UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

DOCTORADO EN CIENCIAS BIOMÉDICAS

INSTITUTO DE ECOLOGÍA

Papel de la transferencia horizontal, recombinación y selección natural en la dinámica evolutiva del genoma de bacterias entéricas silvestres y patógenas

TESIS

QUE PARA OBTENER EL GRADO DE:

DOCTORA EN CIENCIAS

P R E S E N T A:

ANDREA GONZÁLEZ GONZÁLEZ

DIRECTORA DE TESIS:

DRA. VALERIA SOUZA SALDÍVAR, Instituto de Ecología

MIEMBROS DEL COMITÉ TUTOR:

DR. LUIS ENRIQUE EGUIARTE FRUNS, Instituto de Ecología

DR. JOSÉ LUIS PUENTE GARCÍA, Instituto de Biotecnología

México, D.F. Abril 2013

Agradecimientos y reconocimientos académicos

Esta tesis de doctorado se realizó bajo la dirección de la Dra. Valeria Souza Saldívar en el departamento de Ecología Evolutiva del Instituto de Ecología en la Universidad Nacional Autónoma de México.

El comité Tutor que siguió y asesoró el desarrollo de este proyecto de doctorado estuvo conformado por:

Dra. Valeria Souza Saldívar, Instituto de Ecología, UNAM.
Dr. Luis Enrique Eguiarte Fruns, Instituto de Ecología, UNAM
Dr. José Luis Puente García, Instituto de Biotecnología, UNAM

Se agradece y reconoce la disposición y el préstamo de las instalaciones del Laboratorio de Genómica Bacteriana del Departamento de Microbiología y Parasitología de la Facultad de Medicina, en la Universidad Nacional Autónoma de México por parte de la Dra. Rosario Morales para realizar los experimentos mediante los cuales se obtuvieron los datos de esta tesis. Asimismo, se agradece y reconoce la asesoría teórica y técnica de la Maestra en Ciencias Gabriela Delgado, así como la asesoría técnica del Biólogo José Luis Méndez, y el apoyo de Laura Molina González por su desempeño como laboratorista, del Laboratorio de Genómica Bacteriana del Departamento de Microbiología y Parasitología de la Facultad de Medicina, en la Universidad Nacional Autónoma de México durante la estandarización y realización de las diversas técnicas realizadas en este proyecto. Se reconoce el apoyo experimental de la Maestra en Ciencias, Luisa Sandner Miranda también del Laboratorio de Genómica Bacteriana del Departamento de Microbiología y Parasitología de la Facultad de Medicina, en la Universidad Nacional Autónoma de México. Al mismo tiempo, se agradece y reconoce la serotipificación de algunas de las cepas analizadas en este trabajo por parte del Maestro en Ciencias Armando Navarro del Departamento de Salud Pública de la Facultad de Medicina, en la Universidad Nacional Autónoma de México.

Del Instituto de Ecología, se reconoce y agradece la asesoría técnica de la M. En IBB. Laura Espinosa Asuar y la Dra. Erika Aguirre Planter durante el desarrollo de esta tesis. Asimismo se reconoce la asistencia de Silvia Barrientos por su trabajo de laboratorista.

De manera muy especial, se agradece y reconoce el constante y esmerado apoyo tanto teórico como experimental ambos brindados por Luna Sánchez Reyes a lo largo del desarrollo de ésta tesis.

Se agradece y reconoce el apoyo brindado por todos los integrantes del Laboratorio de Evolución Molecular y Experimental del Instituto de Ecología, UNAM. En particular a Jaime Gasca y Enrique Scheinvar por el apoyo en cómputo. A Eria Rebollar, Morena Avitia, Santiago Ramírez, Jorge Valdivia por las discusiones de teóricas suscitadas a lo largo del desarrollo de esta tesis. Asimismo, se reconoce y agradece la asesoría teórica de René Cerritos, Pablo Vinuesa, Olivier Tenaillon y Xavier Didelot.

Este proyecto fue apoyado por el proyecto "Evaluación de marcadores genéticos para un microarreglo diagnóstico de enfermedades diarréicas en el Pacífico Mexicano utilizando metagenómica", DGPA-UNAM PAPIIT (Programa de Apoyo a Proyectos de Investigación e Innovación tecnológica) IN219109. Durante mis estudios de doctorado gocé de una beca otorgada por CONACYT para la realización de esta tesis. Finalmente, agradezco al Programa de Doctorado en Ciencias Biomédicas de la UNAM por el apoyo y oportunidad brindada para la realización de esta tesis.

0. Resumen

Estudiar los mecanismos que diversifican y estructuran a las especies bacterianas es un tema fascinante tanto por las implicaciones que tiene en el entendimiento de la evolución de la vida en la Tierra como por las implicaciones de tipo médico y biotecnológico que tiene para el humano principalmente. Dado que *E. coli* es una especie que habita diferentes nichos ecológicos tanto asociados a un hospedero como de vida libre ya sea como patógeno o como comensal, esta especie bacteriana resulta ser un modelo apropiado para estudiar los mecanismos genéticos y ecológicos que promueven la divergencia poblacional en bacterias. Así, analizamos una muestra compuesta por 128 aislados de *E. coli* provenientes de un amplio rango de hospederos tanto patógenos como comensales. Aplicando análisis de genética de poblaciones clásica, estimación del tamaño del genoma y análisis de genómica de poblaciones encontramos que *E. coli* mantiene una estructura filogenética clara a pesar de los niveles de recombinación encontrados, debido a que los mecanismos de diversificación (mutación puntual y recombinación homóloga) se dan de manera diferencial de acuerdo a los diferentes niveles de organización en los que se agrupa la diversidad genética de esta especie. Proponemos que para *E. coli*, la unidad mínima de evolución son los ecotipos, los cuales al tener una asociación con un nicho ecológico definido, promueven diferenciación simpátrica en esta especie. Asimismo encontramos una gran variación en el tamaño del cromosoma de la muestra lo que sugiere un gran dinamismo en términos del componente del genoma que promueve la adaptación a nuevos nichos y estilos de vida. Finalmente, esta tesis propone que la evolución de un estilo de vida, en particular la patogénesis, no consiste solamente en adquirir por transferencia horizontal, los genes apropiados para explotar nuevos nichos, sino que la evolución del genoma central, así como la regulación de la expresión génica juegan un papel igualmente importante.

0. Abstract

Exploring the mechanisms driving diversification and structure in bacterial species is an outstanding topic because its health and biotechnological implications for human being. *Escherichia coli* is considered a suitable model to study the genetic and ecological mechanisms underlying the divergence among populations because inhabit a wide range of ecological niches and life styles. To study the above-mentioned topics, in this thesis a non-outbreak related host-wide *E. coli* sample was analyzed. Classical population genetics analysis, estimation of genome size and population genomics analysis suggests a clear phylogenetic structure for this enteric species in spite of the homologous recombination levels recovered. In this thesis, an alternative explanation is proposed to this clonal paradigm consisting in the presence of differential genetic diversification mechanisms (homologous recombination and point mutation) associated to the different clustering levels of the genetic diversity harboured for this species. Thus, ecotypes are proposed as the minimum evolutionary unit because its ecological coherence promoting sympatric differentiation. Furthermore, a wide range variation in chromosome size was found suggesting adaptation to new ecological niches and life styles. Finally, this thesis proposes the evolution of a new life style, in particular pathogenesis, consisting in the diversification of core genome and genetic expression besides the acquisition of virulence genes by horizontal gene transfer.

Indice final

Mecanismos de diversificación y diferenciación genética en *Escherichia coli*: historia filogenética y ecología.

1. Introducción

1.1. Ecología bacteriana

La extraordinaria diversidad en las formas de vida actual y pasada es resultado de la evolución, entendida ésta como descendencia con modificación a partir de un ancestro común (Futuyma, 2005). Una de estas formas de vida son las bacterias, las cuales han desempañado un papel fundamental en la evolución climática, geológica, geoquímica y biológica de la Tierra (Xu, 2006). Una de las razones de tal importancia es su gran abundancia. Por ejemplo, se ha estimado el número de bacterias existentes sobre la Tierra en alrededor de 5×10^{33} (Whitman et al., 1998; Balloux, 2010), un trillón (10^{12}) de veces más el número de estrellas en el universo o 10 trillones veces más el número de granos de arena presentes en nuestro planeta (Balloux 2010).

Otra característica sobresaliente de las bacterias es que las podemos encontrar en cualquier nicho ecológico imaginable ya sea asociadas a un hospedero o en forma de vida libre habitando mares, desiertos, hielo y la atmósfera (Hughes et al., 2006; Hanson et al., 2012). Se sabe que al encontrarse asociadas a otras forma de vida, establecen relaciones mutualistas (*Buchnera aphidicola*) o antagónicas con sus hospederos (cualquier especies patógenas causante de alguna enfermedad, por ejemplo *Yersinia pestis*) (Moran y Wernegreen, 2000).

Actualmente es totalmente aceptado que las bacterias de vida libre al igual que los macroorganismos, presentan una amplia gama de distribución o patrones biogeográficos. Así, existen bacterias que se distribuyen ampliamente (cosmopolitas) como *Pseudomonas* o *Nitrosococcus* (Cho y Tiedje, 2000; Ward y O'Mullan, 2002) o que son endémicas y altamente restringidas como *Sulfolobus* y *Synechococcus* (Whitaker et al., 2003; Papke et al., 2003). Qué tanto se distribuyan dependerá tanto de su contingencia historia como de factores ambientales tales como pH, temperatura, nutrientes orgánicos e inorgánicos y barreras geográficas (Hughes et al., 2006)

Además de su gran abundancia y variada distribución, la importancia de las bacterias radica en que llevan a cabo una gran diversidad de rutas metabólicas, gracias a lo cual juegan un papel fundamental en los ciclos biogeoquímicos que mantienen a

todos los ecosistemas del planeta (Pace, 1997; Whitman et al., 1998; Falkowski et al., 2008). Es decir, las reacciones químicas que originan a los elementos básicos de todas las macromoléculas biológicas (H, C, N, O, S y P) se encuentran catalizadas por bacterias.

Asimismo, se han descrito al momento 21 phyla bacterianos (Wu et al., 2009), más de los presentes en plantas. Algunos de estos phyla y los diferentes taxa que los conforman, muestran una coherencia ecológica es decir, que los miembros que conforman a un taxón en particular comparten estrategias de vida generales ó rasgos que los distinguen de los miembros de otros taxa (Fierer et al., 2007; Philippot et al., 2010). Por ejemplo, el phylum α-proteobacteria es uno de los más diversos ecológicamente (Ettema y Andersson, 2009) y se ha detectado diferenciación de nicho entre el orden de las Rickettsiales cuyos miembros se encuentran principalmente en sistemas acuáticos y el orden de las Rhizobiales y Burkholderiales los que son en su mayoría terrestres. Asimismo, la consistencia ecológica puede encontrarse también a nivel de género (Philippot et al., 2010). Y más aún, miembros del género *Prochlorococcus* (bacterias de vida libre) muestran una clara diferenciación genética entre los que habitan la superficie marina y los que habitan aguas más profundas (Rocap et al., 2003; Johnson et al., 2006).

1.2. Evolución bacteriana

Y ahora bien, cómo explicar la existencia de tantos phyla bacterianos?

Una de las consecuencias del proceso evolutivo es la gran diversidad de especies y de formas de vida existentes. Y el proceso evolutivo no es más que cambio en las propiedades de los grupos de organismos a través de las generaciones a partir de un ancestro común. Y para que llegue a darse tal cambio, es necesaria la existencia de variación genética al interior de los grupos de organismos, ya que sin variación genética no hay evolución, no hay diversificación de los linajes ni respuesta a las diferentes condiciones ambientales (Eguiarte, 1999; Futuyma, 2005).

En la siguiente sección se describirán los mecanismos por los cuales se genera la variación genética en las especies bacterianas.

1.2.1. Diversidad genética de las especies bacterianas

A la fecha se sabe que la información genética de cualquier especie bacteriana se encuentra codificada en su pangenoma (Tettelin et al., 2005; Medini et al., 2005). Al comparar diferentes aislados de una misma especie se ha identificado al genoma central consistente de los genes comunes a todas las cepas, los cuales codifican las funciones celulares esenciales. Por otro lado se encuentra el genoma flexible conformado por genes dispensables o específicos a cada aislado (Figura 1). Dichos genes contribuyen a la diversidad de la especie y codifican funciones no esenciales para el crecimiento bacteriano los cuales pueden reflejar la evolución a un nicho específico aumentando la adecuación y adaptación a diferentes ambientes como por ejemplo, los genes de virulencia los cuales permiten explotar diversos nichos ecológicos (Medini et al., 2005; Dobrindt et al., 2010; Mira et al., 2010).

Estos dos componentes del pangenoma diversifican por medio de mutaciones, recombinación homóloga y transferencia horizontal de genes. La frecuencia de tales mecanismos genéticos en cada especie bacteriana define si determinada especie presenta una poza génica muy grande o pequeña (si presenta un pangenoma abierto o cerrado respectivamente). Esto es importante ya que en la medida en la que la poza génica sea mayor, mayor respuesta al cambio habrá y por ende una probabilidad más alta de cambio evolutivo.

1.2.2. Mecanismos de diversificación de los genomas bacterianos

Se sabe que la mutación es en definitiva la principal fuente de variación genética debido a su naturaleza azarosa la cual origina alelos *de novo* (Levin, 1981; Futuyma, 2005; Hedrick, 2000). Debido a que las bacterias se reproducen asexualmente, por mucho tiempo se consideró a la mutación como la única fuente de variación (Levin, 1981). Sin embargo, la descripción de mecanismos de intercambio de información genética entre bacterias de la misma o diferentes especies como la conjugación, la transducción y la transformación sugieren una fuente alternativa de diversificación genética, la recombinación (Thomas y Nielsen, 2005; Narra y Ochman, 2006).

En su modalidad de recombinación homóloga, ésta promueve la diversificación del genoma central al favorecer la formación de nuevos haplotipos o alelos a partir de los ya existentes debido al re-emplazamiento de pequeños segmentos del cromosoma

bacteriano por las regiones homólogas de otro aislado de la misma especie o especies muy cercanas (Figura 2a) (Spratt et al., 2001; Dobrindt et al., 2010).

Por otra parte, la recombinación no homóloga o ilegítima comúnmente llamada transferencia horizontal de genes (THG), es la que se encarga de la diversificación del genoma flexible en cuanto que se adquieren genes *de novo* debido a que promueve el intercambio de información genética entre individuos de diferentes especies (Figura 2b) (Gogarten et al., 2002; Lawrence y Hendrickson, 2003; Nakamura et al., 2004; Ochman y Davalos 2006).

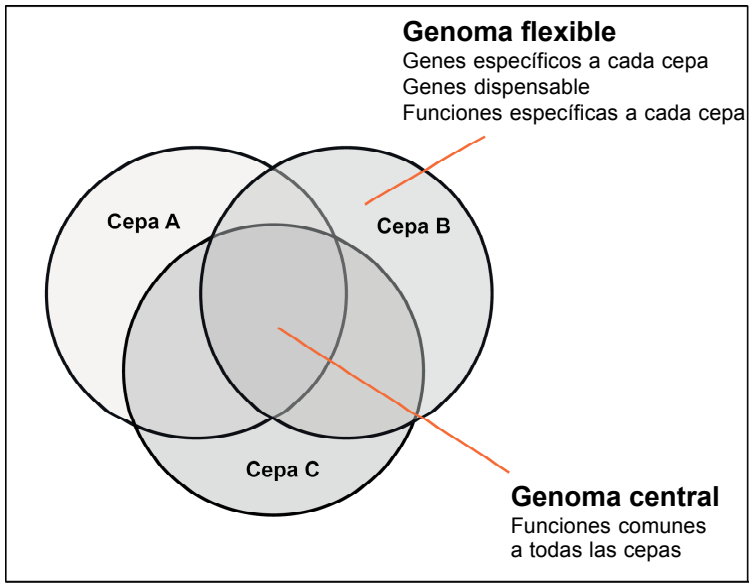

Figura 1. El pangenoma

El pangenoma de una especie consiste en el conjunto total de genes presentes en una especie bacteriana. Se encuentra compuesto por el genoma central el cual consiste en los genes compartidos por todos los miembros de la especie y por el genoma flexible, el cual hace referencia a los genes presentes solamente en ciertos subgrupos o miembros de la especie. Se ha propuesto que estos genes específicos a una cepa reflejan adaptación a determinados nichos (Garrigues et al., 2013; Polz et al., 2013).

Además de los mecanismos arriba mencionados, se han descrito otros como los re-arreglos cromosómicos, la duplicación génica y la erosión genética (Didelot et al., 2007; Hughes, 1999), los cuales junto con la THG, promueven la evolución del tamaño (Gevers et al., 2004; Gregory y DeSalle, 2005) y arquitectura del genoma en bacterias de diferentes especies (Rocha 2008; Koonin, 2009) así como la existencia de grandes diferencias en el tamaño del genoma aún entre aislados de la misma especie (Mira et al. 2001; Polz et al., 2013). Por ejemplo, *Prochlorococcus marinus* la cianobacteria marina

más abundante en los océanos de latitudes medias tiene un rango en el tamaño de sus genomas que va de los 1.69 a los 2.68 Mb (Kettler et al., 2007). Igualmente, el rango del tamaño del genoma de *Yersinia pestis*, enterobacteria facultativa causante de la peste bubónica es de 4.2 a 5.32 Mb (Morelli et al., 2010; Cui et al., 2013). Asimismo, mecanismos como los nombrados anteriormente, dan cuenta de la diferencia en el tamaño del genoma que existe entre bacterias con diferentes estilos de vida (Ochman y Dávalo, 2006). En general, la tendencia sugiere tamaños de genoma grandes (5-10 Mb) para las bacterias de vida libre, intermedios (2-5 Mb) para los patógenos facultativos o recientes y pequeños (0.5-1.5 Mb) para los patógenos obligados o simbiontes (Ochman y Dávalos, 2006).

