para diferentes umbrales (85%, 95% o 100%) en Escherichia&Shigella, Escherichia o Shigella.

La identificación de los genes que forman parte del coregenoma es, a menudo, el primer paso en los estudios de genómica de poblaciones. Normalmente, los genes que forman parte del coregenoma están relacionados con funciones básicas: replicación, translación y mantenimiento de la homeostasis celular [41,49].
El número de genes que componen el coregenoma podría ser un indicador de la diversidad genética: según el coregenoma es más pequeño en una especie dada, su diversidad es mayor [50].
Mira y colaboradores [51], basándose en la cantidad de genes de las 8 primeras cepas secuenciadas de E.coli predijeron que su coregenoma estaba compuesto por unos 2800 genes. En estudios posteriores [52] se calculó que el número de genes nuevos por cada nueva cepa sería de unos 300.
Kaas et al. [48] hallaron un tamaño para el softcoregenoma de E.coli y Shigella de 3051 genes para 186 genomas. Nosotros hemos hallado 2697 genes para 100 genomas y 2672 para 200 (Tabla 2), lo cual se aproxima bastante a esos datos. Hay que tener en cuenta, que Kaas et al. trabajaron solo con E.coli y *Shigella*. En este trabajo se han incluido otras *Escherichias*. Si se toma en cuenta el coregenoma estricto, Kaas et al. encuentran 1702 genes y nosotros, entre 1465 para 100 genomas y 883 para 200, también por debajo de lo encontrado por Kaas et al. Por otro lado, Touchon et al. [39] calculan un tamaño de coregenoma para 20 genomas de *E.coli*, de 1976 genes. Nosotros obtuvimos 2047 genes para 25 genomas de *Escherichia*, también bastante aproximado, pero difícil de comparar por la inclusión en nuestro trabajo de otras especies de *Escherichia*.

Hicimos un cálculo del tamaño del pangenoma y del coregenoma para 13000 cepas (casi el total de las incluidas en este estudio), resultando una media de 461584 genes en el pangenoma y solo 37 en el coregenoma. Se podría concluir que solo hay 37 genes compartidos por todas las cepas.
Cuando calculamos el soft coregenoma, esta cifra asciende a 2644 genes compartidos por el 95 % de las cepas, muy parecido al resto de cálculos para soft coregenoma (Tabla 2). Es decir, hay entre 2500 – 2700 genes que se mantienen estables en el 95 % de las cepas.

Los genes que componen el genoma accesorio son genes relacionados con la supervivencia en determinados nichos. Se suelen relacionar con virulencia o resistencia a antibióticos [51]. Su tamaño se obtiene restando el coregenoma del pangenoma (Tabla 1).

4. Conclusiones

- Se observa que los clústeres de *shigella* se posicionan muy cercanos a los de *escherichia*, a veces confundiéndose con ellos, sugiriendo la gran proximidad genómica entre ambos géneros.
- Más del 90 % de las cepas de Shigella se clasificaron como B1 según el algoritmo de Clermont.
- Las cepas de Shigella, se posicionan a corta distancia de las cepas de Escherichia clasificadas como B1.
- El pangenoma de ambos géneros se clasifica como abierto.
- Las curvas del coregenoma de ambos géneros decrecen según aumenta el número de genomas.
- Hay unos 2500 – 2700 genes comunes al 95 % de las cepas del genoma conjunto de ambos géneros

