Universidad Nacional Autónoma de México

Facultad de Ciencias

Diferencias y similitudes entre cepas de E. coli

Genómica Computacional | Semestre 2026-1

Profesor: M. en C. Sergio Hernández López

Integrantes:

Martínez Oviedo Guillermo

Sánchez Cruz Norma Selene

Sosa Romo Juan Mario

Fecha: 12 de diciembre de 2025

Índice general

Capítulo 1

Introducción

1.1. Contexto biológico

***Escherichia coli* es un microorganismo altamente versátil que forma parte de la microbiota intestinal comensal de humanos y otros mamíferos**, donde generalmente coexiste en un estado de equilibrio y beneficio mutuo con el huésped (Kaper et al., 2004). De hecho, *E. coli* es el anaerobio facultativo más abundante del intestino grueso. No obstante, a pesar de su papel como habitante inofensivo y de su amplio uso como organismo modelo y herramienta biotecnológica, existen múltiples linajes altamente adaptados que actúan como patógenos capaces de causar un amplio espectro de enfermedades, tanto intestinales como extraintestinales, en individuos sanos e inmunocomprometidos.

Las enfermedades diarreicas causadas por cepas patógenas de *E. coli*, denominadas colectivamente *E. coli* diarreagénico (DEC), representan un importante problema de salud pública y constituyen una causa relevante de morbilidad y mortalidad, particularmente en lactantes y niños pequeños en países en desarrollo. La patogenicidad de estas cepas se debe, en gran medida, a la adquisición de atributos específicos de virulencia, frecuentemente mediada por transferencia horizontal de genes, lo que les confiere la capacidad de adaptarse a nuevos nichos ecológicos y de provocar enfermedad.

Estas combinaciones exitosas de factores de virulencia han dado lugar a la emergencia de patotipos bien definidos de *E. coli*, los cuales se clasifican en función de los determinantes de virulencia adquiridos, sus mecanismos moleculares de acción y los cuadros clínicos asociados. Entre los principales patotipos responsables de enfermedades diarreicas se encuentran *E. coli* enteropatógeno (EPEC), enterotoxigénico (ETEC), enteroagregativo (EAEC), enteroinvasivo (EIEC) y enterohemorrágico (EHEC/STEC), entre otros (Gomes et al., 2016).

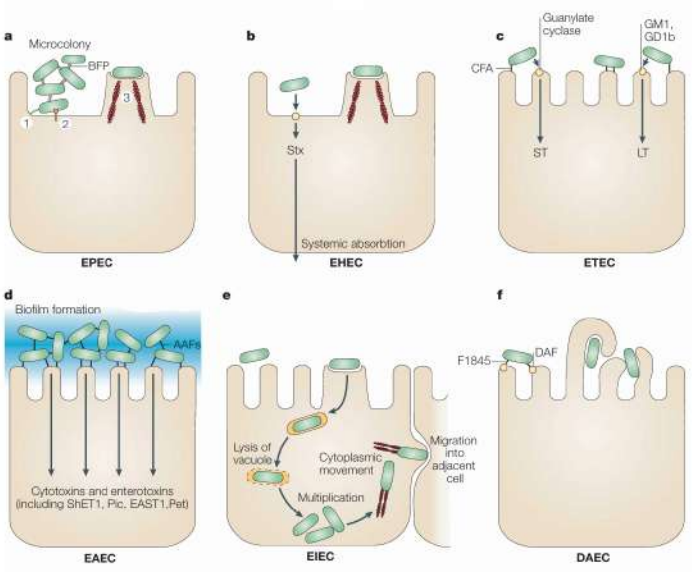


Figura 1.1: Esquema patogénico de E. coli diarreogénica. Fuente: (Kaper et al., 2004)

Desde una perspectiva genómica, *E. coli* presenta un *pangenoma abierto*, lo que implica que el número total de familias génicas continúa incrementándose conforme se incorporan nuevos genomas al análisis comparativo. Estudios previos han estimado que el *genoma núcleo*, definido como el conjunto de genes compartidos por todas las cepas, está compuesto por aproximadamente 2 200 genes, mientras que el *genoma accesorio* es considerablemente más extenso, con un reservorio que supera las 13 000 familias génicas a nivel poblacional (Rasko et al., 2008). Esta fracción accesoria concentra la mayoría de los determinantes de virulencia y adaptación, lo que complica de manera significativa su análisis, ya que la patogenicidad emerge de combinaciones genéticas específicas dentro de un paisaje genómico altamente dinámico y heterogéneo.