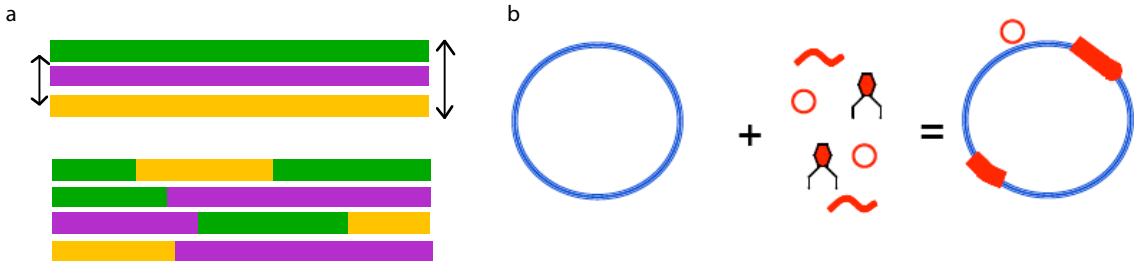

Figura 2. Recombinación homóloga y no homóloga

a. La recombinación homóloga promueve la diversificación del genoma central al generar nuevos alelos mosaico gracias al re-emplazamiento de pequeños segmentos entre cromosoma bacteriano homólogos. **b**. La recombinación no homóloga o transferencia horizontal, promueve la diversificación del genoma flexible y se da mediante la llegada a los genomas de genes o conjunto de genes los cuales permiten la adaptación a nuevos nichos.

1.2.3. Mecanismos evolutivos que actúan en las poblacional bacterianas

A grandes rasgos la evolución biológica consta de dos pasos. El primero consiste en la generación de la variación genética que como ya vimos, en el caso de las bacterias se da mediante la mutación puntual y la recombinación tanto homóloga como no homóloga. En el segundo paso, tanto la deriva génica como la selección natural se encargarán de definir el curso de esta variación. Ambas fases del proceso evolutivo se llevan a cabo a nivel poblacional, es decir, los mecanismos de diversificación genética se llevan a cabo tanto al interior y como entre poblaciones. Asimismo, tanto la deriva génica como la selección natural pueden actuar diferencialmente al interior de las poblaciones, en unos casos fijando determinados alelos, y en otros, eliminando o manteniendo la diversidad genética lo que a la larga promueve la divergencia de las poblaciones.

El componente azaroso de la evolución corresponde a la acción de la deriva génica, la cual cambia las frecuencias alélicas en el tiempo como resultado del muestreo (ó supervivencia) aleatorio de la variación genética que se da de generación en generación. Este proceso estocástico depende del tamaño efectivo poblacional, siendo las poblaciones pequeñas las más fuertemente moldeadas por la deriva génica en comparación con las poblaciones grandes.

Por otro lado, se encuentra la selección natural, que a diferencia de la deriva génica, es un proceso determinista (en términos de aumentar o al menos mantener la adecuación de las poblaciones) el cual describe la supervivencia diferencial y propagación de las variantes génicas dentro de una población.

Es así que el devenir evolutivo de la variación genética de las poblaciones naturales se encuentra modulado por un componente azaroso y otro determinista. Por ejemplo, la variación existente en genes que codifican proteínas y ARNs los cuales llevan a cabo funciones celulares básicas como metabolismo, transcripción y traducción, tiende a ser variación sinónima y generalmente neutral por lo que cualquier cambio en sus frecuencias alélicas es producto de la acción de la deriva génica. En cambio las variantes no sinónimas generalmente resultan en formas deletéreas por lo que la probabilidad de ser eliminados de la población por selección negativa o purificadora es mayor. En contraste, es común encontrar variantes no sinónimas en genes asociados con la adaptación a algún nicho (como factores de virulencia, de colonización, de asimilación, etc.) cuya frecuencia aumenta en las poblaciones gracias a la selección positiva. Finalmente, también es posible que la selección natural favorezca el mantenimiento de múltiples alelos al interior de las poblaciones mediante selección balanceadora o diversificadora.

1.2.4. La naturaleza de las poblaciones bacterianas

Considerando la definición de evolución aquí expuesta, resulta relevante la siguiente pregunta: ¿qué es una población bacteriana? y más aún, ¿cómo definimos a una población bacteriana? Es bien sabido que las poblaciones bacterianas tienen la capacidad de cambiar muy rápido de tamaño, las encontramos compuestas por pocas células bacterianas o constituidas por billones de bacterias. Asimismo pueden sufrir dramáticos cuellos de botella debido al suministro de antibióticos o al cambio de un hospedero (Achtman et al., 1999; Balloux, 2010; Schierup y Wiuf, 2010).

El tipo de relación que guardan las bacterias con un hospedero también es un factor importante que define al tamaño poblacional en bacterias. Por ejemplo, resulta extraordinario que el cuerpo humano sano albergue 10^{14} bacterias principalmente en el tracto digestivo, cantidad que sobrepasa el número de células del cuerpo (Berg, 1996). Tales tamaños poblacionales son gigantes aún después de considerar el hecho de que estas bacterias comensales o simbiontes pertenecen a múltiples especies diferentes.

Con respecto a los patógenos oportunistas, también se esperan tamaños poblacionales inmensos al menos en especies en donde una fracción importante de las cepas son potencialmente patogénicas. Parece ser este el caso de *Staphylococcus aureus*, *Neisseria meningitides* y *Streptococcus mutants*, las cuales se encuentran presentes en gran proporción en la población humana y en donde muchas cepas se vuelven patógena por mero accidente (Herczegh et al., 2008; van Belkum et al., 2009).

En algunos patógenos facultativos (por ejemplo, *Escherichia coli*) solamente una pequeña proporción de las cepas son dañinas por lo que el tamaño de la población infectiva se espera sea mucho más reducida comparado con la población bacteriana global. Finalmente, los tamaños poblacionales de los patógenos obligados se encuentran directamente restringidos por el número de portadores infectados y se espera sean pequeños excepto para aquellas especies que causan enfermedades ampliamente diseminadas como por ejemplo, la tuberculosis (OMS, 2008).

Pero una cosa es el número total de células bacterianas que se estima existen físicamente y otra, el número de células en las que realmente se llevan a cabo los procesos evolutivos, es decir, el tamaño efectivo poblacional. Así, para que haya un cambio evolutivo se requiere de variantes genéticas "raras" que en un principio se dispersen entre los miembros "efectivos" de una población, aumentando su frecuencia de tal manera que la población llega a ser genéticamente diferente a su condición ancestral en algún momento.

1.2.5. Estructura poblacional y divergencia genética

Así como las especies de macroorganismos, las especies bacterianas se encuentran formadas por diferentes poblaciones, las cuales pueden encontrarse divididas o separadas unas de las otras por barreras geográficas, ecológicas y/o temporales. Cuando esto sucede, cada una de las poblaciones puede presentar diferentes niveles de variación genética así como diferentes proporciones de dichas variantes (diferencias en las

frecuencias alélicas y genotípicas) (Hartl y Clark, 1989; Hedrick, 2000), lo que implica que las poblaciones que conforman a determinada especie presenten cierto grado de estructuración genética. Si esta estructuración poblacional es alta, con el tiempo las poblaciones divergen a tal grado que se originan nuevas especies. Sin embargo, no nada más la deriva génica y la selección natural moldean la diferenciación genética de las especies, también la recombinación tanto homóloga como no homóloga mantienen cohesivos a los linajes al promover el flujo génico entre las diferentes poblaciones, al mismo tiempo que las diferencian en la medida en la que actúa la recombinación homóloga al interior de las mismas.

No obstante, durante mucho tiempo prevaleció la idea de que las poblaciones bacterianas eran principalmente clonales, conclusión a la que se llegó después de haber analizado muestras representativas de diferentes especies provenientes de varios continentes a lo largo de los años (15 años en promedio y en su mayoría bacterias patógenas) y utilizando la técnica de enzimas multilocus (MLEE) (Selander et al., 1986; Caugant et al., 1987). Bajo esta metodología, a pesar de que se encontraron niveles de variación genética altos considerando la asexualidad de las bacterias (Whittam et al., 1983; Caugant et al., 1987; Selander et al., 1987; Maynard-Smith, 1993; Souza et al., 1994; Souza et al. 1999); se observó la distribución global de determinados genotipos multilocus para cada especie, sugiriendo así la clonalidad de las poblaciones bacterianas (Maynard-Smith, 2000). Asimismo, otro indicio de clonalidad fueron las observaciones que se tenían sobre la asociación de determinados serotipos con enfermedades particulares (Orskov y Orskov, 1983).

No fue sino hasta la obtención de las primeras secuencias de nucleótidos que la recombinación entró en el escenario evolutivo de las bacterias (O'Rourke et al., 1993), sugiriendo que las genealogías de diferentes genes de la misma especie no eran congruentes y que algunos alelos de dichos genes tenían una estructura en mosaico, lo que indica que diferentes regiones de un gen en particular, pueden ser el resultado de historias evolutivas diferentes debido a la recombinación homóloga (Feil y Spratt, 2001) (Figura 3). Asimismo, estos datos mostraron una asociación aleatoria de los alelos o equilibrio de ligamiento, en donde la presencia de un alelo en un locus determinado, es independiente de la presencia o ausencia de los alelos de otros loci. Lo que también sugiere niveles de recombinación homóloga considerables (Maynard-Smith, 2000).

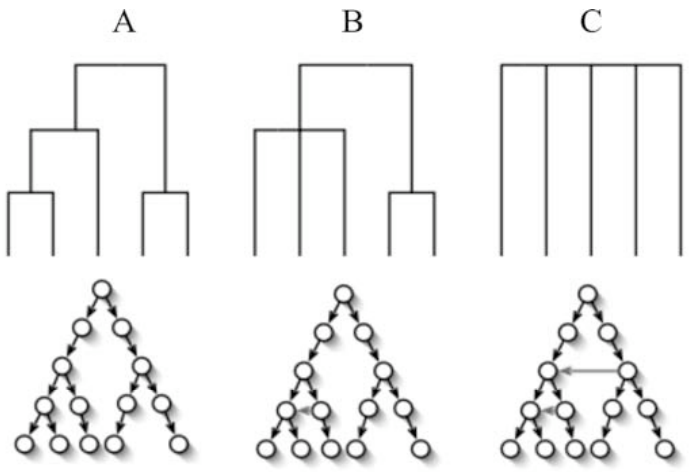

Figura 3. Consecuencias de la recombinación homóloga

La recombinación homóloga puede generar topologías ambiguas. Arriba, topologías esperadas bajo los efectos de la recombinación Abajo, linajes con eventos recombinatiorios marcados con flechas horizontales. (A) Topología resuelta de un linaje clonal. (B) Un evento recombinatorio causa ambigüedad en una de las ramas. (C) Dos eventos de recombinación son suficientes para nublar la historia del linaje (Figura tomada de Salas, 2007).

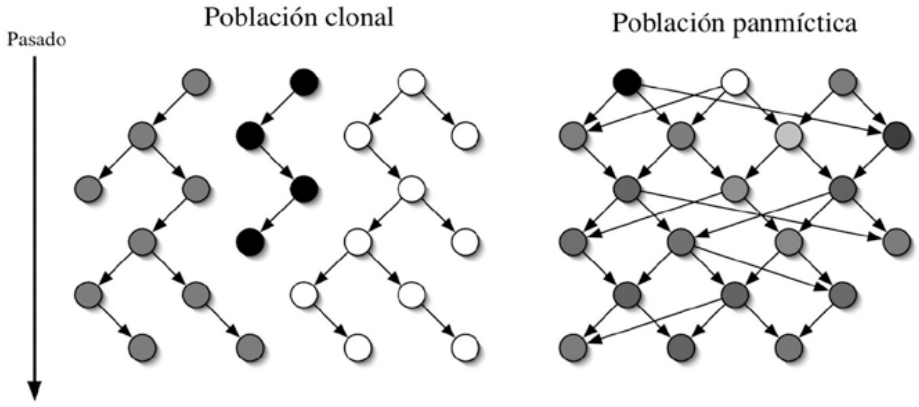

Figura 4. Poblaciones clonales y poblaciones panmícticas

Las poblaciones clonales conservan el mismo genotipo a lo largo del tiempo mientras que las poblaciones panmíticas o recombinantes generan nuevos genotipos por combinación de los alelos de la población (Figura tomada de Salas, 2007).

Años más tarde, se sistematizó el uso de las secuencias de genes constitutivos para el estudio de la diversidad y evolución bacterianas (Maiden et al., 1998). Este secuenciamiento multilocus de genes constitutivos (MLST, por sus siglas en inglés), además de arrojar niveles de variación genética reales, reforzó el papel diferencial que juega la recombinación homóloga en el origen de la diversidad genética y estructuración poblacional asociada a diferentes especies bacterianas (Figura 4) (Feil y Spratt, 2001; Feil, 2004, Hanage, et al., 2006; Vos y Didelot, 2009; Didelot y Maiden, 2010). Así, en la naturaleza podemos encontrar un continuo en la frecuencia de la recombinación (Vos y Didelot, 2009) partiendo de especies altamente clonales en donde dicha frecuencia es sumamente rara como en el caso de *Yersinia pestis* y el complejo de *Mycobacterium* (Achtman et al., 1999; Smith et al., 2003) hasta llegar a especies panmícticas como *Helicobacter pylori* en donde la recombinación es muy frecuente (Suerbaum et al., 1998).

Asimismo, diversos trabajos utilizando esta técnica MLST han arrojado indicios relacionados con los mecanismos que promueven los patrones de estructuración de la diversidad genética. Estos estudios sugieren que la variación genética en bacterias se encuentra organizada mediante mecanismos análogos a los de la especiacion alopátrica (Whitaker et al., 2003), ecológica (Palys et al., 2000; Cohan, 2001) y biológica (Dykhuizen y Green, 1991; Whitaker et al., 2005) propios de los macroorganismos. Por ejemplo, se ha detectado aislamiento geográfico el cual promueve especiación alopátrica, en poblaciones de arqueas (*Sulfolobus islandicus*) y cianobacterias hipertermófilas (*Synechococcus*) a escala intercontinental (Whitaker et al., 2003; Papke et al., 2003). Asimismo, existen barreras ecológicas que llevan a la especiación simpátrica como en el caso de *Vibrio vulnificus*, especie acuática que muestra una marcada diferenciación genética entre las cepas provenientes de humanos infectados con respecto a los aislados de animales marinos y ambientes acuáticos salados (Bisharat et al., 2007). Igualmente, se ha detectado aislamiento sexual el cual promueve especiación biológica análoga a la de eucariontes, en especies tan diferentes como Ferroplasma tipo II y *Salmonella enterica* en donde la frecuencia de recombinación al interior de las poblaciones es mayor que entre ellas debido a barreras mecanicistas propias de la recombinación homóloga como las impuestas por el sistema de reparación pareada de ADN (Tyson et al., 2004 y Didelot et al., 2011 respectivamente).

1.2.6. Dinámica evolutiva de especies clonales

¿Qué sucede cuando la principal fuente de variación genética es la mutación puntual? Las poblaciones bacterianas evolucionan principalmente por "selección periódica (Levin, 1981). En este proceso, surge una mutante la cual si es benéfica aumentará su frecuencia entre las células bacterianas debido a que confiere una mayor supervivencia y éxito reproductivo a la población. Pero al mismo tiempo, la selección natural eliminará a las demás variantes o a aquellas que confieran una adecuación menor a la población por medio de un "barridos selectivo" lo que a la larga disminuye la variación genética. Visto a largo plazo, este modelo implica un ritmo evolutivo a pasos periódicos, en donde cada paso corresponde a la aparición y fijación por un nuevo barrido selectivo, de alguna mutación que confiere una novedad evolutiva dando así origen a un nuevo linaje clonal (Figura 5).

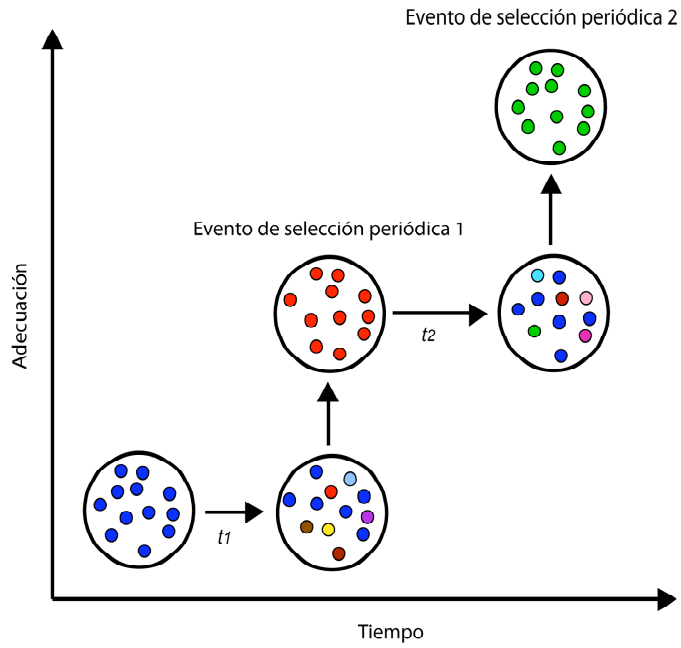

Figura 5. Evolución bacteriana de especies clonales por selección periódica

Una población bacteriana (en color azul) diversifica por mutación puntual promoviendo la aparición de mutantes que confieren una adecuación (éxito y supervivencia). Aquella que confiera la mayor adecuación, será fijada (población roja) eliminándose así toda la variación neutral acumulada mediante un barrido selectivo, esto debido a la ausencia de recombinación en las poblaciones. De nuevo, una población monomórfica diversifica y vuelve a fijarse la mutante adaptativa más exitosa (población verde). Así, la acción de la selección periódica resulta en una disminución dramática de la diversidad genética (Figura modificada de Feil, 2010).

1.2.7. Dinámica evolutiva de especies recombinantes

En cambio, cuando la frecuencia en la recombinación homóloga es alta, la probabilidad de que múltiples mutaciones potencialmente benéficas se dispersen entre las poblaciones aumenta, promoviendo la diversificación y cohesión de una sola y gran poza génica. Además, debido a que la recombinación suscita la aparición de genes y genomas mosaico, se genera cierta independencia entre las diferentes regiones del genoma por lo que tanto la selección natural como la deriva génica actuarán de manera independiente sobre los diferentes alelos y regiones del genoma (Salas, 2007). Pero conforme se van adquiriendo alelos del genoma central y/o genes "flexibles" asociados a diferentes nichos ecológicos y/o historias de vida, las poblaciones que conforman a una especie van diferenciándose una de la otra. Aunado a esto, si la recombinación al interior de cada población que conforma a una especie es mayor con respecto a la que se da entre las poblaciones, la diferenciación entre las mismas aumenta ya que el flujo génico disminuye, promoviendo aún más su divergencia y a la larga, el origen de nuevas especies (Dykhuizen y Green, 1991; Shapiro et al., 2012).