Análisis crítico:
Respecto a la consecución de los objetivos iniciales, se han dejado abiertas cuestiones como la comparativa del viruloma, resistoma y moviloma de los géneros y sub grupos o la profundización en las cepas de shigella, que se apartan de las características comunes a su especie.
El principal problema ha sido la falta de previsión del tiempo consumido en scripts en R y programas, debido al peso de los archivos que se han manejado. A finales de marzo, se dispuso un sistema VPN para poder establecer conexión al servidor de la Unidad de Bioinformática, pero esta conexión se ha caído muy a menudo, a veces en mitad de un trabajo.
Por otro lado, incluso con la conexión, no se mejoró mucho en cuanto al tiempo consumido por R.
A veces se ha tenido que cambiar de metodología y utilizar otra función u otro programa. Es el caso de la creación de clústeres de las secuencias genómicas en R. Se pensó utilizar la función *mclust*. Al cabo de varios intentos la idea se abandonó sin llegar a completarse el script después de más de 24 horas de espera, a favor de *kmeans*, también lenta (horas para completar un gráfico) y mucho menos conveniente para nuestro propósito, como se demostró posteriormente (Figura 5).
También, algunos programas, como *gephi* han consumido mucho tiempo del disponible, ya que se intentaron buscar las mejores condiciones en un compromiso calidad/tiempo. Este programa, aún teniendo grandes ventajas respecto a otros en calidad de visualización de grafos y posibilidad de manipulación de datos y variables, sin embargo adolece de facilidad de uso en cuestiones prácticas, como posibilidad de añadir etiquetas, conservar colores de grupos...etc., por lo que su manejo es bastante lento.
Por otro lado, la autora del trabajo ha adquirido nuevas y variadas habilidades en el campo de la bioinformática: Se ha trabajado con scripts en Perl y Unix. Se

ha empleado software novedoso: Mmseqs2, Mash o Gephi y paquetes y funciones de R, desconocidos hasta ahora para la estudiante, como UMAP e igraph.

Posibles líneas de trabajo futuro:

- Análisis de la matriz de distancias genómicas, para identificar valores atípicos.
- Análisis del conjunto de genes significativamente más frecuentes en cada población.
- Estudio de los genes componentes del genoma accesorio, del coregenoma y del soft coregenoma.
- Identificación del Viruloma, Resistoma y Mobiloma
- Creación de una métrica de virulencia

5. Glosario

UMAP: Uniform Manifold Approximation and Projection.

K-means: Método de agrupamiento, que tiene como objetivo la partición de un conjunto de *n* observaciones en *k* grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano.

Clustering: Agrupación de un conjunto de objetos (secuencias en este caso), de tal manera que los de un mismo grupo (clúster) sean más similares (según la matriz de distancias en este caso) entre sí que a los de otros grupos o clústeres.

Pangenoma: Conjunto de todos los genes de una especie.

Coregenoma: Conjunto de todos los genes compartidos por todos los miembros de una especie.

FASTA: Formato de fichero informático basado en texto, utilizado para representar secuencias bien de ácidos nucleicos, bien de péptido, y en el que los pares de bases o los aminoácidos se representan usando códigos de una única letra.

6. Bibliografía

1. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O. et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 5(1), 2009.

2. Donnenberg, Michael S., ed. *Escherichia coli: virulence mechanisms of a versatile pathogen*. Amsterdam; Boston: Academic press, 2002.
3. Ranjbar R, Dallal MM, Pourshafie MR. Epidemiology of shigellosis with special reference to hospital distribution of *Shigella* strains in Tehran,Iran. J Clin Infect Dis.,3(1), 2008.
4. Hale TL. Genetic basis of virulence in Shigella species. Microbiol Rev 55(2), 1991.

5. Pupo GM, Lan R, Reeves PR. Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics. Proc Natl Acad Sci USA 97 (19), 2000.

6. van den Beld, MJC [illegible], F. A. G. Reubsaet. Differentiation between Shigella, enteroinvasive Escherichia coli (EIEC) and noninvasive Escherichia coli. European journal of clinical microbiology & infectious diseases 31.6, 2012.

7. Brenner DJ, Fanning GR, Skerman FJ, Falkow S. Polynucleotide sequence divergence among strains of Escherichia coli and closely related organisms. J Bacteriol. Mar;109(3), 1972.

8. Friedman D, Court D. Bacteriophage lambda:alive and well ans still doing its thing. Curr.Opin.Microbiol. 4(2), 2001.

9. Beutin L. Emerging enterohaemorrhagic Escherichia coli, causes and effects of the rise of a human pathogen. *J Vet Med B Infect Dis Vet Public Health* **53** (7), 2006.

10. Shakya G, Acharya J, Adhikari S, Rijal N .Shigellosis in Nepal: 13 years review of nationwide surveillance.. J Health Popul Nutr., Nov 4;35(1):36. 2016.

11. Devanga Ragupathi N.K., Muthuirulandi Sethuvel D.P., Inbanathan F.Y. and Veeraraghavan B. Accurate differentiation of Escherichia coli and Shigella serogroups: challenges and strategies. New Microbes and New Infections,, Volume 21 Number C, January, 2018.

