UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

GENÓMICA DE POBLACIONES ASOCIADA A LOS NICHOS ECOLÓGICOS DE *Escherichia coli*

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

B I Ó L O G A

P R E S E N T A:

LUNA LUISA SÁNCHEZ REYES

DIRECTORA DE TESIS:
Biól. ANDREA GONZÁLEZ GONZÁLEZ

2010

Hoja de Datos

1. Datos del Alumno Sánchez Reyes Luna Luisa 53967758 Universidad Nacional Autónoma de México Facultad de Ciencias Biología 404055121
2. Datos del tutor Biól. Andrea González González
3. Datos del sinodal 1 Dr. José Luis Puente García
4. Datos del sinodal 2 Dr. Luis Enrique Eguiarte Fruns
5. Datos del sinodal 3 Dra. Valeria Francisca Eugenia Leopoldina de María de Guadalupe Souza Saldívar
6. Datos del sinodal 4 Dr. Arturo Carlos II Becerra Bracho
7. Datos del trabajo escrito Genómica de Poblaciones asociada a los nichos ecológicos de *Escherichia coli* 116 p 2010

A mis padres, Georgina y Héctor,

por quererme tanto y enseñarme lo que de verdad importa: a ser libre.

A mi hermana Mari,

por tu admiración y cariño increíbles, yo también te admiro y te adoro.

A mis abuelos queridos,

Chata y Mario y Juanita y Antonio,

por todos sus recuerdos de inspiración y consejo.

A P, por ser mi compañero y amigo incansable.

Nos reencontraremos en el jardín de mis sueños.

A Víctor, por compartir conmigo tu maravillosa vida.

A Marie, siempre juntas…

Esta tesis se llevó a cabo en el laboratorio de Evolución Molecular y Experimental del Instituto de Ecología, UNAM.

Gracias al apoyo otorgado por el proyecto DGAPA **IN219109** : "Evaluación de marcadores genéticos para un microarreglo diagnóstico de enfermedades diarréicas en el Pacífico mexicano utilizando metagenómica".

Así como de la beca de ayudante SNI del Dr. Luis E. Eguiarte Fruns dentro del proyecto**:** "Transferencia horizontal en bacterias entéricas diarréicas y la evolución de la patogénesis".

Un agradecimiento especial a la Dra. Rosario Morales, a la M. en C. Gabriela Delgado y al Biol. José Luis Méndez, del Laboratorio de Genómica Bacteriana, Departamento de Microbiología y Parasitología de la Facultad de Medicina, por la ayuda técnica y teórica proporcionada, que impulsó el desarrollo de este trabajo.

Un agradecimiento también al Biól. Tobías Portillo, por el apoyo brindado en el uso del cluster computacional *KanBalan*, DGSCA, UNAM, México, D.F. y en los aspectos bioinformáticos de este trabajo.

Agradezco a los sinodales: Dr. José Luis Puente García, Dr. Luis Enrique Eguiarte Fruns, Dra. Valeria Souza Saldívar, Dr. Arturo Carlos II Becerra Bracho y a mi tutora, la Biól. Andrea González González, quienes con su revisión y sus valiosos comentarios hicieron de esta tesis un trabajo mejor.

AGRADECIMIENTOS

Agradezco a la Biól. Andrea González González, por ser mi guía infatigable, mi tutora, mi compañera y mi amiga. Y principalmente, por creer siempre en mi.

Al Dr. Luis E. Eguiarte Fruns y a la Dra. Valeria Souza Saldívar por la aceptación, la inspiración y las enseñanzas que me han brindado durante estos años.

A la Dra. Rosario Morales y a la M. en C. Gabriela Delgado, por sus consejos, con los que me enseñaron a ver el trabajo de laboratorio con otros ojos. Y a todas las personas del Laboratorio de Medicina, que con su ayuda y compañía, alegraron mi estancia allá.

A la Dra. Susana Magallón, por las valiosas enseñanzas brindadas durante el tiempo que estuve bajo su tutela.

Al Dr. José Guadalupe Pérez Ramírez (Bokhimi) quién me mostró una perspectiva más amplia y más humana de la ciencia.

A los compañeros del LEMyE por su determinación y gran voluntad, con las que me impulsaron a seguir adelante siempre.

A mi mamá, Georgina Reyes López, la mejor maestra que he tenido. Gracias por tú infinito cariño y por tú gran generosidad.

A mi padre, Héctor Sánchez Salgado, quien me enseñó a percibir la vida creativamente.

A mi querida hermana Mari, por ser la compañera del extraordinario universo que hemos creado juntas.

A la familia de mi madre, por apoyarme siempre y simplemente por estar ahí y ser como son.

A la familia de mi padre, por sus maneras tan diversas de ver la vida gracias a lo cual se ha enriquecido la mía.

A la familia Serrano Velázquez, por su gran espíritu. En especial, a la Sra. Judith Velázquez Rojas, por recibirme en su hogar y por aconsejarme, al compartirme siempre sus vivencias.

A todos mis amigos, compañeros de la vida: no hay suficientes páginas para expresar mi cariño, saben que los adoro. Gracias a todos.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El pangenoma de una especie bacteriana está conformado por el genoma central (genes presentes en todos los individuos) y el genoma flexible (genes compartidos por subpoblaciones o presentes en una sola cepa). Debido a que este último se compone principalmente por genes adquiridos horizontalmente que codifican para nuevas funciones metabólicas, surgió la idea de que es el genoma flexible lo que permite la adaptación a nichos nuevos, despreciándose el papel que desempeña el genoma central en el origen y en la evolución de adaptaciones tales como la patogénesis y el comensalismo.

Para examinar este paradigma, realizamos un estudio de genómica de poblaciones en *Escherichia coli*, bacteria con diversos estilos de vida. En primer lugar, se definieron las regiones de análisis ó "loci", no considerando los marcos de lectura de genes, sino sobre el alineamiento del cromosoma de los 12 individuos de la muestra, en ventanas de 1,000 sitios de longitud, las cuales resultaron más informativas que ventanas más grandes (de 10,000 ó 100,000 sitios) ó más pequeñas (de 3 sitios). Al comparar la dinámica evolutiva del genoma central y del genoma flexible, entre los dos "ecogrupos" en los que se dividió la muestra, se encontró que las cepas patógenas (de ave y extraintestinales e intestinales de humano) presentan mayor diversidad genética que las cepas no-patógenas (vida libre y comensales de humano), no solo a nivel del genoma flexible, sino también al del genoma central. Asimismo, pocos loci se encontraron en desequilibrio de ligamiento, lo que indica que *E. coli* es una especie sexual, siendo las cepas no-patógenas ligeramente menos recombinantes que las cepas patógenas. Referente a los patrones de selección natural, las cepas patógenas mostraron señales de selección positiva en genes con funciones variadas (genes de transporte/unión, metabolismo energético y de transcripción), a diferencia de las cepas no-patógenas, en las cuales predominó la selección negativa. Finalmente, un análisis considerando los bloques de sintenia (LCBs) del genoma central reveló que los patrones de diversidad, de clonalidad y de selección fueron heterogéneos a lo largo del cromosoma y sólo algunos LCB's presentaron patrones semejantes en ambos ecogrupos.

Así, concluimos que la adaptación de *E. coli* a diferentes nichos no ocurre solamente por la adquisición horizontal de genes sino que la evolución del genoma central, así como la regulación de la expresión génica de ésta parte del genoma, juegan un papel importante en este proceso.

ABSTRACT

The pan-genome of a bacterial species consists of a core genome (genes shared by all isolates) and a flexible genome (genes shared by subpopulations and strain-specific genes). The latter is composed mainly of horizontally acquired genes encoding new metabolic functions. This promoted the idea that the flexible genome is what allows bacterial adaptation to new niches, thus ignoring the role of the core genome in the origin and evolution of adaptations, such as pathogenesis and commensalism.

To examine this paradigm, we conducted a population genomics study in *Escherichia coli*, a bacterial species with different lifestyles. First of all, we defined the regions of analysis or "loci", not according to reading frames of genes, but according to the alignment of the chromosome of the 12 individuals that constitute the sample, in windows of 1 000 sites length, which were more informative than larger windows (of 100 000 or 10 000 sites) and than smaller windows (3 sites). When comparing the evolutionary dynamics of the core and flexible genome between the two "eco-groups" in which the sample was divided, we found that pathogenic strains (isolated from poultry and from extra-intestinal and intestinal samples in humans) had more genetic diversity than non-pathogenic strains (free-living and commensal in humans), not only within the flexible genome, but also within the central genome. Moreover, few loci were found in linkage disequilibrium, indicating that *E. coli* is a sexual species, in which the non-pathogenic strains are slightly less recombinant than pathogenic strains. Concerning the patterns of natural selection, the pathogenic strains showed signs of positive selection in genes with diverse functions (genes of transport / binding, energy metabolism and transcription), unlike non-pathogenic strains, in which negative selection predominated. Finally, an analysis of diversity among the synteny blocks (LCBs) of the core genome revealed that the patterns of genetic diversity, clonality and selection were heterogeneous along the chromosome. Only a few LCB's showed similar patterns in both ecogroups.

Thus, we conclude that adaptation of *E. coli* to different niches occurs not only by the horizontal acquisition of genes, but also the evolution of the core genome and the regulation of gene expression in this part of the genome, play an important role in this process.

INTRODUCCIÓN

El resultado de la evolución se puede observar en la gran diversidad que existe en los seres vivos. En los numerosos trabajos que realizó en su época, Charles Darwin (1809-1882) explica que la diversidad biológica es el resultado de la adaptación de los seres vivos a la heterogeneidad ambiental. Y propone que la fuerza evolutiva que genera la adaptación es la selección natural.

Con el redescubrimiento de las leyes de la herencia de Mendel y el posterior desarrollo de la Teoría Sintética de la Evolución a principios del siglo XX, se comenzó a estudiar el efecto de otras fuerzas evolutivas en el proceso adaptativo, como son la mutación, la deriva génica y la migración ó flujo génico (Hedrick 2005).

Pero no es hasta principios de los años 50s, con el descubrimiento de la estructura del ADN y el desciframiento del código genético, que se vuelve posible estudiar de manera más directa las bases genéticas de la diversidad, con lo que surgen paradigmas sobre la evolución a nivel molecular y el papel que desempeñan las fuerzas evolutivas a este nivel (Kimura 1969a; Ohta 1976). A pesar de que estos modelos han sido sujeto de múltiples debates y controversias (Crow 2008), de manera general se ha mantenido la visión de que la mayor parte de la diversidad genómica en poblaciones naturales debe ser neutral o casi neutra, es decir generada por un equilibrio entre la mutación y la deriva génica, con un efecto reducido de la selección natural y del flujo génico (Kimura 1983). Recientemente, con el auge de la secuenciación genómica, se ha podido observar directamente la diversidad genética al nivel del genoma, y algunos autores han reportado que un porcentaje importante de las regiones codificantes del genoma, muestra señales de selección natural (10-40% de genes bajo selección en diferentes especies estudiadas; Ellegren et al. 2008).

En este caso, la información de la variación genética a la escala del genoma podría ser útil en el estudio del proceso de adaptación. En particular, pueden ser útiles para estudiar los mecanismos que han dado origen al amplio espectro de diversidad ecológica que se observa en algunas especies (Mira et al. 2002) y especialmente dentro de las especies bacterianas (Cohan 2006), en cuyo caso el análisis de datos genómicos con las herramientas de la genética de poblaciones (genómica de poblaciones), es una aproximación que podrá ayudar al entendimiento de la relación entre la evolución molecular del genoma, la adaptación y la evolución fenotípica al nivel del organismo (Ellegren et al. 2008).

1. Manifestaciones de la variación genética: de los genes a los genomas

La genética de poblaciones, desde su nacimiento con los trabajos de Hardy, Weinberg, Wright, Fisher y Haldane (Eguiarte 1999), ha desarrollado la metodología que permite analizar la distribución de la variación genética y los mecanismos que la mantienen o la cambian en las poblaciones (Kimura 1983; Hedrick 2005) a lo largo de las generaciones.

Los primeros trabajos que realmente se aproximan a describir de una manera más fina la variación a nivel genético en poblaciones son los realizados por Lewontin y Hubby (1966) y Harris (1966), quienes de manera independiente analizan la variación de un número de loci enzimáticos en muestras de *Drosophila pseudoobscura* y de *Homo sapiens* respectivamente. La técnica que utilizan, la electroforesis de proteínas ó MLEE (Multilocus Enzyme Electrophoresis), les permite examinar numerosos caracteres independientes de la función del gen e independientes entre sí. Además cada uno corresponde a un locus diferente y se consideran como un estimado casi directo de la variación genética a nivel genómico (Hedrick 2005). Es con este tipo de marcadores, que finalmente se pueden llevar acabo estudios que abren la discusión hacia la naturaleza evolutiva de las poblaciones bacterianas, siendo *Escherichia coli* uno de los organismos modelo (Milkman 1973).

Sin embargo, estos marcadores moleculares aún corresponden al fenotipo. De tal manera que la variación genética subyacente puede no estar completamente reflejada en los patrones de variación obtenidos con isoenzimas debido a la presencia de sustituciones de aminoácido crípticas (las cuales cambian el aminoácido pero no afectan las cargas de la proteína y por lo tanto el patrón de electroforesis; Ayala 1982) o sustituciones silenciosas (Kreitman 1983).

Ya con el desarrollo de los métodos de secuenciación, primero de proteínas (Lewontin, 1974) y posteriormente de ADN (Kreitman 1983) es que se logra describir la variación genética en las poblaciones de manera directa. Más aún, con el desarrollo de nuevas metodología como la PCR (reacción en cadena de la polimerasa) y la secuenciación automatizada han permitido el uso cada vez más extendido de secuencias de ADN para realizar estudios de genética de poblaciones (Hedrick 2005). Para el caso de bacterias, Maiden et al. (1998) sugieren utilizar al menos 7 genes que se distribuyan a lo largo del genoma y cuyo comportamiento sea neutral, para obtener un estimado no sesgado de la

variación a nivel genómico (Urwin et al. 2003). Esta aproximación, denominada MLST (Multi Locus Sequence Typing), ha sido aplicada con éxito a una gran cantidad de especies microbianas, tanto procariontes como eucariontes (Pérez-Losada et al. 2006). Sin embargo, debido a que se usan genes de mantenimiento, que poseen un alto grado de conservación (identidad nucleotídica) a nivel de secuencia, el MLST algunas veces carece de la suficiente resolución para discriminar de manera efectiva entre los individuos (Nallapareddy et al. 2002; Fakhr et al. 2005).

Con el advenimiento de la secuenciación de genomas completos y de los análisis de genómica comparada se ha observado que la extrapolación de los datos de pocas secuencias, como los obtenidos a partir del MLST, a la escala de todo el genoma es limitada (Allen et al 2007). Por esta razón, en los últimos años cada vez más estudios se dirigen a utilizar un mayor número de loci. Al estudiar una región amplia del genoma es posible conocer mejor los procesos que en éste ocurren (Black et al. 2001), y permite conocer la dinámica de regiones particulares del genoma, como islas genómicas (Anantha et al. 2004) o islas de patogénesis (Castillo et al. 2005) en el caso de bacterias y de cromosomas particulares en el caso de eucariontes, como el género de levadura *Saccharomyces* (Bensasson et al. 2008) y la mosca de la fruta *Drosophila* (Hutter et al. 2007). La aproximación genómica también permite la obtención de una gran cantidad de información evolutiva a partir de una muestra pequeña, pues incluso a partir de los genomas de tan sólo dos individuos, se pueden obtener datos interesantes sobre la dinámica de los genes y otras regiones del genoma, así como de su estructura (Hayashi et al. 2001; Ellegren 2008).

Finalmente, la información de secuencia se ha utilizado, no solamente para entender la dinámica poblacional de una especie, sino para entender procesos adaptativos (Kreitman 1983; Sokurenko et al. 1998) y para el descubrimiento de grupos funcionales ecológicos ó ecotipos (Palys et al. 1997). De la misma manera, los datos de secuencia a nivel genómico pueden ser utilizados para relacionar las características genómicas con la adaptación ecológica (Mira et al. 2002).

2. Genómica y evolución bacteriana

2.1 Genómica comparada de bacterias

En los años 1990, la reducción de costos y el aumento en la eficiencia del proceso de secuenciación trae consigo la posibilidad de obtener la información genética completa del genoma de un organismo. El primer genoma de un procarionte en ser secuenciado fue el de la bacteria *Haemophilus influenzae* (Fleischmann et al. 1995). Con un segundo genoma procarionte liberado, el de *Mycoplasma genitalium* (Fraser 1995), Mushegian y Koonin (1996) llevan a cabo un estudio comparativo para determinar el conjunto de genes que se comparten en estos dos genomas, proponiendo que dicha muestra de genes debió de estar presente en el último ancestro común. Así, este es el primer trabajo que usa la información genómica para tratar de responder una pregunta de interés evolutivo.

En los últimos años, los trabajos de genómica comparada han proliferado (Binnewies et al. 2006). La comparación de genomas entre diferentes especies y dentro de individuos de la misma especie, ha permitido, entre otras cosas, la identificación de algunos mecanismos que dan origen a la variabilidad, modificando el orden y la composición de los elementos genéticos (Mira et al. 2002; Abby y Daubin 2007). En particular, ha permitido la identificación de elementos móviles como islas genómicas, secuencias de inserción y genes sujetos a transferencia horizontal (Ochman et al. 2000), con lo que se confirma (Lawrence y Ochman, 1998; Lan y Reeves, 2000; García-Vallve et al. 2000; Hayashi et al. 2001; Gogarten et al. 2002; Koonin y Wolf 2008) que la recombinación ilegítima y la transferencia horizontal de genes son eventos generalizado en especies procariontes (Maynard-Smith, 1993; Doolittle, 1999a; 199b).

A partir de este tipo de estudios se sabe que los genomas bacterianos en general poseen un gran dinamismo, y se puede considerar que hay dos grandes tipos de diversidad en sus genomas. Aquella que se ve reflejada en diferencias en la estructura genómica, y que es generada por rearreglos, inversiones (Eisen et al. 2000) ó translocaciones (Tillier et al. 2000; Darling et al. 2004). Y aquella que se ve reflejada en diferencias en el repertorio genético entre los individuos, generada por mecanismos que promueven la perdida de genes como la escisión génica ó la recombinación desigual (Anderson y Roth 1981; Bergthorsson et al. 2007), mecanismos que promueven la ganancia de genes nuevos y de genes xenólogos

(homólogos a un gen que ya está presente en la cepa ó linaje que lo recibe) como la recombinación no-homóloga ó ilegítima y la transferencia horizontal (Ochman et al. 2000; Spratt et al. 2001; Fraser 2009), y mecanismos de duplicación génica, con lo que se adquieren genes parálogos (homólogos a otro gen dentro de un mismo individuo o linaje; Jordan et al. 2001).

A partir de la observación de una enorme variación en el repertorio génico, inclusive entre individuos de la misma especie, se desarrolla el concepto de pangenoma (Tettelin et al. 2005), el cual corresponde al total de elementos genéticos que componen el genoma de una especie. El pangenoma se ha definido en función de la presencia/ausencia de genes entre aislados de una misma especie, y se considera que consta de 2 partes: el genoma central, constituido por los genes presentes en todos los individuos de la especie (o al menos de la muestra analizada) y el genoma flexible (también llamado dispensable), el cual es más extenso si se compara con el primero al constar de todos los demás genes, los cuales pueden estar presentes en un sólo individuo ó en submuestras ó subpoblaciones de la especie.

2.2 Evolución del genoma y adaptación en bacterias

Mucho se ha especulado sobre la forma en que estos procesos dan pie a la generación y mantenimiento de la diversidad ecológica de una especie (Mira et al. 2002). En procariontes, los descubrimientos de la genómica comparada, aunados al paradigma clonal de diversidad bacteriana y a las ideas de que los factores genéticos que aumentan la adecuación en ambientes particulares y que confieren nuevas capacidades metabólicas al individuo se encuentran en islas genómicas (Dobrindt et al. 2004), adquiridas por transferencia horizontal (Lawrence 2001) han fomentado el estudio de este fenómeno no sólo como una parte importante en el proceso de evolución del genoma, sino también como un mecanismo de adaptación bacteriana (Stoebel 2005). De tal manera que la transferencia horizontal se ha llegado a considerar como el principal mecanismo causante de dicho proceso evolutivo (Groisman y Ochman, 1996; Bergthorsson y Ochman 1998; Ochman y Moran 2001; Hacker y Kaper, 2000, Lawrence y Roth 1996; Lawrence y Ochman 1998; Ochman et al. 2000; Lawrence 2001).

Inclusive, se ha llegado a atribuir a un único evento de transferencia horizontal (la

obtención del operón *lac* por *E. coli*), el desarrollo de las propiedades metabólicas necesarias para la expansión de una especie a un nuevo nicho, lo que finalmente daría lugar a un evento de especiación/divergencia (entre *E. coli* y Salmonella; Ochman et al. 2000). Por otra parte, la idea de que la plasticidad ecológica y la adaptación en las bacterias se deben al efecto de la transferencia horizontal de genes ha tomado especial fuerza en el ámbito médico, debido a la existencia de bacterias que son generalmente comensales del ser humano, pero que también pueden llegar a ser patógenas, como es el caso de *Escherichia coli* (Nataro y Kaper 1998). Se ha sugerido en numerosas ocasiones que una bacteria comensal puede convertirse en patógena por el mero hecho de adquirir algún factor de virulencia (Kaper et al. 2004; Weintraub 2007; Wiles et al. 2008). Aunque sí se ha logrado experimentalmente insertar un factor de virulencia a una cepa comensal, y al inocular un individuo hospedero, éste desarrolla los síntomas de enfermedad (Wooley et al. 1998; Skyberg et al. 2006), no se sabe con certeza cuanto tiempo puede permanecer esta asociación "artificial" en la naturaleza. Al menos en el caso de virus, se sabe que estas asociaciones no-naturales no son estables temporalmente (Ebert y Bull 2003).

Estudios más profundos sobre la distribución de los factores de virulencia en poblaciones de bacterias patógenas han encontrado que prácticamente ningún factor de virulencia se encuentra presente en muestras de cepas de un mismo patotipo (Kaper et al. 2004; Johnson et al. 2008). Además la comparación de genomas de diferentes patotipos ha confirmado que cepas comensales comparten gran parte de los factores de virulencia (Fricke 2008; Rasko et al. 2008).

La presencia de características que antes se consideraban exclusivamente como marcadores de patogénesis en cepas no-patógenas se ha tratado de explicar de diferentes maneras. Rasko et al. (2008) sugieren que los elementos que antes se consideraban como diagnósticos de la patogénesis en realidad no son estrictamente patogénicos y que pudieran ser utilizados por cepas comensales para procesos de colonización.

Entonces la adquisición de ciertos factores de virulencia no implican forzosamente la transformación inmediata de una cepa avirulenta en patógena y viceversa. Lo que quiere decir que deben de existir otros factores implicados en la evolución de la patogénesis bacteriana (Johnson y Russo 2002).

Entonces ¿las diferencias ecológicas se podrán ver reflejadas en los elementos del genoma

que no están sujetos a transferencia horizontal? ó ¿sólo están determinadas por los elementos que se adquieren mediante recombinación ilegítima?

Las herramientas y conceptos que pudieran ayudarnos a contestar estas preguntas se han desarrollado de manera paralela a la genómica comparada, y constituyen a la "genómica de poblaciones".

2.3. Genómica de poblaciones y el estudio de la adaptación en *E. coli*

Desde hace algunos años, se ha propuesto que para describir las relaciones evolutivas en los seres vivos es más efectivo e informativo analizar más loci y no más individuos (Lewontin 1995). Respondiendo a esta propuesta, la genómica de poblaciones utiliza un número extenso de loci o de regiones genómicos, y preferentemente los genomas completos, para lograr una mejor comprensión de la distribución de la variación genética y el papel relativo de las diferentes fuerzas evolutivas, así como el papel potencial de éstas en la diferenciación ecológica (Black et al. 2001; Li et al. 2008; Nadeau y Jiggins 2010). Se ha propuesto que esta aproximación facilita la disociación de los efectos de las fuerzas evolutivas que actúan de manera homogénea en el genoma, como son la deriva génica y el flujo génico, de aquellos que actúan de manera diferencial, es decir que tienen efectos específicos a locus, como son la selección y la recombinación (Luikart et al. 2003; Li et al. 2008). Pero sobre todo la información proporcionada por la totalidad del genoma bacteriano ha probado ser importante para determinar adecuadamente la historia demográfica y las relaciones filogenéticas entre individuos (Touchon et al. 2009), estudiar procesos adaptativos concretos (Hutter et al. 2007) e inclusive ha permitido el estudio de organismos altamente clonales, como es el caso del agente causante de la úlcera de Buruli *Mycobacterium ulcerans*, donde los marcadores tradicionales no habían detectado la más mínima variación genética (Qi et al. 2009).

Por esta razón la genómica de poblaciones puede ayudarnos en el mejor entendimiento del proceso de adaptación ecológico en las bacterias, al nivel de las regiones del genoma central y no en función del genoma flexible. Sobre todo en aquellas que poseen una estructura poblacional epidémica (Maynard-Smith et al. 1993) y una ecología compleja, como es el caso de *Escherichia coli*, nuestro modelo de estudio.

Como bacteria modelo, el genoma de *E. coli* fue uno de los primeros en ser secuenciado (Blattner et al. 1997) y ha sido ampliamente estudiado a nivel genómico. Pero sólo algunos de trabajos han tratado de correlacionarlo con los estilos de vida de esta bacteria y con el proceso de la adaptación (Chattopadhay et al. 2009; Touchon et al. 2009).

Touchon et al. (2009) analizan una muestra de 20 genomas de *E. coli* y encuentran que a pesar de la estrategia ecológica, de la historia evolutiva de las cepas y de la gran cantidad de flujo génico detectado, la estructura del cromosoma se conserva entre los linajes. Lo que parece sugerir que hay una alta estabilidad de las regiones del genoma central al seleccionarse la organización del genoma con respecto a los procesos celulares como la replicación. Sin embargo, ellos determinan dicha estabilidad en función de la conservación de la posición de regiones donde hay ganancia o perdida de genes, es decir en función del genoma flexible. Por lo que en realidad no están hablando de la dinámica evolutiva dentro de las regiones del genoma central, sino de la dinámica evolutiva de la estructura general del cromosoma.

Por su parte, Chattopadhay et al. (2009) sí estudian directamente la diversidad genética a nivel secuencia en el genoma central de una muestra de 14 cepas de *E. coli* y *Shigella,* sin embargo ellos sólo analizan al nivel de los genes que se comparten entre los individuos de la muestra, dejando de lado el papel potencial de otras regiones genómicas en la evolución.

Así, aún cuando hay numerosos autores que han estudiado a *E. coli* a nivel genómico, estos se enfocan a la evolución de genes, dejando de lado el papel de las regiones intergénicas, en donde se puede encontrar información para la regulación de la transcripción, y otras regiones que no codifican para proteínas, como los ARNs pequeños, que actúan también como reguladores (Gottesman 2004) y que podrían ser de importancia para la evolución de las adaptaciones (Rogozin et al. 2002; Hughes y Friedman 2004), no solo de las regiones del genoma flexible (Pal et al. 2005), sino también del genoma central.

Por lo tanto, a diferencia de otros estudios realizados en *E. coli*, el presente es el primero que tratará de estudiar la dinámica al nivel de las regiones del genoma central, en regiones tanto codificantes como no-codificantes, para determinar la existencia de diferencias en los patrones evolutivos a este nivel del pangenoma así como evidencias de adaptación, con respecto a las estrategias ecológicas o estilos de vida de la bacteria modelo *Escherichia coli*.

OBJETIVO

Analizar la dinámica evolutiva y procesos adaptativos de los componentes del pan-genoma de una especie bacteriana con alta diversidad ecológica: *Escherichia coli*.

OBJETIVOS PARTICULARES

1. Identificar los elementos del genoma central y del genoma flexible presentes en el genoma de la bacteria modelo *Escherichia coli*.
2. Definir una unidad genómica de análisis, que será considerada como locus.
3. Obtener los niveles y patrones de variación genética en *E. coli*, y entre individuos con estrategias ecológicas diferentes.
4. Determinar el efecto de la selección natural y la recombinación en los patrones de variación encontrados.
5. Comparar los valores de diversidad y los patrones evolutivos en el genoma central y en el genoma flexible, entre individuos con diferentes estrategias ecológicas ó estilos de vida de *E. coli*.

HIPÓTESIS NULA

Si el proceso de adaptación de una población bacteriana está determinado solamente por los elementos del genoma flexible, entonces la dinámica evolutiva del genoma central debe ser igual en poblaciones de bacterias con estrategias ecológicas diferentes.

METODOLOGÍA

1. MODELO DE ESTUDIO: LA BACTERIA *ESCHERICHIA COLI*

Escherichia coli fue una de las primeras bacterias en ser aisladas. Fue descrita por primera vez en 1885 por el médico Theodore Escherich, quien la obtuvo de las heces de recién nacidos lactantes.

Durante un tiempo no fue de mayor consecuencia en el ámbito de la biología. Pero debido a los sencillos requerimientos de crecimiento que posee (37º C y una fuente de carbono simple) y a su corto tiempo generacional (cada 20 minutos en fase exponencial), poco a poco fue ganando terreno hasta convertirse en el organismo procarionte modelo por excelencia en biología, y desde mediados del siglo pasado es el caballo de batalla en experimentos de genética, fisiología, biotecnología, y más recientemente en el campo de la evolución, la genética de poblaciones y la genómica (Ho Yoon et al. 2009).

E. coli es una bacteria Gram negativa perteneciente a la clase de las Proteobacterias, de la subclase de las Gamma-proteobacterias, familia Enterobacteriaceae. Esta familia se caracteriza por tener organismos con capacidad de respiración facultativa y porque la mayoría de las especies tienen la capacidad de vivir en asociación con algún hospedero y en el ambiente externo (Logan 1994).

Efectivamente, aunque *E. coli* se aísla típicamente de heces de animales de sangre caliente (Rosebury 1962) como son mamíferos y aves, e inclusive de algunos animales de sangre fría como reptiles (Selander y Levin 1980; Souza et al. 1999) también se puede encontrar comúnmente en ambientes acuáticos y terrestres (Carrillo et al. 1985; Ksoll et al. 2007). Aunque el paradigma nos dice que el ambiente externo es secundario y transitorio para esta especie bacteriana (Savageau 1983), estudios recientes han registrado que las cepas aisladas del ambiente externo, ya sea en regiones contaminadas ó pristinas, pueden persistir en ese ambiente e inclusive crecer (Anderson et al. 2005).

Las cepas asociadas a hospedero son las que se han estudiado con mayor profundidad y se conocen mejor, debido a que en algunos casos son patógenas para el humano. Sin embargo la mayoría de las cepas de *E. coli* que se aíslan de heces fecales, no generan ningún tipo de enfermedad ó merma en la adecuación del hospedero, y puesto que hasta recientemente el

rol que esta bacteria cumple en el colon era poco entendido (Kaper et al. 2004), tradicionalmente son consideradas como simbiontes comensales. Nuevos estudios han demostrado que es pieza clave en el proceso de digestión del hospedero, al proveerles de nutrientes, de rutas de señalización clave para el desarrollo y regulación de reacciones inmunes que protegen al hospedero en contra de diversos patógenos intestinales (Yan y Polk 2004; Schouleur et al. 2009), por lo que podría considerarse de manera más estricta como un simbionte mutualista. Adicionalmente, en el sistema urinario también se pueden encontrar cepas comensales, a las que se les conoce como *E. coli* causantes de bacteriuria asíntomática ó ABU (Asymptomatic Bacteriuria; Hansson et al. 1989; Dobrindt y Hacker 2008). Ya que es capaz de competir con las cepas que generan infecciones en las vías urinarias de manera efectiva desplazándolas, del sistema urinario (Roos et al. 2006), estas cepas podrían también considerarse como mutualistas que protegen al hospedero en contra de patógenos, a cambio de nutrientes y alojamiento. Inclusive, las *E. coli* ABU se han comenzado a usar como tratamiento en contra de infecciones urinarias recurrentes (Wiles et al. 2008).

Por otra parte, se ha descrito que las *E. coli* patógenas pueden generar enfermedad en dos diferentes regiones generales del cuerpo del hospedero: en el sistema gastrointestinal y en la región extra-intestinal, que incluye es el sistema urinario, las meninges, el peritoneo, los pulmones y la región intra-abdominal (Kaper et al. 2004).

Desde el punto de vista médico, los grupos de cepas de *E. coli* que comparten un proceso de patogénesis similar, es decir, que generan un cuadro clínico característico y que comparten un conjunto de características fenotípicas y de factores genéticos (de virulencia), son llamados patotipos (Nataro y Kaper 1998). Las *E. coli* patógenas intestinales son las que se han descrito de manera exhaustiva y en la actualidad se han identificado principalmente 6 patotipos (Eslava et al. 1994; Kaper et al. 2004), los cuales se describen en la Tabla 1.

Las *E. coli* patógenas extra-intestinales ó exPEC se consideran como un patotipo generalista (Russo y Johnson 2000). Aunque originalmente se dividía en varios patotipos, se ha visto que comparten no sólo factores de virulencia (Johnson y Russo 2002; Dobrindt y Hacker 2008), sino también islas genómicas, serogrupos y relaciones filogenéticas (Dziva y Stevens 2008).

Tabla 1. Principales patotipos de *Escherichia coli* y factores de virulencia asociados.

Medio		Patotipo	Factores de virulencia diagnósticos	Referencia
Intestinal	Intestino delgado	EPEC *E.coli* enteropatogénica	& Plásmido del factor de adherencia y esfacelamiento (EAF). & Isla del locus de esfacelamiento de enterocitos (LEE).	Eslava et al. 1994; Nataro y Kaper 1998.
	Intestino delgado	ETEC *E.coli* enterotoxigénica	& Toxinas termoestable (ST) y termolábil (LT). & Antígeno del factor de colonización (CFA).	Eslava et al. 1994; Nataro y Kaper 1998.
	Colon	EHEC *E.coli* enterohemorrágica	& Toxina tipo shiga. & Isla del locus de esfacelamiento de enterocitos (LEE). & Plásmido O157.	Eslava et al. 1994; Nataro y Kaper 1998.
	Intestino delgado y colon	EAEC *E.coli* enteroagregativa	& Toxina termoestable (EASTI). & Adhesinas fimbriales (AAFI y AAFII).	Nataro y Kaper 1998.
		DAEC *E.coli* de adherencia difusa	& Adhesina fimbrial (F1845). & Proteína de membrana externa (AIDA1).	Eslava et al. 1994; Nataro y Kaper 1998.
	Colon	EIEC *E.coli* enteroinvasiva	& Plásmido *Inv.* & Proteínas de membrana *Vir.*	Nataro y Kaper 1998.
Extra-intestinal	Sistema urinario	UPEC *E.coli* uropatogénica	& Polisacárido capsular del grupo II *kpsMT.*	Dobrindt y Hacker 2008.
	Meninges, sangre	MNEC *E.coli* asociada a sepsis-meningitis	& Cápsula *K.*	Dobrindt y Hacker 2008.

Cada uno de estos patotipos despliega un patrón de adherencia y de toxicidad particular, sin embargo existe cierta superposición de los factores de virulencia entre los patotipos (Rasko et al. 2008).

Además, existen cepas atípicas, las cuales generan cuadros clínicos que se pueden asociar un patotipo en particular, pero carecen de uno ó más de los factores de virulencia diagnósticos de dicho patotipo.

Todas estas características hacen de *E. coli* un modelo interesante para tratar de entender los mecanismos a nivel genómico, que dan origen a los diferentes estilos de vida y que permiten la adaptación a distintos nichos ambientales.

2. MUESTRA Y MARCADOR

La muestra consta de 12 individuos de *Escherichia coli*, cuyas características se enlistan en la Tabla 2. Estas cepas son representativas de diferentes estilos de vida de la bacteria *E. coli*. Además, sus genomas se encuentran secuenciados con una cobertura ≥ 5x, y de acuerdo a Blattner et al. (1997) y Rasko et al. (2008), ese es el mínimo necesario para obtener un ensamblado correcto y una mayor seguridad sobre la información de secuencia que se analizará.

Las secuencias de ADN del genoma de estos individuos fueron obtenidos de la base de datos de recursos genomicos del Centro Nacional de Información Biotecnológica de E.U.A., *NCBI* (National Center of Biotechnology Information), a la cual se puede acceder a través del enlace http://www.ncbi.nlm.nih.gov.

Para alcanzar los objetivos del presente estudio, se analizó únicamente el cromosoma completo, cuyo tamaño se encuentra entre los < 4.5 Mb y los < 5.5 Mb (cf. Tabla 2). Los plásmidos fueron descartados del análisis.

Al estudiar el cromosoma completo de *E. coli*, se incluyeron tanto regiones codificantes como no-codificantes en el análisis. Con esto, se buscó abordar, si bien de manera incipiente, el papel de las regiones no-codificantes en la evolución genómica de las bacterias, sobre lo cual se ha especulado (Rogozin et al. 2002; Hughes y Friedman 2004), pero solamente en eucariontes se han hecho algunos estudios al respecto (Andolfatto et al. 2005; Hutter et al. 2007). Asimismo, el análisis del cromosoma completo permitió tener un enfoque no-genecéntrico de la evolución de *E. coli*, y por lo tanto más amplio (Thomas et al. 2005; Didelot et al. 2009), en oposición a la tendencia predominante en los estudios de evolución a nivel genómico en bacterias, los cuales, en su mayoría se limitan a analizar regiones que codifican para proteínas (Castillo et al. 2005; Charlesworth y Eyre-Walker 2006; Chattopadhay et al. 2009), con contadas excepciones (Barrick et al. 2009).

Tabla 2. Características de los genomas de *E. coli* de la muestra de estudio.

Hábitat	Características ecológicas	Cepa/ plásmido	Grupo filogenético (de acuerdo a Wirth et al. 2006)	Serotipo	Tamaño del replicón (Mb)	% codificante y número correspondiente de ORFs ó CDS de acuerdo al NCBI	Contenido GC	Número de plásmidos	Número de acceso NCBI	Referencia
Intestinal	Comensal Adaptada al laboratorio	K12 MG1655	A	N.E. O16	4.639675	85% genes 4493	50%	-	NC_000913	Blattner et al. 1997
		K12 W3110	A	N.E. O16	4.646332	86% genes 4444	50%	-	AC_000091	Hayashi et al. 2006
		K12 ATCC8739	A	N.E. O16	4.746218	86% genes 4409	50%	-	NC_010468	Joint Genome Institute 2008
	Comensal silvestre	HS	A	O9	4.643538	86% genes 4630	50%	-	NC_009800	Rasko et al. 2008
De vida libre	No-patógena	SMS-3-5 pSMS35_130 pSMS35_8 pSMS35_4 pSMS35_3	D	O19:H34	5.068389 0.130440 0.008909 0.004074 0.003565	87% genes 4943 79% genes 166 64% genes 11 84% genes 4 41% genes 4	50% 50% 46% 49% 43%	4	NC_010498	Fricke et al. 2008
Intestinal/ Extra-intestinal	Patógena APEC	APEC-O1 pAPEC-O1-R pAPEC-O1-ColBM	B2	O1:K1:H7	5.082025 0.241387 0.174241	86% genes 4542 82% genes 224 84% genes 199	50% 46% 49%	2	NC_008563	Johnson et al. 2007
Intestinal	Patógena ETEC	E24377A pETEC_80 pETEC_35 pETEC_73 pETEC_6 pETEC_74 pETEC_5	-	O139:H28	4.979619 0.079237 0.034367 0.070609 0.006199 0.074224 0.005033	85% genes 4981 53% genes 87 61% genes 29 63% genes 76 40% genes 5 66% genes 77 37% genes 6	50% 47% 51% 50% 52% 49% 49%	6	NC_009801	Rasko et al. 2008
	Patógena EHEC	EDL933 pO157	-	O157:H7	5.528445 0.092077	87% genes 5441 85% genes 101	50% 47%	1	NC_002655	Perna et al. 2001
	Patógena EHEC	Sakai pO157 pOSAK1	-	O157:H7	5.498450 0.092721 0.003306	85% genes 5372 77% genes 85 52% genes 3	50% 47% 43%	2	NC_002695	Hayashi et al. 2001
Extra-intestinal	Patógena, UPEC	UTI 536	B2	O6:K15:H31	4.938920	87% genes 4780	50%	-	NC_008253	Hochhut et al. 2006
	Patógena, UPEC	UTI CFT073	B2	O6:K2:H1	5.231428	87% genes 5549	50%	-	NC_004431	Welch et al. 2002
	Patógena UPEC	UTI89 pUTI89	B2	O18:K1:H7	5.065741 0.114230	88% genes 5131 79% genes 145	50% 51%	1	NC_007946	Chen et al. 2006

Estos genomas estaban completamente secuenciados y disponibles en el NCBI hasta el mes de mayo del 2008.
N.E. No expresado (Stevenson et al. 1994).

3. GENÓMICA COMPARADA

3.1. Alineación de genomas

Para estudiar e inferir las relaciones evolutivas entre los seres vivos de manera empírica, un primer paso es detectar la variación existente en el grupo de organismos que se está estudiando (Hedrick 2005). Para lograr esto, se usa una aproximación comparativa de rasgos ó caracteres, los cuales pueden ser desde morfológicos hasta moleculares (Elena y Lenski 2003). Puesto que la evolución estudia las relaciones de ancestría-descendencia entre los individuos (Futuyma 2005), solamente caracteres homólogos (que compartan un ancestro común reciente) deberán compararse para detectar la variación en los seres vivos (Dayhoff et al. 1983; Hall 1994).

Sin embargo, muchas veces la detección de caracteres homólogos puede ser complicada, debido a la existencia de procesos que pueden obscurecer las relaciones de ancestría-descendencia (como es la convergencia en los caracteres fenotípicos; Hall 1994). En el caso de los caracteres de secuencia, ya sea de aminoácidos ó de ADN, también hay mecanismos que pueden dificultar la determinación de las relaciones de homología entre los individuos. Estos mecanismos son las mutaciones a pequeña escala ó micromutaciones (mutación puntual por sustitución nucleotídica ó por inserciones ó deleciones *indels*) y las mutaciones a gran escala ó macromutaciones (inversiones, translocaciones, transposiciones, duplicación génica, transferencia horizontal; Ureta-Vidal et al. 2003).

En los estudios que usan como caracteres las secuencias, ya sea de proteínas ó de ADN, el alineamiento es lo que permite comparar y documentar la variación existente en los seres vivos (Waterman 1984; Mount 2004). Para tener un alineamiento "correcto", ó al menos el mejor alineamiento posible, se deben de comparar secuencias homólogas, tomando en cuenta los mecanismos antes mencionados, y los diferentes tipos de homología que pueden existir al nivel de secuencia (Dayhoff et al. 1983; Darling et al. 2004).