1.2.8. Procesos históricos y contemporáneos en las poblaciones bacterianas

Gracias a la disponibilidad de genomas completos de varios aislados de una misma especie bacteriana es que se ha logrado estudiar a fondo el papel que juegan tanto la selección natural como la deriva génica en la estructuración poblacional. Así, la genómica de poblaciones permite identificar y separar los efectos que tienen en la divergencia de las poblaciones bacterianas, los procesos histórico promovidos por la deriva génica, de los contemporáneos o ecológicos, moldeados por la selección natural (Luikart et al., 2003). Por lo que los efectos que se dan a nivel de todo el genoma, nos informan a cerca de la demografía de las poblaciones así como de su historia filogenética, mientras que los efectos que se dan a nivel de loci específicos, nos permiten identificar genes importantes para la adecuación y adaptación a nuevos nichos (Black et al, 2001; Luikart et al., 2003; Nadeau y Jiggins, 2010).

2. Modelo de estudio: *Escherichia coli*

Considerando lo anteriormente expuesto, un modelo de estudio apropiado que nos permita responder preguntas concernientes a los mecanismos evolutivos y ecológicos que promueven la diversificación, estructuración y divergencia de las poblaciones naturales, es aquella especie bacteriana que ocupe un amplio rango de nichos así como diferentes estilos de vida. Es así que elegimos a *Escherichia coli*.

2.1. Historia natural de *E. coli*

E. coli es una bacteria Gram negativa perteneciente a la clase de las Proteobacterias, subclase Gamma-proteobacterias, familia Enterobacteriacea. Esta familia se caracteriza por tener organismos facultativos lo que significa que pueden vivir en asociación con algún hospedero o en el ambiente externo (Logan, 1994).

E. coli se aísla típicamente de heces fecales de animales de sangre caliente como mamíferos con diferentes tipos de dietas y aves (Rosebury, 1962; Souza et al., 1999; Gordon y Cowling, 2003) así como también de animales de sangre fría como reptiles (Selander y Levin, 1980; Souza et al., 1999; Gordon y Cowling, 2003). Sin embargo, también se le ha aislado de ambientes acuáticos y terrestres (Ksoll et al., 2007) y a pesar de que se dice que este ambiente externo es secundario y transitorio (Savageau, 1983), estudios recientes han demostrado que genotipos provenientes de aislados de ambientes externos pueden estar adaptados y bajo la influencia de la selección natural (Power et al., 2005; Walk et al., 2007; Alm et al., 2011).

En general, la mayoría de las cepas de *E. coli* provenientes del tracto intestinal, son consideradas comensales. Asimismo, en el sistema urinario también se pueden encontrar cepas comensales, a las que se les conoce como *E. coli* causantes de bacteriuria asintomática (Dobrindt y Hacker, 2008). Sin embargo, existen cepas patógenas capaces de causar enfermedad en el sistema gastrointestinal así como a nivel extra-intestinal, el cual incluye al sistema urinario, las meninges, el peritoneo, los pulmones y la región intra-abdominal (Kaper et al., 2004).

Desde el punto de vista médico y epidemiológico se han clasificado a las cepas de *E. coli* en diferentes categorías o patotipos los cuales agrupan aislados que comparten un proceso de patogénesis similar, así como la generación de un cuadro clínico característico y un conjunto de factores genéticos también llamados factores de virulencia (Nataro y Kaper, 1998). Al momento, se han descrito principalmente 6

patotipos de *E. coli* patógenas intestinales (Tabla 1) (Kaper et al., 2004) y 4 extra-intestinales (Johnson, 2011).

Tabla 1. Factores de virulencia asociados a los principales patotipos de *Escherichia coli*

Medio	Microambiente	Patotipo	Principales factores de virulencia	Referencias
Intestinal	Intestino delgado	EPEC *E.coli* enteropatogénica	Factores de colonización: Plásmido del factor de adherencia y esfacelamiento (EAF). • Adhesinas: eae, Paa, LPB (Long Polar Bundle) • Pili: BFP (Bundle Forming Pilus) • LifA/Efa Toxinas y efectores: Isla del locus de esfacelamiento de enterocitos (LEE). • Autotransportador: EspC • Efectores tipo III: CifC (Cycle-Inhibiting factor), EspF, EspG, EspH, Map, Tir (Translocated Intimin Efector)	Eslava et al. 1994; Nataro y Kaper 1998; Kaper et al., 2004.
	Intestino delgado	ETEC *E.coli* enterotoxigénica	Factores de colonización: • Adhesinas: CFA(Colonization factor antigen) Toxinas y efectores: • Toxinas termoestable (STa y STb) y Termolábil (LT)	Eslava et al. 1994; Nataro y Kaper 1998; Kaper et al., 2004.
	Colon	EHEC *E.coli* enterohemorrágica	Factores de colonización: • Adhesinas: eae, Paa, ToxB, LifA/Efa, Saa, OmpA • Fimbria: LPF (Long Polar Fimbrial) Toxinas y efectores: Toxina tipo shiga (STx), ureasas Isla del locus de esfacelamiento de enterocitos (LEE). • Autotransportador: EspP • Efectores tipo III: CifC (Cycle-Inhibiting factor), EspF, EspH, Map, Tir (Translocated Intimin Efector), StcE	Eslava et al. 1994; Nataro y Kaper 1998.
	Intestino delgado y colon	EAEC *E.coli* enteroagregativa	Factores de colonización: • Adhesinas fimbriales agregativas: AAFs, *agg*, *aaf*, *aaf3*, *hdc*, *aap*, *shf* Toxinas y efectores: • Toxina termoestable (EASTI) *astA*, *pet*, *pic*	Nataro y Kaper 1998; Kaper et al., 2004; Dudley y Rasko, 2011.
		DAEC *E.coli* de adherencia difusa	Factores de colonización: • Adhesina fimbrial (F1845). Efectores: • Proteína de membrana externa (AIDA1).	Eslava et al. 1994; Nataro y Kaper 1998; Kaper et al., 2004.
	Colon	EIEC *E.coli* enteroinvasiva	Factores de colonización: • virG: nucleación de los filamentos de actina Factores de adecuación: • Sideróforo: aerobactina Toxinas y efectores: • Shigella enterotoxina 1 (ShET1), sepA, sigA, ipaA, ipaB, IPAC, ipaH, ipgD, virA	Nataro y Kaper 1998; Kaper et al., 2004; Bumbaugh y Lacher, 2011.
Extra-intestinal	Sistema urinario	UPEC *E.coli* uropatogénica	Factores de colonización: • Fimbrias: papA, F1C, S, *fimH* • Adhesinas: Dr Toxinas y efectores:	Kaper et al., 2004; Dobrindt y Hacker 2008; Johnson, 2011.

			• Sat, HlyA, CNF-1,-2 (Cytotoxic necrotizing factor), *kpsMT* (Polisacárido capsular del grupo II). Factores de adecuación: • Sideróforos: IreA, IroN • Transporte Heme: Shu	
	Meninges, sangre	MNEC *E.coli* asociada a sepsis-meningitis	Factores de colonización: • Fimbrias: S, ompA, Ibe A, B, C, AslA. • Cápsula K Factores de adecuación: • Shu (Transporte grupo Heme). Toxinas: • CNF-1,-2 (Cytotoxic necrotizing factor)	Kaper et al., 2004; Dobrindt y Hacker 2008.

Sin embargo, estudios sobre la distribución de los factores de virulencia en aislados que causan diferentes cuadros clínicos, sugieren que la frontera que define a cada uno de los patotipos es difusa ya que se han reportado aislados que presentan genes característicos a diferentes patotipos (Kaper et al., 2004; Johnson et al., 2008). Además, se ha visto que cepas comensales pueden albergar algunos de estos factores de virulencia (Fricke, 2008; Rasko et al., 2008). Por lo que únicamente la presencia de los factores de virulencia no implica forzosamente que determinada cepa sea un aislado patogénico. Entonces qué otros factores promueven la evolución de la patogénesis en *E. coli*?

2.2. Dinámica evolutiva de *E. coli*

Durante mucho tiempo, se consideró a *E. coli* como una especie totalmente clonal (Whittam et al., 1983; Maynard-Smith et al., 1993). Trabajos recientes utilizando el genoma central completo, sugieren que a pesar de los niveles de recombinación homóloga detectados, ésta especie permanece aparentemente clonal (Wirth et al., 2006; Touchon et al., 2009; Tenaillon et al., 2010; Denamur et al., 2010). En detalle, la tasa de recombinación es similar a lo largo de todo el genoma a excepción de tres regiones localizadas o "hotspots" (dos regiones alrededor del operon *rfb* involucrado en la síntesis del antígeno O y la otra región alrededor del gen fimA codificante del factor de adherencia fimbria A) en donde los niveles de recombinación son muy altos (Didelot et al., 2012). Esta dinámica en la recombinación favorece la existencia de una señal filogenética clara la cual organiza a los miembros de esta especie en cuatro grandes grupos filogenéticos (A+B1, B2, D y E) independientemente del tipo de muestra (Selander et al., 1987; Goullet y Picard, 1989; Herzer et al., 1990; Escobar-Páramo et

al., 2004; Gordon et al., 2008). No obstante, estudios recientes han descrito la presencia de nuevos grupos filogenéticos menores quedando los siguientes grupos *E. coli sensu stricto* (A+B1, B2, C, D, E y F) y un grupo correspondiente al clado I de *Escherichia* (Figura 6) (Clermont et al., 2013). Este clado I es considerado como un grupo de *E. coli* en términos de la recombinación existente entre las cepas pertenecientes a este linaje críptico y otros miembros pertenecientes a los clados mayores de *E. coli* (Luo et al., 2011).

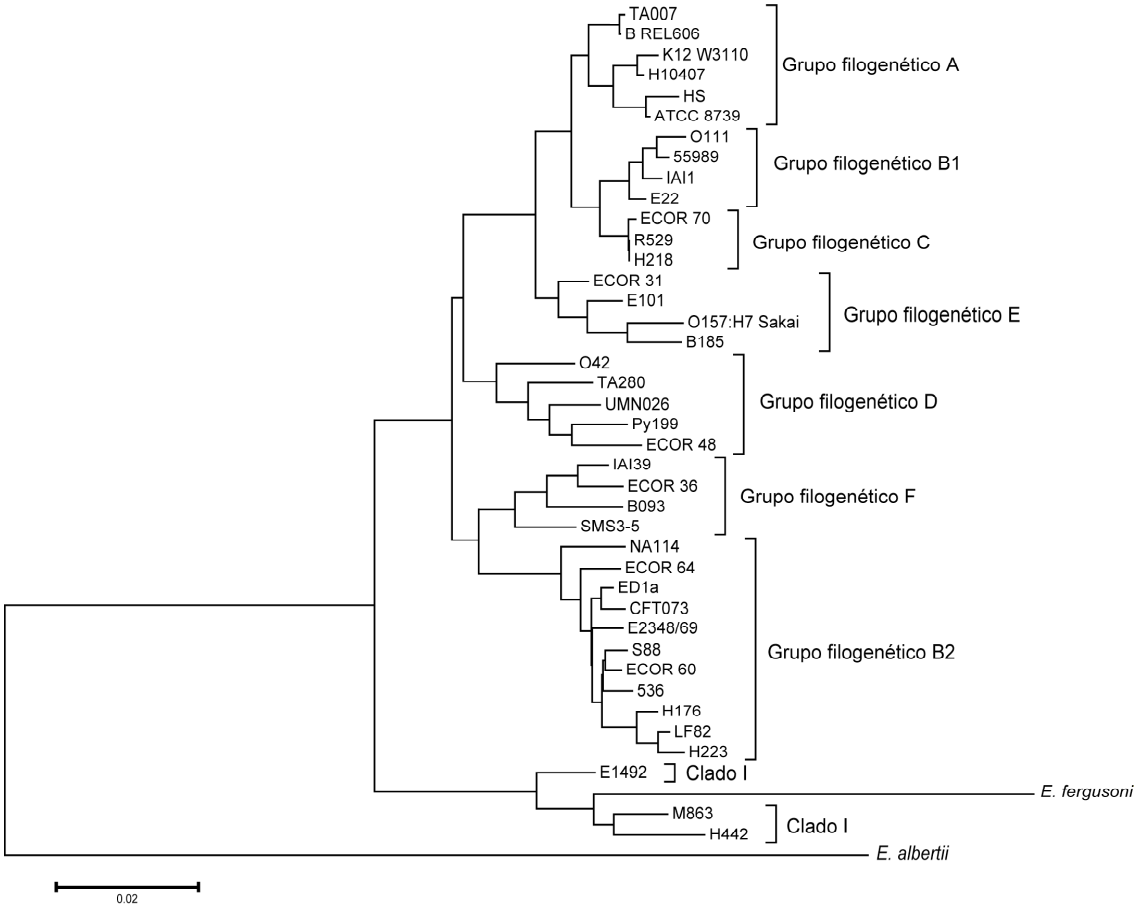

Figura 6. Estructura filogenética de *Escherichia coli*

Reconstrucción arrojada por el método de Máxima Verosimilitud utilizando el modelo evolutivo (GTR+G+I) y basado en la secuencia parcial de 13 genes constitutivos (9819 pb) (Figura tomada de Clermont et al., 2012).

Asimismo, se ha sugiriendo un componente ecológico en la evolución de esta especie entérica debido a que se ha establecido que los principales grupos filogenéticos difieren en sus nichos ecológicos, rasgos de historia de vida así como en la habilidad para causar determinadas enfermedades (Gordon, 2010; Tenaillon et al., 2010, Alm et

al., 2011). Así, cepas responsables de infecciones extra-intestinales es más probable que sean miembros del grupo B2 ó D en lugar del grupo A ó B1 (Picard et al., 1999; Johnson y Stell, 2000). En cambio, cepas clasificados como A ó B1 generalmente son aislados de mamíferos intestinales y comensales (Escobar-Páramo et al., 2006) así como de ambientes externos, peces, ranas y reptiles (Gordon y Cowling, 2003). Igualmente, aislados provenientes de animales de granja es más probable que pertenezcan a los grupos A y B1 que los aislados de animales silvestres (Picard et al., 1999).

Por otro lado, se ha visto que la capacidad de *E. coli* de habitar diferentes nichos y estilos de vida se debe a la alta plasticidad genómica que presenta esta especie la cual está dada por recombinación homóloga así como por la inserción de ADN adquirido horizontalmente (Touchon et al., 2009; Schubert et al., 2009; Didelot et al., 2012, Leimbach y Dobrindt, 2013). Así, esta dinámica del genoma da lugar a un reservorio genético grande lo que a su vez implica la presencia de un pangenoma abierto para esta enterobacteria. Cálculos recientes arrojan más de 18,000 genes, mientras que el genoma típico tiene alrededor de 5,000 genes (Chaudhuri y Henderson, 2012; Touchon et al., 2009; Rasko et al., 2008). Sin embargo, el tamaño del genoma varía en un rango que va de los 4.6 a los 5.6 Mb (Bergthorsson y Ochman, 1998) diferencia que se explica gracias a la adquisición de genes por transferencia horizontal y recombinación homóloga así como por eventos de reducción del genoma (Mira et al., 2001; Gregory y DeSalle, 2005). Asimismo, se ha sugerido una correlación entre el tamaño del genoma y el grupo filogenético al cual pertenece determinada cepa. Así, genomas más chicos se encuentran en los grupos A y B1 y los más grandes en los grupos B2, D y E (Bergthorsson y Ochman, 1998).

3. Justificación

A la fecha, los estudios que se han realizado sobre la diversificación y estructuración genética de *E. coli* se han realizado en muestras exclusivamente de aislados patógenos, o de cepas ambientales, o de cepas comensales de humanos (Escobar-Páramo et al., 2006; Walk et al., 2007; Jaureguy et al., 2008; Okeke et al., 2010;). Asimismo, se han llevado a cabo diversos trabajos con un enfoque filogenético en donde analizan el genoma completo de una muestra limitada de cepas representativas de los diferentes patotipos así como de comensales (Touchon et al., 2009; Leopold et al., 2011). Por otro lado, el estudio de la evolución de la patogénesis y comensalismo en *E. coli* se ha

basado principalmente en la descripción del flujo de información genética entre aislados de diferentes patotipos y con otras especies entéricas al determinar la presencia ó ausencia de factores de virulencia y colonización (Reid et al., 2000; Bumbaugh y Lacher, 2011; Lloyd y Mobley, 2011). Por lo que la información que alberga la gran gama de aislados provenientes de diferentes hospederos y estilos de vida en los que habita esta especie no se ha considerado aún en su totalidad.

El Instituto de Ecología cuenta con una colección de cepas de *E. coli* proveniente de hospederos tan diversos como aislados de ambientes externos como agua, aire, lodo, así como de mamíferos, reptiles, aves, marsupiales, tanto silvestres como domesticados, con diferentes tipos de dieta ya sea comensales o patógenas (Souza et al., 1999; 2002).

De esta muestra sabemos a nivel de genoma central, que los niveles de variación genética son altos (Souza et al., 1999). Asimismo sabemos, que aislados de animales silvestres provenientes de hospederos sanos en algunas ocasiones, cuentan con un genoma flexible dinámico, (correspondiente a la patogénesis) (Sandner et al., 2001). Pero ahora, cómo es que se genera y mantiene esta gran poza génica al mismo tiempo que promueve la presencia de diferentes estilos de vida en esta especie?

Es así que el objetivo general de este proyecto fue determinar los mecanismos de diversificación genética y los patrones de diferenciación poblacional en *Escherichia coli*.

Y en particular:

1.- Estimar el impacto de la recombinación homóloga y mutación puntual en la diversificación del genoma central de *E. coli* así como determinar su papel en la diferenciación genética de las poblaciones que conforman a esta especie.

2.- Analizar el papel del nicho ecológico en la estructuración de la diversidad genética de esta especie.

3.- Evaluar la influencia de la historia filogenética y el nicho ecológico en la diversificación del genoma flexible de *E. coli*.

4.- Determinar la contribución del genoma central en la evolución de la patogénesis en *E. coli*.