1.2. Planteamiento del problema

La identificación taxonómica por sí sola resulta insuficiente para predecir el potencial patogénico de una cepa bacteriana, ya que linajes tradicionalmente considerados *comensales* de *Escherichia coli* pueden adquirir, mediante procesos evolutivos, un comportamiento virulento y transformarse en patógenos. Esta plasticidad fenotípica y genómica limita el valor predictivo de los enfoques basados únicamente en clasificación taxonómica.

El problema central de este estudio consiste en descifrar la *firma genómica* asociada a la virulencia mediante métodos computacionales. En este contexto, surge la siguiente pregunta de investigación: **¿Es posible diferenciar computacionalmente las cepas patógenas de *E. coli* de las comensales basándose únicamente en la presencia o ausencia de marcadores genéticos de virulencia específicos?**

En este trabajo se propone aplicar herramientas y conceptos de genómica computacional para integrar múltiples niveles de análisis comparativo, con el objetivo de caracterizar

de manera sistemática los determinantes de virulencia en cepas patógenas y emergentes de *E. coli*.

1.3. Objetivos

1.3.1. Objetivo general

Analizar comparativamente el contenido genómico de cepas patógenas y comensales de *Escherichia coli* mediante herramientas bioinformáticas, con el fin de identificar patrones de presencia y ausencia de genes que permitan discriminar los principales patotipos diarreagénicos.

1.3.2. Objetivos Específicos

1. Establecer un perfil de referencia de los genes de virulencia canónicos (adhesinas, toxinas y sistemas de secreción) asociados a los patotipos DEC mediante la consulta de bases de datos especializadas.

2. Determinar *in silico* la prevalencia y distribución de dichos marcadores de virulencia en un conjunto de genomas representativos, cuantificando las diferencias entre cepas patógenas y el grupo de control comensal.

3. Evaluar el agrupamiento de las cepas analizadas mediante técnicas de genómica comparativa (como análisis de componentes principales o filogenia basada en genoma accesorio) para verificar si la presencia de estos genes permite reconstruir la clasificación patogénica conocida.

Capítulo 2

Materiales y Métodos

El flujo de trabajo se diseñó siguiendo un enfoque de minería de datos genómicos. Se priorizó el uso de herramientas de línea de comandos para la detección masiva de genes y librerías de ciencia de datos para el análisis estadístico, evitando la anotación manual.

2.1. Selección y adquisición de genomas

Se descargaron las secuencias completas (formato FASTA) de cepas de *Escherichia coli* desde la base de datos NCBI RefSeq. El conjunto de datos incluyó:

- Un grupo control de cepas comensales/no patógenas.
- Cepas representantes de los patotipos DEC (principalmente EPEC y EHEC).
- La cepa de referencia *E. coli* K-12 substr. MG1655 (Hayashi et al., 2006) para normalizar los resultados.

Los metadatos asociados (fecha de aislamiento, hospedero y patotipo reportado) se obtuvieron directamente de los archivos de descripción del NCBI.

2.2. Detección automatizada de factores de virulencia

Para determinar el perfil de virulencia, no se realizó una anotación completa *de novo*, sino un cribado dirigido (*screening*). Se utilizó la herramienta bioinformática **ABRicate** (v1.0.1), la cual permite realizar búsquedas locales tipo BLAST de alto rendimiento.

Los genomas fueron confrontados contra la base de datos de factores de virulencia **VFDB** (Virulence Factor Database) (Chen et al., 2005). Para reducir falsos positivos, se establecieron los siguientes umbrales de filtrado estrictos:

- **Identidad mínima:** 90 % (similitud de secuencia).
- **Cobertura mínima:** 80 % (porcentaje del gen detectado).

2.3. Procesamiento de datos y construcción de matrices

Las salidas crudas del cribado se procesaron mediante scripts personalizados en **Python** (usando la librería *Pandas*). Se transformaron los resultados en una **matriz binaria de presencia/ausencia**, donde las filas representan los genomas analizados y las columnas los genes de virulencia únicos identificados.

$$M_{ij} = \begin{cases} 1 & \text{si el gen 'j' esta presente en el genoma 'i'} \\ 0 & \text{si el gen 'j' esta ausente} \end{cases}$$

Esta matriz constituyó la estructura de datos central para los análisis comparativos posteriores.