Cuando las secuencias por alinear son cortas, la probabilidad de que un evento de macromutación afecte la región es baja (Didelot et al. 2009), en cuyo caso los algoritmos de comparación pareada y múltiple clásicos, como los de Needleman y Wunsch (1970) y Waterman (1986), pueden generar un alineamiento correcto sin dificultad y con poco tiempo de cómputo.

Sin embargo, para alinear secuencias a nivel genómico, estos algoritmos son insuficientes en dos aspectos. En primer lugar, poseen un tiempo de cómputo que aumenta exponencialmente con el tamaño de la secuencia, por lo que la alineación de regiones cuya longitud sobrepase los 10 Kb es prohibitiva con estos métodos (Ureta-Vidal et al. 2003). En segundo lugar, asumen que la información de secuencia es colinear ó sinténica (es decir que conserva su orden enttre los individuos; Brudno et al. 2003), por lo que no toman en cuenta la presencia de eventos de macromutación, los cuales son frecuentes en genomas microbianos (Mira et al. 2002).

En el presente estudio, se utilizó el algoritmo de alineación implementado en el programa MAUVE versión 2.1.1 para Linux (Darling et al. 2004), el cual es de libre distribución y se encuentra disponible en la página asap.ahabs.wisc.edu/mauve/. Se escogió este programa porque corrige las limitaciones de algoritmos anteriores, al poder alinear múltiples secuencias genómicas en un tiempo de cómputo razonable, en la presencia de rearreglos genómicos, de duplicaciones génicas y de transferencia horizontal. El funcionamiento del algoritmo se detalla en el apartado siguiente (3.1.1).

La alineación de las 12 secuencias de ADN cromosómico de *E. coli* con MAUVE se corrió en el *cluster* computacional *KanBalam* de memoria distribuida, que se encuentra en el edificio DGSCA, UNAM, México D.F., con los parámetros de alineación que se explican en la Tabla 3.

El alineamiento generado por MAUVE se divide en varios sub-alineamientos (como se explica en el apartado 3.1.1). Estos sub-alineamientos corresponden a las regiones que han sufrido rearreglos, a los cuales se les llama bloques localmente colineares (LCBs; *Locally Colinear Blocks*). Cada uno de éstos se revisó, para detectar posibles regiones alineadas incorrectamente y realinearlas en los casos que fuera necesario, como se explica en el apartado 3.1.2.

Finalmente, se concatenaron los sub-alineamientos ya revisados, tomando como referencia el orden del genoma de la cepa K12 MG1655 de *E. coli*. De esta manera, se trabajó con un único alineamiento general del cromosoma para realizar los análisis de genómica de poblaciones.

Tabla 3. Parámetros de alineación del programa MAUVE versión 2.1.1.

Parámetro	*Parámetro*	Valor	Descripción del valor
Tamaño de semilla	*Match seed weight*	15	El tamaño mínimo de anclas (sitios homólogos) para empezar un alineamiento es de 15 nucleótidos.
Tamaño mínimo de LCB	*Min LCB weight*	45	El tamaño mínimo de los bloques colineares es de 45 nucleótidos homólogos.
Determinar LCBs	*Determine LCBs*	yes	Sí delimita bloques colineares.
Asumir que los genomas son colineares	*Asume collinear genomes*	no	No asume que los genomas de los individuos de la muestra sean colineares.
Alineamiento completo	*Full alignment*	yes	Sí realiza el alineamiento final.
Extender LCBs	*Extend LCBs*	yes	Sí realiza una búsqueda de anclas para extender el tamaño de los LCBs
Algoritmo de alineación	*Aligner*	MUSCLE 3.6	Utiliza el algoritmo de MUSCLE para realizar el alineamiento final.
Longitud mínima de isla	*Minimum island size*	50	El tamaño mínimo de las regiones del genoma flexible es de 50 sitios no homólogos.
Longitud mínima de gaps en el *backbone*	*Minimum backbone gap size*	50	La extensión mínima de gaps en las regiones del genoma central es de 50 sitios.
Longitud de *backbone* mínima	*Minimum backbone size*	50	El tamaño mínimo de las regiones del genoma central es de 50 sitios homólogos.

3.1.1. Descripción del algoritmo de MAUVE versión 2.1.1

Básicamente, el algoritmo de MAUVE alinea secuencias genómicas usando un método de alineación múltiple, con comparaciones tanto locales como globales. Esto reduce el tiempo de cómputo, y permite identificar regiones conservadas ortólogas, delimitar los puntos de quiebre exactos de rearreglos de secuencia y finalmente, generar un alineamiento global clásico para identificar sustituciones nucleotídicas e *indels*. El proceso de alineación se lleva a cabo en varias partes.

a) El primer paso del algoritmo consiste en una búsqueda heurística y alineación de regiones altamente conservadas ó semillas, a las que se conoce como *anchors* ó anclas. Los métodos heurísticos, funcionan bajo el supuesto de que en la extensión de las secuencias a alinear deben de existir regiones con alto porcentaje de similitud, las cuales son detectadas y alineadas localmente, mediante una comparación múltiple (Delcher et al. 1999). Al permitir la búsqueda de manera no-colinear se detectan anclas que no son sinténicas, con lo que se logra una primera identificación de regiones discordantes entre los genomas, probablemente generadas por eventos de rearreglo genómico ó transferencia horizontal.

Los alineamientos locales que se obtienen se utilizan como puntos de "anclaje" para el

alineamiento subsiguiente. Con esto, el tiempo de cómputo se reduce considerablemente (Ureta-Vidal et al. 2003).

Adicionalmente, el algoritmo identifica anclas presentes en todos los genomas y presentes sólo en una submuestra de genomas, lo que permite alinear también las regiones correspondientes al genoma flexible. En este paso, MAUVE sólo considera anclas ortólogas, por lo cual el autor (Darling et al. 2004) las nombró concordancias múltiples únicas máximas (multiMUMs; *Multiple Maximal Unique Matches*).

b) A continuación el algoritmo calcula una medida de distancia para realizar un árbol filogenético guía con el método de *Neighbor Joining* (NJ; Saitou y Nei 1987).

Como ocurre con otros algoritmos de alineación, como es CLUSTALW, MAUVE usa un árbol para determinar la sucesión óptima de comparación de las secuencias en el proceso de alineamiento. A partir de la información de los MultiMUMs, el programa MAUVE calcula la similitud al nivel de secuencia como el número de núcleotidos compartidos entre dos genomas sobre el número promedio de nucleótidos presentes en ambos genomas. El estimado obtenido se convierte después a una medida de distancia para realizar la matriz con la que se crea el árbol de *Neighbor Joining*. Al calcular un árbol global y no uno por cada región colinear, el tiempo de alineamiento también se reduce considerablemente.

c) El tercer paso es la selección de multiMUMs y la delimitación de bloques de colinearidad local (LCBs; *Locally Colinear Blocks*). Para lograr esto, primero se eliminan multiMUMs espurios, que son aquellas anclas detectadas en el primer paso del algoritmo que nos son verdaderamente homólogas, y que se alinearon debido a similitud por azar. Posteriormente, el algoritmo agrupa los multiMUMS restantes en conjuntos de anclas colineares. Estos grupos de multiMUMs colineares son los llamados bloques localmente colineares (LCBs). La agrupación de multiMUMs se realiza con un algoritmo de eliminación recursiva de puntos de quiebre, el cual identifica un peso (longitud) mínima que deben satisfacer los multiMUMs para ser verdaderamente homólogos. En este punto, el peso mínimo es de 3 kb por omisión.

d) A continuación, se realiza una búsqueda recursiva de anclas para refinar el alineamiento. Debido a que el paso de alineación final sólo se realiza entre anclas que estén separadas por menos de 10 kb, para poder alinear la mayor parte del genoma es necesario identificar anclas en el mayor número de regiones posibles. Para lograr esto, el algoritmo

reduce progresivamente el peso (longitud) mínimo de identificación de anclas. Con esto se logran alinear las regiones menos conservadas del genoma. Al buscar nuevas anclas, tanto al interior de los LCBs como al exterior de éstos, se amplía la extensión de los LCBs, y se logra la alineación de una mayor región del genoma.

e) Finalmente, una vez que se han identificado el mayor número de anclas, se realiza el alineamiento global de las regiones entre anclas en cada LCB, a partir del árbol guía. Para esto se puede usar el algoritmo de CLUSTALW (Thompson et al. 1994) versión 1.8.4 ó el de MUSCLE (Edgar 2004) versión 3.6, según se elija en las opciones de parámetros de inicio. En este estudio elegimos MUSCLE 3.6, el cual se basa en un método de alineación local, y por lo tanto genera alineamientos con mayor rapidez (Do y Katoh 2009). El algoritmo de alineación se ejecuta una sola vez en las regiones entre anclas de cada LCB. En este paso se alinean las regiones parálogas cuya posición se conserve entre los genomas analizados. El resultado final es un alineamiento para cada LCB del genoma.

El programa MAUVE genera varios archivos de salida. Dos de ellos contienen la información del alineamiento genómico completo, 1) uno en formato mauve, el cual es utilizado por el visor de alineamientos para generar la figura y 2) otro en formato Multi-FASTA extendido (XMFA; *Extended Multi-Fasta*). Éste último se desarrolló para alineaciones en las que hay rearreglos, ya que permite el almacenamiento de varios sub-alineamientos en un solo archivo en formato Fasta. En este caso los sub-alineamientos corresponden al alineamiento de cada LCB. Además, en este formato se pueden indicar las coordenadas genómicas de las regiones alineadas, así como la dirección de alineación de las secuencias de cada individuo, es decir, si se alinearon en sentido *forward* (es decir que se alineó en la misma dirección que la secuencia de referencia) ó *reverse* (si se alineó en sentido reverso-complementario a la secuencia de referencia), lo cual permite detectar rápidamente cuales son los individuos que presentan regiones con inversiones.

Los demás archivos de salida contienen: 3) la matriz de identidad nucleotídica, 4) el árbol guía, 5) la matriz del número de permutaciones ó rearreglos por pares de individuos, 6) los límites de los LCBs en cada individuo, 7) las coordenadas de las regiones correspondientes al genoma flexible (referidos como islas en el programa) en cada individuo y 8) las coordenadas de las regiones del genoma central ó *backbone* en cada individuo.

3.1.2. Revisión del alineamiento y realineación

MAUVE 2.1.1 detecta eficientemente la variación genómica generada por rearreglos, inversiones y translocaciones (Darling et al. 2004).

Sin embargo, una deficiencia del algoritmo de MAUVE 2.1.1, es que no siempre alinea correctamente las regiones del genoma flexible (Darling et al. 2010). Esto ocurre debido a que estas regiones están flanqueadas por regiones conservadas del genoma (Figura 1A), las cuales usualmente fungen como anclas. Dado que el programa asume que todas las regiones entre dos anclas adyacentes son homólogas y por lo tanto las alinea, cuando las secuencias del genoma flexible son diferentes entre subgrupos ó individuos de la muestra, el programa genera un alineamiento forzado entre secuencias no-homólogas (Darling et al. 2004). De tal manera que a lo largo del alineamiento se pudieron encontrar regiones alineadas de manera forzada, como ocurre en el ejemplo de la Figura 1B.

Si bien, versiones posteriores de MAUVE ya incluyen un modelo de Markov para identificar alineamientos incorrectos y corregirlos (Darling et al. 2010), el algoritmo de MAUVE que se usó en el presente trabajo no está diseñado para hacer esto. Entonces, para poder detectar la presencia de errores en el alineamiento, se realizó una revisión "a mano" de cada sub-alineamiento, con ayuda del programa BIOEDIT (Hall 1999) versión 7.0.5. Se registraron las coordenadas de las regiones del sub-alineamiento que parecían mal alineadas ó forzadas y se hizo un nuevo archivo de BIOEDIT para cada una de estas regiones, para trabajarlas por separado y evitar modificar las demás regiones bien alineadas.

Para determinar si las regiones estaban efectivamente mal alineadas ó simplemente eran muy divergentes, se utilizó el algoritmo de BLASTn versión 2.2.14 (Altschul et al. 1990; Zhang et al. 2000). Si se obtenía con BLAST una $E \leq 10^{-70}$ y una similitud del 50% en al menos 50% de la secuencia, se consideraba que las secuencias eran homólogas y que el alineamiento era correcto (Didelot et al. 2009). Las regiones que no cumplieron estas características, se corrigieron de la siguiente manera. Primero, se hizo un archivo en formato fasta para cada uno de los grupos de secuencias que sí eran homólogas. Después, cada grupo de secuencias se realineó por separado con el programa CLUSTALW (versión 1.8.4) con los parámetros dados por omisión. Una vez corregidas las alineaciones problemáticas, se volvieron a insertar en la misma posición del sub-alineamiento de donde habían sido extraídas.

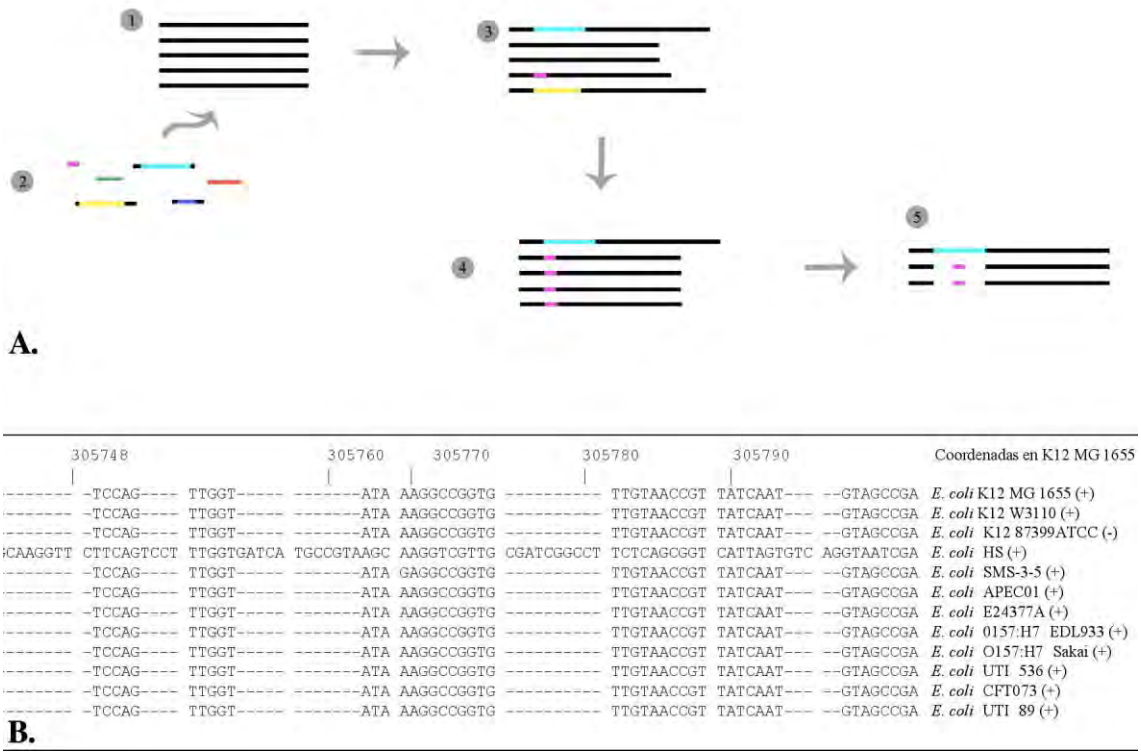


Figura 1. **A.** Procesos que dan origen a la variación en el repertorio genético de una población. Los genomas de una población que evoluciona solamente por mutaciones puntuales, presentan regiones genómicas del mismo tamaño y composición (1). (2) Eventos de transferencia horizontal, recombinación desigual y recombinación ilegítima, entre otros, introducen variación en la población (3). (4) Por el efecto de la selección natural, algunas de estas combinaciones aumentan en frecuencia. Finalmente, (5) al alinear los genomas de una muestra de la población original, sólo las regiones verdaderamente homólogas (ortólogas ó con parálogos posicionalmente conservados) se alinearan correctamente, las demás regiones se alinean de manera forzada como se muestra en B.

B. Muestra de un segmento de alineación cromosómica de *E. coli* incorrecta. La región total tiene una extensión de 24,756 sitios. En este fragmento se puede apreciar que la secuencia de la cepa comensal de *E. coli* HS es diferente a las demás secuencias del alineamiento. La revisión con BLASTN arrojó que esta región de secuencia corresponde a una lipoproteína de fago presente únicamente en la cepa comensal HS. Por su parte, la secuencia presente en las demás cepas, corresponde a los genes conservados *yagW* y *yagX*. Estos tres elementos genéticos se encuentran entre regiones conservadas que funcionaron como anclas para el alineamiento de esta región. Por esta razón el algoritmo de MAUVE asume que la secuencia de fago es ortóloga a la de los genes *yagW* y *yagX* y forza su alineación.

(+) secuencia alineada en sentido *forward*; (-) secuencia en sentido *reverse*, con respecto a la secuencia de referencia.

Otra fuente de error del programa MAUVE, es que sólo aplica el algoritmo de alineación final (CLUSTALW ó MUSCLE) a las secuencias delimitadas dentro de cada LCB (Darling et al. 2004), por lo que las zonas externas a los LCBs, aún cuando pudieran presentar secuencias homólogas a un subgrupo de la muestra, no se alinean.

Para alinear las regiones externas a los LCBs primero se realizó una búsqueda con BLASTn (Altschul et al. 1990; Zhang et al. 2000), usando el mismo criterio ($E \leq 10^{-70}$, similitud 50%, cobertura 50%) sugerido por Didelot et al. (2009). Aquellas que resultaron ser efectivamente homólogas, se alinearon con CLUSTALW (versión 1.8.4) con los parámetros dados por omisión.

3.2. Detección y delimitación de los componentes del "pangenoma" en *E. coli*.

La manera clásica de describir el pangenoma de una especie es a partir de la identificación de proteínas ortólogas y la determinación conservadas en todos los individuos (genoma central) ó presentes únicamente en algunos individuos (genoma flexible) de una muestra, mediante comparaciones recíprocas de tipo BLAST y sus variantes (Tettelin et al. 2005; Rasko et al. 2009; Schoen et al. 2008; D'Auria et al. 2010; Nandi et al. 2010).

A diferencia de estos estudios, en este trabajo se describió el pangenoma de *E. coli* a partir del alineamiento y no a partir de unidades funcionales, como son los genes. Para esto se siguió el método de Darling et al. (2004), quienes definen al genoma central como aquella región del alineamiento que presenta más de 50 columnas sin gaps, que no se encuentren intercaladas por regiones de 50 ó más gaps consecutivos en cualquiera de los genomas. Por exclusión, las regiones del alineamiento que no cumplan con esto, forman parte del genoma flexible.

Sin embargo, hay discrepancia en cuanto a la longitud mínima que pueden tener las regiones del genoma central, y se ha propuesto que pudieran ser de tamaño menor a 50 nucleótidos (Darling et al. 2004; Didelot et al. 2009).

En el presente estudio, se consideró que una región era parte del genoma central si el alineamiento presentaba al menos 10 columnas de nucleótidos sin gaps, que no se encontraran intercaladas por 10 ó más gaps en cualquiera de los individuos. Las regiones que no cumplieran con estas características se consideraron como parte del genoma flexible.

La posición de estas regiones se registró en un archivo de texto de tipo *Block Data File* (BDF). El formato BDF enlista mediante tabulaciones las coordenadas de regiones particulares de un alineamiento, ya sea en relación a una secuencia de referencia ó al alineamiento total (Hutter et al. 2006). Este tipo de archivos sirve para circunscribir un análisis a las regiones delimitadas en el archivo, por lo que son particularmente útiles en estudios genómicos (Rice et al. 2000; Hutter et al. 2006).

4. GENÓMICA DE POBLACIONES

Con la finalidad de entender mejor el papel de las fuerzas evolutivas en la generación de los patrones de diversidad dentro y entre poblaciones, así como la posible presencia de variación en el efecto de dichas fuerzas en diferentes regiones del genoma, la genómica de poblaciones usa las herramientas de la genética de poblaciones para analizar un número grande de loci ó regiones genómicas (Luikart et al. 2003).

Al contar con la información del cromosoma completo y no de ciertas unidades, surge la posibilidad de analizarlo como un todo, ó de estudiarlo por unidades más pequeñas. Para definir las regiones de análisis genómico se delimitaron loci de la manera en que se describe en el apartado siguiente.

4.1. Unidades de estudio genómico: "loci"

Debido a que no se circunscribió la información de secuencia a la posición ó marco de lectura de regiones funcionales, como son los genes y otros elementos de secuencia, primero se describió el comportamiento general del cromosoma a diferentes "escalas" ó "niveles", con el fin de determinar una unidad de estudio genómico, que se tomaría como locus para los análisis subsiguientes. Para esto, se examinó el genoma central de la muestra, por ventanas subsecuentes (corridas y no traslapadas) de longitud constante (abarcando el mismo número de sitios homólogos). Se estimó la diversidad genética y se aplicaron las pruebas de selección y desequilibrio de ligamiento (descritas en el apartado siguiente) en cada ventana (con el programa VARSICAN 2.0.2) y se registraron las ventanas en las cuales las pruebas resultaban significativas. Se repitieron dichas mediciones para seis tamaños

diferentes de ventana: 3, 10, 100, 1 000, 10 000 y 100 000 sitios. De esta manera, al final tuvimos seis niveles de análisis del genoma central.

Para comparar los valores obtenidos en cada nivel de análisis, primero se probó si los datos presentaban una distribución normal, con la prueba D de Kolmogorov-Smirnoff. Así, se decidió usar la prueba no-paramétrica de Friedman, que está diseñada para comparar más de dos grupos no independientes. Ambas pruebas estadísticas se aplicaron con el programa JMP (Instituto SAS, 1997) versión 7.0.1.

Finalmente, para detectar diferencias en los patrones de selección y de clonalidad a las diferentes escalas de análisis, se calcularon las proporciones de ventanas significativas en las 3 pruebas de neutralidad (D de Tajima, D* y F* de Fu-Li) y en la prueba de desequilibrio de ligamiento (Zns).

En el caso de las pruebas de neutralidad, las proporciones se estimaron a partir del número de ventanas con valores significativos ($p<0.05$), negativas ó positivas en cada prueba, sobre el número de ventanas polimórficas (a lo que se llamó N_P ó número de loci polimórficos) en cada una de estas dos categorías. La proporción de ventanas en desequilibrio de ligamiento detectadas con la prueba Zns se estimó dividiendo el número de ventanas con valores significativos (a $p<0.025$) sobre el número de ventanas con sitios informativos para esta prueba, a lo que se denominó N_{DL} (este número solamente puede ser igual ó menor al número de loci polimórficos N_P).

A partir de los datos obtenido con este análisis (explicados en detalle en los Resultados en el apartado 2.1), se analizó el cromosoma en ventanas de tamaño fijo, no traslapadas, con una longitud máxima de 1,000 sitios, al interior de las regiones del genoma central y del genoma flexible (Figura 6). Cada una de estas ventanas ó regiones genómicas se consideraron como locus para realizar los análisis de genética de poblaciones clásicos.

4.2 Estimación de parámetros de genética de poblaciones a nivel genómico

El programa VARISCAN versión 2.0.2 (Vilella et al. 2005), está diseñado para calcular rápidamente estadísticos básicos de genética de poblaciones, como son estimados de diversidad genética y de divergencia, pruebas de desequilibrio de ligamiento y pruebas de selección natural, en un número grande de loci ó de regiones genómicas. Los estimados utilizados en el presente estudio, así como la manera en que son calculados por este

programa se explican más adelante.

Tabla 4. Parámetros de inicio del programa VARISCAN 2.0.2.

	Parámetro	Valor	Significado del valor del parámetro
Posición de inicio del análisis	*StartPos*	1	Comienza el análisis en la posición 1 del alineamiento.
Posición de termino del análisis	*EndPos*	0	Termina el análisis en la última posición del alineamiento.
Posición de inicio y término de acuerdo a la referencia	*RefPos*	0	Los parámetros anteriores corresponden a posiciones en el alineamiento global (y no en el genoma de referencia).
Archivo de datos de análisis por bloque	*BlockDataFile*	C:\Users\...	Analiza por segmentos definidos en el archivo BDF que se encuentra en la dirección indicada en este parámetro.
Nombres de los individuos	*IndivNames*	-	Sólo se requiere para archivos tipo MAF. En nuestro caso el alineamiento se guardó en formato PHY, por lo que se ignoró este parámetro.
Secuencias por analizar	*SeqChoice*	all	Todas las secuencias se incluyen en el análisis.
Grupo externo	*Outgroup*	none	No se define grupo externo para el análisis.
Secuencia de referencia	*RefSeq*	-	No se define individuo de referencia.
Modo de análisis	*RunMode*	12 31	Calcula los parámetros π, θ, D Tajima, D* y F* Fu-Li, para $n \geq 4$ Calcula los parámetros D, D'y r^2promedio, Hd y Fs de Fu para $n \leq 2$
Usar η ó S	*UseMuts*	0	Usa el número total de sitios segregantes S (y no el número total de mutaciones η) para calcular θ y D de Tajima.
Utilizar ó no sitios con gaps	*CompleteDeletion*	1	Todos los sitios con gaps (-) ó nucleótidos ambiguos son excluidos del análisis.
Análisis de un número fijo de secuencias	*FixNum*	-	Se define sólo si se incluyen sitios con gaps ó ambigüedades en el análisis. Como no es el caso se ignora.
Análisis de un número mínimo de secuencias	*NumNuc*	-	Se define sólo si se incluyen sitios con gaps ó ambigüedades en el análisis. Como no es el caso se ignora.
Análisis por ventanas	*SlidingWindow*	1 0	Lleva a cabo un análisis por ventanas corredizas. No realiza análisis por ventanas corredizas.
Longitud de la ventana	*WidthSW*	3 10 100 1000 10000 100000	Se requiere sólo si *SlidingWindow*=1. Determina el número de sitios que abarca la ventana.
Avance de la ventana	*JumpSW*	3 10 100 1000 10000 100000	Se requiere sólo si *SlidingWindow*=1. Determina el número de sitios que avanza para la ventana siguiente. Como WidthSW es igual a JumpSW las ventanas no se traslapan.
Tipo de ventana	*WindowType*	1	El número de sitios corresponde a sitios netos (no cuenta los gaps).
Usar singletons para DL	*UseLDSinglets*	0	Ignora los singletons para el cálculo de desequilibrio de ligamiento DL.

4.2.1 Estimación de la diversidad genética

Para describir la variación en general al nivel de secuencia de ADN existen varias medidas de diversidad. Una de ellas es la proporción de sitios segregantes **ó distancia p_S**, la cual se calcula como la proporción de sitios homólogos que difieren en las secuencias de una población ó **sitios segregantes S** sobre el total de sitios en el alineamiento **N**, de acuerdo a Nei y Kumar (2000):

$$P_S = S/N$$

Esta medida de diversidad se encuentra influenciada fuertemente por la presencia de alelos deletéreos (Hedrick 2004; Castillo 2007), ya que éstos generalmente se encuentran en baja frecuencia en la población. Además el número S no toma en cuenta la frecuencia de los mutantes (Tajima 1989) ni las frecuencias alélicas en general. Por lo tanto, una medida de diversidad poblacional más adecuada es la diversidad nucleotídica **π (pi)** la cual es calculada por el programa VARISCAN 2.0.2 de acuerdo a Nei (1987) y Nei y Miller (1990) como:

$$\pi = \sum_{ij} x_i x_j \pi_{ij}$$

donde

ij es un par de secuencias de la muestra

x_i , es la frecuencia de la secuencia i

x_j , es la frecuencia de la secuencia j

π_{ij}, es el número de diferencias nucleotídicas por sitio entre el par de secuencias ij.

De esta fórmula se deduce que la diversidad nucleotídica corresponde al promedio de las diferencias nucleotídicas entre pares de secuencias, ponderadas por su frecuencia. Por ello el estimado de diversidad π se correlaciona directamente con las frecuencias alélicas. Y si la frecuencia alélica de algunos alelos con respecto a otros es alta, la diversidad será alta. Por el contrario, si la frecuencia alélica es baja, la diversidad será baja. Por esta razón el valor del parámetro π no se ve afectado por la presencia de alelos deletéreos, ya que considera las frecuencias alélicas de los mutantes (Tajima 1989a). Este parámetro tampoco es sesgado por el tamaño de muestra (Tajima 1983).

Otra medida de diversidad es el parámetro de mutación poblacional **θ (theta),** el cual permite conocer la diversidad genética, entendida como la probabilidad de que dos

secuencias de ADN tomadas al azar de una población difieran en algún sitio (Kimura 1968). Éste parámetro se calcula para organismos haploides como:

$$\theta = 2N_e\mu$$

donde $\mathbf{N_e}$ es el tamaño efectivo poblacional y $\boldsymbol{\mu}$ es la tasa de mutación.

Si no se puede conocer el valor de N_e y μ, lo que ocurre con frecuencia en estudios de genética de poblaciones, otra forma de calcular θ de acuerdo a Watterson (1975; Tajima, 1983) es:

$$\theta_W = S/a$$

donde **S** es el número total de sitios segregantes en una muestra de secuencias y $a = \sum 1/k$, desde k = 1 hasta k = n-1, donde n es el número de secuencias de la muestra.

Si el tamaño efectivo es pequeño, θ es influenciado fuertemente por la deriva génica y los alelos deletéreos tienen una frecuencia mayor (Tajima 1983; Tajima, 1989a). Si θ es alto quiere decir que la diversidad correspondiente a los alelos deletéreos es alta.

De la primera fórmula de θ se obtiene que este parámetro sí se ve afectado por el tamaño de muestra. Y si la población ha experimentado un cuello de botella recientemente, un muestreo pequeño puede inflar ó reducir el valor de θ (Tajima 1989a).

Bajo los supuestos de la teoría neutral, estos dos parámetros deben tener el mismo valor (Kimura 1969a; 1969b; Watterson 1975; Li 1977a; 1977b; Nei y Li 1979; Tajima 1983; 1989a). Por lo tanto la existencia de diferencias estadísticamente significativas entre los valores de π y θ en una región de secuencia de ADN de una población determinada se toma como evidencia de selección.

Bajo este principio se han desarrollado algunas pruebas para detectar desvíos de neutralidad en las regiones de ADN. En particular tres de ellas, la D de Tajima (1989), la D* y la F* de Fu-Li (1993) han sido utilizadas en el presente estudio y se describen en el apartado 4.2.3 de la metodología.

4.2.2 Determinación de la diversidad de haplotipos y desequilibrio de ligamiento

Otra manera de describir la variación genética en las poblaciones es la frecuencia de haplotipos. Tradicionalmente, la definición de haplotipo corresponde a la combinación de alelos en varios loci de una población. Al analizar datos de secuencia, tanto de ADN como de aminoácidos, el termino haplotipo indica la asociación de sitios polimórficos en una región en particular del genoma (Hedrick 2005).

Para cuantificar la diversidad de haplotipos se ha definido el parámetro **Hd** (Nei 1987; Depaulis y Veuille 1998) el cual toma en cuenta la distribución de frecuencias haplotípicas de la siguiente manera:

$$Hd = (1 - \Sigma p_i^{\,2})\, n / (n - 1)$$

donde

p_i es la frecuencia de cada haplotipo en la muestra

n es el número de muestra.

Valores pequeños de Hd revelan una estructuración de los sitios polimórficos en pocos haplotipos. Y valores altos cercanos a 1, muestran un exceso de haplotipos (Depaulis y Veuille 1998).

La asociación de los alelos ó sitios puede no ser al azar, en cuyo caso una combinación particular de alelos será más ó menos frecuente en la población. A la asociación no azarosa de alelos se le conoce como desequilibrio de ligamiento. Dado que es inversamente proporcional a la presencia de recombinación (Dawson et al. 2002; Hedrick 2005) se utiliza como una medida del grado de clonalidad que puede existir en una población (Maynard-Smith 1993). Para comprobar la presencia de desequilibrio de ligamiento se han descrito al menos 4 tipos de métodos (Mueller 2004). Las medidas clásicas se realizan con modelos de asociación pareada como son los parámetros D, ′D′, D’, ′D’′ (Lewontin y Kojima 1960, Lewontin 1964a; 1964b; 1995), r y r^2 (Hill y Robertson 1968). Un segundo tipo lo constituyen las medidas de asociación multi-locus como el índice de asociación I_A (Brown et al. 1980; Souza et al. 1992; Maynard-Smith, 1993), la homocigosis de haplotipos H_H (Sabatti y Risch 2002) ó la diferencia normalizada de la entropía ε que es una extensión multi-locus del coeficente r^2 de Hill y Robertson (Nothnagel, Furst y Rohde 2002). El tercer grupo de métodos utilizan el modelo de haplotipo específico, el cual es una extensión de la

homocigosis de haplotipos ó EHH (Mueller y Andreoli 2004). Por último, están los métodos basados en modelos filogenéticos, los cuales asumen modelos de genética de poblaciones definidos con respecto a tasa de mutación, recombinación, migración, para estimar desequilibrio de ligamiento (Morton et al. 2001).

Los métodos basados en asociación pareada y multi-locus, como la r^2 de Hill y Robertson (1968) y el I_A (Maynard-Smith 1993) muestran una metodología clara y directa, y son los más comúnmente utilizados para detectar desequilibrio de ligamiento. Sin embargo los métodos de asociación son altamente sensibles a las frecuencias alélicas (Devlin y Risch 1995), y varían dependiendo de la distancia física entre los sitios analizados. Esto complica la interpretación de los resultados y la comparación entre diferentes regiones del genoma.

Se ha propuesto que los promedios de los parámetros D' y r^2 se relacionan de manera monotónica con la distancia física, lo que los hace más apropiados para estudios de regiones amplias del genoma donde se espera comparar diferentes loci con diferentes frecuencias haplotípicas (Wang et al. 2006). Se sabe que la prueba de neutralidad **Zns** (Kelly 1997) es equivalente al valor promedio del parámetro r^2 (Hill y Robertson 1968) de todas las comparaciones pareadas dentro de un locus ó de una región multilocus (Hutter 2006; Librado y Rozas 2009).

La prueba Zns corresponde a una prueba de neutralidad de clase II (Ramos-Onsins y Rozas 2002), que a diferencia de las pruebas de neutralidad de Tajima y Fu-Li, comparan la distribución de haplotipos en las poblaciones y no la frecuencia de las mutaciones. Las pruebas de neutralidad de clase II han mostrado ser efectivas para detectar desequilibrio de ligamiento entre loci y dentro de loci y se han usado con éxito en estudios de genómica de poblaciones para dicho propósito (Hutter et al. 2007; Bensasson et al. 2008).

La fórmula de Zns de acuerdo a Kelly (1997) es:

$$Zns = [2/S(S-1)] \sum_{i=1} \sum_{j=i+1} \delta_{ij}$$

donde δ_{ij} es una medida estandarizada del desequilibrio de ligamiento.

Matemáticamente δ_{ij} corresponde a la correlación cuadrada de identidad alélica entre dos loci ó sitios (i-j).

De tal manera que si se tienen valores de Zns más elevados de lo esperado, cercanos a 1, implica que existe una asociación no azarosa entre los sitios de la región analizada. Si los valores son más bajos de los esperado, cercanos a 0, sólo puede ser por efecto de

recombinación intralocus (Kelly 1997). El programa VARISCAN 2.0.2 lo calcula como el promedio de r^2 (Hill y Robertson 1968) para todas las comparaciones pareadas Z dentro de una región (Hutter 2006; Librado y Rozas 2009) y se calcula como:

$$Zns = \sum_{ij} r^2_{ij} / Z$$

donde ij corresponde a cada par de sitios polimórficos en la región analizada,

$$r^2_{ij} = \frac{D^2}{p_{i1}\, p_{i2}\, p_{j1}\, p_{j2}}$$

donde

$D = x_{11} - p_{i1}\, p_{j1}$ (Lewontin y Kojima 1960)

donde x_{11} es la frecuencia esperada del haplotipo 11

p_{i1} es la frecuencia del alelo 1 del sitio i

p_{j1} es la frecuencia del alelo 1 del sitio j

p_{i2} es la frecuencia del alelo 2 del sitio i

p_{j2} es la frecuencia del alelo 2 del sitio j

Para determinar el nivel de significancia estadística se utilizó la ecuación 6, tabla 1 de Kelly (1997).

4.2.3 Detección de señales de selección natural

De manera general existen 2 grupos de métodos que permiten estimar la presencia de selección a nivel molecular en un locus. Por un lado tenemos aquellos métodos que se basan en la distribución de la diversidad (frecuencias polimórficas) y por otro lado tenemos los métodos que se basan en modelos de sustitución (Castillo 2007; Hurst 2009).

Los métodos basados en modelos de sustitución utilizan la información de las sustituciones sinónimas y no-sinónimas en las secuencias. Es decir, que calculan la diferencia entre la proporción de mutaciones al nivel de secuencia de ADN que generan un cambio al nivel de proteína (sustitución no-sinónima) y la proporción de mutaciones que no generan cambio al nivel de proteína (sustitución sinónima). Esto implica que tal tipo de pruebas de neutralidad sólo son aplicables a regiones codificantes.

Por el contrario, las pruebas basadas en la distribución de la diversidad comparan la

distribución de la frecuencia de mutación, asumiendo un modelo de sustitución de sitios infinitos (Kimura 1968; Kingman 1982a; Hudson 1990) lo que, aún con sus limitaciones, sí permite analizar y comparar regiones codificantes y no-codificantes (Andolfatto 2005).

Adicionalmente, se debe considerar que la efectividad de los métodos para detectar selección se ve afectada por la naturaleza clonal ó recombinante de la población por analizar (Shapiro et al. 2009). Así, los métodos basados en modelos de sustitución son efectivos en poblaciones tanto clonales como sexuales, y los métodos basados en distribuciones de frecuencia solamente son efectivos para analizar poblaciones sexuales. Es importante tomar esto en cuenta, debido a que el modelo aquí estudiado, *E. coli*, ha sido reiteradamente considerado como una especie no-sexual y prácticamente el organismo modelo del paradigma de diversidad clonal propuesto por (Selander y Levin 1980). Sin embargo, la evidencia con respecto a la naturaleza recombinante de esta especie se ha acumulado en las últimas décadas (Maynard-Smith 1993; Souza et al. 1999; Morris y Drouin 2008), a lo cual se suman los resultados del presente estudio (cf. Tabla 11). Por lo cual, se concluyó que las pruebas de selección basadas en la distribución de frecuencias sí pueden ser aplicadas en el estudio de *E. coli*.

Hay numerosas pruebas de neutralidad de este tipo, como son la D de Tajima (Tajima, 1983), la prueba HKA (Hudson et al. 1987), la D* y F* de Fu-Li (Fu y Li, 1993), la H de Fay-Wu (Fay y Wu 2000), la r_2 (Ramos-Onsins y Rozas 2002), y estadísticos ponderados para análisis multilocus como la D/Dmin (Schaeffer 2002) y H/Hmin (Schmid et al. 2005). En el caso particular de este trabajo se decidió aplicar las pruebas D de Tajima (Tajima 1983), D* y F* de Fu-Li (Fu y Li 1993), pues algunos estudios han demostrado que son las que tienen mayor poder, sobre todo si se usan complementariamente (Fu y Li 1993; Ramos-Onsins y Rozas 2002).

Las tres pruebas se fundamentan en el supuesto de que bajo equilibrio neutro las mutaciones se distribuyen aleatoriamente a lo largo del genoma. Esto generaría un patrón que se vería reflejado en los parámetros de diversidad nucleotídica π y de mutación poblacional θ, y que resultaría en valores semejantes ó iguales de diversidad estimados con ambos parámetros. Entonces, cuando no hay equilibrio neutral, los valores de π y θ deberían diferir entre sí. De tal manera que las pruebas consisten en determinar las diferencias entre los valores de ambos parámetros de diversidad y en establecer si dichas

diferencias son significativas.

La fórmula de la prueba **D de Tajima** se escribe en términos de π y θ (Hedrick 2005) como:

$$D = \frac{\pi - \theta}{\sqrt{V(\pi - \theta)}}$$

Donde V es la varianza.

Bajo el modelo de neutralidad, se espera que la D de Tajima sea igual a cero. Por las propiedades de π y θ explicadas en el apartado anterior, tenemos que si el valor de D es negativo, es debido a que θ es mayor que π y existe un exceso de alelos deletéreos, probablemente debido a la acción de selección purificadora ó negativa; si la D es positiva quiere decir que π es mayor que θ y que hay un exceso de alelos con frecuencias altas, con lo que se asume que debe existir selección balanceadora ó diversificadora. El programa VARISCAN 2.0.2 utiliza en particular la fórmula 38 en Tajima (1989a):

$$D = \frac{\pi - \theta_W}{\sqrt{V(\pi - \theta_W)}}$$

Donde V es la varianza.

Para determinar si las diferencias son estadísticamente significativas se utiliza la ecuación 47 en Tajima (1989a) que corresponde a los intervalos de confianza de la tabla 2 en el mismo trabajo.

Otra de las ventajas de usar esta prueba es que únicamente requiere los datos de polimorfismo, es decir las comparaciones intraespecíficas de diversidad. A diferencia de la prueba HKA y otras pruebas que también requieren de los datos de divergencia, es decir de diversidad interespecífica (Tajima, 1989a), para lo cual se necesita la información de un grupo externo.

Las pruebas de neutralidad **D* y F* de Fu-Li** (1993) difieren de la prueba de Tajima en cuanto a que estiman la distribución de mutaciones a lo largo de un árbol filogenético resuelto (es decir siempre bifurcado). Con esto determinan diferencias significativas en los valores de diversidad entre las ramas más recientes (ramas externas) y las ramas más antiguas (ramas internas). La fórmula general de estas pruebas es descrita por Castillo (2007):

$$G = \frac{\eta_i - \eta_e / a - 1}{\sqrt{V[\eta_i - \eta_e / (a - 1)]}}$$

donde V es la varianza,

η_e = número total de mutaciones en ramas externas,

η_i = número total de mutaciones en ramas internas,

$a = \sum 1/k$, desde k = 1 hasta k = n-1, donde n es el número de secuencias de la muestra (Watterson 1975).

De igual manera que para la prueba de Tajima, estas pruebas asumen que en equilibrio neutro la distribución de mutaciones es aleatoria y debe ser semejante en ramas internas y externas. Cuando hay selección, la diversidad en ramas internas y externas será diferente. Teóricamente, las mutaciones de las ramas externas, es decir, las mutaciones más recientes, corresponderían a alelos deletéreos en baja frecuencia, por lo que un exceso de mutaciones en las ramas externas refleja la presencia de selección purificadora. En este caso, el estadístico resultará con valor negativo. Por el contrario, un exceso de mutaciones en las ramas internas implica que hay un mayor número de mutaciones antiguas que de mutaciones recientes, las cuales sólo podrían estar mantenidas en la población por el efecto de selección positiva ó balanceadora, y en cuyo caso el estadístico tendrá un valor positivo.