12. Ondov, Brian D., et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome biology, 17.1, 2016.
13. https://igraph.org, Marzo, 2019.
14. Leland Mc, Healy J, and Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018.
15. Steinegger M and Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. J. Nature Biotechnology, volume 35, 2017.
16. https://gephi.org, Mayo, 2019.
17. https://www.ncbi.nlm.nih.gov/genbank/, 20-Febrero-2019.
18. https://www.ncbi.nlm.nih.gov/refseq/, 20-Febrero-2019.
19. Broder AZ. COM '00 Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching. London: Springer; 2000. Identifying and filtering near-duplicate documents; pp. 1–10.
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol. 1990 Oct 5; 215(3):403-10.*
21. Chum O, Philbin J, Zisserman A. Near Duplicate Image Detection: min-Hash and tf-idf Weighting. In: Proceedings of the British Machine Vision Conference 2008. Durham, UK: British Machine Vision Association and Society for Pattern Recognition; 2008.
22. Rasheed Z, Rangwala H. 2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum: A Map-Reduce Framework for Clustering Metagenomes IEEE. 2013.
23. Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. Jun; 33(6), 2015.
24. Fan H, Ives AR, Surget-Groba Y, Cannon CH. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. BMC Genomics. Jul 14; 16(), 2015.

25. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. Big Data: Astronomical or Genomical? PLoS Biol. Jul; 13(7), 2015.
26. https://mash.readthedocs.io/en/latest/distances.html, abril, 2019.
27. E.W. Forgy. "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". Biometrics. 21 (3), 1965.
28. http://zzz.bwh.harvard.edu/luna/vignettes/nsrr-umap, Abril, 2019.
29. https://umap-learn.readthedocs.io/en/latest/clustering.html, Abril, 2019.
30. E.W. Forgy. "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". Biometrics. 21 (3), 1965.
31. https://uc-r.github.io/kmeans_clustering#elbow, Abril, 2019.
32. https://github.com/A-BN/ClermonTyping#citing, Mayo, 2019
33. Beghain, Johann, et al. "ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping." *Microbial genomics* 4.7 (2018).
34. http://clermontyping.iame-research.center/. Mayo, 2019.
35. Jacomy, Mathieu, et al. "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software." *PloS one* 9.6, 2014.
36. Kwak, Min-Jung, et al. "Genome sequences of the Shiga-like toxin-producing Escherichia coli NCCP15655 and NCCP15656." *Gut pathogens* 7.1, 2015.
37. Wang, Fu-Sheng, Thomas S. Whittam, and Robert K. Selander. "Evolutionary genetics of the isocitrate dehydrogenase gene (icd) in Escherichia coli and Salmonella enterica." *Journal of bacteriology* 179.21,1997.
38. Escobar-Paramo, Patricia, et al. "The evolutionary history of Shigella and enteroinvasive Escherichia coli revised." *Journal of molecular evolution* 57.2, 2003.
39. Touchon, Marie, et al. "Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths." *PLoS genetics* 5.1, 2009.
40. Sims, Gregory E., and Sung-Hou Kim. "Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs)." *Proceedings of the National Academy of Sciences* 108.20, 2011.
41. H. Tettelin et al., “Genome analysis of multiple pathogenic isolates of Streptococcus agalatciae: Implications for the bacterial ‘pan-genome,’” *PNAS*, 102:13950-55, 2005.
42. https://www.the-scientist.com/features/the-pangenome-are-single-reference-genomes-dead-32458, mayo, 2019.
43. http://www.metagenomics.wiki/pdf/definition/pangenome, mayo, 2019.

44. https://github.com/soedinglab/mmseqs2, Abril, 2019.
45. https://mmseqs.com/latest/userguide.pdf, Abril, 2019.
46. Heaps, H. S. Information Retrieval – Computational and Theoretical Aspects, Academic Press, 1978.
47. van Tonder, Andries J., et al. "Defining the estimated core genome of bacterial populations using a Bayesian decision model." PLoS computational biology 10.8, 2014.