2.4. Análisis de agrupamiento (Clustering)

Para verificar si la presencia de genes de virulencia permite discriminar computacionalmente entre patotipos, se aplicó un análisis de agrupamiento jerárquico (*Hierarchical Clustering*) sobre la matriz binaria. Se utilizó la métrica de distancia de Jaccard (adecuada para datos binarios asimétricos) y el método de enlace promedio (*average linkage*). La visualización se generó mediante mapas de calor (*heatmaps*) utilizando la librería *Seaborn* de Python.

Capítulo 3

Resultados

3.1. Distribución cuantitativa del contenido de virulencia

El cribado genómico masivo contra la base de datos VFDB permitió cuantificar la carga de factores de virulencia en las ocho cepas analizadas. El análisis reveló una disparidad significativa en el número de determinantes genéticos entre los distintos grupos funcionales (Figura 3.1).

Las cepas patógenas dominaron el conteo de genes únicos. El grupo EHEC (cepas Sakai y EDL933) presentó la mayor complejidad genómica con **165 genes** de virulencia cada una, seguido por la cepa EPEC con **130 genes** y la uropatógena UPEC CFT073 con **128 genes**. En el extremo opuesto, la cepa de referencia K-12 mostró un repertorio basal de solo **49 genes**, correspondientes a sistemas de mantenimiento y transporte de hierro no patogénicos.

Un hallazgo notable fue el perfil de la cepa probiótica *E. coli* Nissle 1917, la cual exhibió **123 genes** de virulencia. Esta cifra es comparable a la de los patógenos francos y sugiere que su capacidad probiótica reside en mecanismos agresivos de colonización y exclusión competitiva, a pesar de no causar enfermedad clínica.

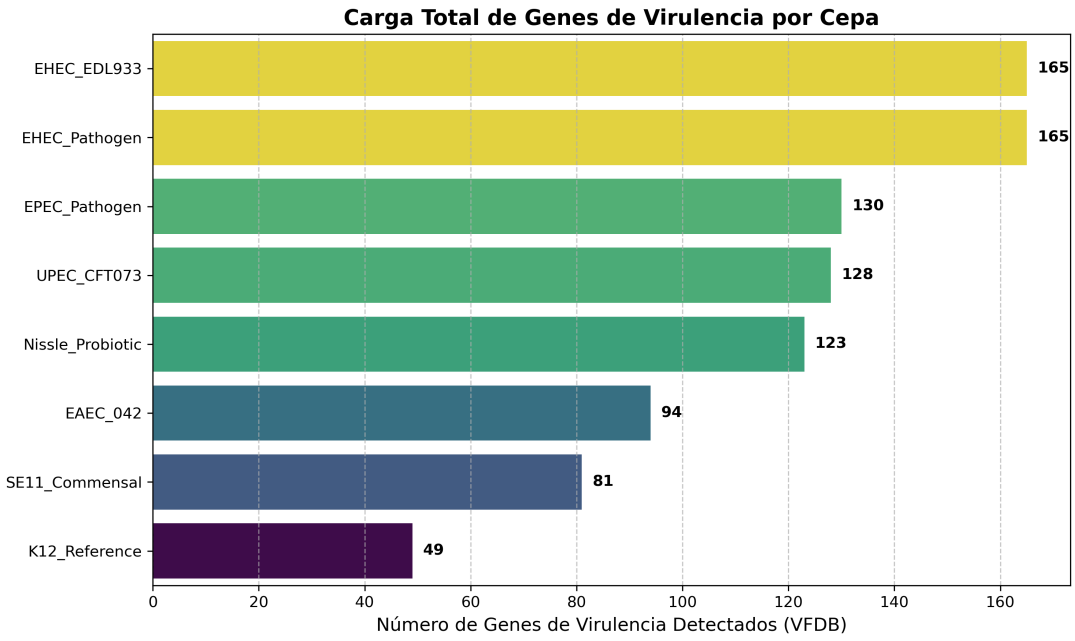


Figura 3.1: Carga total de genes de virulencia únicos identificados por cepa mediante comparación contra VFDB (identidad >90 %). Se observa una clara brecha entre el grupo comensal/laboratorio y los linajes patógenos/probióticos.