Los resultados y conclusiones derivados de los dos primeros objetivos se encuentran plasmados en un primer artículo titulado "Hierarchical clustering of genetic diversity associated to different levels of mutation and recombination in *Escherichia coli*: A study based on Mexican isolates" y se obtuvieron mediante análisis de genética de poblaciones. El tercer objetivo se resolvió determinando el tamaño del cromosoma de la muestra en estudio mediante la técnica de campos pulsados y los resultados se encuentran plasmados en un segundo artículo titulado "Chromosome size variation in *Escherichia coli* is not linked either to phylogeny nor ecological niche". Para llevar a cabo el cuarto objetivo se realizó un análisis de genómica de poblaciones utilizando genomas completos disponibles en las bases de datos y constituye la tesis de licenciatura de la Bióloga Luna Luisa Sánchez Reyes.

Los hallazgos de esta tesis contribuyen de manera significativa al entendimiento de los mecanismos evolutivos y genéticos que promueven la divergencia poblacional y evolución de la patogénesis en *E. coli*. En detalle, la poza génica correspondiente al genoma central de esta especie, se genera principalmente por recombinación homóloga que por mutación puntual. Sin embargo, la frecuencia a la que suceden estos mecanismos genéticos es diferente de acuerdo al nivel de organización de la diversidad genética y su asociación a nicho ecológico y estilo de vida. Así, esta tesis propone una explicación alternativa al paradigma clonal reportado para *E. coli*. Asimismo, los resultados de esta tesis sugieren que el tamaño del genoma en *E. coli* no está determinado por su historia filogenética como se ha sugerido hasta ahora. Y finalmente, los análisis de genómica de poblaciones sugieren que la adaptación de *E. coli* a diferentes nichos no se da solamente por la adquisición horizontal de genes sino que la evolución del genoma central, así como la regulación de la expresión génica juegan un papel importante en la evolución de la patogénesis.

4. Artículo 1

"Agrupamiento jerárquico de la diversidad genética, asociado a diferentes niveles de mutación y recombinación en *Escherichia coli*: un estudio basado en aislados mexicanos.

4.1. Resumen

Escherichia coli es un microorganismos tanto de vida libre como asociado al colon de mamíferos y aves ya sea como patógeno o comensal. A pesar de que la población mexicana de *E. coli* intestinal mantiene altos niveles de diversidad genética, se desconocen al momento los mecanismos mediante los cuales se origina dicha variación. En este trabajo investigamos el papel de la recombinación homóloga y la mutación puntual en la diversificación genética y estructura poblacional de aislados de *E. coli* provenientes de México. Para lo cual obtuvimos la secuencia parcial de siete genes constitutivos (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, *recA*) de una muestra compuesta por 128 aislados provenientes de un amplio rango de hospederos no asociados a brotes epidémicos. En general encontramos que la diversificación de ésta muestra se lleva a cabo principalmente por recombinación homóloga y en menor medida por mutación puntual. Y debido a que la diversidad genética se encuentra organizada genéticamente de acuerdo a la genealogía inferida utilizando los marcadores moleculares previamente descritos, observamos que no existe una tasa de recombinación homogénea sino más bien, que diferentes tasas emergen de acuerdo con los diferentes niveles de agrupación tales como grupo filogenético, linaje y complejo clonal. Además, detectamos una clara señal de subestructura al interior del grupo filogenético A+B1, en donde los aislados que lo componen se diferenciaron en cuatro linajes discretos. El patrón de subestructura se explica por la presencia de varios complejos clonales asociados a un estilo de vida y hospedero particular así como a los diferentes mecanismos de diversificación genética asociados. Por lo que proponemos estos hallazgos como una explicación alternativa al mantenimiento de la señal filogenética de esta especie a pesar de la prevalencia de la recombinación homóloga. Finalmente corroboramos que el utilizar tanto la aproximación filogenética como la de la genética de poblaciones se convierte en una herramienta efectiva para el establecimiento de una vigilancia epidemiológica a la medida de las especificidades ecológicas de cada región geográfica.

Infection, Genetics and Evolution 13 (2013) 187–197

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Hierarchical clustering of genetic diversity associated to different levels of mutation and recombination in *Escherichia coli*: A study based on Mexican isolates

Andrea González-González [a], Luna L. Sánchez-Reyes [a,1], Gabriela Delgado Sapien [b], Luis E. Eguiarte [a], Valeria Souza [a,*]

[a] Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, México D.F., Mexico
[b] Departamente de Microbiología y Parasitología, Facultad de Medicina, Universidad Nacional Autónoma de México, México D.F., Mexico

ARTICLE INFO

Article history:
Received 30 May 2012
Received in revised form 4 September 2012
Accepted 5 September 2012
Available online 18 September 2012

Keywords:
Escherichia coli
Population structure
Clonal complex
Homologous recombination
Point mutation
Epidemiological surveillance

ABSTRACT

Escherichia coli occur as either free-living microorganisms, or within the colons of mammals and birds as pathogenic or commensal bacteria. Although the Mexican population of intestinal *E. coli* maintains high levels of genetic diversity, the exact mechanisms by which this occurs remain unknown. We therefore investigated the role of homologous recombination and point mutation in the genetic diversification and population structure of Mexican strains of *E. coli*. This was explored using a multi locus sequence typing (MLST) approach in a non-outbreak related, host-wide sample of 128 isolates. Overall, genetic diversification in this sample appears to be driven primarily by homologous recombination, and to a lesser extent, by point mutation. Since genetic diversity is hierarchically organized according to the MLST genealogy, we observed that there is not a homogeneous recombination rate, but that different rates emerge at different clustering levels such as phylogenetic group, lineage and clonal complex (CC). Moreover, we detected clear signature of substructure among the A + B1 phylogenetic group, where the majority of isolates were differentiated into four discrete lineages. Substructure pattern is revealed by the presence of several CCs associated to a particular life style and host as well as to different genetic diversification mechanisms. We propose these findings as an alternative explanation for the maintenance of the clear phylogenetic signal of this species despite the prevalence of homologous recombination. Finally, we corroborate using both phylogenetic and genetic population approaches as an effective mean to establish epidemiological surveillance tailored to the ecological specificities of each geographic region.

1. Introduction

It has long been assumed that mutation is the primary mechanism by which prokaryotic populations diversify (Levin, 1981). Given that homologous recombination is fundamental to bacterial evolution (Spratt et al., 2001), it is now widely accepted that most bacterial species do not conform to the clonal model of evolution (Fraser et al., 2007; Vos and Didelot, 2009).

Homologous recombination involves the replacement of small segments of the bacterial chromosome with the homologous region from another isolate of the same, or a closely related species promoting the emergence of new haplotypes (Spratt et al., 2001). The frequency of these localized recombinational events varies broadly among bacteria of different species or genera (Vos and Didelot, 2009). Hence, this gives rise to a range of population structures spanning highly clonal populations, where recombination is extremely rare (such as *Yersinia pestis* and *Mycobacterium* species) (Achtman et al., 1999; Smith et al., 2003) to panmictic structures, where recombination is more frequent (*Helicobacter pylori*) (Suerbaum et al., 1998).

A classical model for the study of bacterial population genetics is *Escherichia coli*. Members of this species can be found inhabiting both water and sediment environments (Savageau, 1983) in addition to the vertebrate gut, where it occurs as either a commensal or pathogenic bacteria (Kaper et al., 2004; Tenaillon et al., 2010). For many years, a clonal population structure has been attributed to this species (Whittam et al., 1983; Maynard-Smith et al., 1993). However, recently published works suggest that the genetic structure of *E. coli* remains seemingly clonal in spite of the recombination levels detected. (Touchon et al., 2009; Tenaillon et al., 2010; Denamur et al., 2010). Thus, a clear phylogenetic signal delineates members of this species into four major phylogenetic groups (A + B1, B2, D and E) (Selander et al., 1987; Goullet and Picard, 1989; Herzer et al., 1990; Escobar-Páramo et al., 2004; Gordon et al., 2008). It has been assumed that each one of these lineages

* Corresponding author. Address: Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Apartado Postal 70-275, México D.F. 04510, Mexico. Tel.: +52 55 56 22 90 06; fax: +52 55 56 22 89 95.
E-mail address: souza@servidor.unam.mx (V. Souza).
[1] Present address: Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, México D.F., Mexico.

1567-1348/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2012.09.003

could be associated with a particular niche, suggesting an ecological component in *E. coli* evolution (Picard et al., 1999; Duriez et al., 2001; Gordon and Cowling, 2003; Power et al., 2005; Gordon et al., 2008). Recent studies however, have questioned the existence of such ecological specificity coupled to phylogenetic origin (Tenaillon et al., 2010; Clermont et al., 2011; Sabarly et al., 2011; White et al., 2011).

One of the widely used approaches into genetic population bacterial studies is MLST. This technique utilizes housekeeping gene nucleotide sequences to account for the mechanisms (recombination and mutation) and the processes (natural selection and genetic drift) that promote bacterial diversification (Maiden et al., 1998; Feil, 2004, 2010). Likewise, this method facilitates the description of bacterial populations as a number of discrete clusters of related genotypes also known as clonal complexes (CCs). Theoretically, CCs are the product of clonal expansions that can occur even within a freely recombining population when their founding genotypes have acquired some selective advantage (Feil, 2004).

In Mexico, high levels of genetic diversity coupled to specific host taxonomic groups have been reported for wild *E. coli* (Souza et al., 1999, 2002b). However, the exact mechanisms that generate this diversity are still unknown. Furthermore, studies on human-associated clinical isolates in this country are limited to the description of virulence genes and do not consider an evolutionary framework (López-Saucedo et al., 2003; Paniagua et al., 2007; Estrada-Garcia et al., 2009; Nicklasson et al., 2010).

Hence, the objective of this study is to determine the phylogenetic composition, as well as the dynamics of homologous recombination and point mutation, in the genetic diversification of non-outbreak strains of *E. coli* in Mexico. For this purpose, we analysed MLST data derived from a host wide range sample.

2. Material and methods

2.1. E. coli strains

A total of 128 strains isolated between 1994 and 2003 were selected exclusively from the Instituto de Ecología Collection of *E. coli* (IECOL), (Souza et al., 1999, 2002a; Sandner et al., 2001) and from the Enteric Pathogen Reference Laboratory collection, Public Health Department, Facultad de Medicina both at the Universidad Nacional Autonóma de México. The strains were serotyped for antigens O and H following the protocol described in Orskov and Orskov (1984), by the Enteric Pathogen Reference Laboratory, Public Health Department, Facultad de Medicina at the Universidad Nacional Autonóma de México. A pathogenic type was assigned for each strain based on serotyping (Supplementary Table S1).

Three groups of *E. coli* strains were analyzed: (a) 39 strains isolated from healthy animals from which, 25 were isolated from wild animals [5 enteropathogenic *E. coli* (EPEC), 4 enterotoxigenic *E. coli* (ETEC) and 16 non-pathogenic strains], 4 strains from captive animals [1 ETEC and 3 non-pathogenic strains] and 10 strains from domesticated animals [2 enterohemorragic *E. coli* (EHEC), 2 EPEC, 1 ETEC and 5 non-pathogenic strains]; (b) 67 human strains from which, 14 strains were isolated from faeces of diarrheagenic babies [12 enteroaggregative *E. coli* (EAEC) and 2 ETEC strains], 10 strains from the faeces of healthy babies [1 EPEC, 2 EAEC, 1 EHEC, 3 ETEC and 3 non-pathogenic isolates] and 43 strains from healthy adults [8 EHEC, 6 enteroinvasive *E. coli* (EIEC), 11 EPEC, 13 ETEC and 5 non-pathogenic strains]; (c) 22 environmental strains, from which 9 were isolated from air (2 EPEC, 7 unknown), 4 from soil (1 EIEC, 3 unknown), 1 from a drainpipe, 6 from mud and 2 isolates from water (Supplementary Table S1).

2.2. Culture conditions and DNA extraction

E. coli strains were grown overnight at 37 °C in 5 ml of Luria Broth (LB) culture with agitation at 200 rpm. Cells were harvested by centrifugation and DNA was isolated using a DNA extraction kit (DNeasy Blood & Tissue kit, Qiagen) according to manufacturer's instructions.

2.3. Clermont group determination

A triplex-PCR method was used to assign the *E. coli* isolates to phylogenetic group A, B1, B2 or D using a dichotomous key approach based on the presence or absence of two genes (*chuA* and *yjaA*) and an anonymous DNA fragment (TSPE4.C2) (Clermont et al., 2000).

2.4. MLST gene sequencing

Gene amplification of seven housekeeping genes conforming to the MLST scheme (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA*) were performed using the primers and protocol specified by the *E. coli* MLST website (http://mlst.ucc.ie/mlst/dbs/Ecoli) (Wirth et al., 2006) with minor modifications for annealing temperatures (Supplementary Table S2). All polymerase chain reaction (PCR) products were sequenced by the High Throughput Genomics Unit at the University of Washington (http://www.htseq.org). Forward and reverse sequences of each isolate were assembled, quality assessed and edited by visual inspection with PHRED/PHRAP/Consed software (Ewing et al., 1998; Gordon et al., 1998) (http://www.phrap.com). Sequences were aligned using the ClustalW program (Thompson et al., 1994) as implemented in BioEdit software package version 7.0.9.0 (Hall, 1999) (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). Sequences of MLST genes were submitted to GenBank (Accession Nos. JQ714288–JQ714385, JQ728594–JQ728623, JQ729124–JQ729251, JQ750344–JQ750471, JQ771211–JQ771296, JQ818824–JQ819249).

2.5. DNA polymorphism analyses

The number of alleles, haplotype diversity, numer of polymorphic sites, synonymous and non-synonymous changes, Tajima's *D* statistics, and the nucleotide diversity per site (π) were calculated using DnaSP v4.0 (Rozas et al., 2003).

2.6. Determination of sequence types (STs) and identification of clonal complexes (CCs)

A unique number was assigned to each distinct housekeeping gene sequence variant (allele) and a specific sequence type (ST) was attributed to each unique combination of alleles. New gene sequences for each locus were submitted to the *E. coli* MLST site (http://mlst.ucc.ie/mlst/dbs/Ecoli/) and new allelic and ST numbers were confirmed and assigned by the curator. These are all publicly available. To localize the new Mexican STs in a global diversity context, we performed a maximum likelihood (ML) phylogenetic inference considering the Mexican sample plus 1876 STs from the *E. coli* MLST database. This analysis was conducted using the RAxML v7.0.4 software (Stamatakis, 2006) implemented in the CIPRES web interface (http://www.phylo.org/news/RAxML). CCs were defined using eBURST V3 (http://eburst.mlst.net) (Feil et al., 2004) as groups sharing at least six identical alleles and bootstrapping with 1000 samplings using the whole MLST database as reference.

2.7. Phylogenetic analysis

Concatenated nucleotide sequences of STs were used to infer the genealogy of the sample using ClonalFrame version 1.2 (Didelot and Falush, 2007). ClonalFrame is a model-based method that reconstructs genealogies considering the effect of homologous recombination events that disrupt a clonal pattern of inheritance. Four independent runs of ClonalFrame were performed each consisting of 450,000 MCMC, where the first 150,000 iterations were discarded as burn-in. Convergence of all parameters was assessed by Gelman–Rubin method (1992). The genealogy was summarized in a 50% majority rule consensus tree using the ClonalFrame GUI (Didelot and Falush, 2007).

2.8. Genetic differentiation and host association

To account for genetic differentiation levels, we calculated the average number of nucleotide differences within and among phylogenetic groups using DnaSP v4.0 (Rozas et al., 2003). In addition, the fixation index (F_{ST}) was calculated using the Arlequin 3.1 software package (Excoffier et al., 2005). Furthermore, this software was also used to test for associations between the genetic data and host or life style (pathogenic or non-pathogenic) by performing an analysis of molecular variance (AMOVA).

2.9. Recombination analyses

To detect homologous recombination within housekeeping genes we calculated two parameters: ρ/θ (the ratio of rates at which recombination and mutation occur) and r/m (the ratio of rates that a nucleotide is changed as the result of recombination relative to point mutation). This was achieved by extracting from the ClonalFrame output the numbers of mutation events, recombination events, and substitutions introduced by recombination for the total sample, for each phylogenetic group, for each lineage and for each CC.

To assess the influence of recombination in the phylogenetic history of the Mexican sample, four individual ClonalFrame runs (150,000 burn-in iterations plus 450,000 sampling iterations) were performed ignoring the role of recombination in the estimation of the genealogy (i.e., the recombination rate ρ was set equal to zero).

To describe the levels of intra-lineage recombination we obtained the number of recombination breakpoints per lineage using the genetic algorithm for recombination detection (GARD) tool. This was implemented in the Datamonkey web interface (http://www.datamoneky.org/GARD/) (Pond and Frost, 2005) with default settings.

To detect the levels of inter-lineage recombination we performed an analysis using the linkage model of the program STRUCTURE (Pritchard et al., 2000). This software permitted us to infer the ancestry of the different lineages identified among the Mexican sample. Four runs of 10,000 iterations of burning, and 20,000 MCMC repeats after burning, were performed independently for each value of the number of populations K ranging from 2 to 7 (20 replicates per value of K). The optimal value was $K = 4$ by comparing the posterior probabilities of the data given each value of K from 2 to 7 (Supplementary Fig. S1). Applying the Evanno method (Evanno et al., 2005) also resulted in the estimate $K = 4$ (Supplementary Fig. S2). Assignment of STs to ancestral populations was performed as described by Wirth et al. (2006) with some modifications regarding unassigned STs. Hybrid STs were allocated as AxB1 or B2xD if the sum of A + B1 and B2 + D ancestral proportion was $\geqslant 0.8$. Otherwise, full hybrids ABD were designated.

To detect if recombination is sufficient to create random-association of alleles, the levels of linkage disequilibrium were obtained. For this, we obtained the standardized index of association (I^S_A) with 10,000 Monte Carlo simulations to confirm statistical significance. Both methods were implemented in LIAN v3.5 software (Haubold and Hudson, 2000).

Table 1
Summary of genetic diversity parameters for the 7 coding loci of 128 mexican *E. coli* strains, for each phylogenetic group and for each lineage.