En particular el programa VARISCAN 2.0.2 utiliza las fórmulas descritas en Fu y Li (1993):

$$D^* = \frac{(n/n-1)\eta - a(\eta_e + \theta_W)}{\sqrt{V\,[(n/n-1)\eta - a(\eta_e + \theta_W)]}}$$

y en Simonsen et al. (1995):

$$F^* = \frac{\pi - (n/n-1)(\eta_e + \theta_W)}{\sqrt{V\,[\pi - (n/n-1)(\eta_e + \theta_W)]}}$$

donde n es el número de secuencias ó tamaño de muestra.

Para determinar si dichas diferencias son estadísticamente significativas se utilizaron los intervalos de confianza en las tablas 2 y 4 en Fu y Li (1993).

Finalmente, debemos considerar que las 3 pruebas de neutralidad que se usaron en el presente estudio, se basan en el parámetro θ, el cual está directamente relacionado con el tamaño efectivo poblacional y por lo tanto puede llegar a reflejar variaciones generadas por fluctuaciones demográficas y no por el efecto de la selección natural (Tajima 1983). De esto surgen dos problemáticas. Primero, si la población ha experimentado un cuello de botella ó una expansión poblacional recientemente, un muestreo pequeño puede inflar ó reducir el valor de θ (Tajima 1989a). Segundo, la presencia de diferencias significativas en

los parámetros π y θ puede deberse a cambios demográficos recientes (expansiones ó reducciones poblacionales), y no al efecto de la selección (Tajima 1989b).

Se ha propuesto que las variaciones en la demografía y el efecto de la deriva génica afectan la diversidad del genoma de manera homogénea (Luikart et al. 2003; Holsinger y Weir 2009).

Por esta razón, al aplicar las pruebas de selección a nivel genómico, se deberían de obtener resultados con tendencias similares en todos los loci analizados. Por lo que los análisis a nivel genómico, en teoría, permiten reconocer si los resultados de este tipo de pruebas de neutralidad se deben a cambios en el tamaño poblacional ó efectivamente a selección natural, aún cuando la muestra sea pequeña (Shapiro et al. 2009).

4.3. Dinámica evolutiva de los "ecogrupos" de *E. coli*

4.3.1 Reconstrucción filogenética

Para reconstruir la filogenia de los individuos de la muestra de *E. coli* se usaron los datos del genoma central de los LCBs determinados con el programa MAUVE. Se infirió un árbol por cada uno de los LCBs. Para esto, se utilizó el programa PHYML versión 3.0 para Linux (Guindon y Gascuel 2003), el cual es de libre distribución y se encuentra disponible en la página http://atgc.lirmm.fr/phyml. El algoritmo de este programa se basa en el principio de Máxima Verosimilitud (ML) desarrollado por Felsenstein (1981), y permite un análisis rápido de una gran cantidad de datos de secuencia, tanto en número de individuos analizados como en la longitud de las secuencias. Siendo por esta última razón de particular interés para el presente estudio.

En cada caso, se usó el modelo de sustitución GTR (Lanave et al. 1984) y se hicieron 1,000 réplicas de *bootstrap*.

4.3.2 Comparaciones entre "ecogrupos"

Para estudiar la evolución del estilo de vida de *E. coli*, la muestra se clasificó de acuerdo a las características ecológicas de los individuos (cf. Tabla 2). Bajo este criterio, dividimos la muestra en dos grupos funcionales, a los que llamamos ecogrupos. Uno fue integrado por

las cepas de *E. coli* no-patógenas: las cepas de K12 MG1655, W3110, y ATCC8973, la cepa silvestre HS y la cepa de vida libre SMS3-5. Y el otro ecogrupo se integró por las cepas de *E. coli* patógenas: de ave APEC O1, intestinales EHEC (O157:H7) Sakai y EDL933 y ETEC E24377 y las extra-intestinales UPEC UTI89, 536 y CFT073.

Los parámetros de genética de poblaciones que se describieron previamente, fueron estimadas para el total de la muestra de *E. coli* y para cada ecogrupo por separado, con el programa VARISCAN 2.0.2 (Vilella et al. 2005).

Las comparaciones de parámetros estimados, entre componentes del pangenoma y entre ecogrupos se realizaron con la prueba de Wilcoxon. Esta prueba permite comparaciones de datos no-paramétricos entre dos grupos no independientes por lo que se puede utilizar para examinar diferentes regiones del genoma dentro de una misma población de acuerdo a Andolfatto (2005). Igualmente, se realizó una corrección de Bonferroni para comparaciones múltiples. Los análisis estadísticos se llevaron a cabo con el programa JMP 7.0.1 (Instituto SAS, 1997).

Finalmente, se determinaron las regiones funcionales presentes en cada uno de los loci bajo selección. En el caso de que se encontraran genes, se les asignó una categoría funcional de acuerdo al sistema de clasificación del JCVI (J. Craig Venter Institute), actualmente disponible en la página http://cmr.jcvi.org (Davidsen et al. 2010), usando como genoma de referencia las cepas de *E. coli* K12 MG1655 y UTI CFT073. Se calcularon las proporciones de genes presentes en cada categoría funcional, para determinar si las diferencias adaptativas entre ecogrupos se acumulaban en algún tipo particular de genes.

RESULTADOS

1. HISTORIA FILOGENÉTICA DE LA MUESTRA DE *E. COLI*

Las relaciones filogenéticas entre los individuos de la muestra de *E. coli*, se muestran en la Figura 2. Los linajes que pertenecen a un mismo estilo de vida ó ecogrupo (no-patógeno y patógeno) parecen ser polifiléticos. La cepa de vida libre se agrupa con las cepas patógenas extra-intestinales y con la cepa patógena de ave. Y las cepas comensales se agrupan con las cepas patógenas intestinales. Sin embargo, al analizar otras regiones del genoma, no siempre se recupera el mismo árbol filogenético (Apéndice 1). Esto da un indicio de que, a pesar de que la historia filogenética se puede recuperar a partir de ciertas regiones del genoma, hay flujo génico constante entre los individuos de esta muestra, como lo revelan los resultados de la prueba de clonalidad, que se presentan más adelante.

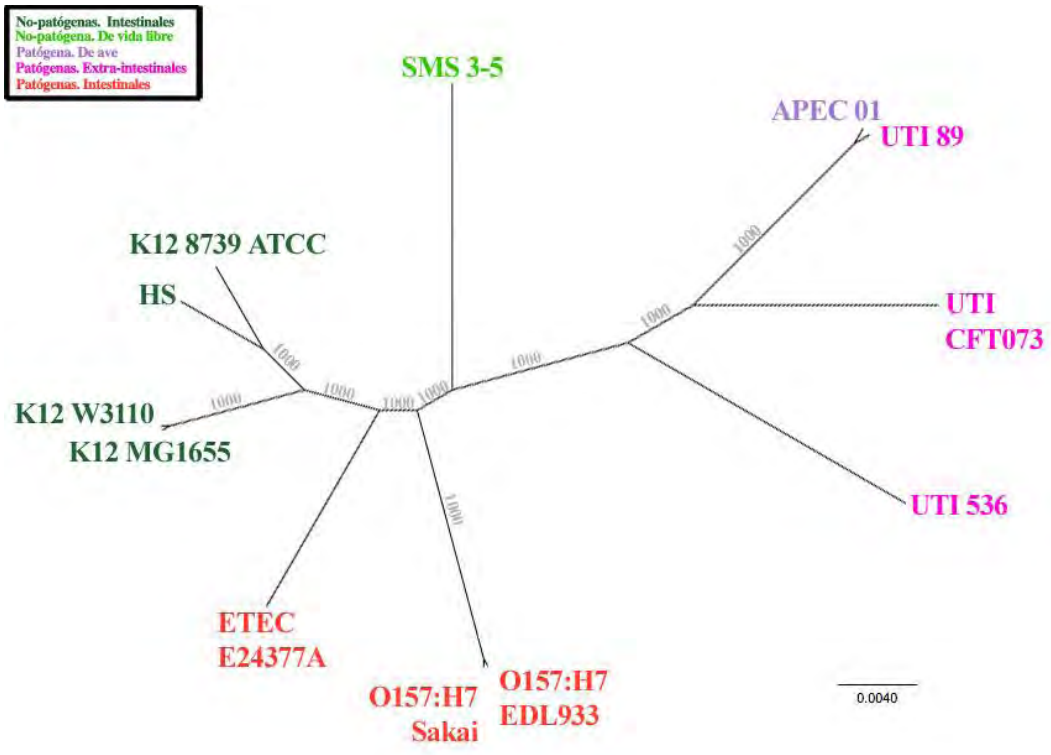


Figura 2. Reconstrucción filogenética de Máxima Verosimilitud (ML), a partir de la información de secuencia del genoma central del LCB 19 de *E. coli*, realizada con el programa PHYML 3.0 (Guindon y Gascuel, 2003). Valores de *bootstrap* obtenidos a partir de 1,000 réplicas.

2. GENÓMICA COMPARADA: DINÁMICA EVOLUTIVA DEL CROMOSOMA DE *E. COLI*

2.1. Variación en la estructura cromosómica de *E. coli*

Los detalles sobre el alineamiento de MAUVE y la revisión del mismo, así como la información complementaria para la generación del alineamiento, se muestran en el Apéndice 1.

La alineación cromosómica con MAUVE identificó 21 bloques de colinearidad local (LCBs) que corresponden a 21 regiones del cromosoma donde la sintenia de regiones conservadas se mantiene. Lo que implica que en el cromosoma de *E. coli* hay al menos 21 regiones sujetas a rearreglo genómico (Figura 3).

El tamaño de los LCBs fue poco homogéneo, abarcando varios órdenes de magnitud (cf. Tabla 6). El más pequeño tuvo una longitud de 144 sitios homólogos (LCB 14) y el más grande fue de casi 3 millones de posiciones nucleotídicas (LCB 19). Los alineamientos de cada LCB concatenados de acuerdo a su orden en el cromosoma de la cepa K12 MG1655 dieron un alineamiento final con una extensión total de 11, 505,324 posiciones nucleotídicas que comprenden 3, 634,829 sitios homólogos a todas las cepas.

En la Figura 3 se puede apreciar la posición de los LCBs en el cromosoma de cada individuo, lo que permite visualizar las regiones donde han ocurrido rearreglos cromosómicos. En esta Figura, las translocaciones son identificadas por las líneas que conectan, y que cruzan entre un cromosoma y otro. Las inversiones son identificadas por aquellos LCBs que se posicionan por debajo de las líneas centrales en cada secuencia.

De esta manera, se observó que la mayoría de los rearreglos se localizan en las regiones de origen y término de replicación, correspondiendo principalmente a inversiones cromosomales. Por ejemplo alrededor del origen de replicación, la región que abarca los LCBs 20 a 21 y 1 a 6 está completamente invertida en la cepa K12 W3110, y los LCBs 2, 5 y 6 están invertidos y translocados en la cepa UTI89.

En la cepa SMS3-5 de vida libre, la región que va del LCB 10 al 18, que incluye la zona del término de replicación, se encuentra invertida. Y se detectó, como ya se había descrito en estudios previos (Darling et al. 2004) la presencia de una inversión cromosómica en la cepa O157:H7 EDL933 con respecto a la cepa K12 MG1655 (Perna et al. 2001), en el LCB 12, justo donde se encuentra el término de replicación.

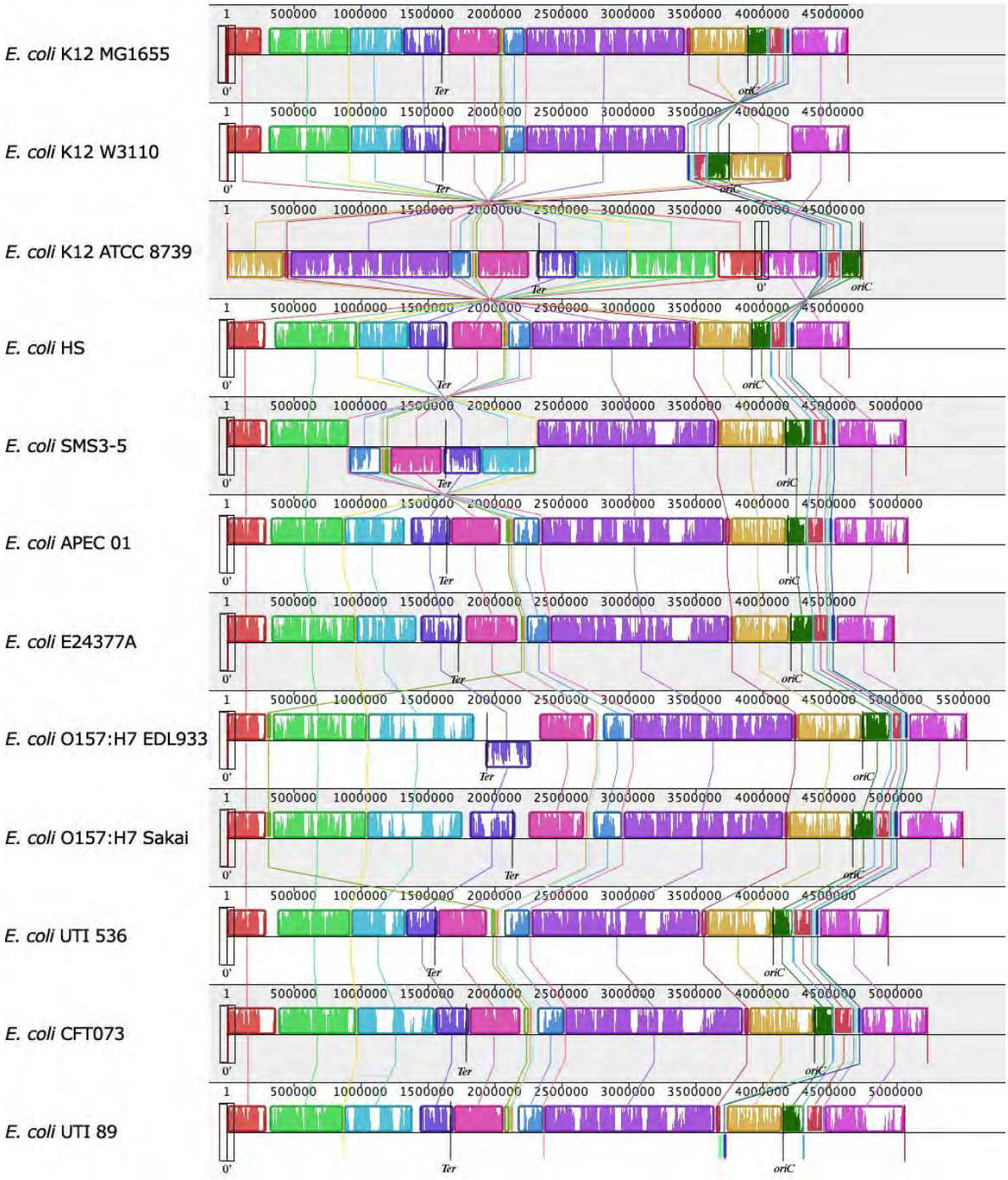


Figura 3. Alineamiento múltiple del cromosoma de *E. coli,* generado con el programa MAUVE. Cada renglón corresponde a la secuencia cromosómica de un individuo. En la parte superior del renglón se indican las coordenadas del cromosoma, donde el 1 corresponde al inicio del reloj cromosomal (*0'*). Cada LCB está representado por un color diferente. Los LCBs homólogos están conectados entre individuos por una línea de su mismo color. Las regiones sin color dentro y entre LCBs representan regiones que son parte del genoma flexible. La altura de las barras de color al interior de los LCBs es proporcional a la similitud de secuencia promedio de la región. *oriC* – posición del origen de replicación; *Ter* – posición del término de replicación.

El número de eventos de rearreglo cromosómico entre pares de individuos se muestra en la Tabla 5. El número de rearreglos detectados entre individuos de un mismo ecogrupo fue desde 1 hasta 4 rearreglos (celdas sombreadas de la Tabla 5). Notablemente, se detectaron rearreglos cromosómicos inclusive entre cepas diferentes de la misma clona como son las 3 cepas de K12 en el ecogrupo de no-patógenas ó las 2 cepas de EHEC O157:H7 en el ecogrupo de cepas patógenas.

En contraste, el número de rearreglos entre cepas de diferente ecogrupo, fue un poco más elevado que entre cepas del mismo ecogrupo (celdas no sombreadas de la Tabla 5). Así, el mayor número de rearreglos se detectó entre la cepa patógena extra-intestinal UTI 89 y la cepa comensal (aunque adaptada al laboratorio) K12 8739ATCC. Notablemente entre la cepa ETEC E24377A patógena y la K12 MG1655 no-patógena, solamente se detectó un rearreglo. También se identificaron rearreglos entre diferentes cepas de la misma clona de *E. coli*. Efectivamente, como se ve en la Figura 3 y se resume en la Tabla 5, se observaron hasta 4 rearreglos cromosómicos entre las cepas de *E. coli* K12, en las cepas MG1655 y W3110 con la cepa 8739ATCC, y solamente 1 evento de inversión entre las cepas Sakai y EDL933 del serotipo O157:H7.

Tabla 5. Número mínimo de rearreglos cromosómicos detectados por el programa MAUVE entre pares de los 12 cromosomas de *E. coli* analizados.

	Cepa	1	2	3	4	5	6	7	8	9	10	11	12
1	**k12 MG1655**	0	-	-	-	-	-	-	-	-	-	-	-
2	**k12 W3110**	1*	0	-	-	-	-	-	-	-	-	-	-
3	**k12 8739ATCC**	4*	4*	0	-	-	-	-	-	-	-	-	-
4	**HS**	3	1	4	0	-	-	-	-	-	-	-	-
5	**SMS-3-5**	1	2	4	2	0	-	-	-	-	-	-	-
6	**APEC-O1**	3	4	7	2	4	0	-	-	-	-	-	-
7	**ETEC E24377A**	1	2	5	2	2	2	0	-	-	-	-	-
8	**O157 EDL933**	4	4	8	5	5	4	3	0	-	-	-	-
9	**O157 Sakai**	3	3	7	4	4	3	2	1*	0	-	-	-
10	**UTI 536**	4	5	7	3	6	1	3	4	3	0	-	-
11	**UTI CFT073**	3	5	7	3	6	1	3	4	3	0	0	-
12	**UTI 89**	5	6	9	4	5	1	3	4	3	2	2	0

*Rearreglos entre cepas distintas de la misma clona

Las celdas sombreadas corresponden al número de eventos de recombinación entre individuos del mismo ecogrupo.

Las regiones sombreadas en la Tabla 5 destacan el número de rearreglos que se identificaron entre individuos de un mismo ecogrupo. Comparando estas cifras, vemos que en ambos ecogrupos el número más grande de rearreglos fue 4. Dentro del ecogrupo de cepas no-patógenas este número de rearreglos se registró entre la cepa K12 ATCC8739 y las demás del mismo ecogrupo. Dentro del ecogrupo de cepas patógenas, el número de rearreglos más grande se presentó entre la cepa O157:H7 EDL933 y el resto de los individuos de dicho ecogrupo.

En contraste, el número de eventos de rearreglo identificados entre individuos de diferentes ecogrupos fue mucho mayor, y se presentaron hasta 9 rearreglos, entre la cepa UTI89 y la cepa K12 ATCC8739. Por otra parte, también se puede ver en la Tabla 5 que el número mínimo de rearreglos cromosómicos entre las cepas de *E. coli* K12 es similar al que existe entre las cepas comensales en general y la cepa ETEC E24377A patógena, lo que implica que al menos en cuanto a estructura cromosómica esta cepa de ETEC es similar a las *E. coli* comensales.

2.2 Variación en el repertorio genético del cromosoma de *E. coli*: componentes del pangenoma

La delimitación de los elementos correspondientes al genoma central y al genoma flexible en el cromosoma de *E. coli* resultó en la identificación de 1681 regiones. De éstas 843 pertenecieron al genoma flexible, ya sea compartidos por un subgrupo de la muestra ó presentes en un solo individuo. Las 841 regiones restantes correspondieron al genoma central.

Al genoma flexible corresponde 7, 870,495 sitios en el alineamiento, lo que representa casi el 70% del pangenoma de esta muestra. En las Figuras 3 y 4a el genoma flexible se representa con regiones en blanco al interior de los LCBs y entre ellos. El genoma flexible consta de secuencias que pueden estar presentes en un solo individuo ó que pueden ser compartidas por un subgrupo de la muestra, como por ejemplo entre los individuos del mismo ecogrupo, como se muestra en la Figura 4b. En esta figura se indica el porcentaje de regiones correspondientes al genoma flexible compartido por los individuos de un mismo ecogrupo.

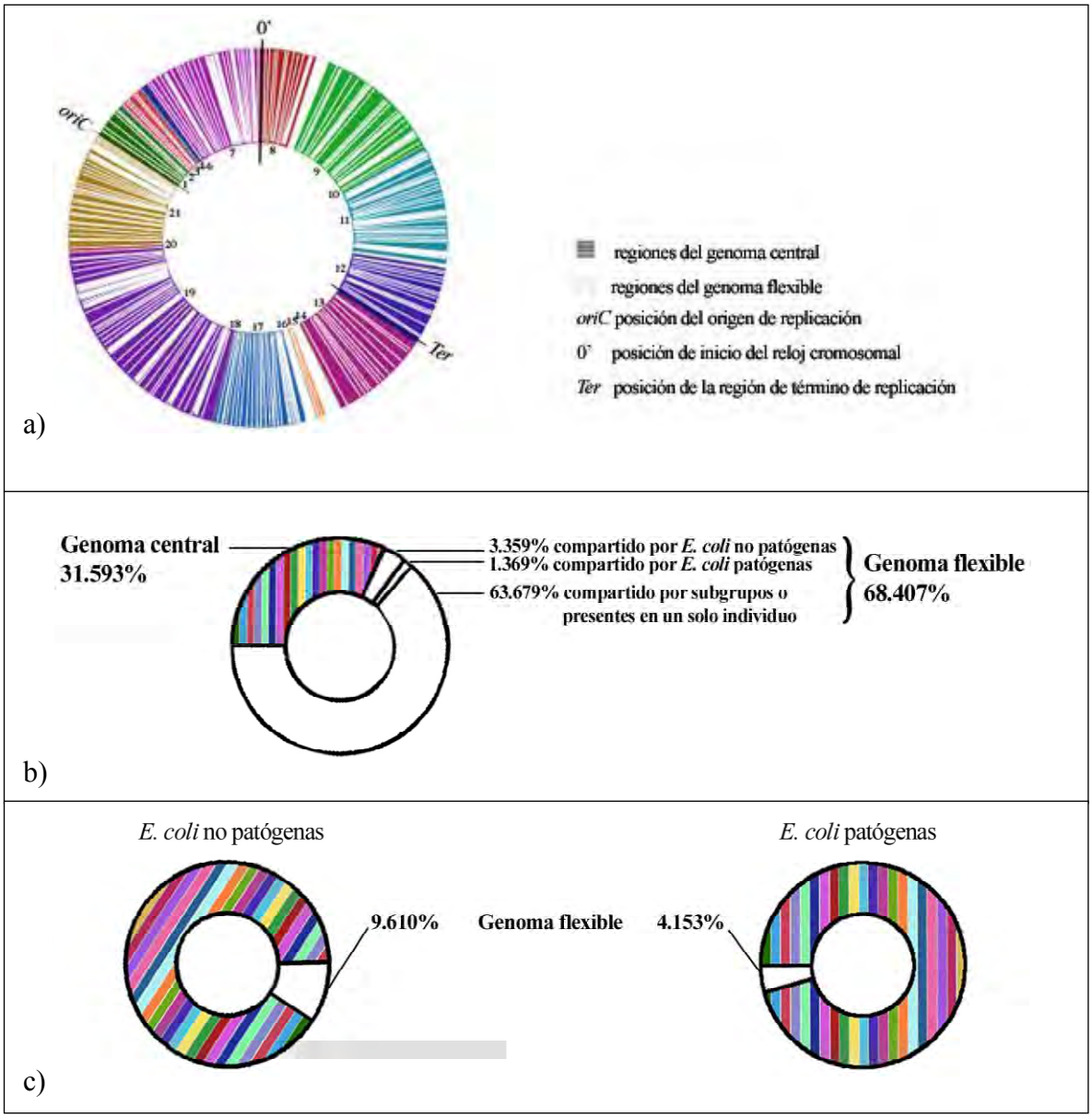


Figura 4. a) Distribución de regiones del genoma central y del genoma flexible en el alineamiento consenso de cromosoma de *E. coli*. Los LCBs se ordenaron de acuerdo al cromosoma de la cepa K12 MG1655. Los colores que representan cada LCB son los mismos que asignó el programa MAUVE, como se ve en la Figura 3. Adicionalmente asignamos un número a cada LCB, el cual se indica en el interior del círculo cromosomal, comenzando por el LCB verde oscuro, donde se encuentra el origen de replicación *oriC*.

La proporción de regiones de los dos componentes del pangenoma en cada LCB se detalla en la Tabla 6.

b) Proporción de sitios del alineamiento correspondientes al genoma central y al genoma flexible compartido por los individuos de un mismo ecogrupo. La distribución de regiones del genoma flexible en los LCB para cada ecogrupo se muestra en la Tabla 7.

c) Proporción de sitios del alineamiento correspondientes al genoma flexible, relativa al número de sitios analizados en cada ecogrupo.

Las regiones del genoma central y del genoma flexible se distribuyeron de manera uniforme en todo el cromosoma, y se distribuyeron aproximadamente al 50% en cada replicón (416 regiones del genoma central y 411 del genoma flexible en el primer replicón vs. 425 regiones del genoma central y 422 del genoma flexible en el segundo replicón).

Adicionalmente, como se puede apreciar en la Tabla 6 los LCBs grandes fueron los más segmentados, como en el LCB 19, el de mayor tamaño, donde se registraron 194 regiones del genoma central y 195 regiones del genoma flexible. En general, las regiones del genoma flexible fueron más numerosas que las del genoma central. Aunque algunos de los LCBs más pequeños no presentaron regiones del genoma flexible, como son los LCBs 14 y 16.

Tabla 6. Tamaño de los 21 bloques localmente colineares (LCBs) identificados en *E. coli*, y proporción de los componentes del pangenoma, genoma central y genoma flexible, en cada uno de ellos.

LCB	Número de regiones			Número y porcentaje de sitios					
	Genoma central	Genoma flexible	Total	Genoma central		Genoma flexible		Total	
oriC 1	33	33	**66**	161819	1.406%	72100	0.627%	**233919**	**2.033%**
2	2	1	**3**	446	0.004%	4100	0.036%	**4546**	**0.040%**
3	23	22	**45**	82973	0.721%	78248	0.680%	**161221**	**1.401%**
4	3	3	**6**	5204	0.045%	375	0.003%	**5579**	**0.048%**
5	2	1	**3**	4376	0.038%	17	0.0001%	**4393**	**0.0381%**
6	7	7	**14**	35455	0.308%	2070	0.018%	**37525**	**0.326%**
7	89	88	**177**	309466	2.690%	1180492	10.260%	**1489958**	**12.950%**
0' 8	63	62	**125**	239647	2.083%	476139	4.138%	**715786**	**6.221%**
9	121	121	**242**	487919	4.241%	494387	4.297%	**982306**	**8.538%**
10	2	3	**5**	144	0.001%	1691	0.015%	**1835**	**0.016%**
11	71	70	**141**	325315	2.828%	1376560	11.965%	**1701875**	**14.792%**
Ter 12	53	52	**105**	200555	1.743%	352503	3.064%	**553058**	**4.807%**
13	50	51	**101**	345495	3.003%	336957	2.929%	**682452**	**5.932%**
14	1	0	**1**	144	0.001%	0	0%	**144**	**0.001%**
15	5	4	**9**	9483	0.082%	47040	0.409%	**56523**	**0.491%**
16	1	0	**1**	4314	0.037%	0	0%	**4314**	**0.037%**
17	35	34	**69**	107165	0.932%	547582	4.759%	**654747**	**5.691%**
18	2	3	**5**	273	0.002%	1489	0.013%	**1762**	**0.015%**
19	194	195	**389**	968897	8.422%	2025999	17.609%	**2994896**	**26.031%**
20	7	7	**14**	31009	0.270%	19709	0.171%	**50718**	**0.441%**
21	77	76	**153**	314730	2.736%	482698	4.195%	**797428**	**6.931%**
interLCB	-	10	**10**	-	-	370339	3.219%	**370339**	**3.219%**
Total	841	843	**1684**	3634829	31.593%	7870495	68.407%	**11505324**	**100%**

El porcentaje es relativo al total de sitios del alineamiento final (11, 505,324 sitios).

oriC - origen de replicación; 0' - inicio del reloj cromosomal; *Ter* - término de replicación.

En la Tabla 7 se muestra la distribución de regiones del genoma flexible de cada ecogrupo. Aproximadamente el 3.4% del genoma flexible es común al ecogrupo de las *E. coli* no-patógenas, y un 1.4% es compartido por las cepas patógenas (Figura 4b). Así, las cepas no-patógenas presentaron casi el doble de secuencias del genoma flexible que las cepas patógenas. En los LCBs 6, 8 y 17 predominan las regiones flexibles compartidas por las cepas patógenas. En contraste, hay mucho más LCBs donde predominan las secuencias del genoma flexible compartidas por cepas no-patógenas (LCBs 1, 3-5, 7, 9-13, 15, 18-21).

Tabla 7. Proporción de regiones del **genoma flexible** de los dos ecogrupos analizados: *E. coli* patógenas y *E. coli* no-patógenas, en los 21 bloques localmente colineares (LCBs) determinados por MAUVE.

LCB		Número de regiones del genoma flexible		Número y porcentaje de sitios del genoma flexible			
		No Patógenas	Patógenas	No Patógenas		Patógenas	
oriC	*1*	6	7	6787	0.059%	2927	0.025%
	2	0	0	0	0%	0	0%
	3	12	3	12047	0.105%	1863	0.016%
	4	2	0	249	0.002%	0	0%
	5	1	0	17	0.0001%	0	0%
	6	1	2	297	0.003%	591	0.005%
	7	34	31	31263	0.272%	6901	0.060%
0'	*8*	10	15	3802	0.033%	6992	0.061%
	9	50	39	42435	0.369%	19572	0.170%
	10	3	3	1628	0.014%	63	0.001%
	11	29	18	35658	0.310%	2926	0.025%
Ter	*12*	27	10	66204	0.575%	8902	0.077%
	13	27	23	33651	0.292%	12320	0.107%
	14	-	-	-	-	-	-
	15	4	0	7461	0.065%	0	0%
	16	-	-	-	-	-	-
	17	15	15	10587	0.092%	35783	0.311%
	18	1	0	91	0.001%	0	0%
	19	93	59	89398	0.777%	22769	0.198%
	20	2	0	580	0.005%	0	0%
	21	30	28	44261	0.385%	28900	0.251%
	interLCB	1	1	50	0.0004%	7000	0.061%
	Total	348	254	386466	3.359%	157509	1.369%

El porcentaje es relativo al total de sitios del alineamiento final (11, 505,324 sitios).

-Estos LCBs sólo tienen regiones del genoma central.

oriC indicado en el LCB 1 donde se encuentra la región de inicio de replicación,

0' indicado en el LCB 8 donde se encuentra el inicio del reloj cromosomal,

Ter indicado en el LCB 12 donde se encuentra la región de término de replicación.

Finalmente, se observó que los niveles de identidad nucleotídica del genoma central, entre pares de cepas, son altos (Tabla 8). La identidad más baja, de 0.972, se presentó entre cepas de EHEC (O157:H7) y UPEC-APEC, y la más alta, 0.999, entre las dos cepas de EHEC (O157:H7) y entre las cepas K12 MG1655 y K12 W3110. Es decir, que inclusive al nivel del genoma central, se pueden detectar diferencias entre clonas de una misma cepa.

Tabla 8. Matriz de identidad nucleotídica de las regiones del genoma central de los 12 cromosomas de *E. coli*.

	Cepa	1	2	3	4	5	6	7	8	9	10	11	12
1	**k12 MG1655**	*1.00*	-	-	-	-	-	-	-	-	-	-	-
2	**k12 W3110**	0.999	*1.00*	-	-	-	-	-	-	-	-	-	-
3	**k12 8739ATCC**	0.992	0.992	*1.00*	-	-	-	-	-	-	-	-	-
4	**HS**	0.99	0.99	0.993	*1.00*	-	-	-	-	-	-	-	-
5	**SMS-3-5**	0.976	0.976	0.976	0.975	*1.00*	-	-	-	-	-	-	-
6	**APEC-O1**	0.973	0.973	0.973	0.973	0.976	*1.00*	-	-	-	-	-	-
7	**ETEC E24377A**	0.987	0.987	0.987	0.988	0.976	0.973	*1.00*	-	-	-	-	-
8	**O157 EDL933**	0.983	0.983	0.983	0.982	0.975	0.972	0.982	*1.00*	-	-	-	-
9	**O157 Sakai**	0.983	0.983	0.983	0.982	0.975	0.972	0.982	0.999	*1.00*	-	-	-
10	**UTI 536**	0.973	0.973	0.973	0.973	0.977	0.991	0.973	0.972	0.972	*1.00*	-	-
11	**UTI CFT073**	0.973	0.973	0.973	0.973	0.976	0.991	0.973	0.972	0.972	0.991	*1.00*	-
12	**UTI 89**	0.973	0.973	0.973	0.973	0.976	0.999	0.973	0.972	0.972	0.991	0.992	*1.00*

En conjunto, estos resultados muestran que hay mucha variación en cuanto a la estructura del cromosoma de esta bacteria, por lo que los eventos de rearrreglo, translocaciones e inversiones y la transferencia horizontal deben ser frecuentes. En contraste, los niveles de identidad nucleotídica encontrados indican que las regiones del genoma central son aparentemente estables.

3. GENÓMICA DE POBLACIONES

3.1 Delimitación de la unidad de análisis: loci

La distribución de frecuencias de los parámetros estimados en los seis diferentes niveles de análisis se muestran en el Apéndice 3. Como fue revelado por las pruebas de Shapiro-Wilk y KSL ningún parámetro, a ninguno de los 6 niveles de análisis se distribuye de manera normal, por lo que se usaron estadísticos de comparación no-paramétricos.

Tabla 9. Diversidad, pruebas de selección natural y desequilibrio de ligamiento en el genoma central del cromosoma de *E. coli*, a diferentes niveles de análisis.

	Nivel de análisis (tamaño de ventana)					
Parámetro	**3**	**10**	**100**	**1,000**	**10,000**	**100,000**
N	*1,209,955*	*362,987*	*36,299*	*3,628*	*363*	*37*
N_P	*184,920*	*138,595*	*33,926*	*3,628*	*363*	*37*
N_{DL}	*6,270*	*25,398*	*26,546*	*3,620*	*363*	*37*
π [1]	0.0217819	0.0217819	0.0217819	0.0217819	0.0217819	0.0217819
(varianza)	(0.00275905)	(0.001101908)	(0.00031559)	(0.00012699)	(4.0778E-05)	(1.0928E-05)
θ [1]	0.0195899	0.0195899	0.0195899	0.0195899	0.0195899	0.0195899
(varianza)	(0.00192943)	(0.000769197)	(0.00022792)	(0.000099802)	(0.000032881)	(8.5048e-6)
Hd [1]	0.0819524	0.2169754	0.7020735	0.9270682	0.9709142	0.9864865
(varianza)	(0.04078275)	(0.085064097)	(0.05129034)	(0.00200145)	(9.7356E-05)	(4.8258E-05)
Zns [2]	0.6593931*	0.5817063*	0.4234298*	0.3536454*	0.3099458	0.2881486
(varianza)	(0.152831973)	(0.143342958)	(0.064099555)	(0.011739918)	(0.003266683)	(0.000498995)
D de Tajima [3]	0.2416228**	0.2506166**	0.3307601**	0.4480764**	0.4842147	0.4964335
(varianza)	(1.04216576)	(1.050689781)	(0.82291876)	(0.29527421)	(0.08358847)	(0.01447664)
D* de Fu-Li [3]	0.1604474**	0.1722623**	0.24169**	0.3306628**	0.3517829	0.3598978
(varianza)	(0.89248152)	(0.916197955)	(0.77201212)	(0.30246348)	(0.09300359)	(0.01646553)
F* de Fu-Li [3]	0.2030854**	0.2159971**	0.2997007**	0.4121553**	0.4414876	0.4520686
(varianza)	(1.06777854)	(1.09879955)	(0.92688641)	(0.36035721)	(0.10934483)	(0.01943732)

[1, 2, 3] Promedio. Estimado sobre [1] número de loci del genoma central (N), [2] número de loci con sitios informativos para la prueba Zns (N_{DL}) y [3] número de loci polimórficos (N_P), de cada nivel de análisis.

Niveles de análisis en los que se registraron ventanas significativas * positivas ó *negativas en las pruebas de neutralidad, y * significativas en la prueba de desequilibrio de ligamiento Zns. La proporción de ventanas significativas en cada nivel de análisis se muestra en la Figura 5.

Los valores promedio de diversidad nucleotídica π fueron iguales en todas las longitudes de ventana, pues al ser una misma muestra analizada, la diversidad es igual en todos los casos (Tabla 9). Sin embargo se observó que la varianza fue significativamente distinta para cada nivel de análisis, aumentando a medida que la longitud de ventana se reducía (Friedman p <0.0001 para ambos parámetros).

En cuanto a la diversidad de haplotipos, ésta fue diferente de acuerdo al tamaño de ventana (Tabla 9). Y a medida que la diversidad de haplotipos es mayor (en las ventanas de mayor longitud), el valor Zns se acerca más a 0, es decir que hay menor desequilibrio de ligamiento, probablemente por el efecto de la recombinación. Congruentemente las varianzas son significativamente distintas, y de la misma manera que ocurrió en los demás parámetros la varianza fue mayor en las ventanas de menor longitud (Friedman p<0.0001).

En la Figura 5 se muestra además la proporción de loci con evidencia de clonalidad. En las ventanas de 10,000 y 100,000 sitios no se presentaron regiones significativamente distintas de 0 en la prueba Zns. Es a partir de la ventana de 1,000 sitios que empiezan a presentarse regiones en desequilibrio de ligamiento y la proporción va aumentando a medida que la longitud de la ventana se reduce. Así, la ventana de 3 sitios es la que presenta mayor proporción de loci clonales, casi 50%. Si calculamos este resultado con respecto al número de sitios que representan estos loci, tenemos que al nivel de análisis más pequeño corresponden 9 960 nucleótidos con evidencias de desequilibrio de ligamiento, lo que corresponde a una fracción muy pequeña del genoma central (9,960 / 3,634,829 sitios totales = 0.0027 sitios en desequilibrio de ligamiento a la escala de 3).

De esta manera, es el nivel de análisis de 100 sitios el que abarca más sitios del genoma central con evidencia de clonalidad (363,400 / 3,634,829 sitios totales = 0.0998).

En cuanto a evidencias de selección, los valores de las pruebas de neutralidad aplicadas al total de regiones del genoma central fueron positivos en las tres pruebas. En todos los casos se aceptó la hipótesis de neutralidad (no significativas a p <0.05).

En este caso, las medias de las 3 pruebas fueron más pequeñas a medida que se reducía la longitud de ventana cuyas varianzas fueron diferentes significativamente (Friedman p <0.0001). En las ventanas de longitud de 10,000 y 100,000 nucleótidos no se presentaron regiones que rechazaran la hipótesis de neutralidad.

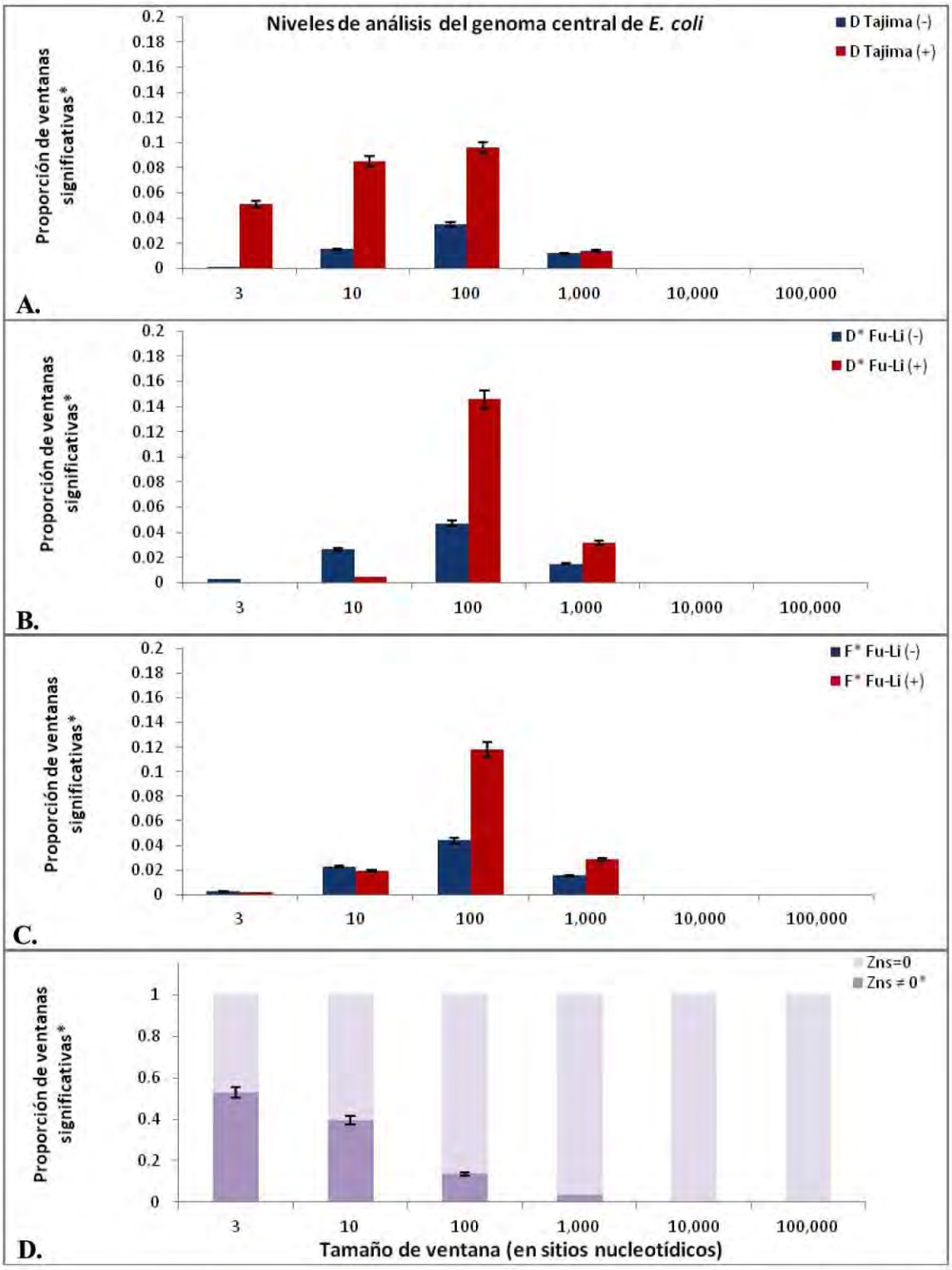


Figura 5. Proporción de ventanas significativas* en las pruebas de A. D de Tajima, B. D* de Fu-Li, C. F* de Fu-Li y D. desequilibrio de ligamiento Zns, para cada nivel de análisis del cromosoma de *E. coli*. Los datos se muestran en el Apéndice 4.