3.2. Agrupamiento jerárquico y validación de patotipos

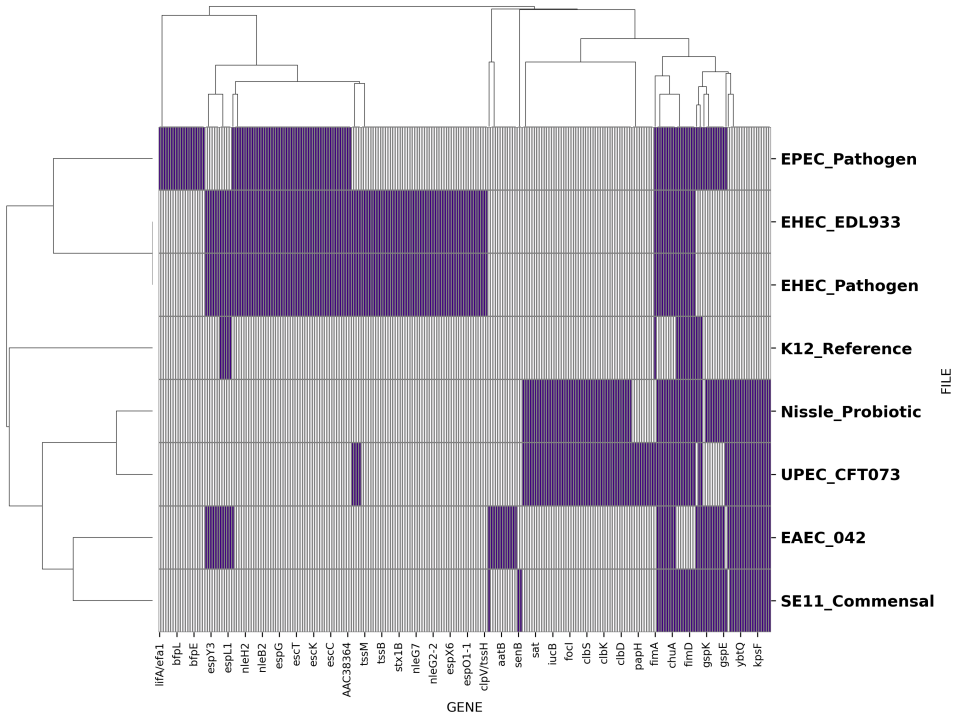
Para evaluar si la presencia o ausencia de estos genes es suficiente para discriminar patotipos, se construyó una matriz binaria y se aplicó un análisis de agrupamiento jerárquico (*Hierarchical Clustering*) utilizando la distancia de Jaccard. El dendrograma resultante (Figura 3.2) reconstruyó con alta fidelidad la clasificación biológica esperada.

El análisis computacional identificó tres clústeres principales:

1. **Clúster LEE-positivo:** Agrupó a las cepas EPEC y EHEC. La similitud entre las dos cepas EHEC (Sakai y EDL933) fue cercana al 100 %, validando la robustez del método. Este agrupamiento se debe a la presencia compartida de la Isla de Patogenicidad del Locus de Borramiento de Enterocitos (LEE).

2. **Clúster Uropatógeno:** La cepa UPEC CFT073 se segregó en una rama independiente, diferenciada por un repertorio de genes únicos asociados a la infección del tracto urinario, ausentes en las cepas entéricas.

3. **Grupo Basal/Comensal:** Las cepas K-12 y SE11 formaron un grupo externo, caracterizado por la ausencia de las grandes islas de patogenicidad.



3.3. Identificación de firmas genómicas discriminantes

El análisis de la matriz permitió extraer los marcadores específicos que actúan como discriminantes computacionales para cada patotipo (Tabla 3.1).

Aunque EPEC y EHEC comparten la maquinaria de secreción tipo III (genes *esc*/*esp*), el análisis detectó que la diferenciación entre estos "patotipos hermanos" recae en dos factores: la toxina Shiga (*stx*), exclusiva de EHEC, y el pili formador de manojos (*bfp*), exclusivo de EPEC. Por otro lado, la cepa enteroagregativa (EAEC 042) mostró un perfil divergente, careciendo de los marcadores clásicos LEE y Stx, lo que explica su posición aislada en el dendrograma.

Tabla 3.1: Matriz resumen de marcadores de virulencia detectados *in silico* que permiten la diferenciación automática de cepas.