		No. of alleles	H[b] ± SD[c]	No. (%) of polymorphic sites	Synonymous changes	Non synonymous changes	Tajima's D[d]	π[e]
Gene (128 isolates)	Fragment size (bp)[a]							
adk	534	21	0.789 ± 0.270	40 (7.49)	40	3	−0.423	0.012
fumC	465	26	0.900 ± 0.015	50 (10.75)	45	6	0.360	0.022
gyrB	459	29	0.890 ± 0.019	32 (6.97)	32	3	−0.762	0.009
icd	516	26	0.920 ± 0.011	43 (8.33)	42	5	0.022	0.015
mdh	450	18	0.890 ± 0.014	21 (4.66)	19	2	0.351	0.009
purA	474	24	0.740 ± 0.040	26 (5.48)	24	2	−1.218	0.006
recA	507	19	0.800 ± 0.020	24 (4.73)	25	0	−0.989	0.006
Phylogenetic group (3405 bp)	No. of isolates							
A + B1	91	58	0.978 ± 0.010	132 (3.87)	124	14	0.239	0.008
B2	15	12	0.971 ± 0.030	102 (2.99)	95	8	−0.221	0.009
D	15	10	0.914 ± 0.060	82 (2.40)	81	2	0.285	0.008
E	7	2	0.476 ± 0.171	1 (0.03)	0	1	0.559	0.0001
Lineage (3405 bp)	No. of isolates							
A	32	18	0.879 ± 0.050	55 (1.61)	49	7	−1.356	0.002
A-ST522	7	6	0.952 ± 0.096	19 (0.55)	17	2	0.727	0.002
B1	45	28	0.971 ± 0.011	67 (1.96)	66	4	0.369	0.005
B1-ST86	5	4	0.900 ± 0.161	3 (0.01)	2	1	−0.174	0.0004
Total Sample (3405 bp)	No. of isolates 128	82	0.986 ± 0.004	236 (6.90)	227	21	−0.336	0.011

[a] Number of base pair of protein coding region amplified.
[b] Haplotype (gene) diversity.
[c] Standard Deviation.
[d] Not significant ($P > 0.05$).
[e] Average number of nucleotide differences per site.

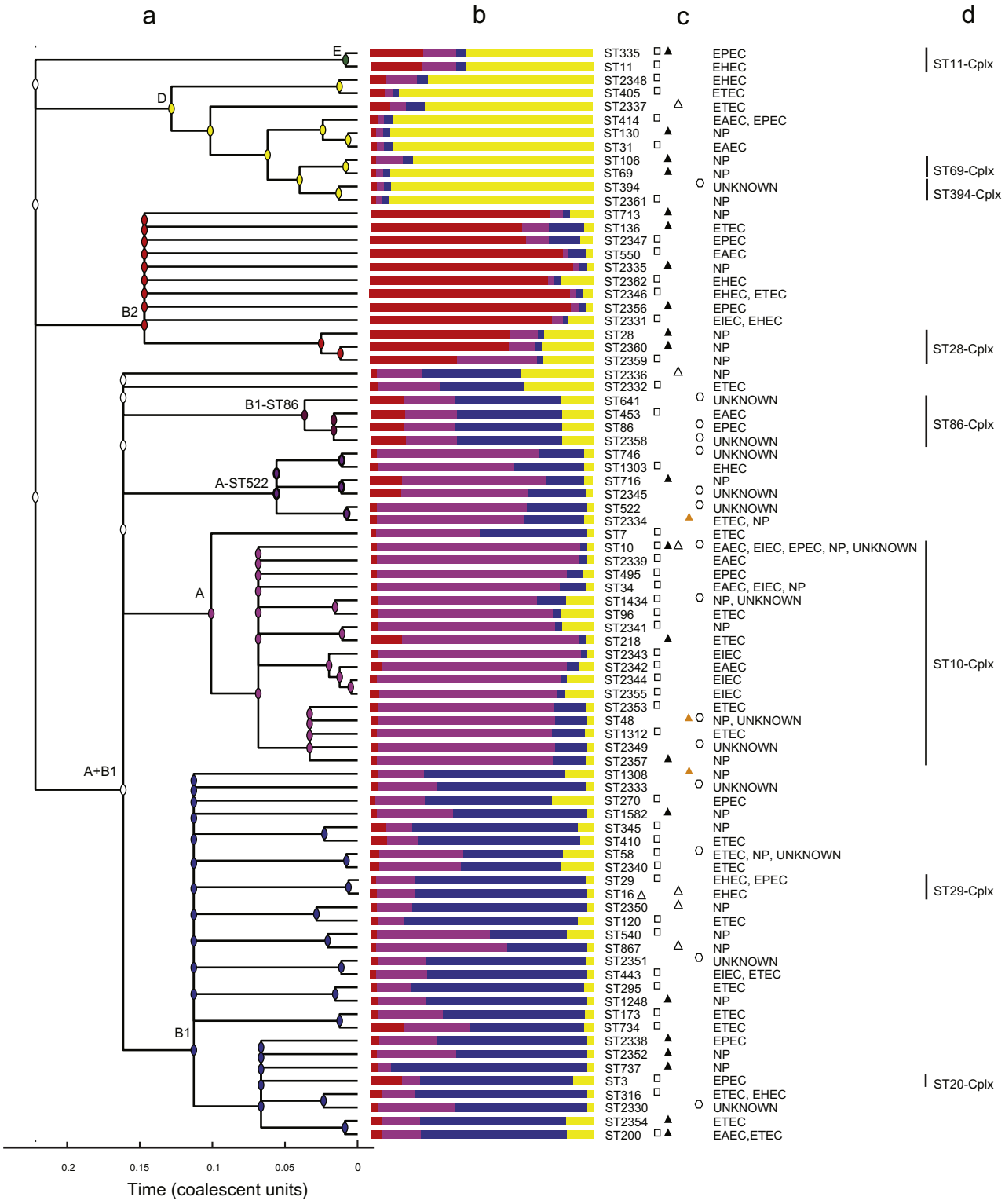

Fig. 1. (a) Unrooted consensus tree displaying the relationship between the 82 STs of the Mexican *E. coli* sample at seven concatenated loci (3423 nucleotides in total). The dendogram was constructed from the combination of five ClonalFrame runs with a cut-off value of 0.5 as a majority rule consensus. Scale is in coalescent units. (b) Proportion of ancestry of each of 82 STs inferred by STRUCTURE assuming $K = 4$ ancestral groups. Each vertical line represents an individual ST and indicates the proportion of ancestry from the four ancestral groups. (c) Host and life style. Human host is indicated by an open square, domesticated animal by an open triangle, wild animal by a black full triangle, captive animal by an orange full triangle and environmental host by an open hexagon. NP: Non-Pathogenic (d) STs that conform the major Clonal Complexes (containing five or more Mexican *E. coli* strains) identified by eBURST are grouped by a solid line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

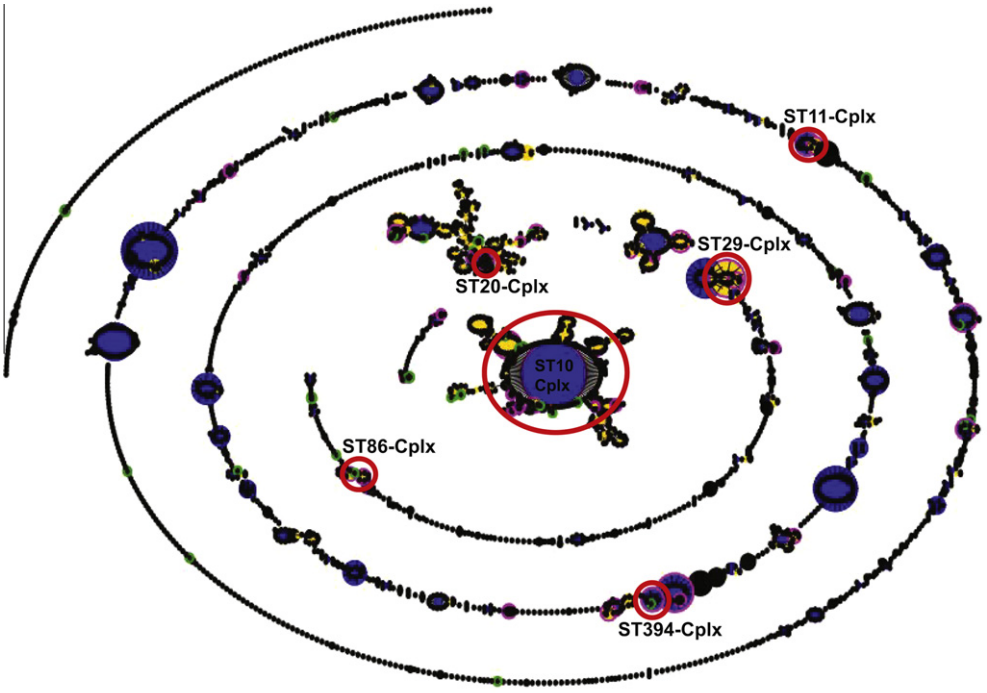

Fig. 2. eBURST diagram displaying the comparative analysis between Mexican sample and *E. coli* MLST database. Clusters of linked isolates correspond to clonal complexes (CC). Primary founders in the cluster are shown in blue and the subgroup founders are shown in yellow. The size of each circle is proportional to the number of isolates with the same ST. STs found in both populations are circled in pink and the STs found only in the Mexican population are shown in green. CCs containing five or more Mexican strains are circled in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Average number of nucleotide differences within and between each phylogenetic group, host types and life styles.

Phylogenetic group		A + B1	B2	D	E
	A + B1	27.8			
	B2	53.2	29.8		
	D	52.9	65.1	26.9	
	E	43.1	56.2	44.4	0.5
Host		Animal	Environment	Human	
	Animal	39.6			
	Environment	38.4	35.4		
	Human	39.5	37.9	39.6	
Life style		Non-pathogenic	Pathogenic		
	Non-pathogenic	39.5			
	Pathogenic	39.8	39.2		

3. Results

3.1. Nucleotide, ST and clonal complex diversity

The nucleotide diversity (π) among Mexican strains of *E. coli* ranged from 0.00588 (*purA*) to 0.02213 (*fumC*) with a mean value of 0.011 in concatenated genes. Across the 128 strains, Tajima's *D* was not significant for any single locus nor for the concatenated sequences (Table 1).

Based on the combination of alleles at the seven housekeeping gene fragments, 82 STs were recovered from the 128 strains. ST10 was the most frequent, represented by 11 isolates (8.6% of the sample). The majority of STs (60 STs representing 46.8% of the total sample) occurred as singletons in the dataset.

The Mexican sample adds 33 new STs to the *E. coli* MLST database (http://mlst.ucc.ie/mlst/dbs/Ecoli). From these new STs, 14 STs were derived from new combinations of previously reported alleles. A further 19 STs, were evident from the sequencing of new gene alleles. These new haplotypes were designated as ST2330 through ST2362 in the MLST database and distributed over the global genetic diversity of the species (Supplementary Fig. S3).

eBURST analysis grouped 85 (66.4%) of the 128 Mexican strains into 21 clonal complexes (CCs) (Supplementary Table S1). Fifty one (39.8%) strains fell into six major CCs (those CCs that cluster at least five isolates) where ST10-Cplx is the dominant CC, clustering 23 strains (18%) followed by ST394-Cplx, ST20-Cplx, ST29-Cplx and ST11-Cplx (Figs. 1 and 2).

3.2. Phylogenetic inference, genetic differentiation and niche association

ClonalFrame analysis clusters the Mexican STs into the four major phylogenetic groups comprising *E. coli* as species as divided into phylogenetic groups A + B1, B2, D and E. The majority of Mexican

Table 3
Recombination rates per phylogenetic group, lineage and clonal complex inferred by ClonalFrame analysis.

	Recombination events	Mutation events	Substitutions introduced by recombination	ρ/θ[a]	r/m[b]
Phylogenetic group					
A + B1	19	118	196	0.16	1.66
B2	8	42	112	0.19	2.7
D	13	25	99	0.52	3.96
E	1	4	14	0.25	3.5
Lineage					
A	7	25	60	0.28	2.4
A-ST522	1	13	16	0.08	1.26
B1	9	67	85	0.13	1.27
B1-ST86	1	6	9	0.16	1.46
Clonal complex					
ST10-Cplx	4	8	30	0.5	3.75
ST165-Cplx	0	5	0	NA[c]	NA
ST20-Cplx	1	2	4	0.5	2
ST29-Cplx	0	3	0	NA	NA
ST28-Cplx	4	4	40	1	10
ST69-Cplx	2	1	9	2	9
ST394-Cplx	3	1	16	3	16

[a] Relative frequency of occurrence of recombination as compared to point mutation in genetic diversification.
[b] Relative impact of recombination as compared to point mutation on the genetic differentiation of population.
[c] NA, not applicable.

strains (71%) cluster in the A + B1 monophyletic group. Mexican *E. coli* is genetically structured as suggested by a significant F_{ST} value of 0.50 ($P < 0.05$) and by the higher than average nucleotide differences observed between phylogenetic groups, rather than within them (Table 2). Furthermore, we did not find a significant ecological niche association with any phylogenetic group, since no genetic differentiation due to host (animal, human and environment) or life style (pathogenic and non-pathogenic) was recovered. This was also pointed out by a similar average number of nucleotide differences among groups (Table 2) and by the AMOVA analysis performed (99.5% and 98.73% of genetic variation was within populations respectively).

Despite the lack of statistical association of ecological traits at phylogenetic group level, we found that at lower phylogenetic levels, such as CCs, a clear host or lifestyle association prevails (Supplementary Table S1). For example, we found a correspondence to open environments in the ST86-Cplx and ST394-Cplx and to specific pathotypes and serotypes such as ST20-Cplx (EPEC O111:H- and O111ab:H-), ST29-Cplx (EHEC and EPEC O26:H-/H11) and ST11-Cplx (EHEC O157:H7, EPEC O55:H7/H-).

3.3. Levels and patterns of homologous recombination

Overall, the MLST nucleotide diversity was generated mainly by homologous recombination events (ρ/θ= 1.5 (95% CI [0.91, 2.58]) producing more nucleotide substitutions than point mutation (r/m= 3.24 (95% CI [2.1, 4.81]).

Furthermore, we found that mutation and recombination estimates are different through all clustering levels (phylogenetic group, lineage and CCs) in which Mexican genetic diversity is organized.

In detail, homologous recombination has a more important role than point mutation in the genetic diversification of phylogenetic groups D and E (r/m = 3.96 and 3.5, respectively) than in lineage B1 (r/m = 1.27) and an intermediate impact in phylogenetic group B2 and lineage A (r/m = 2.7 and 2.4, respectively). Moreover, homologous recombination had a remarkable impact on the generation of genetic diversity of determined CCs, such as ST28-Cplx, ST69-Cplx and ST394-Cplx. This was in contrast to other CCs, such as ST165-Cplx and ST29-Cplx where mutation is accounted for the diversity levels. On the other hand, recombination levels were intermediate in ST10-Cplx and ST20-Cplx (Table 3).

Ignoring recombination in the reconstruction of Mexican genealogy permit us to depict the effects of recombination in the phylogenetic signal (Supplementary Fig. S4). This analysis shows a clear separation between A and B1 sister lineages as well as a merging between D and E into one phylogenetic group due to recombination. Visual inspection of the events inferred by ClonalFrame revealed that substitutions promoting the diversification of the Mexican sample are the result of differential mutation/recombination gene dynamics within each phylogenetic group, lineage and CC. On one hand, substitutions in B2, D and E occur more extensively throughout all genes. In contrast, the pattern present in A + B1 indicates that substitutions occur in specific genes for each lineage (Supplementary Fig. S5).

In addition, intra-lineage recombination (measured as the number of recombination breakpoints per lineage) occurs more frequently in B1 and B2 than in E, and is intermediate in A and D (Table 4).

STRUCTURE analysis revealed different levels of inter-lineage recombination as different admixture levels were found across phylogenetic groups (Fig. 1). In particular, there was higher admixture between lineages A and B1 and between B2 and D. Phylogenetic group E, however, did not exhibit a distinct ancestral source of polymorphism. Furthermore, the genetic diversity of this group is derived mainly from groups B2 and D (contributing 57% and 24% of genetic diversity in Group E, respectively) (Fig. 1 and Supplementary Table S3).

The high levels of homologous recombination previously described were not sufficient to create random-association of alleles, as linkage disequilibrium (LD) was observed for all 128 *E. coli* strains (I^S_A = 0.304, $P < 0.0001$) remaining significant when only STs were considered (I^S_A = 0.201, $P < 0.0001$).

3.4. Substructure in A + B1 phylogenetic group

In detail, four subgroups were identified within phylogenetic group A + B1. Two of these subgroups correspond to A and B1 lineages, each with its own ancestral signature. The other two subgroups are recombinant lineages referred to as A-ST522 and B1-ST86. The first one comprises A Clermont strains with an

Table 4
Recombination breakpoints estimated by GARD algorithm per gene, per lineage, per phylogenetic group and for the total sample.

		Number of point break detected	Position[b]	Δc-AIC[c]
Gene	Length (bp)[a]			
adk	534	1	**333*****	73.16
fumC	465	1	**225*****	108.5
gyrB	459	1	**216*****	134.7
icd	516	1	**354*****	111.8
mdh	450	0	0	0
purA	474	0	0	0
recA	507	0	0	0
Phylogenetic group				
A + B1	3405	4	**534*****, **1185*****, **1974*****, **2784*****	2041.3
B2	3405	6	525, **900*****, **1455*****, **2112*****, 2685, 2865	418.5
D	3405	5	987***, 1416, 1722, **2268*****, **2799*****	217.2
E	3405	NA[d]	NA	NA
Lineage				
A	3405	2	**1554*****, **1921*****	79.60
A-ST522	3405	2	**1554*****, 2121	17.56
B1	3405	4	**915*****, **1470*****, **1977*****, **2742*****	988.90
B1-ST86	3405	0	0	0
Total sample	3405	4	**534*****, **1426*****, **1974*****, **2799*****	3340.8

[a] Number of base pair of protein coding region amplified.
[b] Each number indicates the nucleotide position into the alignment. Bold numbers points out the statistically significant positions by SH test. $^{***}p = 0.01$, $^{**}p = 0.05$.
[c] Δc-AIC indicates the difference in the AIC between the non recombination model (single-tree model) and the best recombination model.
[d] NA, not applicable.

AxB1 ancestry and the second, clusters A and B1 Clermont strains with ABD ancestry. Two ungrouped STs (ST2332 and ST2336) were also found within A + B1, corresponding respectively to A and B1 Clermont strains with an ABD ancestry (Fig. 1 and Supplementary Table S3).