Por el contrario desde la ventana de 1,000 nucleótidos hasta la de 3, sí se encontraron regiones con evidencia de selección en las tres pruebas. Las proporciones relativas de loci significativamente positivos y negativos se muestran en la Figura 4, donde se observa que la ventana de 100 nucleótidos de longitud fue la que presentó mayor proporción de loci con evidencia de selección en las tres pruebas.

Estos análisis nos muestran que las evidencias de selección y clonalidad sólo se encuentran a partir de las ventanas de longitud de 1000 nucleótidos. Y que pudiera ser más informativo tomar como loci regiones de longitud pequeña. Por esta razón para los análisis subsiguientes consideramos los loci dentro de las regiones de genoma central y de genoma flexible respectivamente, y dado que el tamaño de estas regiones fue heterogéneo (desde 10 hasta 32,74 sitios), las regiones de mayor longitud se subdividieron en ventanas de 1,000 sitios, como se muestra en la Figura 6. Así, las regiones del genoma central ó flexible con longitud menor ó igual a 1,000 sitios se analizaban completas. Entonces, el tamaño final de los loci se encontró entre los 10 y los 1,000 sitios.

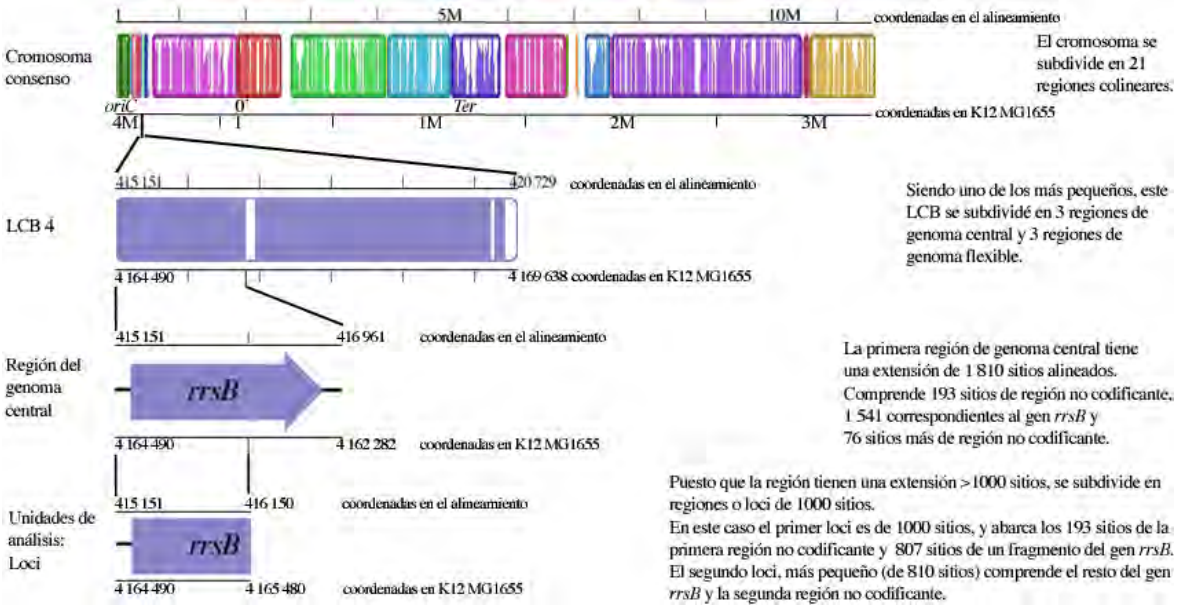


Figura 6. Esquema de las regiones que componen al alineamiento, desde los LCBs, hasta la unidad de análisis final ó loci, que fueron las ventanas de ≤1000 sitios dentro de las regiones correspondientes a genoma central y genoma flexible de cada LCB.

3.2. Diversidad genética en el cromosoma de *E. coli*

Se analizó un total de 4,122 loci del genoma central en *E. coli,* los cuales abarcan 3,629,864 sitiosñ De estos, 4096 loci fueron polimórficos, con 197,251 sitios segregantes, correspondientes a una proporción P_S de 0.0543.

En el grupo de *E. coli* no-patógenas, se analizaron un total de 4,665 loci (4,122 del genoma central y 543 del genoma flexible) con una proporción de loci polimórficos alta (P= 4,535 / 4,665 loci totales = 0.9721) y una proporción de sitios segregantes P_S de 0.0306 (P_S=119,077 / 3, 893,922 sitios totales).

En las cepas patógenas se analizaron 4,376 loci, que corresponden a un poco menos loci que en el grupo de las no-patógenas (igualmente, 4,122 del genoma central, pero 254 del genoma flexible). La proporción de loci polimórficos fue de 0.9835 (P=4,304/4,376 loci totales) y la proporción de sitios segregantes de 0.0418 (P_S=155,535/3,718,841).

Como se puede observar en la Tabla 10, la diversidad del genoma central de toda la muestra fue de $\pi = 0.0217819 \pm 0.0002824$, con $\theta = 0.0195899 \pm 0.0002313$. Estos valores fueron más grandes que los encontrados en cualquiera de los ecogrupos.

Asimismo, dentro de los ecogrupos, la mayor diversidad, estimada con π y con θ se encontró en el genoma flexible (Wilcoxon $p < 0.0001$). En contraste, la diversidad de haplotipos en el genoma flexible fue significativamente menor (Wilcoxon $p < 0.0001$) que el genoma central, en ambos ecogrupos.

Adicionalmente, se estimó la diversidad en cada subgrupo de cepas patógenas, intestinales y extraintestinales.

En las cepas extraintestinales se analizó un total de 4,959 loci, de los cuales 4,652 presentaron 68,758 sitios segregantes, correspondientes a una P_S= 68,758/ 4,033,628 = 0.0170. Del genoma central se analizaron 4,122 loci, 136 sin variación. Los 3,986 loci restantes presentaron 6,725 sitios segregantes.

En las cepas patógenas intestinales el patrón fue similar. En este ecogrupo se analizaron 5,232 loci en total, con una proporción P_S= 55,535/ 4,276,738 = 0.01299. Y la diversidad dada por los parámetros π y θ fue mayor en el genoma flexible. Los estimados de π y θ promedio fueron significativamente más altos en el genoma flexible que en el genoma

central (Wilcoxon p<0.0001). Siendo en general, las cepas extraintestinales más diversas que la cepas patógenas intestinales, tanto en el genoma central ($\pi_{extraintestinal}$= 0.0125736 ± 0.0002437; $\pi_{intestinal}$= 0.0090948 ± 0.0001952; Wilcoxon significativa a p<0.0001), como en el genoma flexible ($\pi_{extraintestinal}$= 0.0200723 ± 0.0009842; $\pi_{intestinal}$= 0.0129991 ± 0.0008608; Wilcoxon significativa a p<0.0001)

Tabla 10. Diversidad, pruebas de selección natural y desequilibrio de ligamiento en los componentes del pangenoma del cromosoma de *E. coli*, en los dos ecogrupos y en la muestra total.

Ecogrupo	Parámetro	Componente del pangenoma: Genoma central	Componente del pangenoma: Genoma flexible (compartido por cada ecogrupo)
No-patógenas	*N* [1]	*4,122*	*543*
	N_P [2]	*4,058*	*477*
	N_{DL} [3]	*2,628*	*231*
	π [1]	0.0148857	0.0251052
	θ [1]	0.0157071	0.0264413
	Hd [1]	0.7501213	0.6338858
	Zns [2]	0.7282182*	0.7732863*
	D de Tajima [3]	**-0.454711****	**-0.467283****
	D* de Fu-Li [3]	**-0.498445****	**-0.5052****
	F* de Fu-Li [3]	**-0.522249****	**-0.523444****
Patógenas	*N*	*4,122*	*254*
	N_P	*4,085*	*219*
	N_{DL}	*3,992*	*177*
	π [1]	0.0208802	0.0444992
	θ [1]	0.0184001	0.0388579
	Hd [1]	0.8732007	0.6882265
	Zns [2]	0.6347493*	0.6694768*
	D de Tajima [3]	**0.7365571****	**0.6695001****
	D* de Fu-Li [3]	**0.5570546****	**0.4874307****
	F* de Fu-Li [3]	**0.6593397****	**0.5790057****
TOTAL	*N*	*4,122*	-
	N_P	*4,096*	-
	N_{DL}	*4,037*	-
	π [1]	0.0217819	-
	θ [1]	0.0195899	-
	Hd [1]	0.9079772	-
	Zns [2]	0.3642279*	-
	D de Tajima [3]	**0.4484568****	-
	D* de Fu-Li [3]	**0.3270199****	-
	F* de Fu-Li [3]	**0.4089824****	-

[1, 2, 3] Promedio. Estimado sobre [1] número de loci del genoma central (N), [2] número de loci con sitios informativos para la prueba Zns (N_{DL}) y [3] número de loci polimórficos (N_P), de cada componente del pangenoma (varianzas de los promedios reportados en el Apéndice 5).

Presencia de loci significativos *positivos ó *negativos en la prueba de neutralidad y * loci significativos en la prueba de desequilibrio de ligamiento Zns. La proporción de loci significativos en cada región del pangenoma, para cada ecogrupo se muestra en la Figura 7.

3.3 Patrones de desequilibrio de ligamiento

Posteriormente se aplicó la prueba de desequilibrio de ligamiento Zns, a los loci informativos para esta prueba (N_{DL}). El valor de Zns en la muestra total fue más cercano a 0 que a uno (Tabla 10), lo que indica de manera general, que el genoma central de *E. coli* no se encuentra en desequilibrio de ligamiento. Efectivamente, en la Figura 7, se muestra que la prueba Zns fue significativamente distinta de 0 en una proporción de loci reducida (solamente en 202 loci, correspondientes al 4.9% del total de loci analizados en el cromosoma).

La comparación de los valores promedio de Zns para el genoma central y el genoma flexible, arrojó que éste último es significativamente más elevado que el genoma central, en ambos ecogrupos (Wilcoxon p= 0.0051). Esto implica que de manera general el genoma central está más cerca del equilibrio de ligamiento que el genoma flexible, es decir que el genoma flexible es más clonal que el genoma central. Efectivamente la proporción de loci significativamente distintos de cero (es decir en desequilibrio de ligamiento) es mayor en el genoma flexible que en el genoma central (Figura 7) en ambos ecogrupos. Estos resultados concuerdan con los valores de diversidad de haplotipo Hd encontrados, en los que se obtuvo que en el genoma flexible la Hd promedio era significativamente menor que en el genoma central (Tabla 10).

Asimismo, el genoma flexible presentó valores de Zns más altos que el genoma central, en el ecogrupo de cepas no-patógenas y de cepas patógenas (Wilcoxon significativa, p= 0.0051 y p = 0.0001 respectivamente). Esto implica que de manera general el genoma central está más cerca del equilibrio de ligamiento que el genoma flexible, es decir que el genoma flexible es más clonal que el genoma central. Efectivamente la proporción de loci significativamente distintos de cero (es decir en desequilibrio de ligamiento) es mayor en el genoma flexible que en el genoma central (Figura 7D). Estos resultados concuerdan con los valores de diversidad de haplotipo Hd encontrados, en los que se obtuvo que en el genoma flexible la Hd es significativamente menor que en el genoma central.

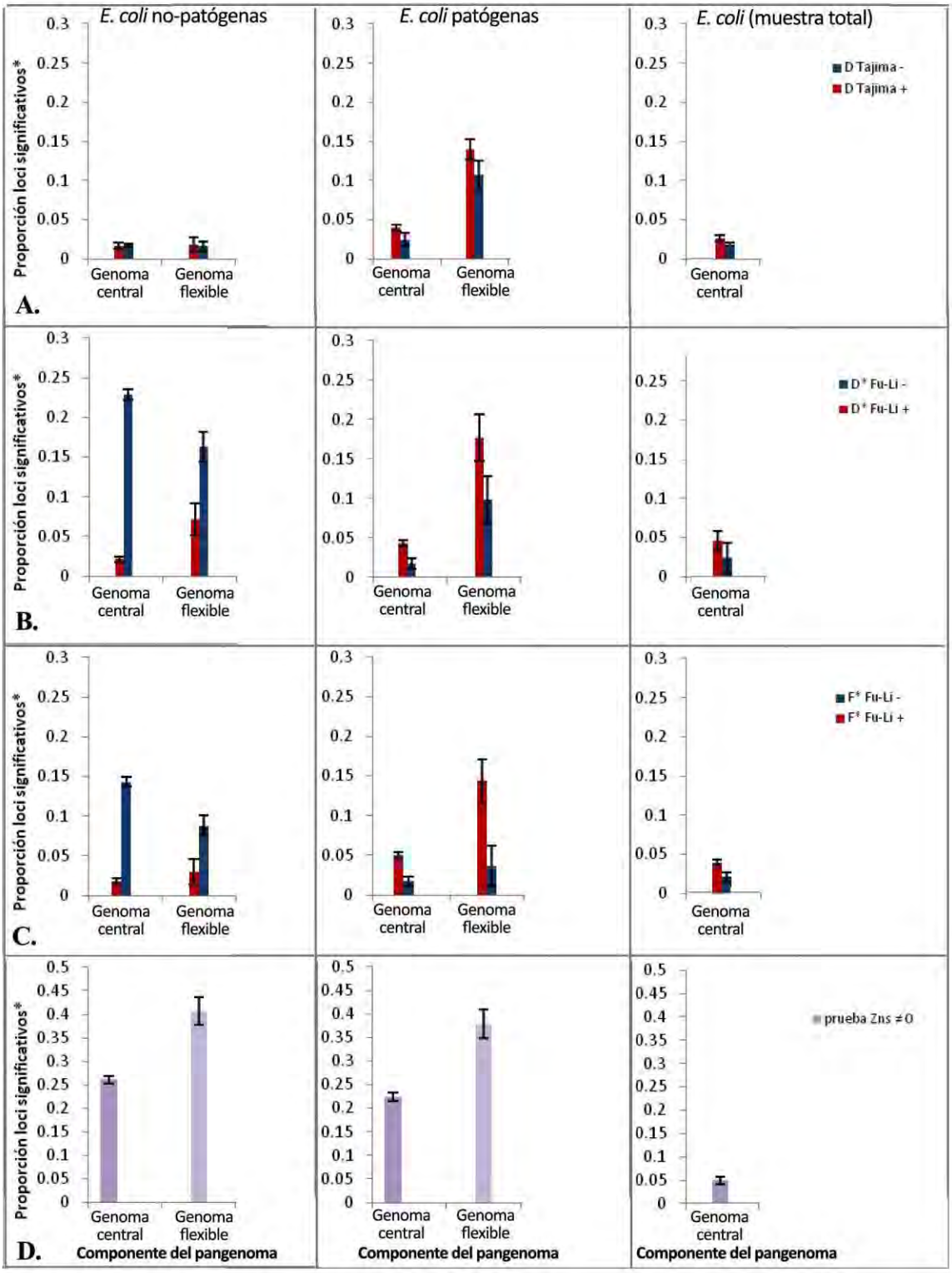


Figura 7. Proporción de loci significativos* en las pruebas de A) D de Tajima, B) D* de Fu-Li, C) F* de Fu-Li y D) desequilibrio de ligamiento Zns, en los ecogrupos y en la muestra total de de *E. coli*.

* Significativos a $p<0.05$ en las pruebas de D de Tajima, D* y F* de Fu-Li, y a $p<0.025$ en la prueba de desequilibrio de ligamiento Zns.

3.4 Señales de selección natural

El análisis de la distribución de frecuencias polimórficas realizado en los 4, 096 loci polimórficos de la muestra total de *E. coli*, mostró que la mayoría de los loci presentan valores positivos (Tabla 10; Figura 7). Se obtuvo un total de 197 loci significativos positivos y 19 loci significativos negativos, en al menos una de las tres pruebas de neutralidad aplicadas. Entre los loci bajo selección positiva, 13 correspondieron a regiones no-codificantes del genoma. En los 184 loci bajo selección positiva restantes se encontraron genes en todas las categorías funcionales del JCVI, pero principalmente entre las categorías de biosíntesis de cofactores y grupos prostéticos, transcripción y procesos celulares (Figura 8). Por el contrario, los genes encontrados en los 19 loci bajo selección negativa no se encontraron en todas las categorías funcionales, faltando en aquellas de biosíntesis de cofactores y grupos prostéticos, elementos móviles y extracromosomales, funciones desconocidas, funciones regulatorias, metabolismo de ADN, metabolismo intermediario central y metabolismo de nucleótidos (Figura 8).

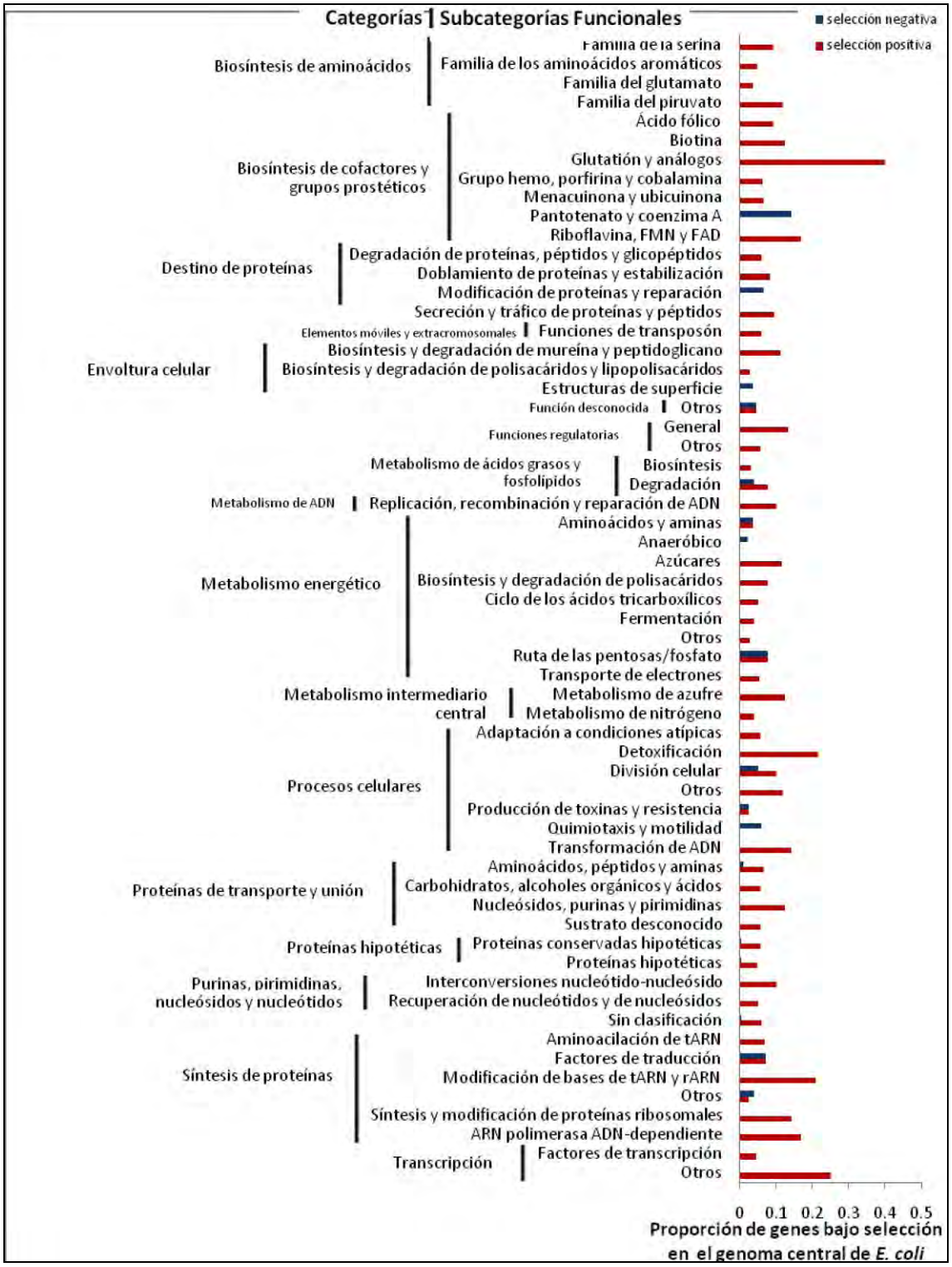
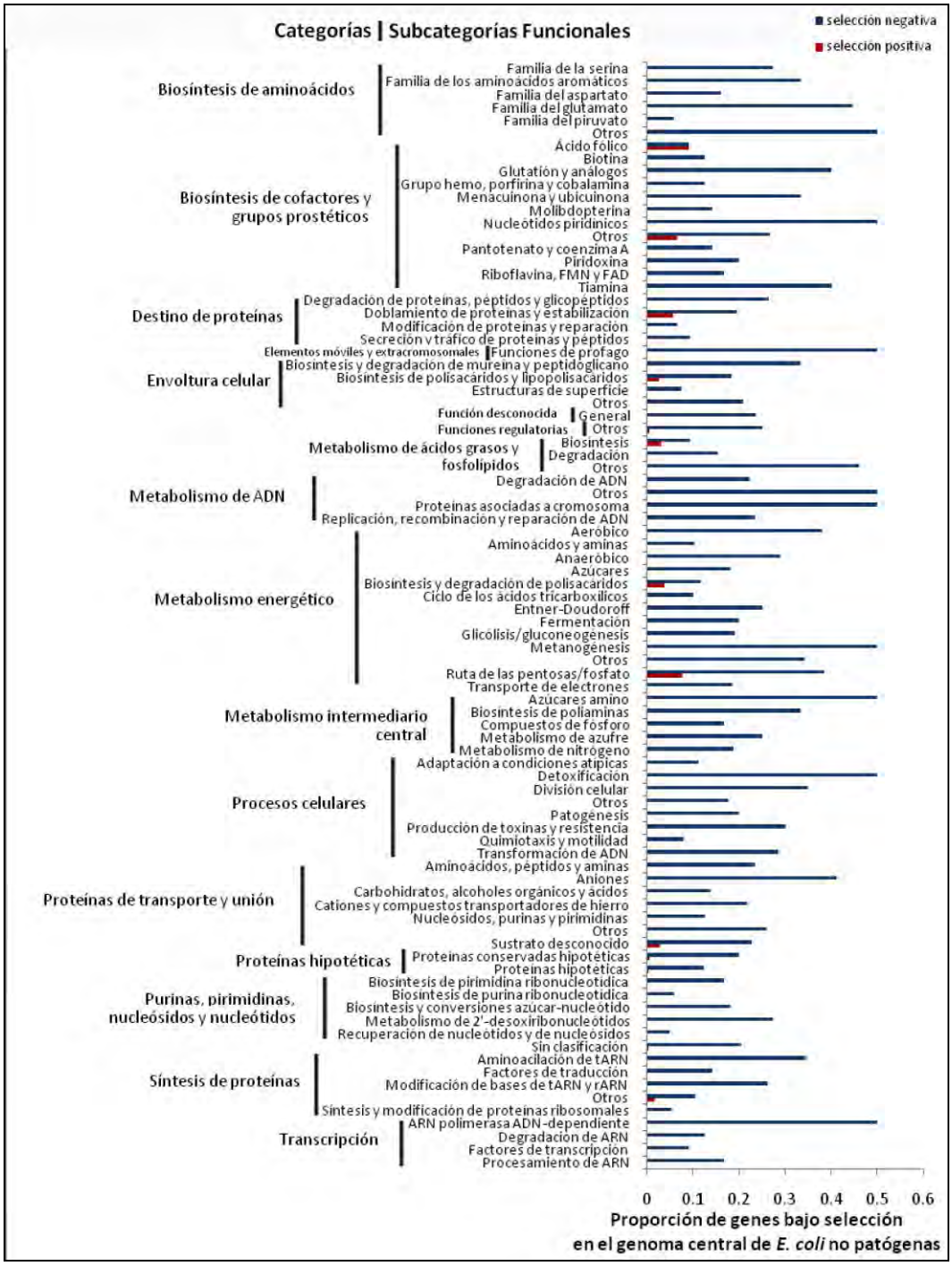


Figura 8. Proporción de genes con evidencia de selección positiva y selección negativa en el genoma central de la muestra total de *E. coli*, dentro de las diferentes categorías funcionales. Se incluyeron genes presentes en loci que fueran significativos (a $p<0.05$) en al menos una de las tres pruebas de neutralidad.

En el ecogrupo de *E. coli* no-patógenas, se encontró un sesgo hacia loci con valores negativos en las tres pruebas, tanto en del genoma central, como en el flexible (Tabla 10). Por lo que no hubo diferencias significativas entre estas dos regiones del cromosoma con ninguna prueba de neutralidad (Wilcoxon no significativa: Tajima p = 0.3091; D* Fu-Li p = 0.3186; F* Fu-Li p = 0.5247). En la Figura 7, se observa el mismo patrón, donde la proporción de loci con valores significativos negativos fue mayor que la proporción de loci con valores significativos positivos, tanto en el genoma central como en el genoma flexible. Así, se obtuvieron al menos 803 loci con evidencia de selección. En proporción, ligeramente más loci bajo selección en el genoma central (735 loci significativos/ 4,058 loci polimórficos =0.181) que en el genoma flexible (61 loci significativos / 477 loci polimórficos=0.142), pero para ambos componentes del pangenoma con predominancia de valores negativos. Dentro de los loci del genoma central con valores significativos encontramos al menos 16 loci no-codificantes con evidencia de selección, 5 bajo selección positiva y 11 bajo selección negativa. En los 719 loci restantes se encontraron 898 genes, 22 de ellos bajo selección positiva y 876 bajo selección negativa. El análisis de las categorías funcionales del JCVI correspondientes a estos genes (Figura 9), mostró que en el genoma central hay genes bajo selección negativa en todas las categorías funcionales, destacando en Biosíntesis de aminoácidos y Funciones regulatorias. En contraste, los pocos genes bajo selección positiva se agruparon en las categorías de Biosíntesis de cofactores y grupos prostéticos – ácido fólico y otros, Destino de proteínas – doblamiento de proteínas y estabilización, Envoltura celular – biosíntesis y degradación de polisacáridos y lipopolisacáridos, Funciones regulatorias, Metabolismo de ácidos grasos y fosfolípidos – biosíntesis, Metabolismo energético – biosíntesis y degradación de polisacáridos y ruta de las pentosas/fosfato, Proteínas hipotéticas, Proteínas sin clasificación y Síntesis de proteínas (Figura 9A). En el genoma flexible encontramos 3 loci no-codificantes, uno con evidencia de selección positiva y 2 con evidencia de selección negativa. En los 58 loci restantes se localizaron 77 genes, 6 bajo selección negativa y 71 bajo selección positiva.

De igual manera que en el genoma central, en el flexible se encontraron genes bajo selección negativa en todas las categorías funcionales, aunque en menor proporción con respecto al genoma central (Figura 9B). Solamente destacan las categorías de Elementos móviles y extracromosomales – funciones de profago, donde hay un exceso de genes bajo

selección positiva. Así, de manera general, este ecogrupo está sujeto a selección negativa ó purificadora.



A.

Figura 9. Proporción de genes con evidencia de selección positiva y selección negativa dentro de las diferentes categorías funcionales, en el ecogrupo de *E. coli* no-patógenas, A. en el genoma central y B. en el genoma flexible. Se incluyeron genes que tuvieran evidencia de selección (significativos a $p<0.05$) en al menos una de las tres pruebas de neutralidad.

B.

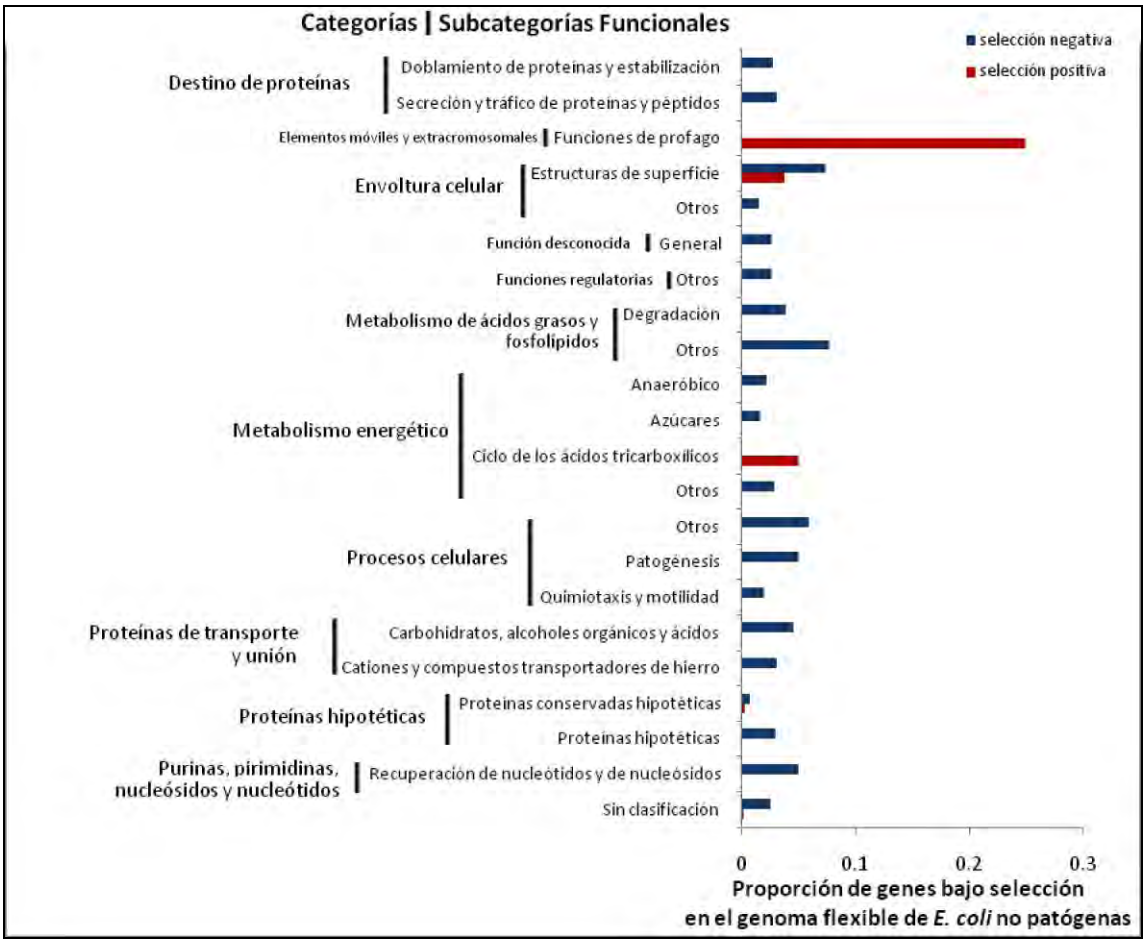


Figura 9. *Continuación.*

En el ecogrupo de *E. coli* patógenas, el patrón de selección fue opuesto (Tabla 10), y se encontró un sesgo hacia loci con valores positivos en las tres pruebas, en el genoma central y en el genoma flexible.

Tampoco se encontraron diferencias significativas en los promedios entre el genoma central y el genoma flexible en ninguna prueba de neutralidad (Prueba de Wilcoxon no significativa: Tajima p = 0.9798; D* Fu-Li p = 0.4607; F* Fu-Li p = 0.5731).

En ambos componentes del pangenoma, la proporción de loci con valores positivos significativos fue mayor que la proporción de loci con valores negativos significativos en cada una de las tres pruebas de neutralidad (Figura 7).

Así, en este ecogrupo, el genoma central presentó 240 loci que partían del modelo neutro en alguna de las tres pruebas de neutralidad (229 con valores positivos y 11 con valores negativos). Entre éstos 25 loci correspondieron regiones no-codificantes. Los 204 loci

restantes comprendieron 283 genes. La distribución de estos genes en las categorías funcionales del JCVI se muestra en la Figura 10. Ahí se puede ver que hay genes bajo selección positiva en todas las categorías funcionales. En particular en Biosíntesis de cofactores y grupos prostéticos – glutatión y análogos, Envoltura celular – otros, Metabolismo energético – Entner Doudoroff y Transcripción. Los genes bajo selección negativa se agrupan en las categorías de Procesos celulares - división celular, Proteínas hipotéticas y Transcripción- factores de traducción (Figura 10A).

En el genoma flexible se encontraron 43 loci con valores significativos en alguna de las tres pruebas de neutralidad aplicadas, 37 loci positivos y 6 negativos. De estos, 15 loci correspondieron a regiones no-codificantes, 14 bajo selección positiva y 1 con evidencia de selección negativa. En los 28 loci se encontraron 31 genes. En este caso, encontramos una mayor proporción de genes bajo selección positiva en las categorías de Destino de proteínas – modificación de proteínas y reparación y Proteínas de transporte y unión – aniones (Figura 10B).

En este ecogrupo, obtuvimos que en el genoma central hay genes bajo selección positiva en todas las categorías funcionales (Figura 10). En particular en Biosíntesis de cofactores y grupos prostéticos – glutatión y análogos, Envoltura celular – otros, Metabolismo energético – Entner Doudoroff y Transcripción.

Los genes bajo selección negativa se agrupan en las categorías de Procesos celulares - división celular, Proteínas hipotéticas y Transcripción- factores de traducción.

En el genoma flexible se encontraron genes bajo selección positiva principalmente en las categorías de Destino de proteínas – modificación y reparación de proteínas y Proteínas de transporte y unión – aniones. Y solamente en las categorías de proteínas conservadas e hipotéticas y de envoltura celular se encontraron genes bajo selección negativa (Figura 10B).

A.

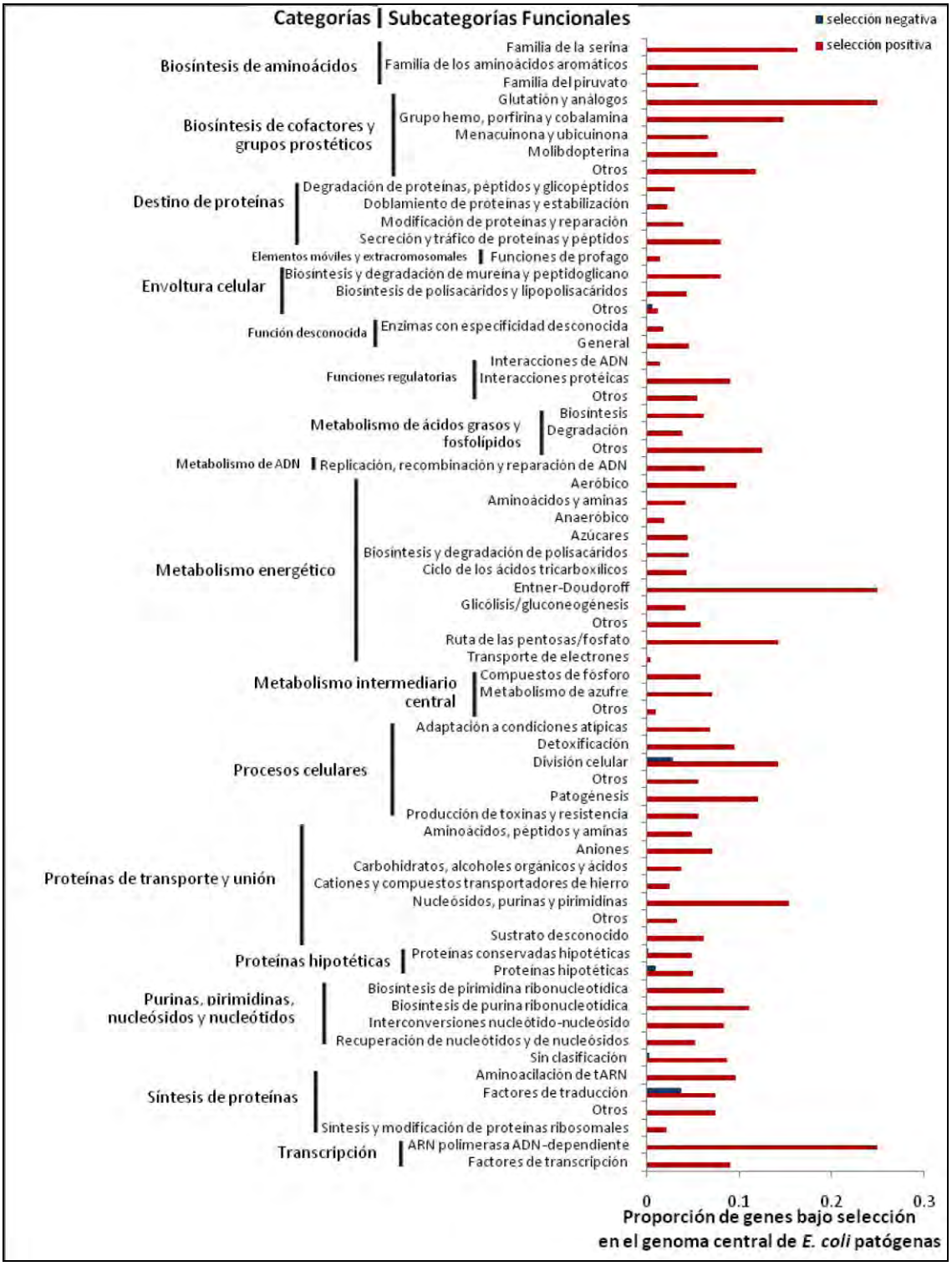


Figura 10. Proporción de genes con evidencia de selección positiva y selección negativa dentro de las diferentes categorías funcionales, en el ecogrupo de *E. coli* patógenas, A. en el genoma central y B. en el genoma flexible. Se incluyeron genes que tuvieran evidencia de selección (significativos a $p<0.05$) en al menos una de las tres pruebas de neutralidad.

B.

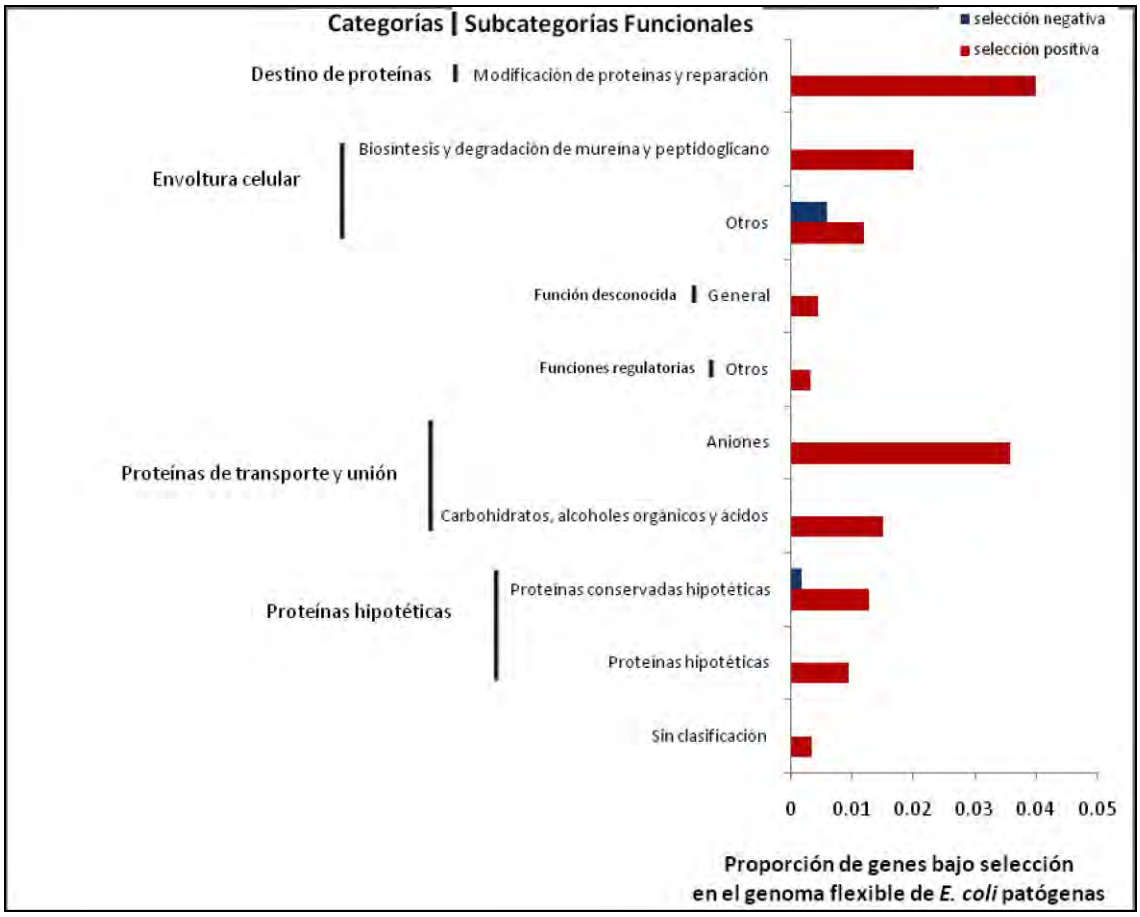


Figura 10. *Continuación.*

4. PATRONES EVOLUTIVOS DEL GENOMA CENTRAL A LO LARGO DEL CROMOSOMA DE *E. COLI*

La distribución de la diversidad π y θ en el genoma central de los bloques localmente colineares (LCBs) se muestra en la Tabla 11. Aquellos con mayor diversidad fueron los LCBs 2, 10 y 18. Estos corresponden a 3 de los 4 LCBs de menor tamaño (LCBs 2, 10, 14 y 18; Tabla 8). El resto de los LCBs presentaron una diversidad nucleotídica baja ($\pi \leq 0.04$). En cuanto a diversidad de haplotipos Hd y desequilibrio de ligamiento Zns, vemos en la Tabla 11 que todos los LCBs presentan una Hd ≥ 0.7. Destacan los LCBs 2 y 18 al presentar un valor de Zns más cercano a 1, lo que podría sugerir clonalidad. Consecuentemente, en ambos LCBs encontramos regiones en desequilibrio de ligamiento significativo (Figura 11).

Tabla 11. Diversidad, pruebas de selección natural y desequilibrio de ligamiento en el genoma central de los bloques localmente colineares (LCBs) del cromosoma de *E. coli*, en los dos ecogrupos y en la muestra total.