48. Kaas, Rolf S., et al. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. *BMC genomics* 13.1, 2012.
49. Medini, Duccio, et al. The microbial pan-genome. Current opinion in genetics & development 15.6, 2005.
50. Lawrence, Jeffrey G., and Heather Hendrickson. Genome evolution in bacteria: order beneath chaos. Current opinion in microbiology 8.5, 2005.
51. Mira, Alex, et al. The bacterial pan-genome: a new paradigm in microbiology, Int Microbiol 13.2, 2010.
52. Konstantinidis, Konstantinos T., and James M. Tiedje. Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences 102.7, 2005.

7. Anexos

1. Script igraph de R

```
#Se lee el archivo, obtenido con mash dist
dos<-read.table("~distred.tsv")

#Conservamos las variables que nos interesan: las parejas de secuencias y los valores de distancia
dosa<-dos[, 1:3]

#Eliminamos secuencias repetidas.
library(dplyr)
n<-distinct(dos, dos$V1)

#Con igraph se agrupan secuencias en grupos, y conservamos una secuencia de cada grupo
```

```
library(igraph)
net<-graph_from_data_frame(d = dosa, vertices = n, directed=T)
E(net)

#Se obtiene el archivo con las secuencias agrupadas
sink("components.txt")
clu<-components(net)
gc<-groups(clu)
sink().
```

2. **Script umap de R**

```
#Se cargan las librerías necesarias.
library("tidyverse")
library("umap")
setwd("~/...")

#Lectura del archivo con las secuencias y sus distancias
genall = read.csv("~/...gensinred.tsv", sep = "\t", header = TRUE)

# Creación de un vector (sp) que adjudica la clase "1" a las secuencias de
Shigella (las primeras 1268). Y, el "2" al resto (Escherichia)
sp <- c(rep(1:2, c(1268,12215)))

#Se preparan los datos
genall2b = genall %>% column_to_rownames("X.query")

#Función umap. Los parámetros pueden verse en el archivo "umap1.defaults".
genumb = umap(genall2b)

#Se preparan los datos para el gráfico
testx <- cbind(genumb$layout[,1], sp)

#Gráfico con la población de secuencias.
plot(genumb$layout, col=testx[,2]) #File umap1

#Se cambia el parámetro n_neighborgs a un valor de 5
Genumb2 <- umap(genall2b, n_neighbors =5)
plot(genumb2$layout, col=testx[,2]) #File umap2
```

3. **Archivo “umap1.defaults”**

umap configuration parameters

n_neighbors: 15
n_components: 2
metric: euclidean
n_epochs: 200
input: data
init: spectral

min_dist: 0.1
set_op_mix_ratio: 1
local_connectivity: 1
bandwidth: 1
alpha: 1
gamma: 1
negative_sample_rate: 5
a: NA
b: NA
spread: 1
random_state: NA
transform_state: NA
knn_repeats: 1
verbose: FALSE
umap_learn_args: NA

4. **Script *k-means* en R**

```
#Se cargan las librerías
library(tidyverse)
library(factoextra)

#Se lee el archivo de secuencias genómicas y preparación
genall = read.csv("~/gensinred.tsv", sep = "\t", header = TRUE)
genall2b = genall %>% column_to_rownames("X.query")
genalldist<-as.dist(genall2b)

#Cálculo del número ideal clusters
fviz_nbclust(genall2b, FUNcluster = kmeans)

#Se calcula kmeans con 2 clusters
fitk <- kmeans(genalldist, centers = 2, nstart = 25 )

#Creación del gráfico
fviz_cluster(fitk, data = genalldist, geom = c("point"))
```