Despite A + B1 clusters the majority of STs, its genetic diversity levels are identical to those present in D and lower than those harboured by B2 phylogenetic groups. Likewise, Mexican A + B1 exhibits complex selective dynamics, as positive and purifying selection is differentially acting upon certain genes inside this phylogenetic group. In particular, the *mdh* gene shows significantly positive Tajima's *D* value in A + B1 (2.41, $P < 0.05$), *fumC* gene displays a significantly negative Tajima's *D* value in A-ST522 (-2.38, $P < 0.01$) and *icd* gene a significantly positive Tajima's *D* value in B1-ST86 (3.02, $P < 0.01$).

4. Discussion

In this study we investigate the genetic diversification mechanisms of non-outbreak related *E. coli* isolates from Mexico. Overall, levels of genetic diversity harboured by multilocus sequences typing (MLST) genes are explained mainly by homologous recombination rather than by point mutation. Since genetic diversity is hierarchically organized according to the MLST genealogy, we observed that there is not a homogeneous recombination rate, but that different rates emerge at different clustering levels such as phylogenetic group, lineage and clonal complex (CC). We propose this finding as an alternative explanation for the maintenance of the phylogenetic signal of this species despite the prevalence of homologous recombination.

4.1. The recombinant nature of diversity in Mexican E. coli

Our results suggest that recombination is the main molecular process involved in the generation of Mexican nucleotide diversity. Levels of polymorphism and recombination are consistent with previously reported values for wild animal *E. coli* isolates from Mexico (Peek et al., 2001). This is also the case among different samples (environmental, different pathogenic types and host-wide range samples) and other housekeeping genes (Reid et al., 2000; Feil et al., 2001; Wirth et al., 2006; Walk et al., 2007; Jaureguy et al., 2008; Touchon et al., 2009; Okeke et al., 2010; Bergholz et al., 2011).

Despite the homologous recombination levels found, linkage disequilibrium (LD) prevails across the sample suggesting a clonal population structure for Mexican *E. coli* (Maynard-Smith et al., 1993). This could be explained by the relative impact of recombination on the genetic diversification as compared to point mutation. Thus, although more nucleotides become substituted by each recombination event in contrast to point mutation, the frequency of homologous recombination events in the Mexican sample is not sufficient to promote a random assortment of alleles at the different loci analyzed. This is not surprising since it has been proposed that recombination must be 10–20 times higher than point mutation in order to reflect linkage equilibrium (Maynard-Smith et al., 1993).

Alternatively, Touchon et al. (2009) address this clonal paradigm applying coalescent simulations and Bayesian calculations to the core genome, proposing that recombination events occur within short fragments (50 bp on average). Thus, an apparent clonal population structure and a clear phylogenetic signal are recovered, despite a higher recombination rate than mutation rate. However, our data and a recent whole genome study suggest an average tract length higher (380 bp long; 95% CI [249.8, 568.98 pb]; this study; 542 bp long; Didelot et al., 2012). This data could suggest that even when the average tract length of recombination is large, a clonal population structure can be depicted. Nevertheless, comparative algorithm analysis should be performed in order to shed more light on this topic.

4.2. Inter and intra-lineage recombination and genetic structure

From a population dynamics perspective, Maynard-Smith et al. (1993) explain that linkage disequilibrium can also be found in a cryptically subdivided population. This occurs when recombination is more frequent within subpopulations than between them, because populations are isolated from each other geographically, ecologically or temporally. Due to this isolation different

populations of the same species may exhibit different genetic substructures (Holmes et al., 1999) associated to different genetic diversification processes.

In general, our study indicates that the diversification of *E. coli* is associated with particular levels of recombination within each phylogenetic group. This result agreed with previous reports corresponding to *E. coli* on a global scale and samples from Nigeria (Wirth et al., 2006; Okeke et al., 2010). Nevertheless, our study also indicates that the different patterns and levels of intra and inter-lineage recombination described in this work, could explain the significant values of genetic structure recovered. Specifically, each phylogenetic group displays its own genetic signature (intra-lineage recombination differentiates the gene pool of the species), while preserving a common genetic background (inter-lineage recombination homogenizes it). Moreover, our results suggest that at phylogenetic level, homologous recombination is both acting as a homogenizing and as diversifying process (Supplementary Fig. S4).

Furthermore, we observed signatures of genetic structure inside the A + B1 phylogenetic group (substructure). To date, evidence of genetic substructure among phylogenetic groups of *E. coli* has only been reported within B2 as nine subgroups correlating to genetic and ecological diversity were detected (Le Gall et al., 2007). The fact that A + B1 consists of fewer lineages and has lower levels of nucleotide diversity than B2 could be attributed to the recent emergence of A + B1, in contrast to the early emergence of B2 (Touchon et al., 2009; Tenaillon et al., 2010; Denamur et al., 2010; Perna, 2011).

4.3. Ecological structure and clonal complexes

It has been suggested that ecological and adaptive mechanisms could represent barriers to recombination (sexual isolation) between bacterial lineages (Koeppel et al., 2008). Our analysis does not support the existence of such ecological sexual isolation between the Mexican *E. coli* phylogenetic groups as also suggested by recent works (Tenaillon et al., 2010; Clermont et al., 2011; White et al., 2011). Nevertheless, ecological barriers in *E. coli* have been proposed at higher phylogenetic levels (Luo et al., 2011).

Hence, determining in a phylogeny the clustering level that corresponds to an ecologically distinct population or ecotype, can be difficult because of the hierarchical nature of genetic diversity with subclusters within clusters and so on (Cohan and Perry, 2007; Koeppel et al., 2008; Cohan and Kopac, 2011). In this work, using a population genetics approach we found that Mexican *E. coli* genetic diversity is hierarchically organized and that CCs are the clustering level corresponding to an ecotype. Previous studies have suggested the existence of ecotypes in *E. coli* (Souza et al., 1999; Gordon and Cowling, 2003; Jaureguy et al., 2008). In our study, we found that such CCs are associated to specific hosts or life styles and different recombination rates. Therefore, CCs might provide evidence for a mechanism of ecological isolation.

In line with the global distribution of *E. coli* (MLST database), we found CCs clearly associated to certain pathotypes and serotypes. In detail, our results suggest that pathogenesis is coupled to different mutation and recombination diversification dynamics. Accordingly, CCs associated to highly pathogenic *E. coli* strains (represented in this sample by ST11-Cplx, O157:H7 and O55:H7/H-) are coupled with higher recombination rates. In contrast to the latter, CCs as ST20-Cplx (O111) and ST29-Cplx (O26), previously associated with less pathogenic *E. coli* strains (Bielaszewska et al., 2008), are affiliated with intermediate and low rates of recombination respectively.

Since serotypes elicit a particular interaction between the host and its immune system (Orskov and Orskov, 1992; Whitfield, 2006), competition between strains within the host might be preventing an interaction among strains that belong to different serotypes. Consequently, serotype can act as an ecological barrier to recombination (Buckee et al., 2008).

We also found evidence of environmental ecotypes in the Mexican sample. In particular, ST86-Cplx includes air and water isolates from Mexico as well as samples from marine intertidal sediment in Hong Kong. As in the fresh water beach ET-1 B1 clade (Walk et al., 2007), natural selection is probably favoring the recombinant nature exhibited by ST86-Cplx (ABD ancestry). Furthermore, our results suggest that homologous recombination and positive selection within this CC are promoting the diversification of the *icd* gene, a gene involved in the transition from an anaerobic to an aerobic niche (Partridge et al., 2006).

In Mexico, we found that the domestication status and diet of animals could be delineating ecotypes in B2 and D groups since CCs associated to wild carnivorous (ST69-Cplx), granivorous and nectarivorous animals (ST28-Cplx) were recovered. Interestingly, these CCs are coupled to the highest recombination rates reported in this study, thus explaining the low levels of LD previously reported for wild animals in Mexico (Souza et al., 2002b). It is important to mention that at a global scale, ST69-Cplx or the *E. coli* clonal group A (CGA) is considered one of the major CCs associated with urinary tract infections (Tartof et al., 2005; Johnson et al., 2006) and an emergent, antimicrobial drug-resistant extra-intestinal pathogen (Johnson et al., 2011). Thus our findings reinforce the idea that wild animals could be a reservoir of antibiotic resistance as previously suggested (Allen et al., 2010). Nevertheless, more detailed studies focused on this human health relevant CC should be performed.

The existence of environmental and wild animal-associated CCs coupled with extreme recombination levels, supports evidence that the evolution of *E. coli* can also occur via recombination linked to adaptive traits other than pathogenesis, such as environment, commensalism or specific host diet.

On the other hand, we found CCs associated to different pathotypes, serotypes, host types and environments such as ST10-Cplx. This CC contains the most common ST (ST10) recovered both in Mexico and at a global scale (Fig. 2; Turner et al., 2006; Lau et al., 2008; Okeke et al., 2010). This worldwide prevalence could suggest some selective advantage, probably coupled to the "generalist" faculty of this CC. Since it has been proposed that intermittent host change can exert decreasing population sizes (Fraser et al., 2007), the role of genetic drift should be also taken into account in further studies.

Considering that any given phylogenetic group consists of many CCs, the status of A strains has been questioned and should be better thought of as a CC (Gordon et al., 2008). However, in this study, the presence of another CC (ST165-Cplx, clustering mainly EIEC human isolates) in addition to ST10-Cplx, confirms the phylogenetic status of group A. Moreover, on a global scale, group A is comprised of additional CCs, as reported in the MLST database (ST46-Cplx, ST168-Cplx, ST184-Cplx, ST226-Cplx).

All the above suggest that we can describe the evolutionary dynamics of our sample by an epidemic population structure, where adaptive CCs are superimposed on a background population made up of recombinant singleton STs. To elucidate in detail the evolutionary dynamics of each CC, consideration of other parameters (besides recombination and mutation frequencies and ecological associations) such as changes in population size and the adaptive value of genetic diversity, is essential. This will facilitate an understanding of whether genetic drift and/or natural selection are responsible for the maintenance and evolution of such CCs.

4.4. Geographic structure and epidemiological surveillance

The genetic structure of *E. coli* Mexican strains is difficult to explain according to geographic barriers. This is because the life

history of the species shows that migration is an important component of its ecology (Kaper et al., 2004; van Elsas et al., 2011).

It has been suggested that due to the high migration rates maintained by *E. coli*, a large proportion of the global diversity can be detected locally, pointing towards the presence of a homogenised global pool (Caugant et al., 1981; Whittam et al., 1983; Maynard-Smith et al., 2000). Nevertheless, a variable prevalence of phylogenetic groups among different hosts and geographic regions (Tenaillon et al., 2010) could suggest that there is a slight degree of ecological structure at large geographical scales.

Likewise, it is possible that each region has a different immunological context. Hence, this explains why *E. coli* serotypes seem to be less virulent in Mexico than in Europe and the USA. For example, obligatory pathogens such as EHEC O157:H7 belonging to E lineage are responsible for hemorrhagic colitis and haemolytic-uremic syndrome in those countries (Nataro and Kaper, 1998; Watanabe et al., 1996). The low recovery and the non-epidemic potential of lineage E in Mexico (this work; Paniagua et al., 2007) could be related to the acquisition of antibodies against serogroup O157 or related *E. coli* lipopolysaccharides during infancy (Navarro et al., 2003). Likewise, it has been shown that Mexican children develop high levels of antibodies against the main EPEC virulence factors via maternal colostrum, supporting the importance of breastfeeding in areas of the world where bacterial diarrhoeagenic disease is endemic (Parissi-Crivelli et al., 2000).

In this context, this work emphasizes the importance of investigations that explore microevolutionary processes at a local scale. Such studies can provide valuable information addressing the details of the evolutionary scenario and genetic reservoirs underlying the emergence of CCs with epidemiological relevance.

5. Conclusions

In this study we investigated the genetic diversification mechanisms of Mexican non-outbreak related *E. coli* strains isolated from a host wide range. Overall, levels of genetic diversity harboured by MLST genes are explained principally by homologous recombination rather than by point mutation. Furthermore, we observed that there is not a homogeneous rate of recombination and mutation, but that different rates emerge at different clustering levels of diversity such as phylogenetic group, lineage and CC.

These findings suggest an alternative explanation for the maintenance of the clear phylogenetic signal of *E. coli* despite the prevalence of homologous recombination.

It is important to point out that all the findings and proposals stated in this study were obtained from an MLST scheme and that a whole genome approach would give more accurate and detailed results. However, a recently published whole genome analysis reports similar rates and patterns of recombination for *E. coli*, supporting our findings (Didelot et al., 2012). Likewise, although there are different MLST schemes developed for typing *E. coli*, it is known that the same phylogenetic groups comprising the species are recovered whatever the MLST scheme used (Escobar-Páramo et al., 2004; Johnson et al, 2006; Wirth et al., 2006; Walk et al., 2007; Jaureguy et al., 2008; Gordon et al., 2008). According to this we decided to use the MLST scheme developed by Wirth et al. (2006) because it allowed us to carry out a more robust epidemiological analysis, since this database is larger and more diverse than the other databases.

Moreover despite the weak signal of substructure among some phylogenetic groups (B2, D and E), the presence of CCs (associated with particular life style and hosts as well as different genetic diversification processes) inside them might indicate differential genetic diversification processes within each phylogenetic group. Such pattern would suggest that all phylogenetic groups of *E. coli* within Mexico are sub-structured. Interestingly, recent whole genome studies (Leopold et al., 2011; Didelot et al., 2012) have demonstrated that intra-lineage recombination is more frequent than inter-lineage genetic exchanges. These results not only supports the patterns of substructure and processes described in this study, but also suggests that it could be extended to the whole species.

Finally, it has been previously observed that using only typing methods can erroneously cluster *E. coli* genotypes (Gordon et al., 2008). Thus, we strongly recommend the combined use of population genetics and phylogenetic approaches in evolutionary and epidemiological studies of local *E. coli* samples.

Acknowledgments

We thank Jaime Gasca, Biól. José Luis Méndez, Laura Espinosa and Dra. Erika Aguirre for technical support during the development of this project. We thank Laboratorio de Genética Bacteriana at Facultad de Medicina, Universidad Nacional Autónoma de México for infrastructure support during the development of this project. We thank Dr. Xavier Didelot for his advice in the use of ClonalFrame software. We thank Christine Rooks, Santiago Ramírez, Eria Rebollar, Morena Avitia, René Cerritos, Rodolfo Salas and Alejandra Vázquez-Lobo for their constructive reviews of the manuscript. This paper is part of the doctoral research of the first author, who thanks the Doctorado en Ciencias Biomédicas (Universidad Nacional Autónoma de México) and CONACYT (Grant No. 169917) for financial support. The project was supported by Grant DGPA-UNAM PAPIIT IN219109.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2012.09.003.

References

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A., Carniel, E., 1999. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. Proc. Natl. Acad. Sci. USA 96, 14043–14048.

Allen, H.K., Donato, J., Wang, H.H., Cloud-Hansen, K.A., Davies, J., Handelsman, J., 2010. Call of the wild: antibiotic resistance genes in natural environments. Nat. Rev. Microbiol. 8, 251–259.

Bergholz, P.W., Noar, J.D., Buckely, D.H., 2011. Environmental patterns are imposed on the population structure of *Escherichia coli* after fecal deposition. Appl. Environ. Microbiol. 77, 129–211.

Bielaszewska, M., Middendorf, B., Köck, R., Friedrich, A.W., Fruth, A., Karch, H., Schmidt, M.A., Mellmann, A., 2008. Shiga toxin-negative attaching and effacing *Escherichia coli*: distinct clinical associations with bacterial phylogeny and virulence traits and inferred in-host pathogen evolution. Clin. Infect. Dis. 47, 208–217.

Buckee, C.O., Jolley, K.A., Recker, M., Penman, B., Kriz, P., Gupta, S., Maiden, M.C.J., 2008. Role of selection in the emergence of lineages and the evolution of virulence in *Neisseria meningitidis*. Proc. Natl. Acad. Sci. USA 105, 15082–15087.

Caugant, D.A., Levin, B.R., Selander, R.K., 1981. Genetic diversity and temporal variation in the *E. coli* population of a human host. Genetics 98, 467–490.

Clermont, O., Bonacorsi, S., Bingen, E., 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic group. Appl. Environ. Microbiol. 66, 4555–4558.

Clermont, O., Olier, M., Hoede, C., Diancourt, L., Brisse, S., Keroudean, M., Glodt, J., et al., 2011. Animal and human pathogenic *Escherichia coli* strains share common genetic backgrounds. Infect. Genet. Evol. 11, 654–662.

Cohan, F.M., Kopac, S.M., 2011. Microbial genomics: *E. coli* relatives out of doors and out of body. Curr. Biol. 21, R587–R589.

Cohan, F.M., Perry, E.B., 2007. A systematics for discovering the fundamental units of bacterial diversity. Curr. Biol. 17, R373–R386.

Denamur, E., Picard, B., Tenaillon, O., 2010. Population genetics of pathogenic *Escherichia coli*. In: Robinson, D.A., Falush, D., Feil, E.J. (Eds.), Bacterial Population Genetics in Infectious Diseases. Wiley-Blackwell, West Sussex, United Kingdom, pp. 269–286.

Didelot, X., Falush, D., 2007. Inference of bacterial microevolution using multilocus sequence data. Genetics 175, 1251–1266.

Didelot, X., Meric, G., Falush, D., Darling, A.E., 2012. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. BMC Genomics 13, 256.

Duriez, P., Clermont, O., Bonacorsi, S., Bingen, E., Chaventré, A., Elion, J., Picard, B., Denamur, E., 2001. Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically distinct human populations. Microbiology 147, 1671–1676.

Escobar-Páramo, P., Clermont, O., Blanc-Potard, A.B., Bui, H., Le Bouguénec, C., Denamur, E., 2004. A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. Mol. Biol. Evol. 21, 1085–1094.

Estrada-Garcia, T., Lopez-Saucedo, C., Thompson-Bonilla, R., et al., 2009. Association of diarrheagenic *Escherichia coli* pathotypes with infection and diarrhea among Mexican children and association of atypical enteropathogenic *E. coli* with acute diarrhea. J. Clin. Microbiol. 47, 93–98.

Evanno, G., Regnaut, S., Goudet, J., 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. Mol. Ecol. 14, 2611–2620.

Ewing, B., Hillier, L., Wendi, M.C., Green, P., 1998. Base-calling of automated sequencer traces using phred I. Accuracy assessment. Genome. Res. 8, 175–185.

Excoffier, L., Laval, G., Schneider, S., 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol. Bioinform. Online 1, 47–50.

Feil, E.J., 2004. Small change: keeping pace with microevolution. Nat. Rev. Microbiol. 2, 483–495.