Ecogrupo	Parámetro	Bloques Localmente Colineares (LCBs)																				
		1 [abc]	2	3 [abc]	4	5	6 [bc]	7 [abc]	8 [abc]	9 [abc]	10	11 [abc]	12 [abc]	13 [abc]	14	15 [bc]	16 [bc]	17 [abc]	18	19 [abc]	20 [bc]	21 [abc]
No-patógenas	π [1]	0.0136	0.1282	0.0148	0.0088	0.0076	0.0107	0.0193	0.0182	0.0157	0.0190	0.0108	0.0147	0.0127	0.0056	0.0131	0.0121	0.0202	0.0126	0.0134	0.0107	0.0171
	θ [1]	0.0143	0.1026	0.0164	0.0080	0.0079	0.0108	0.0196	0.0190	0.0164	0.0214	0.0117	0.0161	0.0132	0.0067	0.0126	0.0127	0.0216	0.0152	0.0144	0.0127	0.0178
	Hd [1]	0.7602	0.7000	0.6833	0.9000	0.7167	0.8325	0.8494	0.8077	0.7615	0.6000	0.6976	0.7466	0.7601	0.4000	0.8833	0.8600	0.7508	0.4000	0.7285	0.4722	0.7260
	Zns [2]	0.7367*	0.9893*	0.7776*	0.5640	0.7941*	0.6994*	0.6641*	0.7272*	0.7505*	-	0.7291*	0.7086*	0.7245*	-	0.8147*	0.8576*	0.7230*	-	0.7506*	0.4468	0.7256*
	D de Tajima [3]	-0.4327*	1.8218*	-0.645*	0.3854	-0.1688	-0.2927*	-0.1474*	-0.3295	-0.4202**	-0.6116	-0.5287**	-0.5555*	-0.437*	-0.9726	0.283	-0.2483	-0.5*	-1.0711	-0.5527**	-0.8918	-0.4783**
	D* de Fu-Li [3]	-0.4597*	1.7346*	-0.6724*	0.3854*	-0.1688	-0.3927**	-0.2203**	-0.4079*	-0.4671**	-0.6116	-0.5566*	-0.5846*	-0.4779**	-0.9726	0.283	-0.2483	-0.5527*	-1.0711	-0.5885**	-0.9296*	-0.5213**
	F* de Fu-Li [3]	-0.4841*	1.8661*	-0.7125*	0.3947	-0.195	-0.3941*	-0.2207**	-0.4202*	-0.4904**	-0.6136	-0.586*	-0.6175*	-0.5005**	-0.9544	0.2896	-0.2621	-0.579*	-1.0826	-0.6195**	-0.9552*	-0.5462**
Patógenas	π [1]	0.0183	0.1260	0.0226	0.0122	0.0143	0.0135	0.0234	0.0204	0.0226	0.0462	0.0178	0.0227	0.0165	0.0073	0.0262	0.0162	0.0304	0.0332	0.0207	0.0164	0.0214
	θ [1]	0.0163	0.0907	0.0198	0.0113	0.0135	0.0122	0.0204	0.0183	0.0198	0.0519	0.0156	0.0198	0.0146	0.0057	0.0223	0.0141	0.0269	0.0460	0.0184	0.0128	0.0187
	Hd [1]	0.8748	0.6905	0.8800	0.8980	0.7302	0.8274	0.8735	0.8695	0.8758	0.9048	0.8656	0.8793	0.8869	0.8095	0.8056	0.8476	0.8813	0.7857	0.8784	0.7513	0.8637
	Zns [2]	0.6214*	1*	0.6366*	0.5485*	0.5619*	0.5456*	0.6374*	0.6370*	0.6304*	0.5819	0.6514*	0.6569*	0.6228*	0.5333	0.7615*	0.6731	0.5968*	0.16	0.6302*	0.6864*	0.6522*
	D de Tajima [3]	0.6959*	2.1622*	0.8278*	0.4304*	0.0101	0.5584*	0.7677*	0.6807*	0.7388*	-0.4867*	0.7305**	0.795**	0.7337*	1.1684	0.9284	0.8914	0.6532*	-1.5262*	0.7138**	0.8953*	0.8323**
	D* de Fu-Li [3]	0.5551	1.5549	0.6087	0.069	-0.2678	0.3898	0.5685*	0.5012*	0.5769	-0.6396*	0.5466*	0.6035*	0.5715	1.1781	0.6834	0.5609	0.5132	-1.5916*	0.5282*	0.6283	0.6526*
	F* de Fu-Li [3]	0.6482	1.8623	0.7258	0.1642	-0.2237	0.4726	0.6766	0.5977*	0.6769	-0.6661*	0.648*	0.7134*	0.6715	1.2596	0.8151	0.7007	0.6016	-1.7311*	0.6291*	0.7534	0.7653*
TOTAL	π [1]	0.0194	0.1211	0.0242	0.0112	0.0138	0.0140	0.0245	0.0222	0.0235	0.0698	0.0179	0.0235	0.0176	0.0078	0.0263	0.0174	0.0309	0.0591	0.0211	0.0155	0.0234
	θ [1]	0.0173	0.0753	0.0212	0.0104	0.0133	0.0130	0.0219	0.0205	0.0207	0.0526	0.0162	0.0210	0.0164	0.0092	0.0215	0.0143	0.0287	0.0373	0.0190	0.0124	0.0210
	Hd [1]	0.9129	0.7879	0.9088	0.9221	0.7652	0.9030	0.9233	0.9078	0.9145	0.9091	0.8944	0.9124	0.9186	0.7879	0.9053	0.9182	0.9155	0.7803	0.9085	0.7534	0.8979
	Zns [2]	0.3538*	0.9114*	0.3720*	0.3228	0.3022	0.3670*	0.3482*	0.3477*	0.3581*	0.4974	0.3701*	0.3755*	0.3519*	0.1	0.4031	0.3434	0.3493*	0.8386*	0.3670*	0.4646*	0.3891*
	D de Tajima [3]	0.6069*	2.6428*	0.6768**	0.9278*	1.2632*	0.6813	0.625**	0.598**	0.5983*	1.3515	0.5378**	0.5549**	0.5006**	0.5399	0.9263	0.9362*	0.5721*	2.4716*	0.5919**	0.7314*	0.6936*
	D* de Fu-Li [3]	0.5841**	1.3821*	0.6013**	0.8401*	1.1599*	0.6843*	0.5826**	0.5993**	0.5364*	0.7204	0.5287*	0.4983**	0.4778**	0.4589	0.8389	0.8164	0.5743*	1.214*	0.5378**	0.6795*	0.6149*
	F* de Fu-Li [3]	0.6565*	1.9471*	0.6914**	0.9784*	1.3521*	0.7604*	0.6661**	0.6708**	0.6173*	1.0032	0.5872*	0.571**	0.5349**	0.5425	0.9815	0.9653*	0.647*	1.7505*	0.6156**	0.772	0.7104**

[1, 2, 3] Promedio. Estimado sobre [1] número de loci del genoma central (N), [2] número de loci con sitios informativos para la prueba Zns (N_{DL}) y [3] número de loci polimórficos (N_P), de cada LCB (N, N_{DL} y N_P se reportan en el Apéndice 6; varianzas de los promedios reportados en el Apéndice 7).

[a,b,c] En estos LCBs los ecogrupos son significativamente diferentes en [a]diversidad genética π y θ (excepto el LCB 3, que no fue distinto en el parámetro θ), [b] en los valores de D de Tajima, D* y F* de Fu-Li y [c] en diversidad Hd y en los valores de desequilibrio de ligamiento Zns (p de Wilcoxon reportada en el Apéndice 8).

- LCBs en los que no se estimó la prueba Zns por falta de sitios informativos.

Presencia de loci significativos *positivos ó *negativos en la prueba de neutralidad; * loci significativos en la prueba de desequilibrio de ligamiento Zns. La proporción de loci significativos en cada LCB se muestra en la Figura 11.

Por otra parte, los LCBs de mayor diversidad también presentaron valores positivos en las pruebas de neutralidad aplicadas (Tabla 11), de éstos, sólo en el LCB 2 dichas pruebas resultaron significativas. Los dos genes bajo selección positiva en el LCB 2 corresponden a *yihT* y *yihW*, con función en el metabolismo energético - ruta de las pentosas/fosfato y función de regulador transcripcional de unión a ADN respectivamente.

En los demás LCBs, menos diversos , pero que también presentaron evidencia de selección positiva (LCBs 4, 6, 9, 16, 17, 20 y 21) predominaron las categorías funcionales de Proteínas hipotéticas y Síntesis de proteínas. En contraste, señales de selección negativa se encontraron en pocos LCBs (Figura 11), siendo el LCB 5 el único en el que se presenta selección negativa y no positiva. En dicho LCB se encuentra la categoría funcional de Síntesis de proteínas. Curiosamente, como se mencionó antes, en el LCB 2 también encontramos evidencias de clonalidad, lo que nos indica la posibilidad de que esta región del genoma posea una dinámica evolutiva favorecida por "hitchhiking".

Por ecogrupos, en las cepas no-patógenas los patrones de diversidad también fueron heterogéneos, y algunos LCBs fueron muy diversos, como el LCB 2 (Tabla 11). En cuanto a desequilibrio de ligamiento, los LCBs 2 y 15 resultaron completamente clonales. En estos LCBs se encuentran genes con funciones de transcripción y biosíntesis de aminoácidos.

En oposición al patrón de selección negativa encontrado de manera general en este ecogrupo, en algunos LCBs predominó la selección positiva, como es el caso de los LCBs 2, 4 y 6, en donde se encuentran genes asociados a procesos celulares como la traducción (LCB 4 y 6) y la regulación de la transcripción (LCB 2).

En el ecogrupo de cepas patógenas, los patrones de diversidad y la distribución de loci con evidencia de clonalidad tampoco es homogénea en los diferentes LCBs. Así, encontramos LCBs, como el 10 y el 18, en los que hay más loci con evidencia de selección negativa que bajo selección positiva, en contraste con el patrón general del ecogrupo. En estos bloques se encuentran genes hipotéticos y regiones no-codificantes.

Finalmente, como se ve en la Tabla 11, hubo algunas regiones del genoma central (6 LCBs) en los que ambos ecogrupos presentaron patrones de diversidad, clonalidad y selección semejantes. Por lo tanto, es en la mayor parte del genoma central que se pueden encontrar diferencias en la dinámica evolutiva entre ecogrupos.

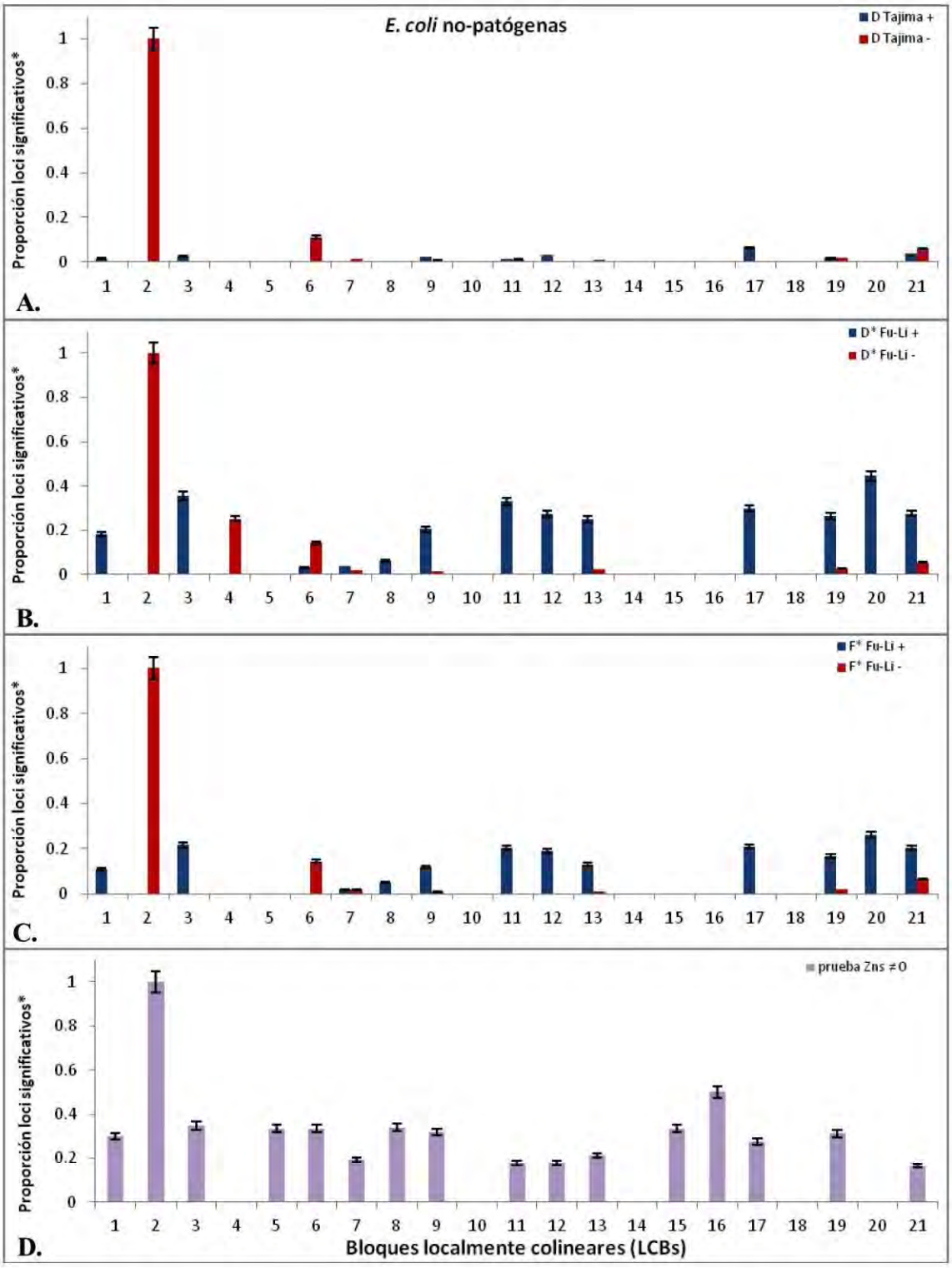


Figura 11. Proporción de loci significativos* en las pruebas de A) D de Tajima, B) D* de Fu-Li, C) F* de Fu-Li y D) desequilibrio de ligamiento Zns, por bloque localmente colinear (LCB) del cromosoma de *E. coli*. Cepas no-patógenas.

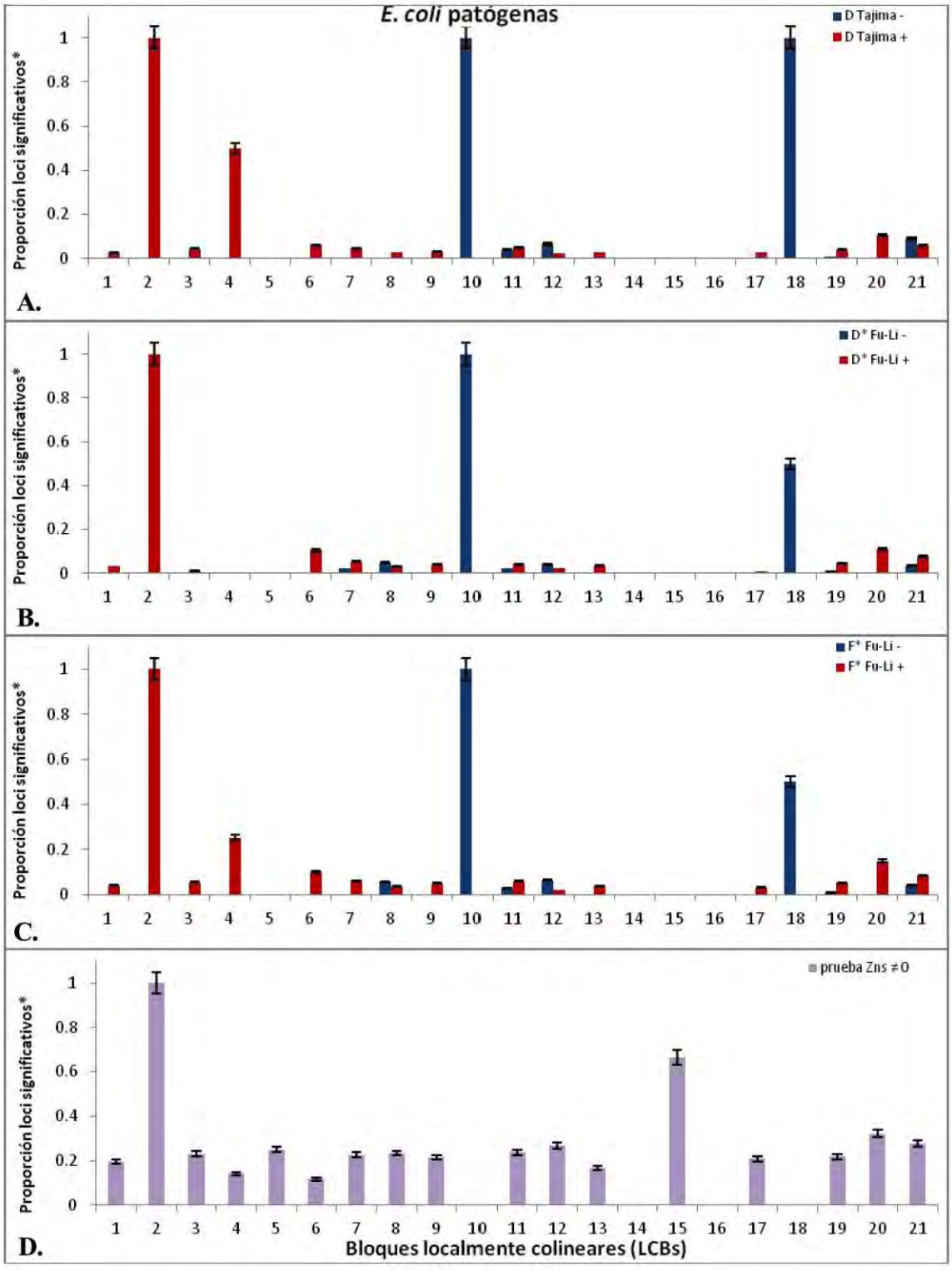


Figura 11. *Continuación*. Cepas patógenas.

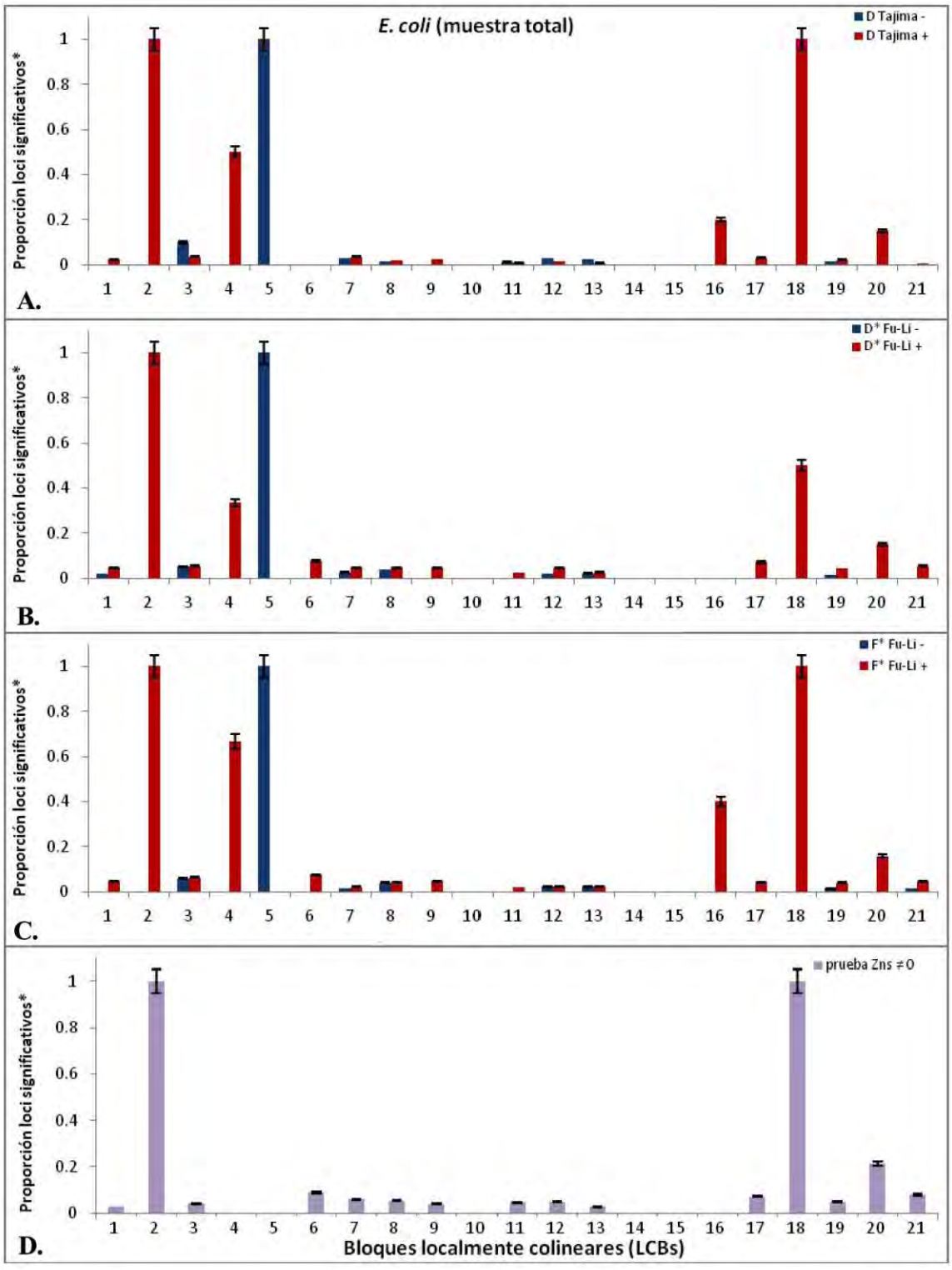


Figura 11. *Continuación*. Muestra total de *E. coli*.

* Significativos a p<0.05 en las pruebas de D de Tajima, D* y F* de Fu-Li, y a p<0.025 en la prueba de desequilibrio de ligamiento Zns. Los datos de donde se obtuvieron estas proporciones se reportan en el Apéndice 9.

DISCUSIÓN

1. HISTORIA FILOGENÉTICA DE *E. COLI*

Las relaciones entre los individuos de la muestra, como se representan en la Figura 2, revelan la existencia de al menos dos grandes linajes en *E. coli*, uno representado por las cepas comensales, de vida libre y patógenas intestinales, y el otro representado por las cepas patógenas de ave y patógenas extra-intestinales. Estos datos sugieren que la patogénesis en *E. coli* es una convergencia evolutiva. Esta topología es similar a lo que encuentran Chattopadhay et al. (2009) y Touchon et al. (2009) con algunos de los mismos genomas analizados en el presente estudio y va de acuerdo con numerosos trabajos que proponen que la patogénesis ha surgido varias veces en la especie *E. coli*, y que por ende su agrupación es polifilética (Pupo et al. 2000; Reid et al. 2000; Sandner et al. 2001; Escobar-Páramo et al. 2003; Bidet et al. 2007).

Sin embargo, aunque es práctico pensar en términos de linajes de cepas dentro de las especies y representar dichas relaciones con árboles, Aunque aparentemente la utilización de todas las regiones del genoma central es la mejor manera de determinar las relaciones filogenéticas relaciones entre individuos (Lerat et al. 2005; Lan y Reeves 2000), al obtenerse valores estadísticos robustos, ya sea de bootstrap ó de probabilidades posteriores (según el método de inferencia filogenética que se use), como ocurre también en el caso del presente estudio (Figura 2), debemos de tomar en cuenta que árboles de este tipo no representarán la verdaderas relaciones entre diferentes regiones del genoma de una especie, sobre todo en grupos recombinantes, en cuyo caso estas relaciones deberían ser representadas por figuras de tipo red y no árbol (Doolittle y Papke 2006) ó presentar diferentes topologías de acuerdo a la región genómica que se analice (Touchon et al. 2009).

Como se puede verificar en el Apéndice 1, las relaciones filogenéticas determinadas a partir de la información de otras regiones del genoma central de nuestra muestra de *E. coli*, están representadas por árboles con diferentes topologías.

Esto nos da una primera idea de que la muestra de análisis pudiera tener una estructura recombinante, con la existencia de flujo génico aún entre individuos de los diferentes linajes representados en la Figura 2.

2. DINÁMICA EVOLUTIVA DEL CROMOSOMA DE *E. COLI*

2.1. Variación en la estructura cromosómica de *E. coli*

El número de regiones que representan eventos de rearreglo cromosómico ó LCBs, encontrados en el presente estudio (21 LCBs) es ligeramente menor a lo que se ha reportado para muestras de *E. coli* y *Shigella* (34 LCBs en 6 cepas; Mau et al. 2006). Esto sugiere que los individuos de *Shigella* han sufrido un mayor número de rearreglos, probablemente debido a la presencia de una mayor cantidad de IS con respecto a *E. coli* (Touchon et al. 2009).

La estructura de rearreglos de esta muestra de *E. coli* (Figura 3) revela que el cromosoma de *E. coli* es plástico en comparación con otras especies de bacterias con hábito patógeno estricto, como son *Mycobacterium tuberculosis* (ningún LCB en 6 cepas; Cubillos-Ruiz et al. 2008), *Burkholderia pseudomallei* (12 LCBs en 11 cepas; Nandi et al. 2010) y el género *Brucella* (10 LCBs en 10 cepas; Wattam et al. 2009). Aunque el número de LCBs detectado en *E. coli* fue menor a lo que se ha descrito en otras bacterias patógenas como *Legionella pneumophila* (16 LCBs en 5 cepas; D'Auria et al. 2010), *Francisella tularensis* (51 bloques de sintenia en 3 cepas; Petrosino et al. 2006), *Neisseria meningitidis* (33 LCBs en 4 cepas; Schoen et al. 2008), el género *Yersinia* (98 LCBs en 11 cepas; Chen et al. 2010; Darling et al. 2008) y en la bacteria de vida libre *Rhodobacter sphaeroides* (382 LCBs en 3 cepas; Choudhary et al. 2007).

El número tan diferente de LCBs reportados en especies de patógenos bacterianos es relativamente bajo con respecto a especies de vida libre. Esto confirma que la evolución por reorganización cromosómica no es una característica exclusiva de las bacterias patógenas (Rasko et al. 2008), como se había sugerido previamente (Hacker y Kaper 2000) y que inclusive pudieran ser más estables en cuanto a estructura cromosómica que las bacterias de vida libre, tal vez debido a presiones selectivas que mantienen cierto orden en algunas regiones genómicas, eliminando aquellas cepas con rearreglos que no son adaptativos (Mira et al. 2002). La ausencia de diferencias en la estructura cromosómica de bacterias patógenas de *E. coli* también puede ser explicado por la estructura epidémica de esta

especie (Maynard-Smith et al. 1993). Así, cuando hay linajes muy clonales, como son las cepas EHEC O157:H7, habrá mayor probabilidad de que el intercambio ó flujo génico, ocurra con individuos idénticos, lo que no generaría ningún tipo de variación.

El número de rearreglos entre individuos de un mismo ecogrupo, resulto ser parecido en ambos ecogrupos (Tabla 5). Esto nos dice que la estructura del cromosoma se conserva dentro de los ecogrupos. Y que el estilo de vida no genera mayores eventos de rearreglo. En contraste, los eventos de rearreglos son mayores cuando se comparan cepas de diferente ecogrupo, y llegan a ser hasta 9 entre las cepas K12 8739ATCC y UTI89 (Tabla 5). La única excepción notable es el número de rearreglos (sólo uno), encontrado entre la cepa ETEC E24377A patógena y la K12 MG1655 no-patógena. Este dato sugiere que ambas cepas son similares, como ya se ha propuesto en trabajos previos (Chen et al. 2006), por lo menos al nivel de la estructura cromosómica.

Los rearreglos encontrados en ambos grupos son de diferente naturaleza. Entre las cepas no-patógenas predominan las inversiones de regiones cromosómicas de tamaño grande (hasta 1 Mb en la cepa SMS-3-5), y se detectaron alrededor del origen y termino de replicación, hecho que se considera como una señal de que hay una tendencia a conservar la simetría en la estructura del cromosoma (Mira et al. 2002) en este grupo de cepas de *E. coli,* probablemente debido a una presión para conservar procesos celulares básicos como la replicación (Darling et al. 2008; Touchon et al. 2009).

Entre las cepas patógenas, los rearreglos corresponden a inversiones en regiones de extensión mucho más pequeña, siendo la inversión más grande la que se encuentra en la cepa O157:H7 EDL933 con un tamaño de 300 kb (Figura 3).

Entonces la estructura del cromosoma es cohesiva dentro de los ecogrupos, pero diferente entre ellos. Es decir que la naturaleza de los eventos de rearreglo cromosómico es similar entre individuos con estilos de vida semejantes. Sin embargo, por el número de eventos de rearreglo detectados en cada ecogrupo (Tabla 5), aparentemente la frecuencia con la que ocurren los eventos de rearreglo es parecida en ambos ecogrupos. Es decir que el estilo de vida no se correlaciona con una mayor ó menor plasticidad estructural del cromosoma en *E. coli*. Esto es contrario a lo que se ha propuesto en trabajos previos donde se compara la

frecuencia de rearreglos entre especies con diferentes estilos de vida (Mira et al. 2002). Sin embargo, dado que el presente estudio describe la estructura cromosómica dentro de una misma especie, la cohesividad estructural está reflejando la cohesividad de la especie, como ha sido sugerido por Touchon et al. (2009).

En contraste con la aparente homogeneidad en la estructura del cromosoma entre individuos del mismo ecogrupo, los rearreglos observados inclusive entre clonas de una misma cepa, como se encontró en el presente estudio (entre las cepas K12 MG1655 y K12 W3110 del ecogrupo de *E. coli* no-patógena y entre las cepas O157:H7 EDL933 y O157:H7 Sakai del ecogrupo patógeno (Figura 3; Tabla 5), y en el trabajo de Ferenci et al. (2009), entre las cepas de K12 analizadas aquí y una cepa proveniente de un linaje distinto de K12, la cepa MC4100, nos indica que el cromosoma de estas bacterias posee un alto potencial de diversidad estructural.

2.2 Variación en el repertorio genético del cromosoma de *E. coli*

El tamaño de la región correspondiente al genoma central de la muestra (<3.6 Mb) es del mismo orden de magnitud con respecto a lo que se ha descrito en trabajos previos en la especie *E. coli* (3.4 Mb en 6 genomas; Mau et al. 2006). Sin embargo en el presente trabajo la región del genoma central es ligeramente más extensa probablemente debido a que no se incluyeron cepas de *Shigella* en este análisis.

Esto se explica por el hecho de que mientras más divergentes sean dos individuos ó grupos de individuos, la proporción de genes que compartirán será menor (Mushegian y Koonin 1996). Dado que las cepas de *Shigella* constituyen un subgrupo de *E. coli* parafilético (Pupo et al. 2000) y altamente divergente del resto de las cepas de esta especie (Fukushima et al. 2002), se espera que la comparación de contenido genómico resulte en una menor cantidad de genes compartidos entre *E. coli* y *Shigella* que entre individuos de *E. coli* solamente.

Por otro lado, encontramos que las regiones del genoma flexible, producto de eventos de transferencia horizontal ó de deleción genómica, son abundantes y representan la mayor parte del pangenoma de esta muestra de *E. coli*. Esto, aunado a los altos niveles de identidad nucleotídica encontrados en el genoma central (≥ 0.96; Tabla 8) apoyan la idea de

que la transferencia horizontal y el flujo génico son los principales procesos que generan la diversidad en la especie (Darling et al. 2004).

Al comparar los elementos del genoma flexible compartidos por los individuos de cada ecogrupo, tenemos que las cepas no-patógenas tienen mayor número de regiones del genoma flexible que las patógenas (Figura 4). Este resultado parece contradecir a la teoría clásica de evolución de la patogénesis, pues se ha sugerido que son los elementos del genoma flexible los que han permitido a las bacterias la adaptación al nicho patógeno. En ese caso las cepas patógenas deberían poseer una poza de genes flexibles más amplia que aquellas cepas que carecen de potencial patogénico. Sin embargo, no encontramos tal patrón. Esto pudiera ser explicado por dos motivos. Por una parte los elementos genéticos que permiten la patogénesis en *E. coli* son muy diversos y las combinaciones de estos que permiten la colonización de diferentes regiones de un hospedero de manera exitosa son muy grandes (Kuhnert et al. 2000; Bekal et al. 2003; Wu et al 2007; Tenaillon et al. 2010), por lo que en teoría no se necesita un repertorio genético particular para que las bacterias posean capacidad patogénica. Ya otros trabajos sobre la patogénesis en *E. coli* encontraban que prácticamente no comparten ningún gen de virulencia (Mokady et al. 2005). Esto ha llevado a proponer que las cepas patógenas pueden usar diferentes genes con funciones similares para el proceso de infección.

Por otra parte, hay numerosos estudios de genómica comparada y de evolución experimental que han encontrado que las cepas comensales comparten una gran parte de factores genéticos asociados al proceso de infección y patogénesis (Rasko et al. 2008; Touchon et al. 2009; Tenaillon et al. 2010), por lo que se ha propuesto que probablemente algunos de los factores antes conocidos como de virulencia más bien funcionan como elementos de adecuación para ambos nichos (Levin et al. 1996; Le Gall et al. 2007). De esta manera, las cepas comensales podrían estar funcionando como un reservorio genético de elementos con potencial patogénico (Rasko et al. 2008). Estas ideas constituyen un primer marco teórico para proponer que la patogénesis no sólo está determinada por la adquisición ó perdida de genes mediante transferencia horizontal.

Entonces, probablemente las cepas patógenas de nuestra muestra comparten menos factores genéticos del genoma flexible que las no-patógenas debido a que el número de genes que

pueden estar involucrados en el proceso de patogénesis general es muy grande y variable (Johnson et al. 2006b).

Ahora, si los factores antes conocidos como de virulencia se encuentran también en las cepas comensales (Sandner et al. 2001; Rasko et al. 2008), debemos preguntarnos si existen otras regiones del genoma donde se esté acumulando la diversidad y permitiendo el desarrollo de la patogénesis en *E. coli*.

3. GENÓMICA DE POBLACIONES: PATRONES DE DIVERSIDAD GENÉTICA Y ADAPTACIÓN ECOLÓGICA EN *ESCHERICHIA COLI*

3.1. Diversidad genética en el cromosoma de *E. coli*

La diversidad nucleotídica promedio del genoma central de *E. coli* con valor de $\pi = 0.0217819 \pm 0.0002824$, fue mayor en relación a lo reportado para regiones codificantes del genoma de otras especies de bacterias patógenas como *Staphylococcus aureus* ($\pi = 0.00010$; Nübel et al. 2008), *Mycobacterium tuberculosis* ($\pi = 0.00024$; Dos Vultos et al. 2008), *Salmonella typhi* ($\pi = 0.00006$; Roumagnac et al. 2009) e inclusive de una muestra diferente de genomas de *E. coli* (8 individuos) y *Shigella* (6 individuos) que solamente analizan regiones codificantes ($\pi = 0.015$; Chattopadhay et al. 2009).

En bacterias, la inexistencia de un proceso de recombinación sexual acoplado a la reproducción, el paradigma de diversidad clonal (Selander y Levin 1980) y la extensiva diversidad en el repertorio genético, generada principalmente por eventos de transferencia horizontal de genes, han promovido la idea de que la diversidad en el genoma central bacteriano está determinada por la mutación y la deriva génica, y por lo tanto los loci centrales deben comportarse de manera neutral, tener baja ó nula recombinación (ser clonales) y no encontrarse bajo presiones selectivas. Adicionalmente, dado que se considera que la adaptación en bacterias está determinada por la naturaleza y función de los genes que han sido adquiridos ó perdidos por transferencia horizontal y que forman parte del genoma flexible, el genoma central debería estar conformado por elementos genéticos esenciales para la realización de las funciones biológicas básicas, por lo que se espera que su diversidad sea más ó menos homogénea y reducida en relación al genoma flexible

(Hochhut et al. 2006).

En el presente trabajo se encontró que la diversidad nucleotídica de los loci del genoma central de *E. coli* fue más bien heterogénea y presentó una varianza grande (V(π)=0.00032863). Los valores abarcaron un rango muy amplio, desde diversidad nula (π = 0) en varios loci tanto codificantes como no-codificantes, hasta una diversidad de π = 0.224242 en una región del gen *tynA* que codifica para una oxidasa activada en condiciones de anaerobiosis, cuya diversidad es similar a lo que se ha registrado a nivel genómico entre individuos de diferentes especies, como el género *Yersinia* en donde se reporta una diversidad genómica de π = 0.27 (Chen et al. 2010) y a lo que se ha descrito para genes de virulencia en *E. coli*, como por ejemplo en la isla de patogénesis del locus de esfacelamiento enterocítico LEE, donde la diversidad máxima se reporta para el gen *sepZ* y es de π = 0.24 (Castillo et al. 2005).

Esta amplia variación en la diversidad a lo largo del cromosoma ya había sido descrita para eucariontes como es el humano (Hellmann et al. 2003), el tomate salvaje *Solanum spp*. (Roselius et al. 2005) y la mosca de la fruta *Drosophila* (Begun et al. 2007), pero no para el caso de bacterias. En eucariontes, la existencia de loci con mayor ó menor variación en el genoma con respecto a los loci aledaños se explica por el efecto de una mayor ó menor recombinación respectivamente en conjunto al efecto de la selección (Charlesworth 2009), de tal manera que las regiones más clonales presentarían menor diversidad que aquellas más recombinantes y bajo selección.

El genoma central de *E. coli* presenta un patrón de diversidad aproximadamente similar al que presentan los genomas de especies eucariontes, con presencia de loci de muy baja diversidad ("coldspots" de diversidad) y de muy alta diversidad ("hotspots" de diversidad). Entonces podemos pensar que al igual que en eucariontes el efecto de la recombinación homóloga varía a lo largo del genoma central de *E. coli*, y hay regiones donde es más frecuente, regiones de clonalidad, así como regiones bajo el efecto de la selección.

3.2 Patrones de desequilibrio de ligamiento: estructura clonal ó panmíctica?

Efectivamente, encontramos que la prueba de desequilibrio de ligamiento Zns tuvo un valor promedio de 0.3642279 ± 0.0020288, lo que nos indica que la mayoría de los loci analizados obtuvieron valores más cercanos a 0 que a 1, es decir que la tendencia general

de la muestra fue hacia la ausencia de desequilibrio de ligamiento. En consecuencia, de los 4,122 loci del genoma central analizados, solamente 202 loci fueron significativamente diferentes de 0. Estos 202 loci representan el 4.9% de loci del genoma central. Es decir que solamente el <5% del genoma central está en desequilibrio de ligamiento.

Aunque sería necesario conocer las tasas de recombinación en cada loci para tener una medida más exacta de esta fuerza a nivel genómico, se ha visto que hay una asociación inversa entre los estimados de recombinación y desequilibrio de ligamiento (Dawson et al. 2002).

Por lo que, el resultado obtenido nos da una idea del comportamiento de *E. coli* y de la naturaleza de la diversidad en su genoma, indicando que la recombinación homóloga probablemente sea un fenómeno más frecuente que la mutación, como ya había sido propuesto por algunos autores (Desjardins et al 1995; Souza et al. 1999; Brown et al 2001). Y va en acuerdo con los autores que han estimado que la tasa de recombinación es al menos un orden de magnitud mayor que la tasa de mutación en esta especie (Souza et al. 2002).

Más recientemente en una muestra de 6 genomas enterobacterianos, 4 de *E. coli* (la K12 MG1655, dos cepas EHEC O157:H7 Sakai y O157:H7 EDL933 y la cepa CFT073 de UTI) y 2 de *Shigella*, se ha presentado evidencia de recombinación homóloga en un porcentaje apreciable de las regiones del genoma central (7.5% de 3.4 Mb, correspondiente a 251 kb), en regiones no-codificantes y entre fragmentos de genes involucrados en los procesos de recombinación, transporte, quimiotaxis y motilidad, principalmente (Mau et al. 2006). Sus resultados parecen oponerse a lo encontrado en el presente estudio. Pero, debemos considerar que casi la mitad de su muestra de *E. coli* consta de individuos que son clonas del patotipo EHEC O157:H7.

Por otra parte, los resultados obtenidos con marcadores de MLST siguen apoyando el paradigma clonal en *E. coli*. Por ejemplo Pérez-Losada et al. (2005) calculan tasa de recombinación en genes de mantenimiento de 15 especies de patógenos bacterianos, y encuentran que *E. coli* tiene una tasa de recombinación prpacticamente nula, en oposición a las especies con valores de recombinación más altas como *Helycobacter pylori*, *Streptococcus pneumoniae*, *Neisseria gonorrhoeae*, *N. meningitidis*, *S. agalactiae* y *Haemophilus influenzae*. Estos resultados contrastantes se pueden explicar con el modelo de estructura epidémica previamente descrito por Maynard-Smith (1993), en donde *E. coli*

es una especie con recombinación frecuente, pero en la cual ciertos genotipos en condiciones particulares proliferan de manera acelerada reflejando ciertos niveles de diversidad determinada principalmente por mutación (clonalidad).

3.3 Selección natural a nivel molecular

Bajo el modelo de evolución neutral de Kimura (1989a), se predice que la mayor parte de la diversidad del genoma debe ser generada por mutación y eliminada ó mantenida por la deriva génica. Por lo que solamente algunas regiones estarán afectadas por la selección negativa (purificadora), y la selección positiva (diversificadora) debe ser prácticamente nula.

En nuestra muestra, efectivamente sólo identificamos algunos loci en los cuales las pruebas de selección aplicadas resultaron significativas, (215 loci correspondientes al 5.2% del total de loci). Este porcentaje de regiones bajo selección en el genoma central es similar a lo encontrado por Chattopadhay et al. (2009), aunque ellos se limitan a analizar regiones codificantes (300 genes ortólogos correspondientes al 5.6% del repertorio génico total) en una muestra de *E. coli* y *Shigella*.

A pesar de que los loci bajo selección fueron en general escasos, la tendencia de éstos fue hacia valores positivos (196 loci correspondientes al 4.8% del total de loci), y la fracción de loci bajo selección negativa fue muy pequeña (19 loci correspondientes al 0.4% del total).

Este patrón, que contradice lo que se espera bajo neutralidad, ha sido reportado en estudios de selección en regiones codificantes a nivel genómico, en diferentes muestras de genomas de *E. coli*: 5 genomas de *E. coli* de diferentes patotipos y un genoma de K12 (Charlesworth y Eyre-Walker 2006), 3 genomas de UPEC (Chen et al. 2006), 3 genomas de *E. coli* patógenas, una de *E. coli* K12 comensal y 2 de *Shigella* (Petersen et al. 2007), 2 genomas de *E. coli* comensal, 6 de diferentes patotipos de *E. coli* y 6 de *Shigella* (Chattopadhay et al. 2009), en la bacteria patógena *Listeria monocytogenes* (Orsi et al. 2008) y en los endosimbiontes de insectos *Wolbachia sp.* (Brownlie et al. 2007), *Buchnera sp.* y *Blochmannia sp.* (Toft et al. 2009), así como en eucariontes como la mosca de la fruta *Drosophila* (Shapiro et al. 2007; Sawyer et al. 2007; Begun et al. 2007), el humano *Homo sapiens* (Voight et al. 2006; Akey 2009) y en el grupo de los mamíferos (Resch et al. 2007).

La baja señal de clonalidad (que implica un efecto extendido de la recombinación en el genoma) y la mayor proporción de regiones bajo selección positiva que de regiones bajo selección negativa nos inclinan a conjeturar que el cromosoma de *E. coli* podría no estar siguiendo el modelo neutro de evolución genómica.