Cepa / Patotipo	**Isla LEE** (*eae*)	**Toxina Shiga** (*stx*)	**Pili BFP**	**Fimbria P** (*pap*)	**Hemolisina** (*hly*)
EHEC (Sakai/EDL933)	+	+	-	-	+
EPEC (E2348/69)	+	-	+	-	-
UPEC (CFT073)	-	-	-	+	+
EAEC (042)	-	-	-	-	-
Nissle 1917	-	-	-	-	-
K-12 Reference	-	-	-	-	-

Capítulo 4

Discusión

El análisis comparativo de los ocho genomas de *Escherichia coli* mostró que la presencia y ausencia de genes de virulencia es suficiente para recuperar, en gran medida, la estructura biológica conocida de los principales patotipos diarreagénicos. En particular, las cepas EHEC y EPEC formaron un clúster definido asociado a la presencia de la isla de patogenicidad LEE y la maquinaria de secreción tipo III, mientras que CFT073 se agrupó de manera independiente debido a factores característicos de uropatógenos, como fimbrias P y hemolisinas. Las cepas comensales K-12 y SE11 presentaron un repertorio de virulencia reducido y se ubicaron en la base del dendrograma, lo cual es consistente con su biología no patógena. Por otro lado, la cepa probiótica Nissle 1917 mostró un número de genes de virulencia comparable al de cepas patógenas, lo cual coincide con reportes previos que indican que su potencial colonizador depende de determinantes que, fuera de contexto, se asocian a patogenicidad.

Sin embargo, una limitación importante fue que nuestro algoritmo **no detectó correctamente a la cepa enteroagregativa (EAEC 042)**. Aunque EAEC carece de marcadores clásicos como LEE o Stx, el análisis no logró agruparla de manera coherente con su patotipo, ubicándola en una rama aislada de interpretación ambigua. Esto sugiere que varios de los factores clave de patogenicidad enteroagregativa, como reguladores plasmídicos o adhesinas específicas, no fueron detectados bajo los umbrales estrictos de identidad y cobertura utilizados, o bien no están representados de manera uniforme en la base de datos VFDB. La falla en su clasificación indica que un subconjunto limitado de genes de virulencia puede ser insuficiente para capturar la complejidad de determinados patotipos.

En cuanto a la métrica utilizada, la distancia de Jaccard fue adecuada para matrices binarias dispersas, ya que mide la proporción de genes presentes compartidos entre dos cepas. Es importante aclarar que esta métrica **ignora los pares (0,0)** en los que ambos genomas carecen de un gen determinado. Esto no implica que la ausencia de genes carezca de importancia biológica; simplemente significa que, para efectos computacionales, compartir la ausencia de un gen no aporta información útil para estimar similitud entre cepas. Jaccard está diseñada específicamente para resaltar el impacto de los genes efectivamente presentes en cada comparación.

Este estudio también presenta varias limitaciones metodológicas. El tamaño de muestra fue reducido y no incluyó suficiente variación intra-patotipo, lo que dificulta la gene-

ralización de los resultados. Además, el análisis dependió exclusivamente de VFDB y de parámetros estrictos de búsqueda, lo cual puede haber provocado una subdetección de genes divergentes o mal anotados. Asimismo, la codificación binaria de presencia/ausencia elimina información relevante, como variación puntual, número de copias o regulación de la expresión. Finalmente, solo se emplearon métodos no supervisados de agrupamiento, que pueden ser insuficientes para clasificar patotipos como EAEC cuya señal genética es más sutil o distribuida.

Para investigaciones futuras, sería recomendable ampliar el conjunto de genomas, incorporar bases de datos especializadas en enteropatógenos, y ajustar los umbrales de búsqueda para capturar variantes más divergentes. También sería valioso explorar otras métricas de distancia para datos binarios, así como métodos de aprendizaje supervisado entrenados con paneles curados de genes de virulencia. La integración de información transcriptómica o de variantes podría mejorar la resolución funcional y permitir la clasificación más precisa de patotipos complejos.

En conjunto, nuestros resultados muestran que la minería computacional de genes de virulencia es útil para distinguir varios patotipos de *E. coli*, pero también revelan que ciertos linajes, como EAEC, requieren un marco analítico más amplio y marcadores específicos para ser identificados correctamente.