Feil, E.J., 2010. Linkage, selection, and the clonal complex. In: Robinson, D.A., Falush, D., Feil, E.J. (Eds.), Bacterial Population Genetics in Infectious Diseases. Wiley-Blackwell, West Sussex, United Kingdom, pp. 19–35.

Feil, E.J., Holmes, E.C., Bessen, D.E., Chan, M.S., Day, N.P., Enright, M.C., Goldstein, R., Hood, D.W., Kalia, A., Moore, C.E., et al., 2001. Recombination within natural populations of pathogenic bacteria: short term empirical estimates and long-term phylogenetic consequences. Proc. Natl. Acad. Sci. USA 98, 182–187.

Feil, E.J., Li, B.C., Aanensen, D.M., Hanage, W.P., Spratt, B.G., 2004. EBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J. Bacteriol. 186, 1518–1530.

Fraser, C., Hanage, W.P., Spratt, B.G., 2007. Recombination and the nature of bacterial speciation. Science 315, 476–480.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Statist. Sci. 7, 457–472.

Gordon, D., Abajian, C., Green, P., 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8, 195–202.

Gordon, D.M., Clermont, O., Tolley, H., Denamur, E., 2008. Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the triplex method. Environ. Microbiol. 10, 2484–2496.

Gordon, D.M., Cowling, A., 2003. The distribution and genetic structure of *Escherichia coli* Australian vertebrates: host and geographic effects. Microbiology 12, 3575–3586.

Goullet, P., Picard, B., 1989. Comparative electrophoretic polymorphism of esterases and other enzymes in *Escherichia coli*. J. Gen. Microbiol. 135, 135–143.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41, 95–98.

Haubold, B., Hudson, R.R., 2000. LIAN 3.0: detecting linkage disequilibrium in multilocus data. Bioinformatics 16, 847–849.

Herzer, P., Inouye, S., Inouye, M., Whittam, T.S., 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. J. Bacteriol. 172, 6175–6181.

Holmes, E.C., Urwin, R., Maiden, M.C., 1999. The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. Mol. Biol. Evol. 16, 741–749.

Jaureguy, F., Landraud, L., Passet, V., Diancourt, L., Frapy, E., Guigon, G., Carbonnelle, E., Lortholary, O., Clermont, O., Denamur, E., et al., 2008. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. BMC Genomics 9, 560–573.

Johnson, J.R., Menard, M.E., Lauderdale, T.-L., Kosmidis, C.H., Gordon, D., Collignon, P., Maslow, J.N., Tambic-Andrasevic, A., Kuskowski, M.A.Trans-Global Initiative for Antimicrobial Resistance Analysis Investigators, 2011. Global distribution and epidemiologic associations of *Escherichia coli* clonal group A, 1998–2007. Emerg. Infect. Dis. 17, 2001–2009.

Johnson, J.R., Owens, K.L., Clabots, C.R., Weissman, S.J., Cannon, S.B., 2006. Phylogenetic relationships among clonal groups of extraintestinal pathogenic *Escherichia coli* as assessed by multi-locus sequence analysis. Microbes Infect. 8, 1702–1713.

Kaper, J.B., Nataro, J.P., Mobley, H.L., 2004. Pathogenic *Escherichia coli*. Nat. Rev. Microbiol. 2, 123–140.

Koeppel, A., Perry, E.B., Sikorski, J., Krizanc, D., Warner, W.A., Ward, D.M., Rooney, A.P., Brambilla, E., Connor, N., Ratcliff, R.M., et al., 2008. Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. Proc. Natl. Acad. Sci. USA 105, 2504–2509.

Lau, S.H., Reddy, S., Cheesbrough, J., Bolton, F.J., Willshaw, G., Cheasty, T., Fox, A.J., Upton, M., 2008. Major uropathogenic *Escherichia coli* strain isolated in the northwest of England identified by multilocus sequence typing. J. Clin. Microbiol. 46, 1076–1080.

Le Gall, T., Clermont, O., Gouriou, S., Picard, B., Nassif, X., et al., 2007. Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains. Mol. Biol. Evol. 24, 2373–2384.

Leopold, S.R., Sawyer, S.A., Whittam, T.S., Tarr, P.I., 2011. Obscured phylogeny and possible recombinational dormancy in *Escherichia coli*. BMC Evol. Biol. 11, 183–191.

Levin, B.R., 1981. Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. Genetics 99, 1–23.

López-Saucedo, C., Cerna, J.F., Villegas-Sepulveda, N., Thompson, R., Velazquez, F.R., Torres, J., Tarr, P.I., Estrada-García, T., 2003. Single multiplex polymerase chain reaction to detect diverse loci associated with diarrheagenic *Escherichia coli*. Emerg. Infect. Dis. 9, 127–131.

Luo, C., Walk, S.T., Gordon, D.M., Feldgarden, M., Tiedje, J.M., Konstantinidis, K.T., 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. Proc. Natl. Acad. Sci. USA 108, 7200–7205.

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B.G., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. USA 95, 3140–3145.

Maynard-Smith, J.M., Feil, E., Smith, N., 2000. Population structure and evolutionary dynamics of pathogenic bacteria. BioEssays 22, 1115–1122.

Maynard-Smith, J.M., Smith, N.H., O'Rourke, M., Spratt, B.G., 1993. How clonal are bacteria? Proc. Natl. Acad. Sci. USA 90, 4384–4388.

Nataro, J.P., Kaper, J.B., 1998. Diarrheagenic *Escherichia coli*. Clin. Microbiol. Rev. 11, 142–201.

Navarro, A., Eslava, C., Hernandez, U., Navarro-Henze, J.L., Aviles, M., Garcia-de la Torre, G., Cravioto, A., 2003. Antibody responses to *Escherichia coli* O157 and other lipopolysaccharides in healthy children and adults. Clin. Diagn. Lab. Immunol. 10, 797–801.

Nicklasson, M., Klena, J., Rodas, C., Bourgeois, A.L., Tores, O., et al., 2010. Enterotoxigenic *Escherichia coli* multilocus sequence types in Guatemala and Mexico. Emerg. Infect. Dis. 16, 143–146.

Okeke, I.N., Wallace-Gadsden, F., Simons, H.R., Matthews, N., Labar, A.S., Hwang, J., Wain, J., 2010. Multi-locus sequence typing of enteroaggregative *Escherichia coli* isolates from Nigerian children uncovers multiple lineages. PLoS One 5, e14093.

Orskov, F., Orskov, I., 1984. Serotyping of *Escherichia coli*. Methods Microbiol. 14, 43–112.

Orskov, F., Orskov, I., 1992. *Escherichia coli* serotyping and disease in man and animals. Can. J. Microbiol. 38, 699–704.

Paniagua, G.L., Monroy, E., García-González, O., Alonso, J., Negrete, E., Vaca, S., 2007. Two or more enteropathogens are associated with diarrhoea in Mexican children. Ann. Clin. Microbiol. Antimicrob. 6, 17–24.

Parissi-Crivelli, A., Parissi-Crivelli, J., Girón, J., 2000. Recognition of enteropathogenic *Escherichia coli* virulence determinants by human colostrum and serum antibodies. J. Clin. Microbiol. 38, 2696–2700.

Partridge, J.D., Scot, C., Tang, T., Poole, R.K., Green, J., 2006. *Escherichia coli* transcriptome dynamics during the transition from anaerobic to aerobic conditions. Journal of Biological Chemistry 281, 2786–27815.

Peek, A.S., Souza, V., Eguiarte, L.E., Gaut, B.S., 2001. The interaction of protein structure, selection, and recombination on the evolutio of the type-1 fimbrial major subunit (fimA) from *Escherichia coli*. J. Mol. Evol. 52, 193–204.

Perna, N.T., 2011. Genomics of *Escherichia* and *Shigella*. In: Wiedman, M., Zhang, W. (Eds.), Genomics of Foodborne Bacterial Pathogens. Springer, New York, USA, pp. 119–139.

Picard, B., Garcia, J.S., Gouriou, S., Duriez, P., Brahimi, N., Bingen, E., Elion, J., Denamur, E., 1999. The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. Infect. Immun. 67, 546–553.

Pond, S.L., Frost, S.D., 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21, 2531–2533.

Power, M.L., Littlefield-Wyer, J., Gordon, D.M., Veal, D.A., Slade, M.B., 2005. Phenotypic and genotypic characterization of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes. Environ. Microbiol. 7, 631–640.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Reid, S.D., Herbelin, C.J., Bumbaugh, A.C., Selander, R.K., Whittam, T.S., 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. Nature 406, 64–67.

Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X., Rozas, R., 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19, 2496–2497.

Sabarly, V., Bouvet, O., Glodt, J., Clermont, O., Skurnik, D., Diancourt, L., et al., 2011. The decoupling between genetic structure and metabolic phenotypes in *Escherichia coli* leads to continuous phenotypic diversity. J. Evol. Biol. 24, 1559–1571.

Sandner, L., Eguiarte, L.E., Navarro, A., Cravioto, A., Souza, V., 2001. The elements of the locus of enterocyte effacement in human and wild mammal isolates of *Escherichia coli*: evolution by assemblage or disruption? Microbiology 147, 3149–3158.

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

Selander, R.K., Caugant, D.A., Whittam, T.S., 1987. Genetic structure and variation in natural populations of *Escherichia coli*. In: Ingraham, J.L., Low, K.B., Magasanik, B., Schaechter, M., Umbarger, H.E. (Eds.), *Escherichia coli* and *Salmonella typhimurirum*: Cellular and Molecular Biology. ASM Press, Washington, DC, pp. 1625–1648.

Smith, N.H., Dale, J., Inwald, J., Palmer, S., Gordon, S.V., Hewinson, R.G., Maynard-Smith, J., 2003. The population structure of *Mycobacterium bovis* in Great Britain: clonal expansion. Proc. Natl Acad. Sci. USA 100, 5271–15275.

Souza, V., Castillo, A., Eguiarte, L.E., 2002a. The evolutionary ecology of *Escherichia coli*. Am. Sci. 90, 332–341.

Souza, V., Rocha, M., Valera, A., Eguiarte, L.E., 1999. Genetic structure of natural populations of *Escherichia coli* in wild hosts on different continents. Appl. Environ. Microbiol. 65, 3373–3385.

Souza, V., Travisano, M., Turner, P., Eguiarte, L.E., 2002b. Does experimental evolution reflect patterns in natural populations? Comparison of *E. coli* strains from long-term studies to those from wild isolates. Antoine van Leeuwenhoek. 81, 143–153.

Spratt, B., Hanage, W., Feil, E., 2001. The relative contributions of recombination and point mutation to the diversification of bacterial clones. Curr. Opin. Microbiol. 4, 602–606.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.

Suerbaum, S., Smith, J.M., Bapumia, K., Morelli, G., Smith, N.H., Kunstmann, E., Dyrek, I., Achtman, M., 1998. Free recombination within *Helicobacter pylori*. Proc. Natl. Acad. Sci. USA 95, 12619–12624.

Tartof, S.Y., Solberg, O.D., Manges, A.R., Riley, L.W., 2005. Analysis of a uropathogenic *Escherichia coli clonal* group by multilocus sequence typing. J. Clin. Microbiol. 43, 5860–5864.

Tenaillon, O., Skurnik, D., Picard, B., Denamur, E., 2010. The population genetics of commensal *Escherichia coli*. Nat. Rev. Microbiol. 8, 207–217.

Thompson, J., Higgins, D., Gibson, T., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl. Acids Res. 22, 4673–4680.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., et al., 2009. Organised genome dynamics in *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genetics 5, e1000344.

Turner, S.M., Chaudhuri, R.R., Jiang, Z.D., DuPont, H., Gyles, C., et al., 2006. Phylogenetic comparisons reveal multiple acquisitions of the toxin genes by enterotoxigenic *Escherichia coli* strains of different evolutionary lineages. J. Clin. Microbiol. 44, 4528–4536.

van Elsas, J.D., Semenov, A.V., Costa, R., Trevors, J.T., 2011. Survival of *Escherichia coli* in the environment: fundamental and public health aspects. ISME J. 5, 173–183.

Vos, M., Didelot, X., 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 3, 199–208.

Walk, S.T., Alm, E.W., Calhoun, L.M., Mladonicky, J.M., Whittam, T.S., 2007. Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. Environ. Microbiol. 9, 2274–2288.

Watanabe, H., Wada, A., Inagaki, Y., Itoh, K., Tamura, K., 1996. Outbreaks of enterohaemorrhagic *Escherichia coli* O157:H7 infection by two different genotype strains in Japan, 1996. Lancet 348, 831–832.

White, A.P., Sibley, K.A., Sibley, C.D., Wasmuth, J.D., Schaefer, R., Surette, M.G., Edge, T.A., Neumann, N.F., 2011. Intergenic sequence comparison of *Escherichia coli* isolates reveals lifestyle adaptations but not host specificity. Appl. Environ. Microbiol. 77, 7620–7632.

Whitfield, C., 2006. Biosynthesis and assembly of capsular polysaccharides in *Escherichia coli*. Annu. Rev. Biochem. 75, 39–68.

Whittam, T.S., Ochman, H., Selander, R.K., 1983. Geographic components of linkage disequilibrium in natural populations of *Escherichia coli*. Mol. Biol. Evol. 1, 67–83.

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L., Karch, H., Reeves, P., Maiden, M., Ochman, H., Achtman, M., 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol. Microbiol. 60, 1136–1151.

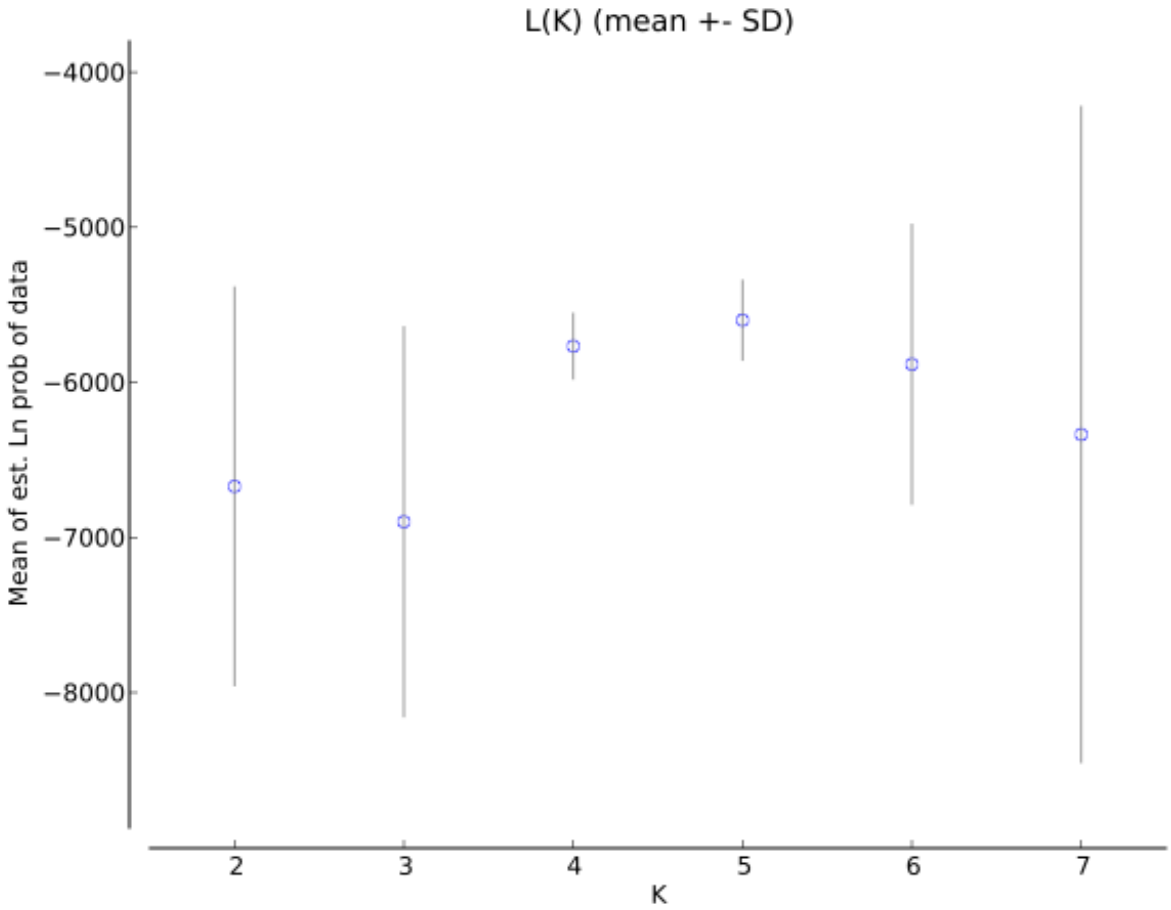

Fig. S1. Posterior probability from the mean estimated log-likelihood of *K* for sets 2-7

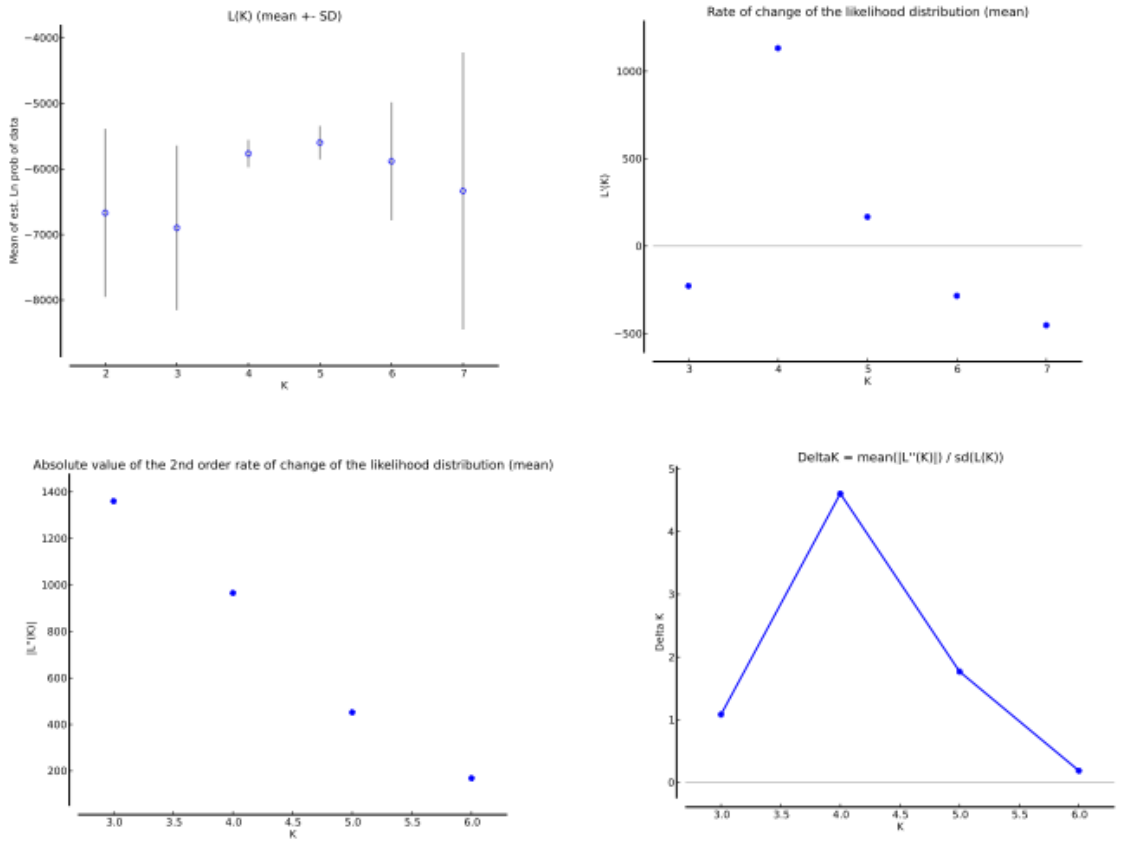

Fig. S2. Description of the four steps for the graphical method allowing detection of the true number of groups K inferred by STRUCTURE software (Evanno et al., 2005).