Sabemos que bajo el modelo neutro sólo la deriva génica y la mutación darían cuenta de la diversidad a nivel del genoma (Kimura 1989). Y estas fuerzas evolutivas deben actuar de manera homogénea en todo el genoma, a menos que haya selección ó recombinación. De tal manera que si los patrones de diversidad no son homogéneos, y esto se considera como evidencia suficiente de que la deriva génica y la mutación no están en equilibrio, entonces las otras fuerzas deben estar desequilibrándolo, ya sea la selección ó la recombinación. Dado que en nuestro caso no fueron muchas regiones las que presentaron evidencia de selección, lo más probable es que el cromosoma no se esté comportando neutralmente, no tanto porque haya mucha selección, sino porque hay mucha recombinación.

Una posible explicación a estos descubrimientos la da Charlesworth (2009), quien ha sugerido que el tamaño efectivo puede llegar a variar de una región a otra del genoma. Como sabemos, el tamaño efectivo va directamente relacionado con la deriva génica. Entonces si el tamaño efectivo puede variar, el equilibrio deriva - mutación también y pudiera ser que las diferencias en diversidad observadas no están generadas por el efecto de la selección ó de la recombinación solamente. Finalmente, las interacciones entre las diferentes fuerzas evolutivas probablemente sean más complejas de lo que se ha pensado, sobre todo a nivel genómico, y se requiera de un modelo diferente para poder describirlas.

Entre las múltiples funciones de los genes bajo selección positiva en el genoma central de *E. coli* (Figura 9), destacan las categorías relacionadas con el proceso de transcripción y con la toxicidad, en particular con el estrés oxidativo. En cuanto a las categorías de genes involucrados en la transcripción tenemos la modificación de bases de tARN y rARN, la síntesis de ARN polimerasa ADN-dependiente y reguladores de la transcripción (más específicamente el represor *hdfR* que se encuentra en la categoría funcional de Transcripción – Otros). Y asociados a la toxicidad celular tenemos la categoría de biosíntesis de cofactores - glutatión y análogos - riboflavina, FMN y FAD y la categoría de

procesos celulares – detoxificación. En esta última categoría se encuentran los genes de la catalasa hidroperoxidasa I *katG*, de la bacterioferritina *bfr* que aparentemente confiere resistencia frente a hidroperóxidos (Abdul-Tehrani et al. 1999) y el represor transcripcional *arsR* del operón *ars*, que confiere resistencia a antimonio y arsénico (Carlin et al. 1995).

Por un lado, la presencia de selección diversificadora en los genes asociados al manejo de la toxicidad puede indicarnos la existencia de un proceso de adaptación a las diferencias en concentración de oxígeno que existen entre el medio interno del hospedero y el ambiente externo, entre los cuales transita esta especie (Savageau 1983). Por otro lado, la existencia de diversidad adaptativa en genes asociadas a la transcripción, y el número grande de regiones, aparentemente, no-codificantes también con evidencia de selección positiva (13 loci correspondientes al 0.3% del total), podrían apoyar la idea de que uno de los elementos clave en el proceso adaptativo, es la evolución de la expresión génica y sus redes regulatorias (King y Wilson 1975; Carroll et al. 2001). Evidencia de selección positiva en regiones no-codificantes del genoma ha sido presentada de manera amplia en el género *Drosophila*, donde se ha visto que algunos tipos de regiones no-codificantes son más diversas que el resto del genoma, encontrándose bajo selección positiva, lo que se ha interpretado como prueba de que la regulación transcripcional es un elemento importante en la evolución de este género de insectos (Andolfatto 2005).

Alm et al. (2006) estudiaron la adquisición histórica de genes asociados a la transducción de señales (histidin-cinasas), involucrados en la regulación de la expresión genética, y encontraron que en diversas especies bacterianas, la introducción de nuevos genes de histidin-cinasas a los genomas (ya sea por transferencia horizontal de genes ó por expansión de familias génicas dentro de los linajes) se asocia a eventos de expansión poblacional. Consideran que el aumento en el tamaño de la población es evidencia de que éstas se han adaptado a nuevas condiciones debido a la adquisición de nuevas funciones de regulación de la expresión génica y de nuevos patrones metabólicos ó fisiológicos.

La evolución al nivel de estos genes es muy rápida, pues una sola mutación en uno de estos genes modifica la expresión de más de 100 genes (Giraud et al. 2008). Incluso hay evidencia experimental de adaptación generada por una sola mutación en regulación-cis en *Salmonella* (Osborne et al. 2009).

En cuanto a los genes con evidencia de selección negativa, se esperaba que correspondieran a funciones del metabolismo básico y a la biosíntesis de estructuras celulares esenciales para la supervivencia. Efectivamente, los encontramos agrupados en las categorías funcionales de metabolismo energético - aminoácidos y aminas, anaeróbico y ruta de las pentosas/fosfato; procesos celulares – división celular, quimiotaxis y motilidad y producción de toxinas y resistencia; proteínas de transporte y unión – aminoácidos, péptidos y aminas; biosíntesis de cofactores - pantotenato y coenzimaA; destino de proteínas - modificación de proteínas y reparación; envoltura celular – otros (donde se encuentra el gen *fhuA* de proteína de membrana externa para transporte de ferricromo, el gen *skp* de chaperona periplásmica y el gen de proteína de membrana externa *ompN*); y la categoría de síntesis de proteínas – factores de traducción y otros (donde se encuentran los ARN ribosomales) (Figura 9).

4. PATRONES DE DIVERSIDAD ENTRE INDIVIDUOS DE *E. COLI* CON ESTILOS DE VIDA DIFERENTES

Los ecogrupos analizados, conformados por las cepas de *E. coli* no-patógenas y las cepas de *E. coli* patógena, presentaron una estructura cromosómica diferente entre sí, aunque conservada al interior de cada grupo. En contraste la poza flexible específica al nicho fue reducida en comparación con el resto del genoma. Esto pudiera decirnos que los elementos adquiridos horizontalmente no son el único factor que diferencia ecológicamente a las bacterias, y que pudieran existir diferencias en las regiones conservadas de *E. coli*.

Efectivamente, como explicamos en los resultados, la diversidad fue diferente entre los ecogrupos de *E. coli* estudiados (Tablas 10 y 11). El ecogrupo de cepas patógenas registró mayor diversidad en los dos componentes del pangenoma que el ecogrupo de cepas no-patógenas.

Se obtuvo un rango similar de valores de diversidad nucleotídica π en los loci del genoma central en estos dos ecogrupos, abarcando de 0-0.0004 hasta 0.20571 en *E. coli* no-patógena y de 0-0.0002857 hasta 0.20635 en *E. coli* patógena. Estos rangos son mucho más amplios de lo que se ha descrito para genes de mantenimiento en la especie (π = 0.004 a

0.013; Wirth et al. 2006), y los valores máximos inclusive se encuentran dentro del rango de diversidad que se ha descrito para algunos genes asociados a virulencia (π = 0.03 a 0.24; Castillo et al. 2005). En el genoma flexible, los rangos de diversidad π estimada fueron más amplios, yendo de 0-0.0013106 a 0.214285 en las no-patógenas, y de 0-0.001476 a 0.36 en las patógenas.

Los valores máximos, tanto de genoma central como de genoma flexible de ambos ecogrupos, correspondieron a regiones no-codificantes, adyacentes a regiones del genoma flexible, en el caso del genoma central, y a regiones variables en el caso del genoma flexible. Se ha encontrado que la mayor diversidad dentro del genoma se da en las regiones adyacentes a regiones variables, sujetas a transferencia horizontal. Por esta razón se cree que dichos puntos funcionan como anclas a la recombinación, tanto homóloga como ilegítima, lo cual promueve la acumulación de nuevos alelos y una mayor diversidad que en las demás regiones (Touchon et al. 2009).

Se ha descrito previamente que las cepas patógenas de *E. coli* tienen menor diversidad que las cepas comensales, pero tales patrones se han observado en muestras de patotipos particulares y asociadas a una sola especie de hospedero, como por ejemplo en cepas causantes de septicemia (Maslow et al. 1995), de meningitis neonatal NMEC (Bingen et al. 1998), y cepas del patotipo ETEC en cerdos (Wu et al. 2007).

Por lo tanto los valores de diversidad en cepas patógenas encontrados en este estudio responden a un muestreo más amplia en dos aspectos, primero con respecto a la muestra que corresponde a cepas de diferentes patotipos, y en segundo con relación al marcador que consta de numerosos loci cromosomales. Pudiera ser entonces que, por una parte la diversidad de todos los patotipos en conjunto es efectivamente superior a la de las cepas comensales y de vida libre. Y por otra parte, pudieran ser las regiones no-codificantes, que sí se analizan en este estudio, las que están aumentando la diversidad en las cepas patógenas. Para probar esto, más adelante tendríamos que analizar de manera más fina la diversidad en regiones codificantes y no-codificantes a nivel genómico.

Pero hay otros fenómenos que pudieran explicar porque la diversidad nucleotídica en las cepas patógenas resultó ser mayor, entre ellos la recombinación homóloga (Perna et al. 1998), la evolución por duplicación génica (Jordan 2002) ó el efecto de bacteriófagos

(Inouye et al. 1991; Groisman y Ochman 1996; Rodríguez-Valera et al. 2009), los cuales pueden imponer un efecto de selección dependiente de la frecuencia en *E. coli,* lo que favorecería el mantenimiento de alelos de restricción-modificación raros para contrarrestar a los bacteriófagos, lo que a su vez promovería la recombinación tanto legítima como ilegítima (Levin 1981).

Efectivamente, en este trabajo, el ecogrupo de *E. coli* patógenas fue significativamente menos clonal que el ecogrupo de *E. coli* no-patógenas (Tabla 10 y 11; Figuras 7 y 11). Estos resultados concuerdan con lo previamente reportado en la literatura. Y hay varios estudios en los que se reporta que las variantes patógenas de una misma especie son más recombinantes que las menos patógenas (Parkhill et al. 2003; Hendrickson y Lawrence 2006). Esto se puede explicar desde un punto de vida adaptativo, si consideramos que la recombinación permite generar nuevas combinaciones alélicas que pudieran ser beneficiosas en nichos nuevos, aún cuando fueran ligeramente menos adaptativas en el nicho original. Al menos experimentalmente, se ha visto que la recombinación es capaz de acelerar la adaptación en *E. coli* (Cooper 2007).

Por otra parte, la presencia de selección se considera como señal de la ocurrencia de un proceso adaptativo. Por lo que diferentes patrones en diferentes grupos de individuos se puede considerar como un indicador de que las presiones ambientales están jugando un papel en el moldeo de la diversidad a nivel molecular. Dado que encontramos grandes diferencias en los patrones de selección entre cepas no-patógenas y cepas patógenas de *E. coli*, podemos pensar que es lo que está generando diferentes patrones en la diversidad del genoma central.

La mayoría de los estudios clásicos (usando como marcador isoenzimas), encuentran que el componente principal de la diversidad en *E. coli* es la variación al interior de los hospederos (40% en babuinos amarillos, Routman et al. 1985; 49% en aves de corral, Whittam 1989), y estudios más recientes lo confirman.

Dixit et al. (2004), también usando como marcador isoenzimas y secuencias de ADN de 18 factores de virulencia, encontraron que en cuatro diferentes regiones del tracto gastrointestinal de cerdos de granja (duodeno, íleo, colon y colon distal) hay grupos de genotipos característicos de cada región fisiológica, y que la variación al interior de cada

hospedero (es decir entre las regiones del tracto gastrointestinal) representa el 27% de la diversidad total con estos marcadores.

Entonces, deben ser las diferencias ecológicas dadas por la diversidad fisiológica de los microambientes al interior del hospedero, lo que está promoviendo la diversificación de *E. coli*, por lo que la adaptación al hospedero también debe jugar un papel importante en su estructura poblacional (Souza et al. 1999).

Por ejemplo, Anderson et al. (2006) encuentran que la variación en los ribotipos de un mismo individuo, en hospedero humano, caballo y ganado, durante 1 día es muy alta. Esto sugiere que debe de haber competencia entre estos linajes por los recursos. Se sabe que un factor que puede favorecer la aparición, mantenimiento y transición de bacterias comensales/oportunistas a patógenas/virulentas en un ambiente determinado es la disponibilidad elevada de recursos en el ambiente (Wedekind et al. 2010), sobre todo si el tamaño efectivo es grande (Stevens et al. 2007). En el caso de *E. coli* la capacidad de colonizar y vivir en ambientes nuevos distintos al colon, en donde hay menor competencia por los recursos y por lo tanto mayor disponibilidad de ellos, es lo que probablemente favoreció el surgimiento de las cepas patógenas y lo que ha permitido que permanezcan evolutivamente, como grupos diferentes de las comensales.

Efectivamente, las cepas de *E. coli* diarréicas se establecen al nivel del intestino delgado, y ahí es donde causan principalmente los efectos patógenos (Nataro y Kaper 1998; Kaper et al. 2004), en cambio las cepas comensales habitan en el intestino grueso (Savageau 1983). Por su parte, las cepas extra-intestinales, como su nombre lo indica, sólo son patógenas al nivel del sistema urinario, del sistema nervioso ó de los pulmones, pues, al menos en el caso de las cepas tipo UPEC, se ha visto que pueden habitar normalmente en el aparato gastrointestinal sin causar ningún problema al hospedero y no es sino hasta que llegan al órgano blanco que comienzan a tener un efecto patógeno (Johnson y Russo 2002).

En cuanto a las cepas de EIEC y de *Shigella*, éstas habitan al interior de las células, lo cual también impone nuevos retos y es un ambiente diferente que moldea de manera diferente la diversidad en las bacterias en general, como ocurre en simbiontes (Andersson 2008), en los cuales se ha promovido la reducción del genoma, al utilizar los productos metabólicos del hospedero celular, con lo que ya no requiere gastar energía en cierto tipo de maquinaria genética, y al deshacerse de ella puede aumentar su adecuación.

Así que en su medio las comensales compiten contra las patógenas al adherirse más eficientemente a la mucosa del colon y a las glicoproteínas salivares (Sokurenko et al. 1995, 1998), lo que además les permite utilizar mejor los nutrientes y la producción antimicrobianos (como las colicinas), lo que ha generado las presiones selectivas que promueven que las cepas de *E. coli* menos adaptadas al colon exploren ambientes nuevos, en los cuales ellas son las de mayor adecuación al no tener competencia, y en los cuales, de manera incidental actúan como patógenos para el hospedero (Le Gall et al. 2007).

Este patrón es revelado por el gen *fimH* de adhesión a manosa (Weissman et al. 2006; 2007) el cual se encuentra bajo selección positiva y recombinación, lo que genera numerosas variantes que corresponden a mutaciones adaptativas en el nicho patógeno (mutaciones patoadaptativas; Sokurenko et al. 1999; Weissman et al. 2007). Estas dos fuerzas evolutivas promueven entonces la aparición constante de nuevas variantes no solamente en genes de adhesión celular sino en genes de mantenimiento.

Este proceso constituye un trueque funcional (Weissman et al. 2007), en el cual se sacrifica la adecuación en el nicho de las cepas parentales, en este caso las cepas comensales, por la posibilidad de una mayor adecuación en un nuevo nicho. De tal manera que los mutantes generados en este proceso tienen una mayor ventaja en la colonización de hábitats secundarios, al "sacrificar" la adecuación en su hábitat original. En el caso de *E. coli*, el cambio de hábitat genera el nicho patógeno, el cual a su vez aumenta el potencial de éxito de transmisión de hospedero a hospedero.

Adicionalmente, dentro del ecogrupo de las cepas patógenas analizamos la diversidad del subgrupo de cepas patógenas intestinales y del subgrupo de patógenas extra-intestinales por separado. Encontramos que la diversidad fue mayor en el grupo de *E. coli* patógenas extra-intestinales ($\pi = 0.0125736 \pm 0.0002437$) que en las cepas patógenas intestinales ($\pi = 0.0090948 \pm 0.0001952$). Chattopadhay et al. (2009) reportan valores de diversidad promedio de genes en los genomas de cepas de UPEC (CFT, 536, UTI89) y *E. coli* patógena de ave (APEC01) de 0.004 ± 0.001, menor a lo que nosotros encontramos en todo el genoma central, que incluye no solamente genes sino también regiones no-codificantes, solamente en las cepas CFT, 536 y UTI89.

De acuerdo a la clasificación filogenética tradicional de *E. coli* (Selander et al 1986), las

cepas extra-intestinales se encuentran en el grupo B2 (Johnson et al. 2006A, 2006b; Moulin-Schouleur et al. 2007; Touchon et al. 2009), y las 3 cepas que usamos en el presente estudio están asignadas a dicho grupo, como se muestra en la Tabla 2 en la metodología (Welch et al. 2002; Hochhut et al. 2006; Chen et al. 2006). De acuerdo a estudios recientes (Lecointre et al, 1998; Escobar-Paramo et al. 2004; Jaureguy et al. 2008; Touchon et al. 2009), este linaje B2 sería el más antiguo, lo que puede explicar porque las cepas de este grupo de *E. coli* patógena son las que tienen la mayor diversidad acumulada. Además de esta explicación "histórica", cabe señalar que las cepas urinarias poseen un gran número y diversidad de islas de patogénesis, factores de virulencia y adecuación provenientes hasta de otras especies a lo largo de su genoma (Bingen et al. 1999; Le Gall et al. 2007; Rasko et al. 2008). Esto parece indicar que en estas cepas hay una mayor frecuencia ó tasa de eventos de transferencia horizontal, es decir, que probablemente son más recombinantes que las cepas intestinales. Sería interesante saber si la recombinación es algo que ya estaba presente en estas cepas antes de que comenzaran a invadir el nicho de patogénesis extra-intestinal, ó sí es una característica que surgió en consecuencia de la colonización y adaptación a ese nuevo nicho.

Finalmente, cabe señalar que los resultados obtenidos representan sólo un fragmento de la diversidad total de la especie y es solamente un vistazo a la historia evolutiva de una fracción de las poblaciones de *E. coli*. Dado que el grupo de cepas patógenas analizado es en su mayoría representativo de los patotipos de naturaleza epidémica, los cuales se dispersan clonalmente, como son las EHEC y las UPEC (Johnson et al. 2002), las características de la muestra y en específico de las cepas patógenas aquí presentadas podrían estar reflejando la dinámica evolutiva de un grupo particular de cepas, cuyo genotipo, al originarse y expandirse en el tiempo y en el espacio (clonas epidémicas) presenta un comportamiento diferente al de la población en general (Maynard-Smith 1993). Posteriormente tendríamos que verificar si las diferencias en el patrón de diversidad en el genoma central se mantienen aún agregando cepas de patotipos no-clonales, como las ETEC, las EPEC ó las EAEC, y más cepas comensales y de vida libre de *E. coli*.

5. PERSPECTIVAS

Podemos comentar varios aspectos del presente estudio que se podrían estudiar más detalladamente.

En primer lugar la muestra que analizamos fue pequeña. Una muestra pequeña puede sesgar de varias maneras los resultados. Ya que podría no representar toda la diversidad de la especie. A éste respecto Rasko et al. (2008) calculan que con los genomas de al menos 17 individuos se puede obtener el total de genes que conforman el genoma central de la especie. Sin embargo, también se estimó que cada nueva cepa de *E. coli* posee aproximadamente 300 nuevos genes del genoma flexible. Lo que implica que la diversidad génica de la especie aumentará con cada nuevo genoma secuenciado. Es decir que talvez no haya una muestra que pueda representar efectivamente la diversidad total del pangenoma de la especie. Así que, mientras más individuos de la especie podamos analizar será más informativo, sin embargo lo ideal sería tener al menos 17 genomas secuenciados.

Basándonos en el trabajo de González-González (en preparación) realizado sobre una muestra de individuos de la colección del Instituto de Ecología (IECOL) de *E. coli*, ya se han considerado individuos que podrían ser interesantes en varios aspectos. Por una parte tenemos individuos que presentaron un número de copias de secuencias ribosomales diferente al de la cepa K12 MG1665 (7 copias en el genoma), como son comensales de humano (cepa 6879; 8 ribosomales), de demonio de Tasmania (cepa 3497; 9 ribosomales) y aisladas del ambiente (cepas 3720 y 3885; 8 ribosomales). Algunos otros individuos se destacaron por presentar un tamaño del replicón cromosómico muy grande, como son comensales de manatí (cepa 1735; 5.725 Mb) y aisladas del ambiente (cepa 3658, 3712 y 5058; 5.698 Mb, 5.663 Mb y 5.614 Mb respectivamente) e individuos patógenos aislados de coyote (cepa 830; 5.603 Mb), de humano (cepa 3622; 5.646 Mb) y del ambiente (cepas 3651, 3659, 3691; 5.696 Mb, 5.606 Mb; 6.097 Mb respectivamente). Asimismo, la cepa 43221, del patotipo ETEC de humano, es interesante, al presentar un patrón muy divergente, en cuanto a la combinación alélica dada por los genes del MLST.

Adicionalmente, la muestra debería de representar mejor, no sólo la diversidad genética sino la diversidad ecológica de la especie. Es decir con varios individuos de todos los nichos y hábitats en los que se puede encontrar la especie: en el ambiente (acuático y

terrestre), asociados a hospedero de diferentes especies (mamíferos y reptiles), comensales y patógenos (intestinales y extra-intestinales). A partir de González-González (en preparación), sugerimos algunos individuos del patotipo EAEC aislados de humano (cepas 3607, 3609 y 3622), del patotipo EPEC y EHEC, aisladas de cerdos (cepas 4952, 4953 y 5014), del patotipo ETEC aisladas de murciélagos y de ratón (cepas 33 y 75) y algunas comensales aisladas de animales salvajes de diferentes partes del mundo, como de águila, pecarí y mapache de México (cepas 55, 2055 y 2064) y de potoro de Australia (cepa 2284).

Por otra parte, una muestra más amplia nos permitiría hacer otro tipo de análisis, más detallados, como de estructuración genética, de tasas de recombinación y otras pruebas de selección, basados en modelos de coalescencia, que arrojarían resultados más robustos.

Finalmente, sería importante delimitar los análisis a unidades funcionales codificantes y no-codificantes. Esto permitirá una interpretación más fina de los resultados, pues en el presente estudio realmente no podemos saber si las señales encontradas en los loci se deben a las regiones correspondientes a genes ó a las regiones no-codificantes. Asimismo, sería conveniente hacer un análisis con ventanas corredizas, para conocer mejor la amplitud y definir exactamente las regiones bajo selección ó desequilibrio de ligamiento.
Y, una vez definida la naturaleza evolutiva de cada región del genoma, la información se puede utilizar para determinar cuales son los marcadores más efectivos para hacer genética de poblaciones, definir las relaciones filogenéticas entre individuos, ó detectar patrones geográficas de distribución de la variación, en muestras más grandes.

CONCLUSIONES

- En *Escherichia coli*, la diversidad genética de los elementos que constituyen el genoma central, donde se encuentran los genes esenciales y de mantenimiento, los cuales deberían estar muy conservados, es alta ($\pi = 0.0217819 \pm 0.0002824$) con respecto a otras especies de bacterias. Y aunque fue menor a lo encontrado en regiones del genoma flexible ($\pi = 0.0251052 \pm 0.0017452$ en cepas no patógenas y $\pi = 0.0444992 \pm 0.0036672$ en cepas patógenas), esto nos indica que la diversidad del genoma no está limitada a los elementos adquiridos por transferencia horizontal.

- Las regiones de genoma central presentaron poca evidencia de desequilibrio de ligamiento (4.9% de loci del cromosoma resultaron clonales). Esto sugiere que la mayor parte de la diversidad del genoma central de *E. coli* pudiera estar influenciada por el efecto de la recombinación homóloga. Aunque si consideramos el espectro de diversidad poblacional en bacterias desde clonal hasta panmixis, *E. coli* probablemente se encuentre en el límite más cercano a clonal, con respecto a otras especies. De cualquier manera, aún cuando la muestra utilizada es pequeña y conformada por individuos que son clonas (particularmente en el caso de las cepas de K12 y de O157:H7), *E. coli* es más recombinante de lo que se había propuesto tradicionalmente.

- Los patrones de diversidad del genoma central de *E. coli* no concuerdan en su totalidad con el modelo de evolución neutral del genoma. En primer lugar porque aún cuando fue un porcentaje relativamente pequeño de regiones bajo selección (5.2%), la mayor proporción de éstas se encontró bajo selección positiva (4.8%), lo que contradice al modelo neutro. En segundo lugar, porque la diversidad a lo largo del genoma no fue homogénea, por un lado debido al efecto de la recombinación homóloga y por otro, al efecto de la selección natural.

- La regulación de la expresión genética es posiblemente un proceso importante en la historia evolutiva de *E. coli*, ya que encontramos evidencia de selección positiva en regiones no-codificantes, donde se encuentran sitios de reconocimiento de promotores,

represores e inductores de la transcripción y en numerosos genes asociados al proceso de transcripción.

- El proceso de adaptación de *E. coli* no está determinada únicamente por los elementos del genoma flexible, dado que se encontraron evidencias de selección positiva a este nivel, en toda la muestra. En particular, las diferencias en los patrones evolutivos en los elementos del genoma central, debido a la acción de la selección natural y de la recombinación entre cepas patógenas y no-patógenas, parecen sugerir que no todo el proceso de adaptación ecológica se debe a los cambios en el repertorio del genoma flexible ni a la diversidad acumulada a ese nivel del genoma, sino que el genoma central también participa en dicho proceso.

- En el ecogrupo de cepas patógenas de *E. coli*, la mayor diversidad se encuentra en las cepas extra-intestinales, tanto en el genoma central ($\pi = 0.0125736$) como en el genoma flexible ($\pi = 0.0200723$), en comparación con las cepas patógenas intestinales que presentan menor diversidad en el genoma central ($\pi = 0.0090948$) y en el genoma flexible ($\pi = 0.0129991$). Esta alta diversidad puede ser explicada debido a que las cepas extra-intestinales analizadas en este estudio son parte de uno de los linajes más antiguos de *E. coli*. Asimismo, las presiones selectivas de un ambiente diferente al del colón podrían estar promoviendo un aumento en la diversidad de dichas cepas (poblaciones).

REFERENCIAS

Abby S, V Daubin. 2007. Comparative genomics and the evolution of prokaryotes. Trends Microbiol. (3):135-41.

Abdul-Tehrani H, Hudson AJ, Chang YS, Timms AR, Hawkins C, Williams JM, Harrison PM, Guest JR, Andrews SC. 1999. Ferritin mutants of *Escherichia coli* are iron deficient and growth impaired, and fur mutants are iron deficient. J Bacteriol 181(5);1415-28.

Akey JM. 2009. Constructing genomic maps of positive selection in humans: Where do we go from here? Genome Res. 19: 711-722.

Allen EE, GW Tyson, RJ Whitaker, JC Detter, PM Richardson, JF Banfield. 2007. Genome dynamics in a natural archaeal population. PNAS 104(6):1883–1888

Alm E, K Huang, A Arkin. 2006. The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. PLoS Comput Biol. 2(11):e143

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Anantha RP, AL McVeigh, LH Lee, MK Agnew, FJ Cassels, DA Scott, TS Whittam, SJ Savarino. 2004. Evolutionary and Functional Relationships of Colonization Factor Antigen I and Other Class 5 Adhesive Fimbriae of Enterotoxigenic *Escherichia coli*. Infection and Immunity 72(12):7190-7201

Anderson KL, JE Whitlock, VJ Harwood. 2005. Persistence and differential survival of fecal indicator bacteria in subtropical waters and sediments. Appl Environ Microbiol. 71(6):3041-8.

Anderson MA, JE Whitlock, VJ Harwood. 2006. Diversity and distribution of *E. coli* genotypes and antibiotic resistance phenotypes in feces of humans, cattle, and horses. Appl. Environ. Microbiol. 72 (11):6914–6922.

Anderson P, J Roth. 1981. Spontaneous tandem genetic duplications in *Salmonella typhimurium* arise by unequal recombination between rRNA (rrn) cistrons. Proc Natl Acad Sci USA. 78:3113–3117.

Andersson DI. 2008. Shrinking Genomes: Former skeptics recognize that the genomes of microbial parasites and symbionts are subject to dynamic downsizing. American Society for Microbiology.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. Nature. 437(7062):1149-52.

Atwood KC, LK Schneider, FJ Ryan. 1951. Periodic selection in *Escherichia coli*. Proc Natl Acad Sci U S A. 37(3):146-55.

Ayala FJ. 1982. Genetic variation in natural populations: problem of electrophoretically cryptic alleles. Proc. Natl. Acad. Sci. USA 79:550-554.

Barrick JE, Kauth MR, Strelioff CC, Lenski RE. 2010. *Escherichia coli* rpoB mutants have increased evolvability in proportion to their fitness defects. Mol Biol Evol. 27(6):1338-47.

Begun DJ, AK Holloway, K Stevens, LW Hillier, YP Poh, MW Hahn, PM Nista, CD Jones, AD Kern, CN Dewey, L Pachter, E Myers, CH Langley. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biol. 5(11):e310.

Bekal S, R Brousseau, L Masson, G Prefontaine, J Fairbrother, J Harel. 2003. Rapid identification of *Escherichia coli* pathotypes by virulence gene detection with DNA microarrays. J. Clin. Microbiol. 41:2113–2125.

Bensasson D, M Zarowiecki, A Burt, V Koufopanou. 2008. Rapid evolution of yeast centromeres in the absence of drive. Genetics. 178(4):2161-7.

Bergthorsson U, H Ochman. 1998. Distribution of chromosome length variation in natural isolates of *Escherichia coli*. Mol Biol Evol 15: 6–16.

Bidet P, Mahjoub-Messai F, Blanco J, Blanco J, Dehem M, Aujard Y, Bingen E, Bonacorsi S. 2007. Combined multilocus sequence typing and O serogrouping distinguishes *Escherichia coli* subtypes associated with infant urosepsis and/or meningitis. J Infect Dis. 196(2):297-303.

Bingen E, B Picard, N Brahimi, S Mathy, P Desjardins, J Elion, E Denamur. 1998. Phylogenetic analysis of *Escherichia coli* strains causing neonatal meningitis suggests horizontal gene transfer from a predominant pool of highly virulent B2 group strains. J. Infect. Dis. 177:642–650.

Binnewies TT, Y Motro, PF Hallin, O Lund, D Dunn, T La, DJ Hampson, M Bellgard, TM Wassenaar, DW Ussery. 2006. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. Funct Integr Genomics. 6(3):165-85.

Black WC 4th, CF Baer, MF Antolin, NM DuTeau. 2001. Population genomics: genome-wide sampling of insect populations. Annu Rev Entomol. 46:441-69. Review.

Blattner FR, GIII Plunkett, CA Bloch, NT Perna, V Burland, M Riley, J Collado-Vides, JD Glasner, CK Rode, GF Mayhew, J Gregor, NW Davis, HA Kirkpatrick, MA Goeden, DJ Rose, B Mau, Y Shao. 1997. The complete genome sequence of *Escherichia coli* K-12. Science. 277: 1453–1462.

Brown AH, Feldman MW, Nevo E. 1980. Multilocus Structure of Natural Populations of *Hordeum spontaneum*. Genetics. 96(2):523-536.

Brown EW, JE LeClerc, B Li, WL Payne, TA Cebula. 2001. Phylogenetic evidence for horizontal transfer of mutS alleles among naturally occurring *Escherichia coli* strains. J Bacteriol. 183(5):1631-44.

Brownlie JC, M Adamski, B Slatko, EA McGraw.2007. Diversifying selection and host adaptation in two endosymbiont genomes. BMC Evolutionary Biology 7:68.

Brudno M, CB Do, GM Cooper, MF Kim, E Davydov, ED Green, A Sidow, S Batzoglou, NISC Comparative Sequencing Program. 2003. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. 13:721–731.

Carlin A, Shi W, Dey S, Rosen BP. 1995. The ars operon of *Escherichia coli* confers arsenical and antimonial resistance. J Bacteriol. 177(4):981-6.

Carrillo M, Estrada E, Hazen TC. 1985. Survival and enumeration of the fecal indicators *Bifidobacterium adolescentis* and *Escherichia coli* in a tropical rain forest watershed. Appl Environ Microbiol. 50(2):468-76.

Carroll SB, Grenier JK, Weatherbee SD. 2001. From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design. Blackwell Science, Malden, Massachusetts.

Castillo A, LE Eguiarte, V Souza. 2005. A genomic population genetics analysis of the pathogenic enterocyte effacement island in *Escherichia coli*: the search for the unit of selection. Proc Natl Acad Sci U S A. 102(5):1542-7.

Castillo A. 2007. La selección natural a nivel molecular. Capítulo 1 en Eguiarte LE, V Souza y X Aguirre. Ecología Molecular. Secretaría de Medio Ambiente y Recursos Naturales, Instituto Nacional de Ecología, Universidad Nacional Autónoma de México, Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. 592 pp.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 10(3):195-205.

Charlesworth J, A Eyre-Walker. 2006. The rate of adaptive evolution in enteric bacteria. Mol Biol Evol. 23(7):1348-56.

Chattopadhyay S, Weissman SJ, Minin VN, Russo TA, Dykhuizen DE, Sokurenko EV. 2009. High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection. Proc Natl Acad Sci U S A. 106(30):12412-7.

Chen PE, C Cook, AC Stewart, N Nagarajan, DD Sommer, M Pop, B Thomason, MP Kiley Thomason, S Lentz, N Nolan, S Sozhamannan, A Sulakvelidze, A Mateczun, L Du, ME Zwick, TD Read. 2010. Genomic characterization of the *Yersinia* genus. Genome Biol. 11(1): R1.

Chen Q, SJ Savarino, MM Venkatesan. 2006. Subtractive hybridization and optical mapping of the enterotoxigenic *Escherichia coli* H10407 chromosome: isolation of unique sequences and demonstration of significant similarity to the chromosome of *E. coli* K-12. Microbiology. 152(Pt 4):1041-54.

Choudhary M, X Zanhua, YX Fu, S Kaplan. 2007. Genome Analyses of Three Strains of *Rhodobacter sphaeroides*: Evidence of Rapid Evolution of Chromosome II. Journal of Bacteriology 189(5):1914–1921.

Cohan FM. 2006. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. Philos Trans R Soc Lond B Biol Sci. 361(1475):1985-96.

Crow JF. 2008. Mid-century controversies in population genetics. Annu Rev Genet. 42:1-16.

Cubillos-Ruiz A, Morales J, Zambrano MM. 2008. Analysis of the genetic variation in *Mycobacterium tuberculosis* strains by multiple genome alignments. BMC Res Notes. 1:110.

Darling ACE, B Mau, FR Blattner, NT Perna. 2004. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Res. 14: 1394-1403.

Darling AE, B Mau, NT Perna. 2010. ProgressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. PLoS ONE 5(6): e11147.

Darling AE, I Miklós, MA Ragan. 2008. Dynamics of Genome Rearrangement in Bacterial Populations. PLoS Genet. 4(7): e1000128.

D'Auria, N Jiménez-Hernández, F Peris-Bondia, A Moya, A Latorre. 2010. *Legionella pneumophila* pangenome reveals strain-specific virulence factors. BMC Genomics. 11: 181.

Davidsen T, Beck E, Ganapathy A, Montgomery R, Zafar N, Yang Q, Madupu R, Goetz P, Galinsky K, White O, Sutton G. 2010. The comprehensive microbial resource. Nucleic Acids Res. 01(38): D340-5.

Dawson E, GR Abecasis, S Bumpstead, Y Chen, S Hunt, DM Beare, J Pabial, T Dibling, E Tinsley, S Kirby, D Carter, M Papaspyridonos, S Livingstone, R Ganske, E Lõhmussaar, J Zernant, N Tõnisson, M Remm, R Mägi, T Puurand, J Vilo, A Kurg, K Rice, P Deloukas, R Mott, A Metspalu, DR Bentley, LR Cardon, I Dunham. 2002. A first-generation linkage disequilibrium map of human chromosome 22. Nature. 418(6897):544-8.

Dayhoff MO, WC Barker, LT Hunt. 1983. Establishing homologies in protein sequences. Methods Enzymol 91:524–545.

Delcher AL, S Kasif, RD Fleischmann, J Peterson, O White, SL Salzberg. 1999. Alignment of whole genomes. Nucleic Acids Res. 27: 2369–2376.

Depaulis F, M Veuille. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. 15:1788-1790.

Desjardins P, B Picard, B Kaltenbock, J Elion, E Denamur. 1995. Sex in *Escherichia coli* does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction fragment length polymorphism. J Mol Evol. 41:440-448.

Devlin B, Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322.

Didelot X, A Darling, D Falush. 2009. Inferring genomic flux in bacteria. Genome Res. 19:306-317

Didelot X, D Falush. 2007. Inference of Bacterial Microevolution Using Multilocus Sequence Data. Genetics 175:1251–1266

Dixit SM, Gordon DM, Wu XY, Chapman T, Kailasapathy K, Chin JJ. 2004. Diversity analysis of commensal porcine *Escherichia coli* - associations between genotypes and habitat in the porcine gastrointestinal tract. Microbiology. 150(Pt 6):1735-40.

Do CB, K Katoh. 2009. Protein Multiple Sequence Alignment en Thompson JD, C Schaeffer-Reis, M Ueffing (Ed.). Methods in Molecular Biology, vol. 484: Functional Proteomics: Methods and Protocols. Pp. 379-413. Nueva York: Humana Press

Dobrindt U, B Hochhut, U Hentschel, J Hacker. 2004. Genomic Islands in Pathogenic and Environmental Microorganisms. Nature Reviews 2:414

Dobrindt U, J Hacker. 2008. Targeting virulence traits: potential strategies to combat extraintestinal pathogenic *E. coli* infections. Curr Opin Microbiol. 11(5):409-13.

Doolittle WF, RT Papke. 2006. Genomics and the bacterial species problem. GenomeBiology 7:116

Doolittle WF. 1999a. Lateral genomics. Trends Cell Biol. 9(12):M5-8

Doolittle WF. 1999b. Phylogenetic classification and the Universal Tree. Science. 284(5423):2124-2128.

Dos Vultos T, O Mestre, J Rauzier, M Golec, N Rastogi, V Rasolofo, T Tonjum, C Sola, I Matic, B Gicquel. 2008. Evolution and Diversity of Clonal Bacteria: The Paradigm of Mycobacterium tuberculosis. PLoS ONE 3(2): e1538.

Durfee T, R Nelson, S Baldwin, G Plunkett III, V Burland, B Mau, JF Petrosino, X Qin, DM Muzny, M Ayele, RA Gibbs, B Csorgo, G Pósfai, GM Weinstock, FR Blattner. 2008. The Complete Genome Sequence of *Escherichia coli* DH10B: Insights into the Biology of a Laboratory Workhorse. Journal of Bacteriology 190(7):2597-2606.

Dziva F, Stevens MP. 2008. Colibacillosis in poultry: unravelling the molecular basis of virulence of avian pathogenic *Escherichia coli* in their natural hosts. Avian Pathol. 37(4):355-66. Review.

Ebert D, JJ Bull. 2003. Challenging the trade-off model for the evolution of virulence: is virulence management feasible? Trends Microbiol. 11:15 - 20.

Eguiarte LE. 1999. Una Guía para Principiantes a la Genética de Poblaciones en J Núñez Farfán y LE.Eguiarte. La Evolución Biológica. Mexico, D.F., UNAM pp: 35-50.

Eisen JA, Heidelberg JF, White O, Salzberg SL. 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol, 1:RESEARCH0011.

Elena SF, Lenski RE. 2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nature Rev. Genet. 4, 457–469.

Ellegren H. 2008. Comparative genomics and the study of evolution by natural selection. Mol Ecol. 17(21):4586-96. Review.

Escobar-Paramo P, A Sabbagh, P Darlu, O Pradillon, C Vaury, E Denamur, G Lecointre. 2004. Decreasing the effects of horizontal gene transfer on bacterial phylogeny: the *Escherichia coli* case study. Mol Phylogenet Evol 30: 243–250.

Eslava C, J Mateo J, A Cravioto. 1994. Cepas de *Escherichia coli* relacionadas con la diarrea en Giono S, A Escobar y JL Valdespino. Diagnóstico de laboratorio de infecciones gastrointestinales. Secretaria de Salud. México, 1994: 251 pp.

Fakhr MK, NoLan LK, Logue CM. 2005. Multilocus sequence typing lacks the discriminatory ability of pulsed-field gel electrophoresis for typing *Salmonella* enterica serovar Typhimurium. J Clin Microbiol. 43(5): 2215–9.

Falush D. 2009. Toward the use of genomics to study microevolutionary change in bacteria. PLoS Genet. 5(10):e1000627.

Fay JC, CI Wu. 2000. Hitchhiking under positive Darwinian selection. Genetics. 155(3):1405–1413.

Felsenstein NJ . 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Ferenci T, Zhou Z, Betteridge T, Ren Y, Liu Y, Feng L, Reeves PR, Wang L 2009. Genomic Sequencing Reveals Regulatory Mutations and Recombinational Events in the Widely Used MC4100 Lineage of *Escherichia coli* K-12. J. Bacteriol. 191: 4025-4029.

Fleischmann R, M Adams, O White, R Clayton, E Kirkness, A Kerlavage, C Bult, J Tomb, B Dougherty, J Merrick. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269 (5223):496–512

Fraser CM, JD Gocayne, O White, MD Adams, RA Clayton, RD Fleischmann, CJ Bult, AR Kerlavage, G Sutton, JM Kelley, RD Fritchman, JF Weidman, KV Small, M Sandusky, J Fuhrmann, D Nguyen, TR Utterback, DM Saudek, CA Phillips, JM Merrick, JF Tomb, BA Dougherty, KF Bott, PC Hu, TS Lucier, SN Peterson, HO Smith, CA Hutchison, JC Venter. 1995. The minimal gene complement of *Mycoplasma genitalium*. Science. 270(5235):397-403.

Fricke WF, MS Wright, AH Lindell, DM Harkins, C Baker-Austin, J Ravel, R Stepanauskas. 2008. Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate *Escherichia coli* SMS-3-5. J Bacteriol. 190(20):6779-94.

Fu YX, WH Li. 1993. Statistical tests of neutrality of mutations. Genetics 133:693-709.

Fukushima M, Kakinuma K, Kawaguchi R. 2002. Phylogenetic analysis of *Salmonella*, *Shigella*, and *Escherichia coli* strains on the basis of the gyrB gene sequence. J Clin Microbiol. 40(8):2779-85.

Futuyma DJ. 2005. Evolution. Sunderland, Massachusetts: Sinauer Associates, Inc.

Garcia-Vallve S, Romeu A, Palau J. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res 10:1719 - 1725.

Giraud A S Arous, M De Paepe, V Gaboriau-Routhiau, JC Bambou, S Rakotbe, AB Lindner, F Taddei, N Cerf-Bensussan. 2008. Dissecting the Genetic Components of Adaptation of *Escherichia coli* to the Mouse GutPLOS Genetics 4(1):e2.