Conclusiones

La metodología de genómica computacional empleada en este estudio, centrada en el análisis de presencia/ausencia de factores de virulencia, cumplió con los objetivos planteados al permitir la discriminación sistemática de la mayoría de los principales patotipos de *Escherichia coli*. Para empezar nos dio la respuesta a nuestra pregunta de investigación, entonces: **¿Es posible diferenciar computacionalmente las cepas patógenas de *E. coli* de las comensales basándose únicamente en la presencia o ausencia de marcadores genéticos de virulencia específicos?** Sí, la minería computacional de genes de virulencia es una herramienta eficaz para distinguir varios de los principales patotipos de *E. coli* (EHEC, EPEC, UPEC) de las cepas comensales.

4.1. Respuesta a objetivos:

El **Objetivo general**, consistente en analizar comparativamente el contenido genómico para identificar patrones discriminantes, se logró con éxito al aplicar el análisis de agrupamiento jerárquico sobre la matriz binaria de virulencia. Esta técnica logró **reconstruir con alta fidelidad la clasificación biológica esperada para las cepas patógenas entéricas y extraintestinales**.

En respuesta al **primer objetivo específico**, se estableció un perfil de referencia *in silico* de genes canónicos (adhesinas y toxinas) al confrontar los genomas contra la base de datos VFDB. Esto permitió abordar el **segundo objetivo específico**, demostrando que las cepas patógenas, como EHEC, exhiben una carga de virulencia significativamente mayor (165 genes) que las cepas de referencia comensales (49 genes). Las diferencias clave que permiten esta clasificación incluyen la presencia del Locus de Borramiento de Enterocitos (LEE) compartido por EPEC y EHEC, y la toxina Shiga (stx), exclusiva de EHEC. El **tercer objetivo ospecífico**, que se buscaba era evaluar el agrupamiento y esto se vio con la formación de clústeres definidos: uno LEE-positivo (EHEC/EPEC) y otro uropatógeno (UPEC), ambos segregados claramente del grupo basal comensal (K-12/SE11).

4.2. Hallazgo más importante y relevancia:

El hallazgo más importante es que *el perfil de presencia y ausencia de genes de virulencia actúa como una firma genómica predictiva eficaz para la mayoría de los linajes de E. coli.* El método demostró ser robusto para clasificar patotipos definidos por grandes islas de patogenicidad (como LEE en EHEC/EPEC) y para diferenciar cepas extraintestinales

(UPEC) basadas en sus repertorios genéticos. No obstante, también reveló una limitación crucial: la **falla en clasificar correctamente al patotipo enteroagregativo (EAEC 042)**, lo que subraya que la complejidad de algunos patotipos requiere el uso de un marco analítico más amplio y marcadores altamente específicos que no siempre están representados en bases de datos generales o son detectables bajo umbrales estrictos. En conjunto, el análisis computacional de la carga de virulencia es una estrategia fundamental para descifrar la base molecular de la patogenicidad de E. coli en un contexto de un pangenoma altamente dinámico y heterogéneo.

Bibliografía

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., & Jin, Q. (2005). VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Research*, *33*(suppl_1), D325-D328. https://doi.org/10.1093/nar/gki008

Gomes, T. A., Elias, W. P., Scaletsky, I. C., Guth, B. E., Rodrigues, J. F., Piazza, R. M., Ferreira, L. C., & Martinez, M. B. (2016). Diarrheagenic Escherichia coli. *Brazilian Journal of Microbiology*, *47*, 3-30. https://doi.org/10.1016/j.bjm.2016.10.015

Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., Ohtsubo, E., Baba, T., Wanner, B. L., Mori, H., et al. (2006). Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Molecular Systems Biology*, *2*(1), 2006-0007. https://doi.org/10.1038/msb4100049

Kaper, J. B., Nataro, J. P., & Mobley, H. L. (2004). Pathogenic Escherichia coli. *Nature Reviews Microbiology*, *2*(2), 123-140. https://doi.org/10.1038/nrmicro818

Rasko, D. A., Rosovitz, M. J., Myers, G. S., Mongodin, E. F., Fricke, W. F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N. R., Chaudhuri, R., Henderson, I. R., Sperandio, V., & Ravel, J. (2008). The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *Journal of Bacteriology*, *190*(20), 6881-6893. https://doi.org/10.1128/JB.00619-08