5. **Script Clermont en R**

```
library(tidyverse)
library(igraph)
library(mclust)
library(umap)
library(stringi)
library(readr)
install.packages("BiocManager")
BiocManager::install("ggplot2")
library(ggplot2)

#Se lee la matriz de distancias de secuencias genómicas, creada con mash
Matrix = read_delim("gensinred", col_names = TRUE, delim = "\t")

#Transformación de la matriz unimodal en un objeto igraph, en el que nos
quedamos con la mitad de la imagen especular que forma la matriz cortada por
la diagonal.
gr = graph_from_adjacency_matrix(Matrix %>% column_to_rownames("query")
%>% as.matrix(), weighted = TRUE, diag = FALSE, mode = "upper")

#Se crea un dataframe de los datos
TablE = as_data_frame(gr)

#Estructura del data.frame:
str(TablE, 2)
#'data.frame':          90888897 obs. of  3 variables
#Algunos datos estadísticos:
```

```
M = mean(TablE$weight)  #M=0.02540488
SD = sd(TablE$weight)   #SD=0.02191903
max(TablE$weight)       #Max=1
min(TablE$weight)       #Min= 2.38274e-05

#Creación de un archivo, al que llamamos "SemiNet---.tsv" con los 10, 30, 60, 100 o 300 pesos más bajos de la tabla
#Para 100:
TablE  %>% group_by(from) %>%  top_n(-100,weight) %>%
  rename(Source = from, Target = to) %>% mutate(weight = 2-weight) %>%
mutate(Type = "Undirect") %>%
  write.table("SemiNet100.tsv",sep = "\t", quote = FALSE, row.names = FALSE)

#Lectura de el archivo "Clermont.tsv, creado con la identidad de la secuencia,el grupo filogenético al que pertenece, la Distancia, pvalue y el sketch (otra manera de medir distancias)
Clermont = read.table("Clermont.tsv")
colnames(Clermont) = c("ID","group","Dist","pvalue","sketch")
head(Clermont, 2)
#                                   ID                group
#1  shig-GCF_000006925.2_ASM692v2_genomic.fna fasta/327_20_F_.fasta
#2 shig-GCF_000012005.1_ASM1200v1_genomic.fna fasta/327_20_F_.fasta
#Dist pvalue   sketch
#1 0.0302852      0 360/1000
#2 0.0313762      0 349/1000

#Se da forma al archivo, eliminando algunos prefijos y definiendo la variable "ID" como "especie"  + "Accesion"
library(stringi)
Clermont.full = Clermont %>% group_by(ID) %>% top_n(-1,Dist) %>% filter(Dist < 0.05) %>%
  mutate(group = gsub("fasta/","",group))%>%
  mutate(group = gsub(".fasta","",group)) %>%
  mutate(group = stri_replace_last(group,"",regex = "_")) %>%
  mutate(group = stri_replace_last(group," ",regex = "_")) %>%
  separate(group,c("kk","phylogroup"), sep =" ") %>% select(-kk) %>%
  mutate(phylogroup = gsub("B2","B2-",phylogroup)) %>%
  separate(ID, c("Specie","Accession"), sep ="-", remove = FALSE)

#Se guarda el archivo resultante como "NodosClermont100.tsv"
write.table(Clermont.full,"NodosClermont100.tsv", row.names = FALSE, quote = FALSE, sep = "\t")

#Umap-grupos filogenéticos
#Preparamos nuestros datos y aplicamos la función "umap" con atributos por defecto
Matrix = Matrix %>% column_to_rownames("query")
library(umap)
UM = umap(Matrix)
```

```
UM.table = UM$layout %>% as.data.frame() %>% rownames_to_column("ID")
%>% left_join(Clermont.full) %>% separate(ID, c("Species","Accession"), sep
="-", remove = FALSE)

#Creamos el gráfico
library(ggplot2)
UM.table %>% ggplot(aes(x= V1, y = V2, color = phylogroup)) +
  geom_point() +  ggtitle("Phylogroups Distances") +
theme(plot.title = element_text(size = 20, face = "bold")) +
  theme(legend.key.size = unit(.8, "cm"))
```