Gogarten JP, WF Doolittle, JG Lawrence. 2002. Prokaryotic Evolution in Light of Gene Transfer. Mol. Biol. Evol. 19(12):2226–2238.

Gottesman S. 2004. The small RNA regulators of *Escherichia coli*: roles and mechanisms. Annu Rev Microbiol. 58:303-28. Review.

Groisman EA, H Ochman. 1996. Pathogenicity islands: bacterial evolution in quantum leaps. Cell 87: 791–794.

Guindon S, O Gascuel. 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52(5):696-704.

Hacker J, JB Kaper. 2000. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol. 54:641–679.

Hall BK. 1994. Homology: the hierarchical basis of comparative biology. Salt Lake City, Ut: Academic Press. Pp. 483

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41:95-98.

Hansson S, D Caugant, U Jodal, C Svanborg-Eden. 1989. Untreated asymptomatic bacteriuria in girls: I-stability of urinary isolates. BMJ 298:853-855.

Hardy GH. 1908. Mendelian proportions in a mixed population. Science 28: 49–50.

Harris H. 1966. Enzyme polymorphisms in man. Proceedings of the Royal Society of London, Series B, Biological Sciences. 164 (955): 298-310.

Hartl DL, DE Dykhuizen. 1984. The population genetics of *Escherichia coli*. Annu. Rev. Genet. 18:31-86.50.

Hayashi T, K Makino, M Ohnishi, K Kurokawa, K Ishii, K Yokoyama, CG Han, E Ohtsubo, K Nakayama, T Murata, M Tanaka, T Tobe, T Iida, H Takami, T Honda, C Sasakawa, N Ogasawara, T Yasunaga, S Kuhara, T Shiba, M Hattori, H Shinagawa. 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res. 8(1):11-22. Erratum en: DNA Res 2001 Apr 27;8(2):96.

Hedrick PW. 2005. Genetics of Populations. 3 ed. Jones & Bartlett Publishers, Sudbury, Massachusetts.

Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. Am. J. Hum. Genet. 72:1527–1535.

Hendrickson H, JG Lawrence. 2006. Selection for chromosome architecture in bacteria. J Mol Evol. 62(5):615-29.

Hill WG, A Robertson. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38(6): 226-231.

Ho Yoon S, H Jeong, SK Kwon, JF Kim. 2009. Genomics, biological features and biotechnological applications of *Escherichia coli* B: "Is B for better". En Lee SY (ed.). Systems Biology and Biotechnology of *Escherichia coli*. Pp. 1-18. Springer Science Business Media BV.

Hochhut B, C Wilde, G Balling, B Middendorf, U Dobrindt, E Brzuszkiewicz, G Gottschalk, E Carniel, J Hacker. 2006. Role of pathogenicity island-associated integrases in the genome plasticity of uropathogenic *Escherichia coli* strain 536. Mol Microbiol. 61(3):584-95.

Hudson RR, M Kreitman, M Aguade. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics. 116:153–159.

Hudson RR. 1990. Gene genealogies and the coalescent process. Oxf. Sur v. Evol. Biol. 9:1-44.

Hughes AL, R Friedman. 2004. Patterns of sequence divergence in 5' intergenic spacers and linked coding regions in 10 species of pathogenic Bacteria reveal distinct recombinational histories. Genetics 168: 1795-1803.

Hurst LD. 2009. Fundamental concepts in genetics: genetics and the understanding of selection. Nat Rev Genet. 10(2):83-93.

Hutter S, AJ Vilella, J Rozas. 2006. Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics. 7:409.

Hutter S, H Li, S Beisswanger, D De Lorenzo, W Stephan. 2007. Distinctly different sex ratios in African and European populations of Drosophila melanogaster inferred from chromosomewide single nucleotide polymorphism data. Genetics. 177(1):469-80.

Inouye S, Sunshine MG, Six EW, Inouye M. 1991. Retronphage phi R73: an *E. coli* phage that contains a retroelement and integrates into a tRNA gene. Science. 1991 May 17;252(5008):969-71.

Instituto SAS Inc., Cary, NC. JMP, Versión 7. 1989-2007.

Jaureguy F, L Landraud, V Passet, L Diancourt, E Frapy, G Guigon, E Carbonnelle, O Lortholary, O Clermont, E Denamur, B Picard, X Nassif, S Brisse. 2008. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. BMC Genomics. 9: 560.

Johnson JR, AR Manges, TT O'Bryan, LR Riley. 2002. A disseminated multidrug-resistant clonal group of uropathogenic *Escherichia coli* in pyelonephritis. Lancet 359:2249–2251.

Johnson JR, TA Russo. 2002. Extraintestinal pathogenic *Escherichia coli*: "The other bad E coli". Journal of Laboratory and Clinical Medicine. 139(3):155-162.

Johnson TJ, Johnson SJ, NoLan LK. 2006a. Complete DNA sequence of a ColBM plasmid from avian pathogenic *Escherichia coli* suggests that it evolved from closely related ColV virulence plasmids. J Bacteriol. 188(16):5975-83.

Johnson TJ, S Kariyawasam, Y Wannemuehler, P Mangiamele, SJ Johnson, C Doetkott, JA Skyberg, AM Lynne, JR Johnson, LK Nolan. 2007. The genome sequence of avian pathogenic *Escherichia coli* strain O1:K1:H7 shares strong similarities with human extraintestinal pathogenic *E. coli* genomes. J Bacteriol. 189(8):3228-36. Erratum in: J Bacteriol. 2007 Jun;189(12):4554.

Johnson TJ, Wannemeuhler YM, Scaccianoce JA, Johnson SJ, NoLan LK. 2006b. Complete DNA sequence, comparative genomics, and prevalence of an IncHI2 plasmid occurring among extraintestinal pathogenic *Escherichia coli* isolates. Antimicrob Agents Chemother. 50(11):3929-33.

Johnson TJ, Y Wannemuehle, SJ Johnson, AL Stell, C Doetkott, JR Johnson, KS Kim, L Spanjaard, LK Nolan. 2008. Comparison of extraintestinal pathogenic *Escherichia coli* strains from human and avian sources reveals a mixed subset representing potential zoonotic pathogens. Appl Environ Microbiol. 74(22):7043-50.

Jordan K, Makarova KS, Spouge JL, Wolf YI, Koonin EV. 2001. Lineage specific gene expansions in bacterial and archaeal genomes. Genome Res 10:555-565.

Kaper JB, JP Nataro, HL Mobley. 2004. Pathogenic *Escherichia coli*. Nat. Rev. Microbiol. 2:123-40.

Kelly JK. 1997. A test of neutrality based on interlocus associations. Genetics. 146:1197-1206.

Kimura M, T Ohta. 1969a. The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. Genetics. 61(3):763-771.

Kimura M, T Ohta. 1969b. The average number of generations until extinction of an individual mutant gene in a finite population. Genetics. 63(3):701-9.

Kimura M. 1968. Evolutionary rate at the molecular level. Nature. 217(5129):624-6.

Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics. 61(4):893-903.

Kimura M. 1983. The neutral theory of Molecular Evolution. Cambridge University Press, Cambridge, Massachusetts.

Kimura M. 1989. The neutral theory of molecular evolution and the world view of the neutralists. Genome. 31(1):24-31.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science 188:107–116.

Kingman JFC. 1982. On the genealogy of large populations. J. Appl. Prob. 19A:27–43.

Koonin EV, YI Wolf. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 36(21):6688-719. Review.

Kreitman M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *D. melanogaster*. Nature (London). 304:412-417.

Ksoll WB, S Ishii, MJ Sadowsky, RE Hicks. 2007. Presence and sources of fecal coliform bacteria in epilithic periphyton communities of Lake Superior. Appl Environ Microbiol. 73(12):3771-8.

Kuhnert P, P Boerlin, J Frey. 2000. Target genes for virulence assessment of *Escherichia coli* isolates from water, food and the environment. FEMS Microbiol. Rev. 24:107–117.50.

Lan R, PR Reeves. 2000. Intraspecies variation in bacterial genomes: the need for a species genome concept. Trends Microbiol. 8: 396–401.

Lanave C, G Preparata, C Saccone, G Serio. 1984. A new method for calculating evolutionary substitution rates. Journal of Molecular Evolution 20:86–93.

Lawrence JG, H Ochman. 1998. Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci USA. 95:9413-7.

Lawrence JG. 2001. Catalyzing bacterial speciation: correlating lateral transfer with genetic headroom. Syst. Biol. 50:479- 496.

Le Gall T, Clermont O, Gouriou S, Picard B, Nassif X, Denamur E, Tenaillon O. 2007. Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains. Mol Biol Evol. 24(11):2373-84.

Lecointre G, L Rachdi, P Darlu, E Denamur. 1998. *Escherichia coli* molecular phylogeny using the incongruence length difference test. Mol Biol Evol 15: 1685–1695.

Lerat E, H Ochman. 2005. Recognizing the pseudogenes in bacterial genomes. Nucleic Acids Res. 33(10):3125-32.

Levin BR. 1981. Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. Genetics. 99(1):1-23.

Levin BR. 1996. The evolution and maintenance of virulence in microparasites. Emerg Infect Dis. 2:93–102.

Lewontin R C. 1974. The analysis of variance and the analysis of causes. American Journal of Human Genetics. 26: 400-411.

Lewontin RC, JL Hubby. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics 54:595-609.

Lewontin RC, K Kojima. 1960. The evolutionary dynamics of complex polymorphisms. Evolution 14:458-472.

Lewontin RC. 1964a. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics. 49(1):49-67.

Lewontin RC. 1964b. The interaction of selection and linkage II. Optimum models. Genetics. 50:757-82.

Lewontin RC. 1995. The detection of linkage disequilibrium in molecular sequence data. Genetics. 140(1):377-88.

Li WH. 1977a. Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. Genetics. 85(2):331-7.

Li WH. 1977b. Maintenance of genetic variability under mutation and selection pressures in a finite population. Proc Natl Acad Sci U S A. 74(6):2509-13.

Librado P, J Rozas. 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 25:1451-1452.

Logan NA. 1994. Bacterial systematics. Blackwell scientific publications, Oxford, 263pp.

Luikart G, PR England, D Tallmon, S Jordan, P Taberlet. 2003. The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet. 4(12):981-94.

Maiden MC, JA Bygraves, E Feil, G Morelli, JE Russell, R Urwin, Q Zhang, J Zhou, K Zurth, DA Caugant, IM Feavers, M Achtman, BG Spratt. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 95(6):3140-5.

Maslow JN, Whittam TS, Gilks CF, et al. 1995. Clonal relationships among bloodstream isolates of *Escherichia coli*. Infect Immun. 63: 2409-27.

Mau B, JD Glasner, AE Darling, NT Perna. 2006. Genome-wide detection and analysis of homologous recombination among sequenced strains of *Escherichia coli*. Genome Biol. 7(5):R44.

Maynard Smith J, NH Smith, M O'Rourke, BG Spratt. 1993. How clonal are bacteria? Proc Natl. Acad. Sci. USA 90:4384–4388.

Milkman R, MM Bridges. 1990. Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. Genetics. 126(3):505-17. Erratum in: Genetics 1990 Dec;126(4):1139.

Milkman R. 1973. Electrophoretic variation in *Escherichia coli* from natural sources. Science. 182(116):1024-6.

Mira A, L Klasson, SG Andersson. 2002. Microbial genome evolution: sources of variability. Curr Opin Microbiol. 5(5):506-12.

Mokady D, U Gophna, EZ Ron. 2005. Extensive gene diversity in septicemic *Escherichia coli* strains. J. Clin. Microbiol. 43:66–73.

Morris RT, G Drouin. 2008. Similar ectopic gene conversion frequencies in the backbone genome of pathogenic and nonpathogenic *Escherichia coli* strains. Genomics 92:168–172.

Morton NE, Zhang W, Taillon-Miller P, Ennis S, Kwok PY, Collins A . 2001. The optimal measure of allelic association. Proc Natl Acad Sci USA 98:5217–5221.

Moulin-Schouleur M, M Reperant, S Laurent, A Bree, S Mignon-Grasteau, P Germon, D Rasschaert, C Schouler. 2007. Extraintestinal pathogenic *Escherichia coli* strains of avian and human origin: link between phylogenetic relationships and common virulence patterns. J. Clin. Microbiol. 45:3366–3376.

Mount DW. 2004. Bioinformatics: sequence and genome analysis. 2nd Ed. Cold Spring Harbor: Laboratory Press. Pp 676.

Mueller JC, C Andreoli. 2004. Plotting haplotype-specific linkage disequilibrium patterns by extended haplotype homozygosity. Bioinformatics. 20:786-787.

Mueller JC. 2004. Linkage disequilibrium for different scales and applications. Brief Bioinform. 5(4):355-64.

Mushegian AR, EV Koonin. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci U S A. 93(19):10268-73.

Nadeau NJ, CD Jiggins. 2010. A golden age for evolutionary genetics? Genomic studies of adaptation in natural populations. Trends Genet. Sept 17.

Nallapareddy SR, Duh RW, Singh KV, Murray BE. 2002. Molecular typing of selected *Enterococcus faecalis* isolates: pilot study using multilocus sequence typing and pulsed-field gel electrophoresis. J Clin Microbiol. 40(3): 868–76.

Nandi T, Ong C, Singh AP, Boddey J, Atkins T, Sarkar-Tyson M, Essex-Lopresti AE, Chua HH, Pearson T, Kreisberg JF, Nilsson C, Ariyaratne P, Ronning C, Losada L, Ruan Y, Sung WK, Woods D, Titball RW, Beacham I, Peak I, Keim P, Nierman WC, Tan P. 2010. A genomic survey of positive selection in *Burkholderia pseudomallei* provides insights into the evolution of accidental virulence. PLoS Pathog. 6(4):e1000845.

Nataro JP, JB Kaper. 1998. Diarrheagenic *Escherichia coli*. Clin Microbiol Rev 11:142-201.

Needleman SB, CD Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–53

Nei M y S Kumar. 2000. Molecular evolution and phylogenetics. Oxford University Press, New York.

Nei M, JC Miller. 1990. A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. Genetics. 125(4):873-9.

Nei M, WH Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. PNAS 76:5269-5273.

Nei M. 1987. Molecular Evolutionary Genetics. Columbia Univ. Press, New York.

Nothnagel M, R Furst, K Rohde. 2002. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. Human Heredity. 54:186-198.

Nübel U, Roumagnac P, Feldkamp M, Song JH, Ko KS, et al. 2008. Frequent emergence and limited geographic dispersal of methicillin-resistant Staphylococcus aureus. Proc Natl Acad Sci USA. 105:14130–14135.

Ochman H, JG Lawrence, EA Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature. 405(6784):299-304.

Ochman H, NA Moran. 2001. Genes lost and genes found: the molecular evolution of bacterial pathogenesis and symbiosis. Science 292: 1096-1098.

Ohta T. 1976 Role of slightly deleterious mutations in molecular evolution and polymorphism, Theor. Popul. Biol. 10:254–275.

Orsi RH, Q Sun, M Wiedmann. 2008. Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of Listeria monocytogenes. BMC Evol Biol. 8: 233.

Osborne SE, D Walthers, AM Tomljenovic, DT Mulder, U Silphaduang, N Duong, MJ Lowden, ME Wickham, RS Waller, LJ Kenney, BK Coombes. 2009. Pathogenic adaptation of intracellular bacteria by rewiring a cis-regulatory input function. PNAS 106(10):3982-

3987.

Pál C, B Papp, MJ Lercher. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet. 37(12):1372-5.

Palys T, LK Nakamura, FM Cohan. 1997. Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data. Int J Syst Bacteriol. 47(4):1145-56.

Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden MT, Churcher CM, Bentley SD, Mungall KL, Cerdeno-Tarraga AM, Temple L, James K, Harris B, Quail MA, Achtman M, Atkin R, Baker S, Basham D, Bason N, Cherevach I, Chillingworth T, Collins M, Cronin A, Davis P, Doggett J, Feltwell T, Goble A, Hamlin N, Hauser H, Holroyd S, Jagels K, Leather S, Moule S, Norberczak H, ONeil S, Ormond D, Price C, Rabbinowitsch E, Rutter S, Sanders M, Saunders D, Seeger K, Sharp S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Unwin L, Whitehead S, Barrell BG, Maskell DJ (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat Genet 35:32–40

Pérez-Losada M, EB Browne, A Madsen, T Wirth, RP Viscidi, KA Crandall. 2006. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect Genet Evol. 6(2):97-112. Epub 2005 Mar 24.

Perna NT, G 3rd Plunkett, V Burland, B Mau, JD Glasner, DJ Rose, GF Mayhew, PS Evans, J Gregor, HA Kirkpatrick, G Pósfai, J Hackett, S Klink, A Boutin, Y Shao, L Miller, EJ Grotbeck, NW Davis, A Lim, ET Dimalanta, KD Potamousis, J Apodaca, TS Anantharaman, J Lin, G Yen, DC Schwartz, RA Welch, FR Blattner. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature. 409(6819):529-33. Erratum en: Nature 2001 Mar 8;410(6825):240.

Perna NT, GF Mayhew, G Pósfai, S Elliott, MS Donnenberg, JB Kaper, FR Blattner. 1998. Molecular Evolution of a Pathogenicity Island from Enterohemorrhagic *Escherichia coli* O157:H7. Infect Immun. 1998 August; 66(8): 3810–3817.

Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R. 2007. Genes under positive selection in *Escherichia coli*. Genome Res 17: 1336-1343.

Petrosino JF, Xiang Q, Karpathy SE, Jiang H, Yerrapragada S, Liu Y, Gioia J, Hemphill L, Gonzalez A, Raghavan TM, Uzman A, Fox GE, Highlander S, Reichard M, Morton RJ, Clinkenbeard KD, Weinstock GM. 2006. Chromosome rearrangement and diversification of Francisella tularensis revealed by the type B (OSU18) genome sequence. J Bacteriol. 188(19):6977-85.

Piñero D, Martinez E, Selander RK. 1988. Genetic diversity and relationships among isolates of Rhizobium leguminosarum biovar phaseoli. Appl Environ Microbiol. 54(11):2825-32.

Pupo GM, Lan R, Reeves PR. 2000. Multiple independent origins of Shigella clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proc Natl Acad Sci U S A. 97(19):10567-72.

Qi W, M Käser, K Röltgen, D Yeboah-Manu, G Pluschke. 2009. Genomic Diversity and Evolution of *Mycobacterium ulcerans* Revealed by Next-Generation Sequencing. PLoS Pathog. 5(9): e1000580

Ramos-Onsins SE, Rozas J. 2002. Statistical properties of new neutrality tests against population growth. Mol Biol Evol. 19(12):2092-100. Erratum in: Mol Biol Evol. 2006 Aug;23(8):1642.

Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J. 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J Bacteriol. 190(20):6881-93.

Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS. 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. Nature. 406(6791):64-7.

Resch AM, Carmel L, Marino-Ramirez L, Ogurtsov AY, Shabalina SA, et al. 2007. Widespread positive selection in synonymous sites of mammalian genes. Mol Biol Evol 24: 1821–18.

Rice P, I Longden, A Bleasby. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16(6):276-277.

Rodríguez-Valera F. 2002. Approaches to prokaryotic biodiversity: a population genetics perspective. Environ Microbiol. 4(11):628-33.

Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, Wolf YI, Yin J, Koonin EV. 2002. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. Nucleic Acids Res. 30(19):4264-71.

Roos V, GC Ulett, MA Schembri, P Klemm. 2006. The asymptomatic bacteriuria *Escherichia coli* strain 83972 outcompetes

uropathogenic *E. coli* strains in human urine. Infect Immun. 74:615–624.

Rosebury T. 1962. Microorganisms indigenous to man. McGraw-Hill, New York.

Roselius K, Stephan W, Stadler T. 2005. The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. Genetics. 171:753-763.

Roumagnac P, Weill FX, Dolecek C, Baker S, Brisse S, Chinh NT, Le TA, Acosta CJ, Farrar J, Dougan G, Achtman M. 2006. Evolutionary history of *Salmonella* typhi. Science. 314(5803):1301-4.

Routman E, RD Miller, J Philips-Conroy, DL Hartl. 1985. Antibiotic resistance and population structure in *Escherichia coli* from free-ranging African yellow baboons. Appl. Envir. Microbiol. 50:749-754.

Russo TA, JR Johnson. 2000. Proposal for a new inclusive designation for extraintestinal pathogenic isolates of *Escherichia coli*: ExPEC. J. Infect. Dis. 181:1753-1754.

Sabatti C, N Risch. 2002. Homozygosity and linkage disequilibrium. Genetics. 160:1707-1719.

Saiki RK, Scharf S, Faloona F, Mullis KB, Erlich HA, Arnheim N. 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230(4732): 1350–4.

Saitou N, M Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Sandner L, L Eguiarte, A Navarro, O Rodríguez, A Cravioto, V Souza. 2001. An analysis of the elements of the locus of enterocyte effacement in human and wild mammal isolates of *Escherichia coli*: evolution by assemblage or deconstruction? Microbiology 147: 3149-3158.

Savageau MA. 1983. *Escherichia coli* habitats, cell types, and molecular mechanisms of gene control. Am Nat 122:732–744.

Sawyer SA, Parsch J, Zhang Z, Hartl DL. 2007. Prevalence of positive selection among nearly neutral amino acid replacements in Drosophila. Proc Natl Acad Sci USA. 104:6504-6510.

Schaeffer S. 2002. Molecular population genetics of sequence length diversity in the Adh region of *Drosophila pseudoobscura*. Genet. Res. 80:163–175.

Schmid KJ, SE Ramos-Onsins, H Ringys-Beckstein, B Weisshar, T Mitchell-Olds. 2005. A multilocus sequence survey in Arabidopsis thaliana reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics, 169:1601-1615.

Schoen C, Blom J, Claus H, Schramm-Glück A, Brandt P, Müller T, Goesmann A, Joseph B, Konietzny S, Kurzai O, Schmitt C, Friedrich T, Linke B, Vogel U, Frosch M. 2008. Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in *Neisseria meningitidis*. Proc Natl Acad Sci U S A. 105(9):3473-8.

Schouler C, A Taki, I Chouikha, M Moulin-Schouleur, P Gilot. 2009. A genomic island of an extraintestinal pathogenic *Escherichia coli* Strain enables the metabolism of fructooligosaccharides, which improves intestinal colonization. J Bacteriol. 191(1):388-93.

Selander RK, BR Levin. 1980. Genetic diversity and structure in *Escherichia coli* populations. Science 210:545-547.

Selander RK, Caugant DA, Ochman H, Musser JM, Gilmour MN, Whittam TS. 1986. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl Environ Microbiol. 51(5):873–884.

Selander RK, DA Caugant, TS Whittam. 1987. Genetic structure and variation in natural populations of *Escherichia coli*. En Neidhardt FC, Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE, eds. *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology. American Society for Microbiology. Washington, D.C. EUA. pp. 1625-1648.

Shapiro BJ, LA David, J Friedman, EJ Alm. 2009. Looking for Darwin's footprints in the microbial World. Trends in Microbiology 17(5):196-204.

Shapiro JA, Huang W, Zhang C, Hubisz MJ, Lu J, Turissini DA, Fang S, Wang HY, Hudson RR, Nielsen R, Chen Z, Wu CI. 2007. Adaptive genic evolution in the *Drosophila* genomes. Proc Natl Acad Sci U S A. 104(7):2271-6.

Simonsen KL, GA Churchill, CF Aquadro. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141:413-429.

Skyberg JA, TJ Johnson, JR Johnson, C Clabots, CM Logue, LK Nolan. 2006. Acquisition of avian pathogenic *Escherichia coli* plasmids by a commensal *E. coli* isolate enhances its abilities to kill chicken embryos, grow in human urine, and colonize the murine kidney. Infection and Immunity. 74:6287-6292.

Sokurenko EV, DL Hasty, DE Dykhuzien. 1999. Pathoadaptive mutations: gene loss and variation in bacterial pathogens. Trends Microbiol. 5:191–195.

Sokurenko EV, HS Courtney, J Maslow, A Siitonen, DL Hasty. 1995. Quantitative differences in adhesiveness of type 1 fimbriated *Escherichia coli* due to structural differences in fimH genes. J. Bacteriol. 177:3680–3686.

Sokurenko EV, V Chesnokova, DE Dykhuzien, I Ofek, XR Wu, KA Krogfelt, C Struve, M A Schembri, DL Hasty. 1998. Pathogenic adaptation of *Escherichia coli* by natural variation of the FimH adhesin. Proc. Natl. Acad. Sci. USA 95:8922–8926.

Souza V, Eguiarte L, Avila G, Cappello R, Gallardo C, Montoya J, Piñero D. 1994. Genetic Structure of Rhizobium etLi biovar phaseoLi Associated with Wild and Cultivated Bean Plants (Phaseolus vulgaris and Phaseolus coccineus) in Morelos, Mexico. Appl Environ Microbiol. 60(4):1260-1268.

Souza V, Nguyen TT, Hudson RR, Piñero D, Lenski RE. 1992. Hierarchical analysis of linkage disequilibrium in *Rhizobium* populations: evidence for sex? Proc Natl Acad Sci USA. 89(17):8389-93.

Souza V, Rocha M, Valera A, Eguiarte LE. 1999. Genetic structure of natural populations of *Escherichia coli* in wild hosts on different continents. Appl Environ Microbiol. 65(8):3373-85.

Spratt BG, Maiden MC. 1999. Bacterial population genetics, evolution and epidemiology. Philos Trans R Soc Lond B Biol Sci. 354(1384):701-10. Review.

Stevens MHH, M Sanchez, J Lee, SE Finkel. 2007. Diversification Rates Increase With Population Size and Resource Concentration in an Unstructured Habitat. Genetics. 177(4): 2243–2250.

Stoebel DM. 2005. Lack of Evidence for Horizontal Transfer of the lac Operon into *Escherichia coli*. Mol Biol Evol. 22(3):683-690.

Tajima F. 1983. Evolutionary relationship of ADN sequences in finite populations. Genetics 105:437-460.

Tajima F. 1989a. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Tajima F. 1989b. The effect of change in population size on DNA polymorphism. Genetics 123:597–601.

Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal *Escherichia coli*. Nat Rev Microbiol. 8(3):207-17.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, AngiuoLi SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, ScarselLi M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJ, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, RappuoLi R, Fraser CM. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 102(39):13950-5. Erratum in: Proc Natl Acad Sci U S A. 2005 Nov 8;102(45):16530.

Thomas DC, RW Haile, D Duggan. 2005. Recent Developments in Genomewide Association Scans: A Workshop Summary and Review. Am J Hum Genet. 77(3): 337–345.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22(22):4673-80

Tillier ERM, Collins RA. 2000. Genome rearrangement by replication directed translocation. Nat Genet, 26:184-186.

Toft C, Williams TA, Fares MA. 2009. Genome-Wide Functional Divergence after the Symbiosis of Proteobacteria with Insects Unraveled through a Novel Computational Approach. PLoS Comput Biol 5(4): e1000344.

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Médigue C, Rocha EP, Denamur E. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet. 5(1):e1000344.

Ureta-Vidal A, Ettwiller L, Birney E. 2003. Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet. 4(4):251-62.

Urwin R, Maiden MC. 2003. Multi-locus sequence typing: a tool for global epidemiology. Trends Microbiol. 11(10): 479–87.

Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. 2005. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics 21:2791-2793.

Voight BF, KudaravalLi S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:e72.

Wang Y, Zhao LP, Dudoit S. 2006. A fine-scale linkage-disequilibrium measure based on length of haplotype sharing. Am J Hum Genet. 78(4):615-28.

Waterman MS. 1984. General methods of sequence comparison. Bull. Math. Biol. 46:473–500.

Waterman MS. 1986. Multiple sequence alignment by consensus. Nucleic Acids Res. 14(22):9095-102.

Wattam AR, KP Williams, EE Snyder, NF Almeida, M Shukla, AW Dickerman, OR Crasta, R Kenyon, J Lu, JM Shallom, H Yoo, TA Ficht, RM Tsolis, C Munk, R Tapia, CS Han, JC Detter, D Bruce, TS Brettin, BW Sobral, SM Boyle, JC Setubal.2009. Analysis of Ten Brucella Genomes Reveals Evidence for Horizontal Gene Transfer Despite a Preferred Intracellular Lifestyle. J Bacteriol. 2009 June; 191(11): 3569–3579.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7:256-276.

Wedekind C, MO Gessner, F Vazquez, M Maerki, D Steiner. 2010. Elevated resource availability sufficient to turn opportunistic into virulent fish pathogens. Ecology 91(5):1251-6.

Weintraub A. 2007. Enteroaggregative *Escherichia coli*: epidemiology, virulence and detection. J Med Microbiol. 56(Pt 1):4-8.

Weissman SJ, Beskhlebnaya V, Chesnokova V, Chattopadhyay S, Stamm WE, Hooton TM, Sokurenko EV. 2007. Differential stability and trade-off effects of pathoadaptive mutations in the *Escherichia coli* FimH adhesin. Infect Immun. 75(7):3548-55. Erratum in: Infect Immun. 2009 Apr;77(4):1720.

Weissman SJ, Chattopadhyay S, Aprikian P, Obata-Yasuoka M, Yarova-Yarovaya Y, Stapleton A, Ba-Thein W, Dykhuizen D, Johnson JR, Sokurenko EV. 2006. Clonal analysis reveals high rate of structural mutations in fimbrial adhesins of extraintestinal pathogenic *Escherichia coli*. Mol Microbiol. 59(3):975-88.

Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci U S A. 2002 Dec 24;99(26):17020-4.

Whittam TS. 1989. Clonal dynamics of *Escherichia coli* in its natural habitat. Antonie Leeuwenhoek 55:23–32.

Wiles TJ, Kulesus RR, Mulvey MA. 2008. Origins and virulence mechanisms of uropathogenic *Escherichia coli*. Exp Mol Pathol. 85(1):11-9.

Williams JGK, AR Kubelik, KJ Livak, J A Rafalski, S V Tingey. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers Nucleic Acids Research 18: 6231-6235.

Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H, Achtman M. 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol. 60(5):1136-51.

Wooley RE, PS Gibbs, TP Brown, JR Glisson, WL Steffens, JJ Maurer. 1998. Colonisation of the chicken trachea by an avirulent avian *Escherichia coli* transformed with plasmid pHK11. Avian Diseases. 42:194-198.

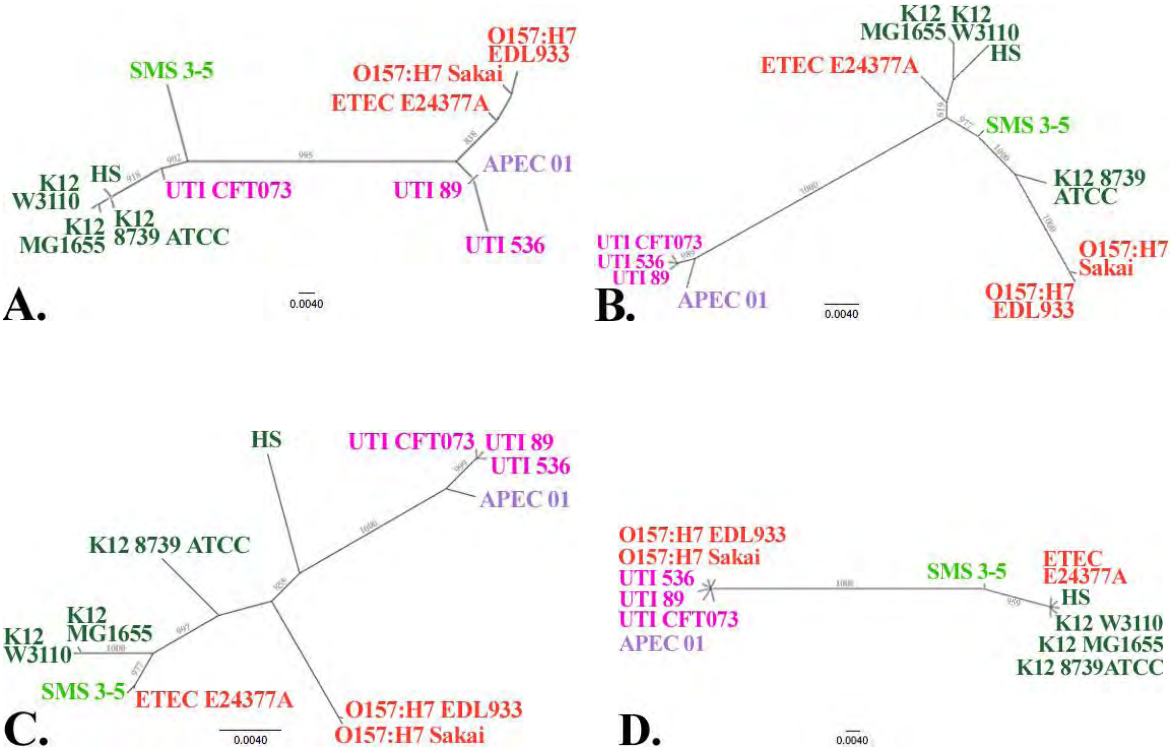
Wu XY, Chapman T, Trott DJ, Bettelheim K, Do TN, Driesen S, Walker MJ, Chin J. 2007. Comparative analysis of virulence genes, genetic diversity, and phylogeny of commensal and enterotoxigenic *Escherichia coli* isolates from weaned pigs. Appl Environ Microbiol. 73(1):83-91.

Yan F, DB Polk. 2004. Commensal bacteria in the gut: learning who our friends are. Curr Opin Gastroenterol. 20:565–571.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol. 7(1-2):203-14.

APÉNDICE 1

Reconstrucciones filogenéticas de Máxima Verosimilitud (ML), a partir de las secuencias del genoma central de **A.** el LCB 10, **B.** el LCB 15, **C.** el LCB 16 y **D.** el LCB 18, de *E. coli*, realizada con el programa PHYML 3.0 (Guindon y Gascuel, 2003). Valores de *bootstrap* obtenidos a partir de 1,000 réplicas.

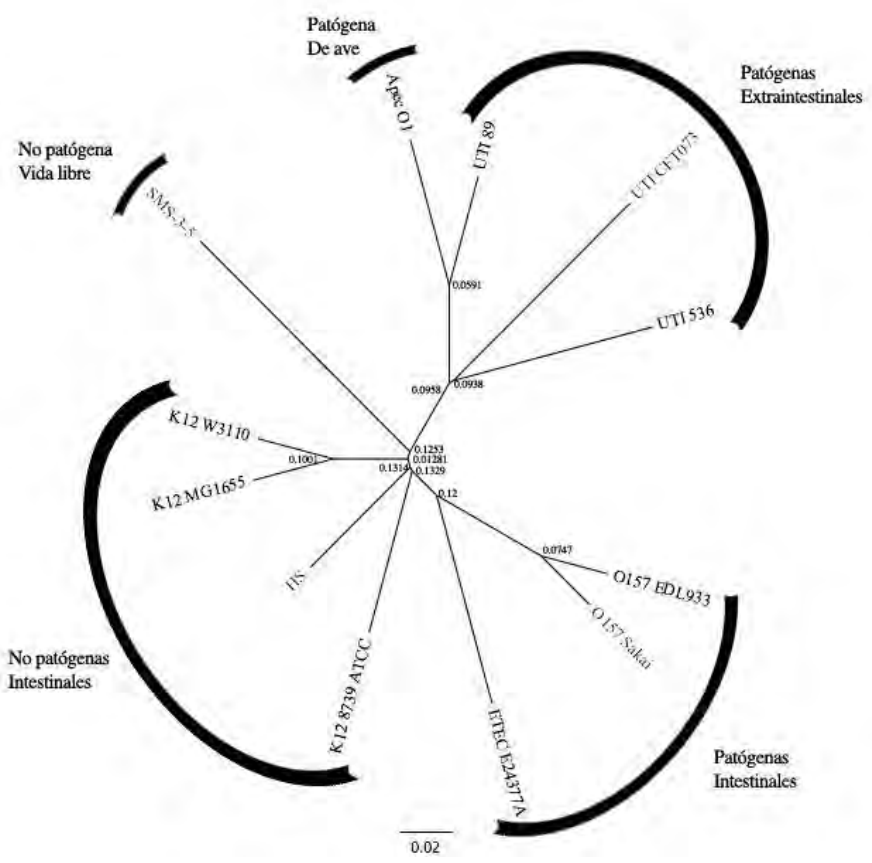


APÉNDICE 2

A. Detalles del alineamiento cromosómico de *E. coli* generado por Mauve 2.1.1.

El programa MAUVE computó el alineamiento múltiple de secuencias de DNA cromosómico de *E. coli*, en ~1 minuto con el equipo *Kan Balam*, DGSCA, UNAM, México, D.F. La revisión de los alineamientos de cada LCB identificó 448 regiones problemáticas, similares al ejemplo de la Figura 1.

B. Árbol guía que relaciona los 12 individuos de la muestra de *E. coli* generado por el programa MAUVE, con el algoritmo de *Neighbor Joining* (NJ). Se realiza a partir del alineamiento de una sub-muestra de *anchors* ó anclas, que son secuencias cromosómicas, con un alto porcentaje de similitud, las cuales son identificadas en el primer paso del algoritmo. Este árbol es utilizado en los siguientes pasos del algoritmo para la realización del alineamiento múltiple final. Los números en los nodos corresponden al índice de distancia entre las ramas.



Apéndice 2. *Continuación*.

C. Número de regiones del genoma central y del genoma flexible identificados inicialmente por el programa MAUVE, antes de la revisión del alineamiento.

	Número de regiones		Número de sitios
LCB	**Genoma central**	**Genoma flexible**	**Genoma central**
1	36	35	164626
2	1	0	5288
3	22	22	83384
4	1	0	5288
5	1	0	4393
6	4	5	35768
7	88	89	316447
8	51	53	243783
9	131	129	498108
10	2	1	223
11	58	58	405685
12	59	60	204193
13	52	51	348284
14	1	0	144
15	14	12	11301
16	1	0	4311
17	61	61	116264
18	2	1	309
19	190	190	990205
20	6	5	31708
21	98	97	324490
Total	879	869	3794202

D. Matriz de identidad nucleotídica de las regiones del genoma central delimitadas inicialmente por MAUVE, antes de la revisión del alineamiento.

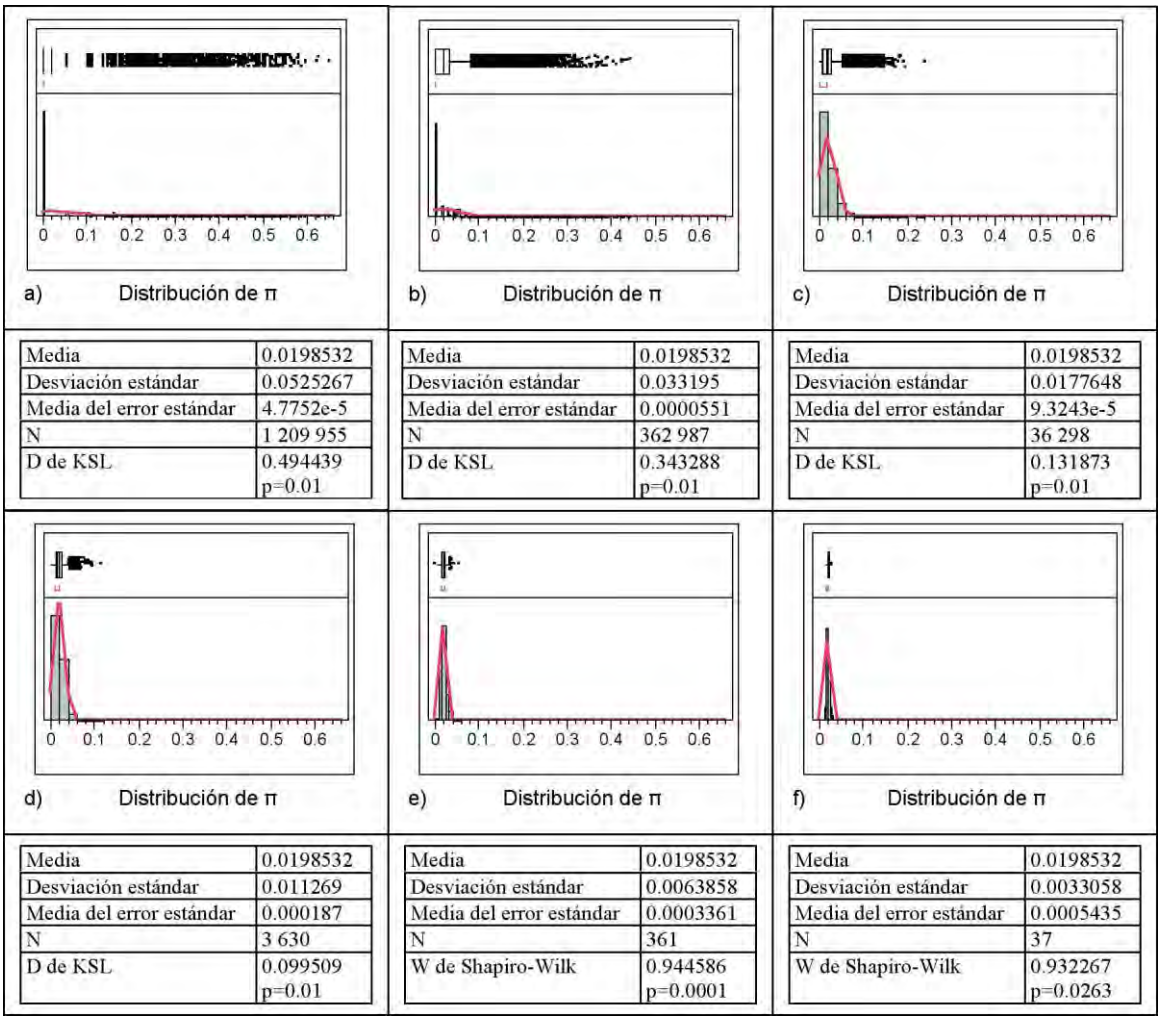
	Cepa	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	**k12 MG1655**	*1.00*	-	-	-	-	-	-	-	-	-	-	-
2	**k12 W3110**	0.916	*1.00*	-	-	-	-	-	-	-	-	-	-
3	**k12 8739ATCC**	0.825	0.82	*1.00*	-	-	-	-	-	-	-	-	-
4	**HS**	0.82	0.822	0.844	*1.00*	-	-	-	-	-	-	-	-
5	**SMS-3-5**	0.684	0.686	0.691	0.699	*1.00*	-	-	-	-	-	-	-
6	**APEC-O1**	0.644	0.645	0.647	0.647	0.64	*1.00*	-	-	-	-	-	-
7	**ETEC E24377A**	0.699	0.7	0.711	0.708	0.626	0.581	*1.00*	-	-	-	-	-
8	**O157 EDL933**	0.742	0.744	0.739	0.737	0.686	0.628	0.683	*1.00*	-	-	-	-
9	**O157 Sakai**	0.74	0.742	0.737	0.735	0.685	0.627	0.681	0.861	*1.00*	-	-	-
10	**UTI 536**	0.696	0.698	0.705	0.703	0.701	0.703	0.635	0.679	0.678	*1.00*	-	-
11	**UTI CFT073**	0.654	0.655	0.658	0.658	0.65	0.677	0.59	0.636	0.635	0.712	*1.00*	-
12	**UTI 89**	0.689	0.69	0.694	0.692	0.682	0.724	0.617	0.671	0.67	0.752	0.722	*1.00*

APÉNDICE 3.