. **Script Pangenoma en R**

```
#Cargamos las librerías
library(tidyverse)
#Lectura del archivo con los clústeres
Pangenomas = read_tsv("...TablaAnalysysPangenomas.tsv", col_names =
FALSE)
#Se pone nombre a las variables
colnames(Pangenomas) = c("Cluster","Genoma","Prot")
#Se conservan los genomas únicos
Genomas = Pangenomas %>% select(Genoma) %>% distinct()
## Tamaño pangenoma en este caso para 5 genomas
pang<-function(i) {Pangenomas %>% semi_join(sample_n(Genomas, 5)) %>%
select(Cluster)%>%  distinct() %>% count()
}
#Se repite la operación anterior 10 veces
Resultpansind5 <- sapply(1:10, pang)

#Se hace lo mismo para 10, 25, 50, 100, 200, 300, 400, 800 y 1000 genomas
#Se calcula la media, la sd y el std error para 5 genomas
result5pansindf <- as.data.frame((unlist(resultpansind5)))
colnames(result5pansindf) <- "pan5sin"
mean5pansin<-round(mean(result5pansindf$pan5sin), 2)
sd5pansin<-round(sd(result5pansindf$pan5sin), 2)
library(plotrix)
se5pansin<-round(std.error(result5pansindf$pan5sin,na.rm), 2)

#Se hace lo mismo para 10, 25, 50, 100, 200, 300,400, 800 y 1000 genomas

## Tamaño core-genoma en este caso para 5 genomas
pangcore<-function(i) {Pangenomas %>% semi_join(sample_n(Genomas,5))
%>%
   group_by(Cluster) %>%
   summarise(ClusterSize = n()) %>% filter(ClusterSize >= 5) %>% count()
}
#Calculamos 10 veces
Resultcorsind5 <- sapply(1:10, pangcore)
#Se hace lo mismo para 10, 25, 50, 100,200, 300,400, 800 y 1000 genomas
```

```
#Se calcula la media, la sd y el std error para 5 genomas
result5corsindf <- as.data.frame((unlist(resultcorsind5)))
colnames(result5corsindf) <- "cor5sin"
mean5corsin<-round(mean(result5corsindf$cor5sin), 2)
sd5corsin<-round(sd(result5corsindf$cor5sin), 2)
se5corsin<-round(std.error(result5corsindf,na.rm), 2)

#Se hace lo mismo para 10, 25, 50, 100,200, 300, 400 , 800 y 1000 genomas

#Se crean variables con todas las medias, sd y se tanto de pan como de core, obtenidas
#---------- Mean
meanpansin2 = c(mean5pansin,mean10pansin,mean25pansin,mean50pansin,
            mean100pansin, mean200pansin, mean300pansin, mean400pansin,
            mean800pansin, mean1000pansin)
meancorsin2 = c(mean5corsin,mean10corsin,mean25corsin,mean50corsin,
            mean100corsin, mean200corsin, mean300corsin, mean400corsin,
            mean800corsin, mean1000corsin )
meansin2 = c(meanpansin2, meancorsin2)
meansin2 <- as.data.frame(meansin2)
#---------- SD
sdpansin2 = c(sd5pansin,sd10pansin,sd25pansin, sd50pansin, sd100pansin,
sd200pansin,
          sd300pansin, sd400pansin, sd800pansin, sd1000pansin )
sdcorsin2 = c(sd5corsin,sd10corsin,sd25corsin,sd50corsin, sd100corsin,
sd200corsin,
          sd300corsin, sd400corsin, sd800corsin, sd1000corsin)
sd2 = c(sdpansin2, sdcorsin2)
#-------- SE
sepansin = c(se5pansin,se10pansin,se25pansin, se50pansin,se100pansin,
se200pansin,
          se300pansin, se400pansin)
secorsin = c(se5corsin,se10corsin,se25corsin,se50corsin, se100corsin,
se200corsin,
          se300corsin, se400corsin)
se = c(sepansin, secorsin)
#----------,
#Se crea la variable "grupo2"
grupo2 = c(rep("Pangenoma", 10), rep("Core", 10))
# "N2" es la variable que representa el nº de genomas en cada uno de los grupos
N2<-c(5,10,25,50,100, 200,300, 400,800, 1000,5,10,25,50,100,200, 300,
400,800, 1000)

# data.frame con las variables
 df_sin3 <- cbind(meansin2, sd2, N2, grupo2)
 df_sin3<-as.data.frame(df_sin3)

#Se crea el gráfico con ggplot
library(ggplot2)
```

```
ggplot(df_sin3, aes(x=N, y=meansin2, colour=grupo2, group=grupo2))+
  geom_line()+
  geom_point()+
  geom_errorbar(data=df_sin3, mapping=aes(x=N, ymin= meansin2 - sd2,
                                          ymax= meansin2 + sd2),
                width=.2, colour="black", position=position_dodge(0.05)) +
   labs(title = "Pangenoma_Escherichia_Shigella") +
  scale_y_log10()+
  scale_y_log10()+
  ylab("log Mean ? sd") +
  xlab("N?mero_de_muestras")
```