Distribución de frecuencias de los parámetros A. diversidad nucleotídica π; B. mutación poblacional θ; C. diversidad haplotípica Hd; D. prueba de desequilibrio de ligamiento Zns; E. D de Tajima; F. D* de Fu-Li, y G. F* de Fu-Li, estimados en seis tamaños de ventana diferentes a)3, b)10, c)100, d)1000, e)10,000 y f)100,000 del genoma central del cromosoma de *E. coli*.

La línea roja representa el ajuste de distribución normal para cada grupo de datos; se realizó la prueba de KSL y Shapiro-Wilk con la cual se rechazó la hipótesis de normalidad en cada caso. En la parte superior de cada gráfica se muestra el diágrama de caja con valores extremos para cada grupo de datos; los extremos de la caja muestran los cuartiles, la línea central la mediana y el diamante ubica la media muestral con un intervalo del 95% de confianza; el corchete rojo indica el intervalo donde se distribuyen el 50% de las observaciones totales.

A. Diversidad nucleotídica π



Media	0.0198532
Desviación estándar	0.0525267
Media del error estándar	4.7752e-5
N	1 209 955
D de KSL	0.494439 p=0.01

Media	0.0198532
Desviación estándar	0.033195
Media del error estándar	0.0000551
N	362 987
D de KSL	0.343288 p=0.01

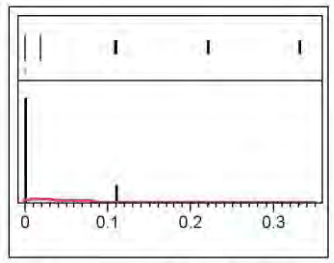
Media	0.0198532
Desviación estándar	0.0177648
Media del error estándar	9.3243e-5
N	36 298
D de KSL	0.131873 p=0.01

Media	0.0198532
Desviación estándar	0.011269
Media del error estándar	0.000187
N	3 630
D de KSL	0.099509 p=0.01

Media	0.0198532
Desviación estándar	0.0063858
Media del error estándar	0.0003361
N	361
W de Shapiro-Wilk	0.944586 p=0.0001

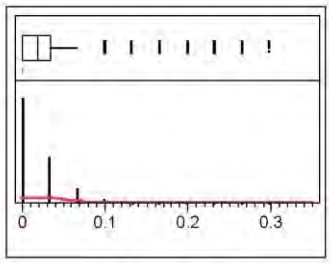
Media	0.0198532
Desviación estándar	0.0033058
Media del error estándar	0.0005435
N	37
W de Shapiro-Wilk	0.932267 p=0.0263

B. Mutación poblacional θ



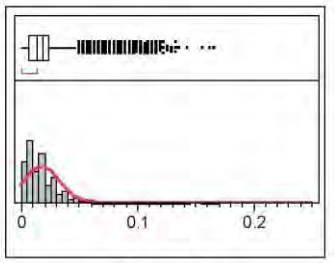
a) Distribución de θ

Media	0.0179945
Desviación estándar	0.0439253
Media del error estándar	3.9933e-5
N	1 209 955
D de KSL	0.506141 p=0.01



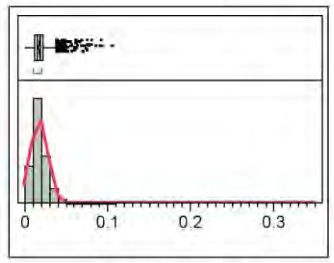
b) Distribución de θ

Media	0.0179945
Desviación estándar	0.0277344
Media del error estándar	4.6033e-5
N	362 987
D de KSL	0.359952 p=0.01



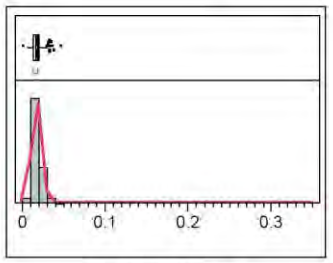
c) Distribución de θ

Media	0.0179948
Desviación estándar	0.0150969
Media del error estándar	7.924e-5
N	36 298
D de KSL	0.145712 p=0.01



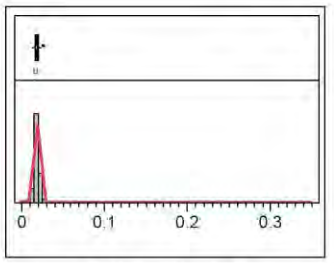
d) Distribución de θ

Media	0.0179903
Desviación estándar	0.0099901
Media del error estándar	0.0001658
N	3 630
D de KSL	0.107355 p=0.01



e) Distribución de θ

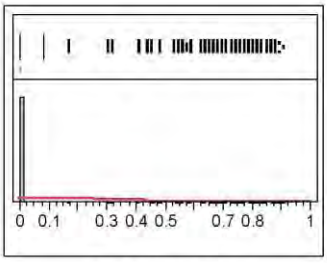
Media	0.0179635
Desviación estándar	0.0057342
Media del error estándar	0.0003018
N	361
W de Shapiro-Wilk	0.932058 p=0.0001



f) Distribución de θ

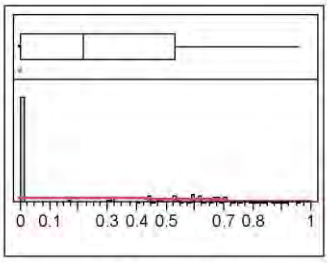
Media	0.0180872
Desviación estándar	0.0029163
Media del error estándar	0.0004794
N	37
W de Shapiro-Wilk	0.922908 p=0.0136

C. Diversidad haplotípica Hd



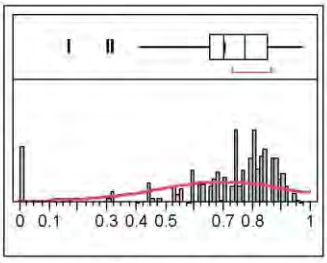
a) Distribución de Hd

Media	0.0819524
Desviación estándar	0.2019474
Media del error estándar	0.0001836
N	1 209 955
D de KSL	0.504727 p=0.01



b) Distribución de Hd

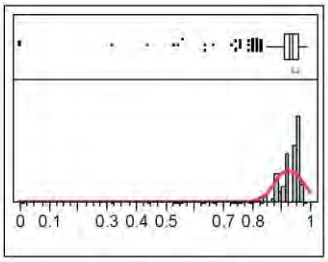
Media	0.2169754
Desviación estándar	0.2916575
Media del error estándar	0.0004841
N	362 987
D de KSL	0.389725 p=0.01



c) Distribución de Hd

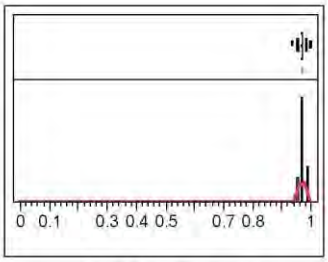
Media	0.7020735
Desviación estándar	0.2264737
Media del error estándar	0.0011887
N	36 298
D de KSL	0.200602 p=0.01

C. *Continuación*



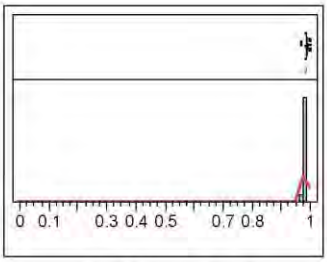
d) Distribución de Hd

Media	0.9270682
Desviación estándar	0.0447376
Media del error estándar	0.0007425
N	3 630
D de KSL	0.221650 p=0.01



e) Distribución de Hd

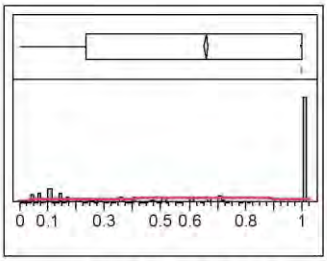
Media	0.9709142
Desviación estándar	0.0098669
Media del error estándar	0.0005193
N	361
W de Shapiro-Wilk	0.801866 p=0



f) Distribución de Hd

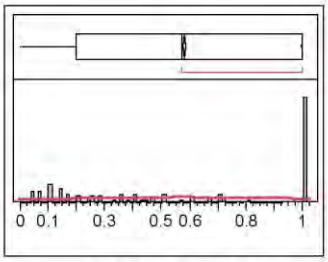
Media	0.9864865
Desviación estándar	0.0069468
Media del error estándar	0.001142
N	37
W de Shapiro-Wilk	0.623105 p=0.0001

D. Prueba de desequilibrio de ligamiento Zns



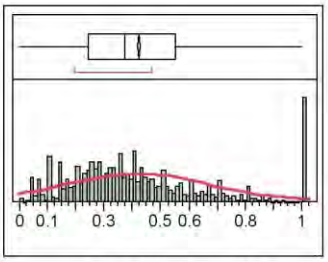
a) Distribución de Zns

Media	0.6593931
Desviación estándar	0.3909373
Media del error estándar	0.0049371
N	6270
D de KSL	0.337699 p=0.01



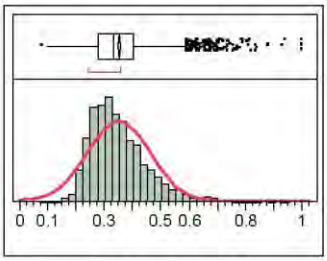
b) Distribución de Zns

Media	0.5817063
Desviación estándar	0.3786066
Media del error estándar	0.0023757
N	25 398
D de KSL	0.258011 p=0.01



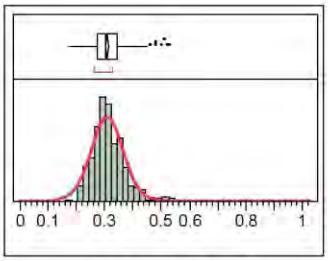
c) Distribución de Zns

Media	0.4234298
Desviación estándar	0.2531789
Media del error estándar	0.0015539
N	26 545
D de KSL	0.103112 p=0.01



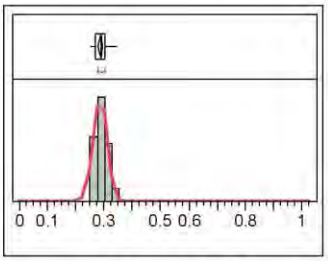
d) Distribución de Zns

Media	0.3536454
Desviación estándar	0.1083509
Media del error estándar	0.0018009
N	3 620
D de KSL	0.091062 p=0.01



e) Distribución de Zns

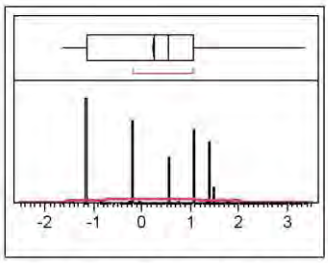
Media	0.3099458
Desviación estándar	0.0571549
Media del error estándar	0.0030082
N	361
W de Shapiro-Wilk	0.957119 p=0.0001



f) Distribución de Zns

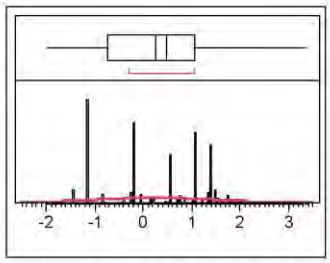
Media	0.2881486
Desviación estándar	0.0223382
Media del error estándar	0.0036724
N	37
W de Shapiro-Wilk	0.977328 p=0.0638

E. Prueba de neutralidad D de Tajima



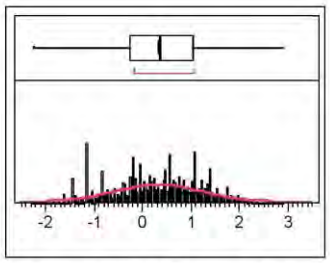
a) Distribución de D de Tajima

Media	0.2416228
Desviación estándar	1.0208652
Media del error estándar	0.002374
N	184 920
D de KSL	0.188002 p=0.01



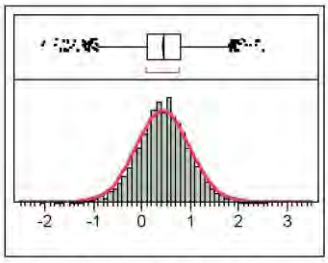
b) Distribución de D de Tajima

Media	0.2506166
Desviación estándar	1.0250316
Media del error estándar	0.0027534
N	138 595
D de KSL	0.149057 p=0.01



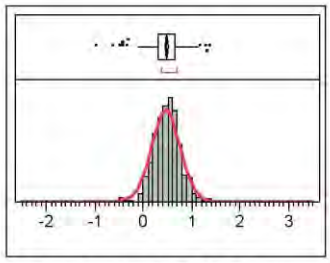
c) Distribución de D de Tajima

Media	0.3307601
Desviación estándar	0.9071487
Media del error estándar	0.0049251
N	33 925
D de KSL	0.032932 p=0.01



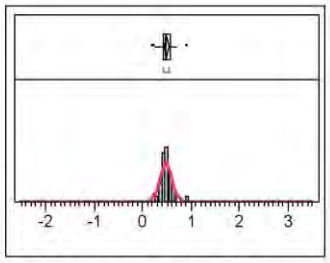
d) Distribución de D de Tajima

Media	0.4480764
Desviación estándar	0.5433914
Media del error estándar	0.0090215
N	3 628
D de KSL	0.021272 p=0.01



e) Distribución de D de Tajima

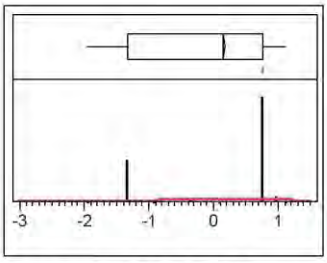
Media	0.4842147
Desviación estándar	0.2891167
Media del error estándar	0.0152167
N	361
W de Shapiro-Wilk	0.971904 p=0.0001



f) Distribución de D de Tajima

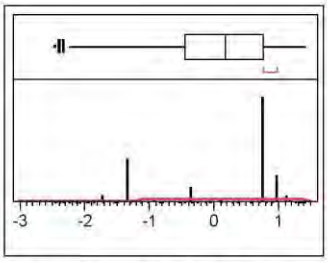
Media	0.4964335
Desviación estándar	0.1203189
Media del error estándar	0.0197803
N	37
W de Shapiro-Wilk	0.936862 p=0.0367

F. Prueba de neutralidad D* de Fu-Li



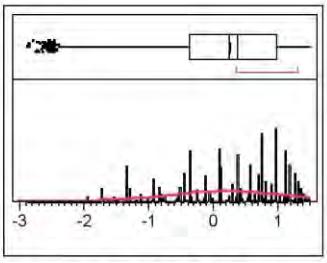
a) Distribución de D* de Fu-Li

Media	0.1604474
Desviación estándar	0.9447124
Media del error estándar	0.0021969
N	184 920
D de KSL	0.440486 p=0.01



b) Distribución de D* de Fu-Li

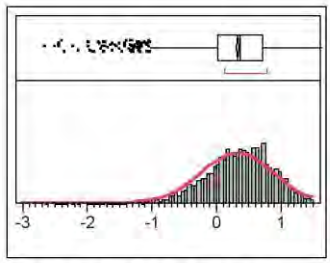
Media	0.1722623
Desviación estándar	0.9571823
Media del error estándar	0.0025711
N	138 595
D de KSL	0.387464 p=0.01



c) Distribución de D* de Fu-Li

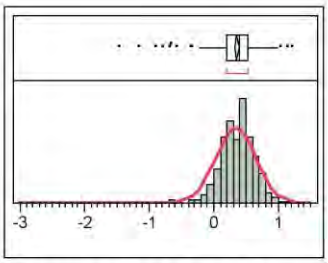
Media	0.24169
Desviación estándar	0.8786422
Media del error estándar	0.0047704
N	33 925
D de KSL	0.122448 p=0.01

F. *Continuación*



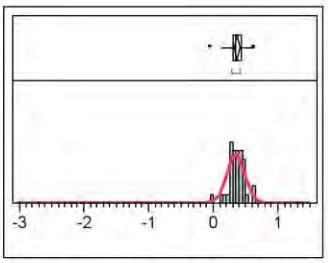
d) Distribución de D* de Fu-Li

Media	0.3306628
Desviación estándar	0.5499668
Media del error estándar	0.0091307
N	3 628
D de KSL	0.045122 p=0.01



e) Distribución de D* de Fu-Li

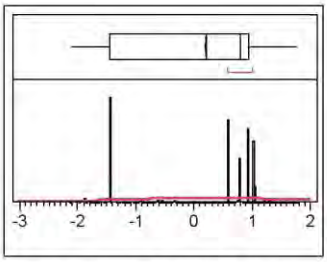
Media	0.3517829
Desviación estándar	0.3049649
Media del error estándar	0.0160508
N	361
W de Shapiro-Wilk	0.918889 p=0.0001



f) Distribución de D* de Fu-Li

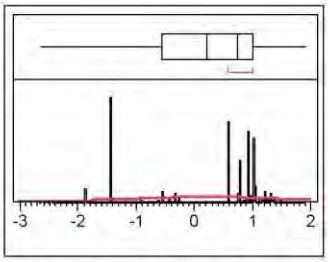
Media	0.3598978
Desviación estándar	0.1283181
Media del error estándar	0.0210954
N	37
W de Shapiro-Wilk	0.953212 p=0.1225

G. Prueba de neutralidad F* de Fu-Li



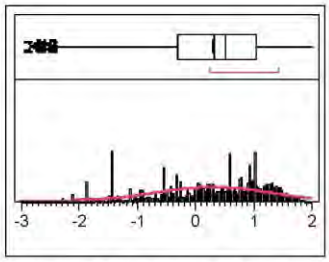
a) Distribución de F* de Fu-Li

Media	0.2030854
Desviación estándar	1.0333337
Media del error estándar	0.002403
N	184 920
D de KSL	0.349679 p=0.01



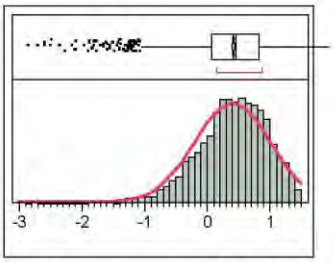
b) Distribución de F* de Fu-Li

Media	0.2159971
Desviación estándar	1.0482364
Media del error estándar	0.0028157
N	138 595
D de KSL	0.299527 p=0.01



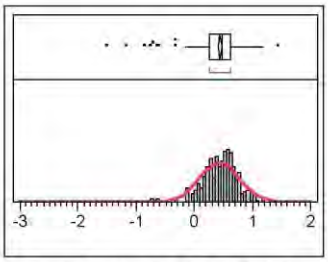
c) Distribución de F* de Fu-Li

Media	0.2997007
Desviación estándar	0.9627494
Media del error estándar	0.005227
N	33 925
D de KSL	0.092847 p=0.01



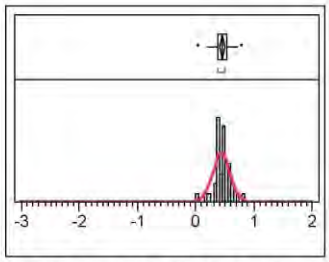
d) Distribución de F* de Fu-Li

Media	0.4121553
Desviación estándar	0.6002976
Media del error estándar	0.0099663
N	3 628
D de KSL	0.043591 p=0.01



e) Distribución de F* de Fu-Li

Media	0.4414876
Desviación estándar	0.3306733
Media del error estándar	0.0174039
N	361
W de Shapiro-Wilk	0.922795 p=0.0001



f) Distribución de F* de Fu-Li

Media	0.4520686
Desviación estándar	0.1394178
Media del error estándar	0.0229201
N	37
W de Shapiro-Wilk	0.961378 p=0.0223

APÉNDICE 4

Distribución de resultados de las pruebas de selección natural y desequilibrio de ligamiento, en los seis niveles de análisis del genoma central del cromosoma de *E. coli*.

Prueba			Nivel de análisis (tamaño de ventana) 3	10	100	1,000	10,000	100,000
D de Tajima	Número de loci	-	88,035	64,211	12,483	680	13	0
		+	96,885	74,384	21,443	2948	350	37
	Número de loci significativos*	-	129	973	433	8	0	0
		+	4924	6333	2050	41	0	0
	Proporción de loci significativos[1]	-	0.0015	0.0152	0.0347	0.0118	0	0
		+	0.0508	0.0851	0.0956	0.0139	0	0
D* de Fu-Li	Número de loci	-	54,018	44,199	11,466	864	30	1
		+	130,902	94,396	22,460	2764	333	36
	Número de loci significativos*	-	150	150	1161	541	0	0
		+	0	0	431	3271	0	0
	Proporción de loci significativos[1]	-	0.0028	0.0263	0.0472	0.0150	0	0
		+	0	0.0046	0.1456	0.0315	0	0
F* de Fu-Li	Número de loci	-	53,909	44,081	11,374	779	22	0
		+	131,011	94,512	22,493	2848	341	37
	Número de loci significativos*	-	129	129	999	501	0	0
		+	239	239	1805	2650	0	0
	Proporción de loci significativos[1]	-	0.0024	0.0227	0.0440	0.0154	0	0
		+	0.0018	0.0191	0.1178	0.0284	0	0
Zns	Número de loci (N_{DL})		6270	25,398	26,546	3620	363	37
	Número de loci significativos*		3 320	10 071	3 634	129	0	0
	Proporción de loci significativos[1]		0.5295	0.3965	0.1369	0.0356	0	0

* Significativos a $p<0.05$ en las pruebas de D de Tajima, D* y F* de Fu-Li, y a $p<0.025$ en la prueba de desequilibrio de ligamiento Zns.

[1]Corresponde al número de loci significativos sobre el número total de loci de cada categoría.

APÉNDICE 5

Varianzas de los promedios de diversidad genética, pruebas de selección y de desequilibrio de ligamiento en los componentes del pangenoma, en el total de la muestra de *E. coli* en los ecogrupos.

Ecogrupo	Parámetro	Componente del pangenoma	
		Genoma central	Genoma flexible (compartido por el ecogrupo)
No-patógenas	**π**	0.00023691	0.00165391
	θ	0.00023686	0.00176313
	D de Tajima	0.49553448	0.62526817
	D* de Fu-Li	0.4552464	0.58570353
	F* de Fu-Li	0.52177638	0.63775397
	Hd	0.0385404	0.08744742
	Zns	0.03774402	0.04048663
Patógenas	**π**	0.0003498	0.00341584
	θ	0.00024091	0.00223057
	D de Tajima	0.32346542	0.82410647
	D* de Fu-Li	0.27444596	0.66231699
	F* de Fu-Li	0.33373914	0.81071566
	Hd	0.01189487	0.09503138
	Zns	0.02121841	0.04753281
TOTAL	**π**	0.00032863	-
	θ	0.00022061	-
	D de Tajima	0.35698247	-
	D* de Fu-Li	0.35685307	-
	F* de Fu-Li	0.42624368	-
	Hd	0.00981326	-
	Zns	0.01661609	-

APÉNDICE 6

Número total de loci (N), loci polimórficos (N_P) y loci con sitios informativos para la prueba Zns (N_{DL}), del genoma central en cada bloque localmente colinear (LCB) del cromosoma de *E. coli*, en los ecogrupos y en la muestra total.

		Bloques Localmente Colineares (LCBs)																				
Ecogrupo	**Parámetro**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
Todos	**N**	181	2	96	7	6	40	362	274	559	2	368	232	373	1	12	5	126	2	1080	36	358
No-patógenas	**N_P**	178	2	95	7	5	39	358	269	551	2	363	229	370	1	12	5	124	2	1063	30	353
	N_{DL}	118	2	49	6	3	24	319	221	378	0	192	135	236	0	9	4	73	0	636	5	218
Patógenas	**N_P**	179	2	96	7	5	38	359	271	555	2	364	231	371	1	12	5	126	2	1073	33	353
	N_{DL}	173	2	95	7	4	34	354	265	545	1	355	225	367	1	12	5	125	1	1052	28	341
TOTAL	**N_P**	179	2	96	7	5	38	359	271	555	2	364	231	371	1	12	5	126	2	1073	33	353
	N_{DL}	173	2	95	7	4	34	354	265	545	1	355	225	367	1	12	5	125	2	1052	28	341

APÉNDICE 7

Varianzas de los promedios de diversidad genética, pruebas de selección y de desequilibrio de ligamiento estimados en los bloques localmente colineares (LCB) del cromosoma de *E. coli*, en los ecogrupos y en la muestra total.

		Bloques Localmente Colineares (LCBs)																			
Ecogrupo	**Parámetro**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
No-patógenas	**π**	0.00015	0.00243	0.00026	2.25E-05	6.93E-05	0.00014	0.00040	0.00048	0.00021	2.11E-05	4.34E-05	0.00018	0.00017	-	8.84E-05	6.51E-06	0.00027	7.65E-05	0.00015	0.00046	0.00040
	θ	0.00016	0.00156	0.00036	1.09E-05	7.46E-05	0.00012	0.00040	0.00049	0.00021	5.50E-05	5.18E-05	0.00025	0.00015	-	8.19E-05	8.58E-06	0.00030	0.00011	0.00014	0.00066	0.00034
	D de Tajima	0.47873	0.01256	0.40513	1.24948	0.82812	0.55258	0.37933	0.32757	0.50183	0.38171	0.58653	0.46394	0.52664	-	0.62787	0.36864	0.48725	0.00103	0.46379	0.26580	0.53283
	D* de Fu-Li	0.46081	0.00013	0.37853	1.24948	0.82812	0.48644	0.35182	0.31019	0.45343	0.38171	0.54690	0.43103	0.48415	-	0.62787	0.36864	0.43091	0.00103	0.42742	0.24380	0.47821
	F* de Fu-Li	0.52292	0.00494	0.43073	1.36303	0.80585	0.54128	0.39858	0.34971	0.52107	0.38419	0.62284	0.49640	0.55617	-	0.69227	0.41130	0.50562	0.00189	0.49184	0.28045	0.55647
	Hd	0.03441	0.02000	0.04330	1.44E-32	0.12567	0.02738	0.01652	0.02650	0.03423	0.08000	0.04100	0.03782	0.03563	-	0.00152	0.00800	0.03660	0	0.04050	0.08663	0.04299
	Zns	0.03872	0.00023	0.02963	0.03253	0.03297	0.06757	0.04048	0.05211	0.03843	-	0.02617	0.03283	0.03050	-	0.02506	0.03796	0.03267	-	0.03925	0.02334	0.02526
Patógenas	**π**	0.00031	0.00119	0.00215	4.53E-05	0.00014	0.00081	0.00054	0.00072	0.00120	0.00075	0.00023	0.00070	0.00014	-	0.00032	2.22E-05	0.00071	0.00019	0.00030	0.00095	0.00036
	θ	0.00021	0.00066	0.00123	2.72E-05	8.24E-05	0.00043	0.00036	0.00064	0.00069	0.00093	0.00016	0.00045	0.00010	-	0.00022	1.67E-05	0.00048	0.00030	0.00026	0.00050	0.00025
	D de Tajima	0.36755	0.00741	0.27912	1.52362	2.09251	0.62505	0.36084	0.43926	0.28378	2.58039	0.31312	0.29823	0.26626	-	0.14934	0.16816	0.30755	0.00145	0.33980	0.68628	0.39382
	D* de Fu-Li	0.31668	0.00324	0.20857	0.84539	1.64317	0.47700	0.29375	0.40094	0.24360	2.32566	0.27198	0.24165	0.21755	-	0.15613	0.09976	0.31453	0.00484	0.28663	0.51544	0.32169
	F* de Fu-Li	0.38482	0.00778	0.26182	1.17649	2.11906	0.59960	0.35818	0.48290	0.29388	2.85806	0.32840	0.29625	0.26394	-	0.18255	0.11107	0.36727	0.00303	0.34887	0.62736	0.39597
	Hd	0.02544	0.02834	0.00406	0.00410	0.13303	0.04098	0.02056	0.01937	0.01681	0.27287	0.02000	0.01180	0.00943	-	0.00885	0.00272	0.02134	0.02834	0.01570	0.07452	0.02642
	Zns	0.02092	0	0.02050	0.05732	0.07009	0.03100	0.02412	0.02432	0.01972	-	0.02045	0.02069	0.01772	-	0.02247	0.00151	0.03129	-	0.02177	0.04552	0.02123
TOTAL	**π**	0.00023	0.00129	0.00050	4.40E-05	0.00016	0.00013	0.00041	0.00031	0.00064	2.15E-07	0.00017	0.00036	0.00010	-	0.00028	2.12E-05	0.00040	0.00044	0.00023	0.00070	0.00025
	θ	0.00014	0.00054	0.00033	2.76E-05	8.59E-05	0.00011	0.00027	0.00030	0.00035	1.36E-05	0.00011	0.00019	0.00010	-	0.00017	6.62E-06	0.00035	0.00020	0.00016	0.00033	0.00023
	D de Tajima	0.19533	0.04480	0.23706	0.65128	0.29892	0.21842	0.19511	0.18870	0.20851	0.08426	0.15474	0.16278	0.16091	-	0.11827	0.54108	0.18254	0.00339	0.18614	0.35881	0.24247
	D* de Fu-Li	0.17874	3.53E-06	0.15617	0.36334	0.90522	0.17505	0.16680	0.18866	0.12714	0.06797	0.12951	0.15614	0.14085	-	0.08661	0.20706	0.16468	0.03751	0.14932	0.22507	0.17221
	F* de Fu-Li	0.21959	0.00563	0.21702	0.54587	0.91388	0.22557	0.20037	0.22227	0.17238	0.09149	0.15920	0.18214	0.17020	-	0.09739	0.35295	0.18063	0.01842	0.18542	0.29498	0.21343
	Hd	0.00765	0.02938	0.00313	0.00210	0.14130	0.00719	0.00834	0.01413	0.00712	0.00184	0.01295	0.00739	0.00528	-	0.00451	0.00191	0.00410	0.00287	0.00785	0.06165	0.01434
	Zns	0.01337	0.00265	0.01411	0.00728	0.03212	0.02382	0.01964	0.01555	0.01555	0.05199	0.01480	0.01578	0.01023	-	0.00823	0.00499	0.02014	6.57E-06	0.01526	0.06295	0.02154

- LCBs en los que solo se analizó un loci, y por lo tanto no se estimó un promedio.

APÉNDICE 8

Valores de p de la prueba de Wilcoxon para comparar la diversidad genética, los valores de las pruebas de selección y de desequilibrio de ligamiento, de los bloques localmente colineares (LCBs) entre los ecogrupos de *E. coli*.

	Bloques Localmente Colineares (LCBs)																			
Parámetro comparado	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
π	<0.0001*	0.8669	<0.0001*	0.6578	0.6908	0.1414	<0.0001*	<0.0001*	<0.0001*	0.1017	<0.0001*	<0.0001*	<0.0001*	-	0.0339	0.1249	<0.0001*	0.1017	<0.0001*	0.1443	<0.0001*
θ	0.003*	0.5647	0.0024	0.3999	0.6438	0.4746	0.0002*	0.0094*	<0.0001*	0.1801	<0.0001*	<0.0001*	<0.0001*	-	0.0735	0.5945	<0.0001*	0.1561	<0.0001*	0.4278	<0.0001*
D de Tajima	<0.000*1	0.6514	<0.0001*	0.4832	0.4806	<0.0001*	<0.0001*	<0.0001*	<0.0001*	0.1712	<0.0001*	<0.0001*	<0.0001*	-	0.0141*	0.0614*	<0.0001*	0.1712	<0.0001*	<0.0001*	<0.0001*
D* de Fu-Li	<0.0001*	0.1017	<0.0001*	0.0481	-	<0.0001*	<0.0001*	<0.0001*	<0.0001*	0.1801	<0.0001*	<0.0001*	<0.0001*	-	0.0495*	0.0265*	<0.0001*	0.1017	<0.0001*	<0.0001*	<0.0001*
F* de Fu-Li	<0.0001*	0.1017	<0.0001*	0.2801	-	<0.0001*	<0.0001*	<0.0001*	<0.0001*	0.1801	<0.0001*	<0.0001*	<0.0001*	-	0.131*	0.014*	<0.0001*	0.1017	<0.0001*	<0.0001*	<0.0001*
Hd	<0.0001*	0.6514	<0.0001*	0.2051	-	<0.0001*	<0.0001*	<0.0001*	<0.0001*	0.1801	<0.0001*	<0.0001*	<0.0001*	-	0.0434*	0.0185*	<0.0001*	0.1017	<0.0001*	<0.0001*	<0.0001*
Zns	<0.0001*	0.1229	<0.0001*	0.0084	-	<0.0001*	<0.0001*	<0.0001*	<0.0001*	1	<0.0001*	<0.0001*	<0.0001*	-	<0.0001*	0.008*	<0.0001*	0.2207	0*	0.0009*	<0.0001*

* p significativa.

- Casos en los que no se aplicó la prueba porque n < 2, en alguno de los ecogrupos.

APÉNDICE 9.

Distribución de resultados de las pruebas de selección natural y desequilibrio de ligamiento, en los bloques localmente colineares (LCBs) del cromosoma de *E. coli*, en los ecogrupos y en la muestra total.

				Bloques Localmente Colineares (LCBs)																		
Ecogrupo	**Prueba**			**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
No-patógenas	**D de Tajima**[1]	Número de loci	-	128	0	76	3	3	30	221	195	397	2	270	178	262	1	3	3	92	2	845	27	264
			+	45	2	17	4	2	9	135	69	153	0	90	50	106	0	9	2	31	0	209	3	84
		Número de loci significativos*	-	2	0	2	0	0	0	0	1	9	0	4	6	0	0	0	0	6	0	14	0	10
			+	0	2	0	0	0	1	2	0	1	0	1	0	1	0	0	0	0	0	4	0	5
		Proporción de loci significativos[1]	-	0.0156		0.0263	0	0	0	0	0.0051	0.0227	0	0.0148	0.0338	0	0	0	0	0.0652	0	0.0166	0	0.0379
			+	0	1	0	0	0	0.1111	0.0148	0	0.0065	0	0.0111	0	0.0094	0	0	0	0	0	0.0191	0	0.0595
	D* de Fu-Li	Número de loci	-	131	0	79	3	3	32	241	209	412	2	276	180	277	1	3	3	98	2	861	27	272
			+	42	2	14	4	2	7	114	56	137	0	83	48	90	0	9	2	24	0	194	3	76
		Número de loci significativos*	-	24	0	28	0	0	1	9	13	84	0	91	49	69	0	0	0	29	0	227	12	75
			+	0	2	0	1	0	1	2	0	2	0	0	0	2	0	0	0	0	0	5	0	4
		Proporción de loci significativos[1]	-	0.1832		0.3544	0	0	0.0313	0.0373	0.0622	0.2039	0	0.3297	0.2722	0.2491	0	0	0	0.2959	0	0.2636	0.4444	0.2757
			+	0	1	0	0.25	0	0.1429	0.01754	0	0.01460	0	0	0	0.0222	0	0	0	0	0	0.0258	0	0.0526
	F* de Fu-Li	Número de loci	-	131	0	79	3	3	32	240	208	411	2	276	179	270	1	3	3	96	2	856	27	270
			+	42	2	14	4	2	7	115	57	138	0	84	49	97	0	9	2	27	0	199	3	78
		Número de loci significativos*	-	14	0	17	0	0	0	4	10	49	0	56	34	35	0	0	0	20	0	141	7	55
			+	0	2	0	0	0	1	2	0	1	0	0	0	1	0	0	0	0	0	4	0	5
		Proporción de loci significativos[1]	-	0.1069		0.2152	0	0	0	0.0167	0.0481	0.1192	0	0.2029	0.1899	0.1296	0	0	0	0.2083	0	0.1647	0.2593	0.2037
			+	0	1	0	0	0	0.1429	0.01739	0	0.0072	0	0	0	0.0103	0	0	0	0	0	0.0201	0	0.0641
	Zns	Número de loci (N_{DL})		118	2	49	6	3	24	319	221	378	0	192	135	236	0	9	4	73	0	636	5	218
		Número de loci significativos*		35	2	17	0	1	8	61	75	120	0	34	24	50	0	3	2	20	0	198	0	36

		Proporción de loci significativos[1]		0.73671	0.98932	0.77759	0.56404	0.79409	0.69941	0.66409	0.72716	0.75046	0	0.72913	0.70863	0.72449	0	0.81474	0.85762	0.72304	0	0.7506	0.44676	0.72556
Patógenas	D de Tajima	Número de loci	-	20	0	3	3	3	5	30	28	33	1	26	15	21	0	0	0	15	2	102	5	22
			+	158	2	93	4	2	33	329	243	522	1	337	216	349	1	12	5	110	0	968	28	331
		Número de loci significativos*	-	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	2	1	0	2
			+	4	2	4	2	0	2	15	7	16	0	17	5	10	0	0	0	3	0	37	3	20
		Proporción de loci significativos[1]	-	0	0	0	0	0	0	0	0	0	1	0.0385	0.0667	0	0	0	0	0	1	0.0098	0	0.0909
			+	0.0253	1	0.0430	0.5	0	0.0606	0.0456	0.0288	0.0307	0	0.0504	0.0231	0.0287	0	0	0	0.0272	0	0.0382	0.1071	0.0604
	D* de Fu-Li	Número de loci	-	23	0	6	4	3	8	45	41	55	1	45	24	43	0	1	0	26	2	161	6	29
			+	156	2	90	3	2	29	314	230	500	1	319	207	326	1	11	5	100	0	909	27	322
		Número de loci significativos*	-	0	0	0	0	0	0	1	2	0	1	1	1	0	0	0	0	0	1	1	0	1
			+	5	2	1	0	0	3	17	7	21	0	13	5	11	0	0	0	1	0	41	3	25
		Proporción de loci significativos[1]	-	0	0	0	0	0	0	0.0222	0.0488	0	1	0.0222	0.0417	0	0	0	0	0	0.5	0.0062	0	0.0345
			+	0.0321	1	0.0111	0	0	0.1034	0.0541	0.0304	0.042	0	0.0408	0.0242	0.0337	0	0	0	0.01		0.0451	0.1111	0.0776
	F* de Fu-Li	Número de loci	-	22	0	4	3	3	7	39	36	44	1	37	16	35	0	0	0	24	2	141	6	26
			+	156	2	92	4	2	30	318	234	509	1	324	212	334	1	11	5	101	0	926	27	326
		Número de loci significativos*	-	0	0	0	0	0	0	0	2	0	1	1	1	0	0	0	0	0	1	1	0	1
			+	6	2	5	1	0	3	19	8	25	0	19	4	12	0	0	0	3	0	46	4	27
		Proporción de loci significativos[1]	-	0	0	0	0	0	0	0	0.0556	0	1	0.0270	0.0625	0	0	0	0	0	0.5	0.0071	0	0.0385
			+	0.0385	1	0.0543	0.25	0	0.1	0.0597	0.0342	0.0491	0	0.0586	0.0189	0.0359	0	0	0	0.0297		0.0497	0.1481	0.0828
	Zns	Número de loci (N_{DL})		118	2	49	6	3	24	319	221	378	0	192	135	236	0	9	4	73	0	636	5	218
		Número de loci significativos*		35	2	17	0	1	8	61	75	120	0	34	24	50	0	3	2	20	0	198	0	36
		Proporción de loci significativos[1]		0.2966	1	0.3469	0	0.3333	0.3333	0.1912	0.3394	0.3175		0.1771	0.1778	0.2119	0	0.3333	0.5	0.2740		0.3113	0	0.1651
TOTAL	D de Tajima	Número de loci	-	42	0	10	3	2	10	70	68	98	0	87	35	89	1	0	0	28	0	210	13	55
			+	137	2	86	4	3	28	289	203	457	2	277	196	282	0	12	5	98	2	863	20	298

	Número de loci significativos*	-	0	0	1	0	2	0	2	1	0	0	1	1	2	0	0	0	0	0	3	0	0
		+	3	2	3	2	0	0	10	4	11	0	2	3	2	0	0	1	3	2	18	3	2
	Proporción de loci significativos[1]	-	0	0	0.1	0	1	0	0.0286	0.0147	0	0	0.0115	0.0286	0.0225	0	0	0	0	0	0.0143	0	0
		+	0.0219	1	0.0349	0.5	0	0	0.0346	0.0197	0.0241	0	0.0072	0.0153	0.0071	0	0	0.2	0.0306	1	0.0209	0.15	0.0067
D* de Fu-Li	Número de loci	-	51	0	20	4	2	12	80	79	120	0	99	55	98	1	0	0	27	0	278	13	68
		+	128	2	76	3	3	26	279	192	435	2	265	176	273	0	12	5	99	2	795	20	285
	Número de loci significativos*	-	1	0	1	0	2	0	2	3	0	0	0	1	2	0	0	0	0	0	4	0	0
		+	6	2	4	1	0	2	13	9	20	0	6	8	7	0	0	0	7	1	34	3	15
	Proporción de loci significativos[1]	-	0.0196	0	0.05	0	1	0	0.025	0.0380	0	0	0	0.0182	0.0204	0	0	0	0	0	0.0144	0	0
		+	0.0469	1	0.0526	0.3333	0	0.0769	0.0466	0.0469	0.0460	0	0.0226	0.0455	0.0256	0	0	0	0.0707	0.5	0.0428	0.15	0.0526
F* de Fu-Li	Número de loci	-	46	0	17	4	2	11	69	77	111	0	94	45	91	1	0	0	27	0	250	14	65
		+	133	2	79	3	3	27	290	194	444	2	270	186	280	0	12	5	99	2	823	19	288
	Número de loci significativos*	-	0	0	1	0	2	0	1	3	0	0	0	1	2	0	0	0	0	0	3	0	1
		+	6	2	5	2	0	2	6	8	20	0	5	4	6	0	0	2	4	2	33	3	13
	Proporción de loci significativos[1]	-	0	0	0.0588	0	1	0	0.0145	0.0390	0	0	0.0000	0.0222	0.0220	0	0	0	0	0	0.0120	0	0.0154
		+	0.0451	1	0.0633	0.6667	0	0.0741	0.0207	0.0412	0.0450	0	0.0185	0.0215	0.0214	0	0	0.4	0.0404	1	0.0401	0.1579	0.0451
Zns	Número de loci (N_{DL})		173	2	95	7	4	34	354	265	545	1	355	225	367	1	12	5	125	2	1052	28	341
	Número de loci significativos*		5	2	4	0	0	3	21	14	21	0	16	11	10	0	0	0	9	2	51	6	27
	Proporción de loci significativos[1]		0.0289	1	0.0421	0	0	0.0882	0.0593	0.0528	0.0385	0	0.0451	0.0489	0.0272	0	0	0	0.072	1	0.0485	0.2143	0.0792

* Significativos a $p<0.05$ en las pruebas de D de Tajima, D* y F* de Fu-Li, y a $p<0.025$ en la prueba de desequilibrio de ligamiento Zns.

[1]Corresponde al número de loci significativos sobre el número total de loci de cada categoría.