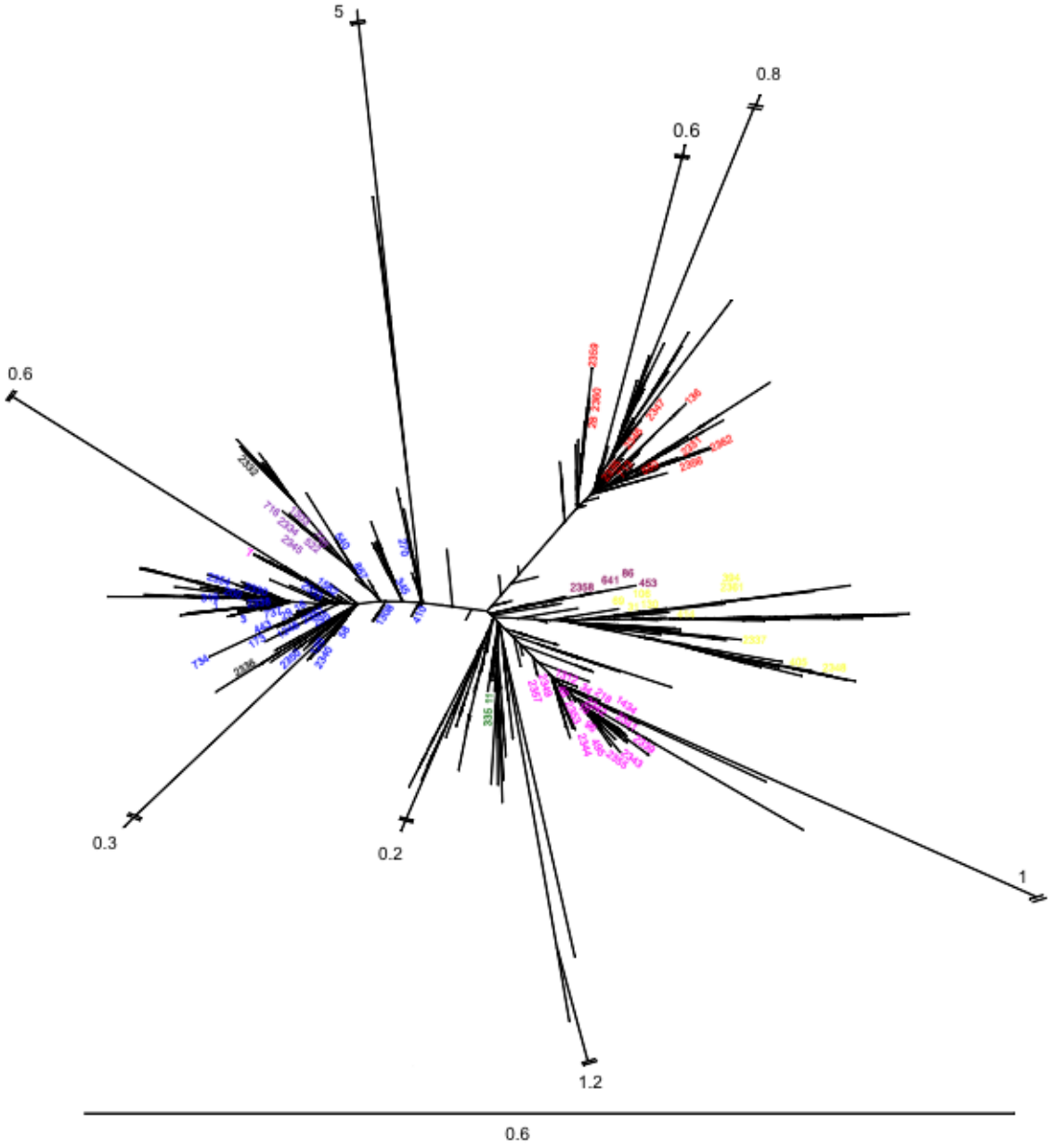

Fig. S3. *E. coli* MLST database maximum likelihood (ML) analysis. Mexican STs coloured according their belonging to each phylogenetic group or lineage detected by ClonalFrame: A (pink), B1 (blue), purple (A-ST522), magenta (B1-ST86), B2 (red), D (yellow), E (green), Ungrouped (black).

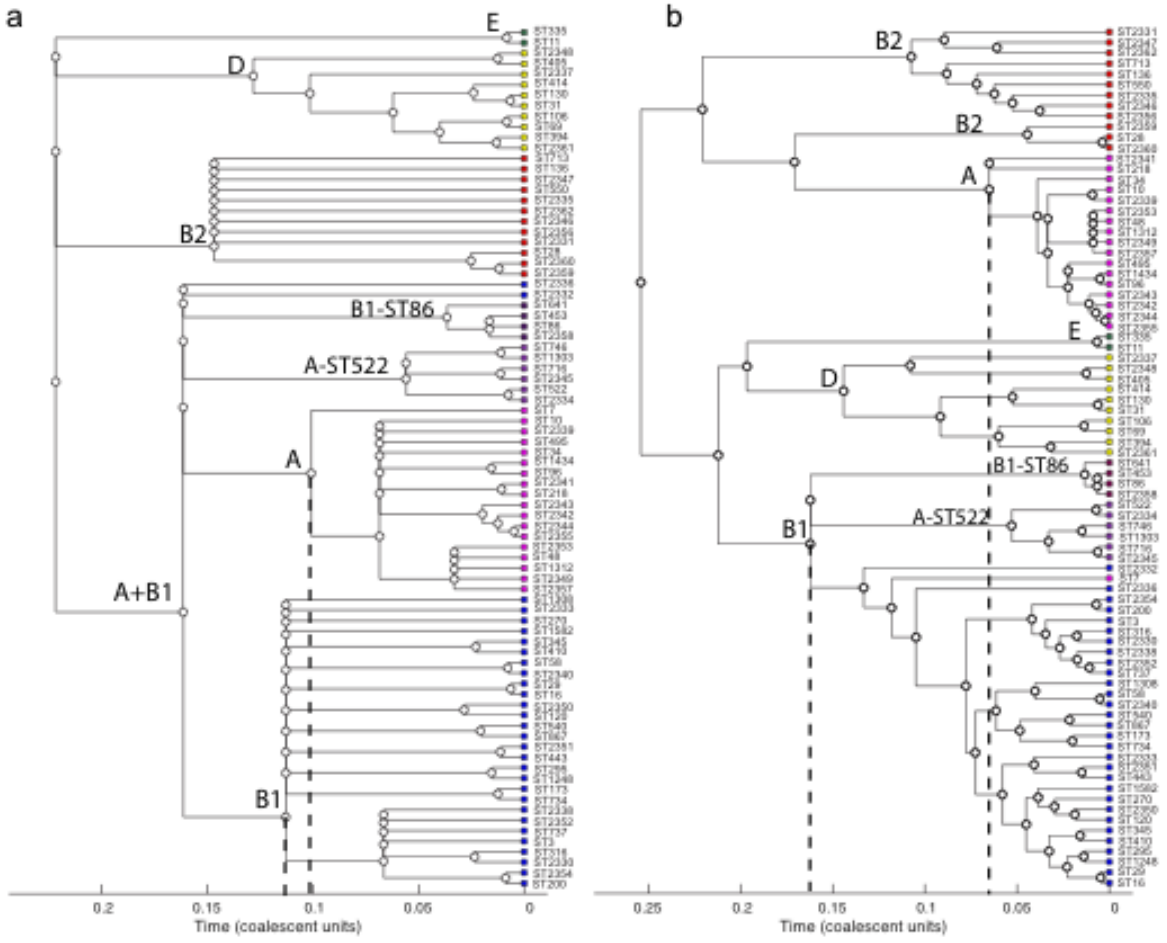

Fig. S4. 50% majority-rule consensus trees based on ClonalFrame output analysis for 82 unique STs considering (a) and ignoring (b) the role of homologous recombination in the estimation of the genealogy. The rulers indicate the time in coalescent units. Dashed black lines show the estimated time to the most recent common ancestors of A and B1 lineages.

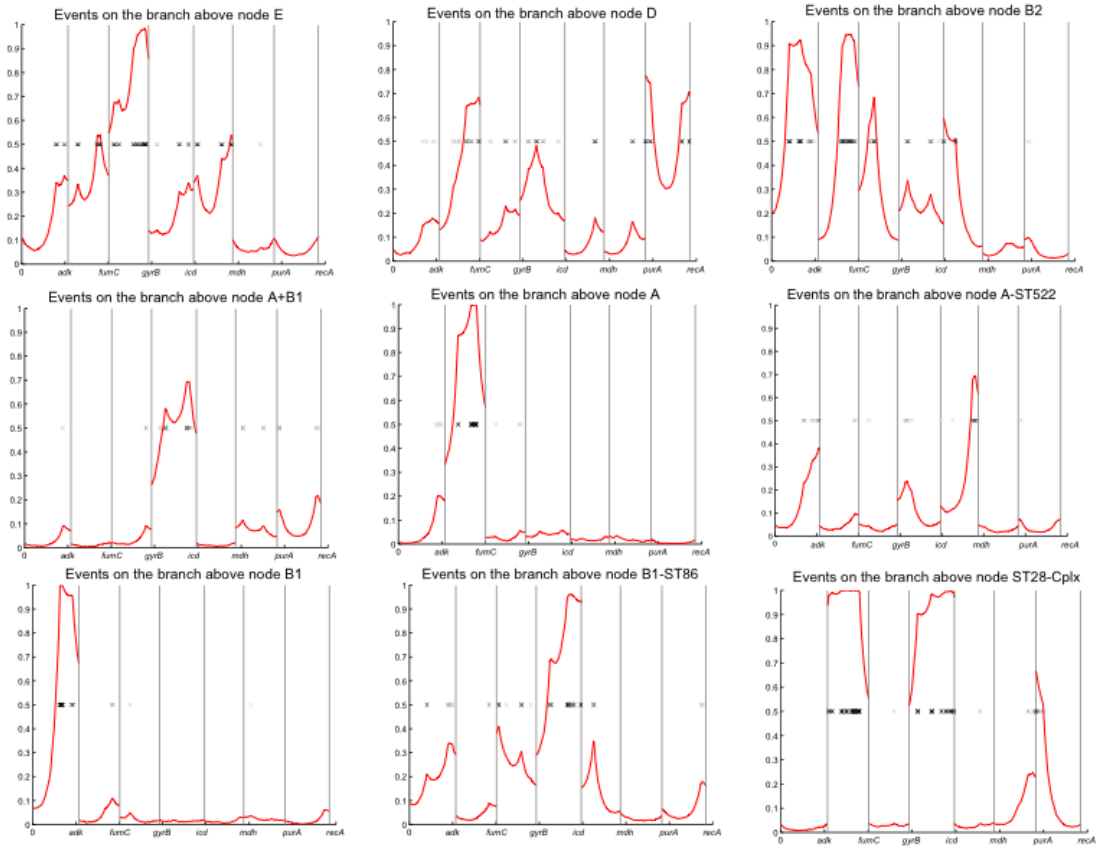

Fig. S5. Evolutionary events inferred by ClonalFrame on the branches of each Mexican *E. coli* phylogenetic group, lineages and CCs as indicated in the Figure 1. Each inferred substitution caused by either mutation or recombination is indicated by a cross, the intensity indicates the posterior probability for that substitution. The red line indicates the probability of recombination.

Table S1. Mexican *Escherichia coli* isolates analyzed. Traits of isolates: Clonal Complex, Clermont group, ancestral group (Structure), MLST allele profiles, ST, host, pathotype, serotype

Strain	Tamaño de Chromosoma Kb's	Asignación ClonalFrame	Grupo Ancestral [a]	CC [b]	GC [c]	ST [d]	Hospedero	Especie del hospedero	Estilo de vida	PT [e]	Serotipo	Enfermedad	Año	Ciudad	País	Laboratorio de origen
1	5392	B1	AxB1	Ninguno	A0	867	Animal domesticado	Perro *Canis lupus familiaris*	No patógeno	Ninguno	O?:H-	Ninguna	1994	México, DF.	México	Souza, V. Instituto de Ecología, UNAM
3	4766	B1	B1	Ninguno	B1	2350	Animal domesticado	Perro *Canis lupus familiaris*	No patógeno	Ninguno	O8:H49	Ninguna	1994	México, DF.	México	Souza, V. Instituto de Ecología, UNAM
19	4841	B1	AxB1	ST40 Cplx	B1	200	Animal silvestre	Ratón *Liomys pictus*	Patógeno	ETEC	O22:H28	Ninguna	1994	Chamela, Jal.	México	Souza, V. Instituto de Ecología, UNAM
33	5181	B2	B2	Ninguno	A0	136	Animal silvestre	Murciélago *Leptonycteris nivalis*	Patógeno	ETEC	O8:H-	Ninguna	1994	Tepoztlán, Mor.	México	Souza, V. Instituto de Ecología, UNAM
53	4813	A	A	ST10 Cplx	A1	10	Animal domesticado	Parrot *Aratinga ocularis*	No patógeno	Ninguno	O78:H2	Ninguna	1994	México, DF.	México	Souza, V. Instituto de Ecología, UNAM
55	5069	D	D	ST69 Cplx	D1	69	Animal silvestre	Aguila *Aquila chrysaetos*	No patógeno	Ninguno	O?:H6	Ninguna	1994	México, DF.	México	Souza, V. Instituto de Ecología, UNAM
68	5143	B1	B1	Ninguno	B1	737	Animal silvestre	Ratón *Sigmodon mascotensis*	No patógeno	Ninguno	037:H21	Ninguna	1994	Chamela, Jal.	México	Souza, V. Instituto de Ecología, UNAM
75	5195	B1	AxB1	ST40 Cplx	B1	2354	Animal silvestre	Ratón *Baiomys musculus*	Patógeno	ETEC	O103:H21	Ninguna	1994	Chamela, Jal.	México	Souza, V. Instituto de Ecología, UNAM
88	4739	A	A	ST10 Cplx	A1	218	Animal silvestre	Ratón *Habromys sp.*	Patógeno	ETEC	O103:H-	Ninguna	1994	Zacualpan, Edomex.	México	Souza, V. Instituto de Ecología, UNAM

90	5046	D	D	ST69 Cplx	D1	106	Animal silvestre	Coyote *Canis latrans*	No patógeno	Ninguno	O77:H18	Ninguna	1994	Omiltemi, Gro.	México	Souza, V. Instituto de Ecología, UNAM
95	4916	B2	B2xD	ST28 Cplx	B22	28	Animal silvestre	Ratón *Peromyscus megalops*	No patógeno	Ninguno	O?:H6	Ninguna	1994	Omiltemi, Gro.	México	Souza, V. Instituto de Ecología, UNAM
270	5096	B1	B1	Ninguno	B1	2338	Animal silvestre	Jaguar *Panthera onca*	Patógeno	EPEC	O159:H46	Ninguna	1994	Chiapas	México	Souza, V. Instituto de Ecología, UNAM
271	5072	B1	B1	Ninguno	B1	2338	Animal silvestre	Jaguar *Panthera onca*	Patógeno	EPEC	O159:H46	Ninguna	1994	Chiapas	México	Souza, V. Instituto de Ecología, UNAM
272	4883	D	D	ST31 Cplx	D1	130	Animal silvestre	Zorro *Urocyon cinereoargenteus*	No patógeno	Ninguno	O77:H18	Ninguna	1994	Chiapas	México	Souza, V. Instituto de Ecología, UNAM
288	4863	B1	AxB1	Ninguno	B1	2352	Animal silvestre	Ratón *Dipodomys merriami*	No patógeno	Ninguno	O132:H28	Ninguna	1994	Mapimí, Dgo.	México	Souza, V. Instituto de Ecología, UNAM
807	5131	Ungrouped	ABD	ST1250 Cplx	B1	2336	Animal domesticado	Caballo *Equus caballus*	No patógeno	Ninguno	O19:H-	Ninguna	1994	Calpan, Puebla	México	Souza, V. Instituto de Ecología, UNAM
814	4677	B2	B2	Ninguno	B23	713	Animal silvestre	Cacomixtle *Bassariscus astutus*	No patógeno	Ninguno	O145:H34	Ninguna	1994	México, DF.	México	Souza, V. Instituto de Ecología, UNAM
815	4713	B2	B2	Ninguno	B23	713	Animal silvestre	Cacomixtle *Bassariscus astutus*	No patógeno	Ninguno	O145:H34	Ninguna	1994	México, DF.	México	Souza, V. Instituto de Ecología, UNAM
825	4761	E	B2xD	ST11 Cplx	D1	335	Animal silvestre	Coyote *Canis latrans*	Patógeno	EPEC	O55:H7	Ninguna	1994	Mapimí, Dgo.	México	Souza, V. Instituto de Ecología, UNAM
830	5603	E	B2xD	ST11 Cplx	D1	335	Animal silvestre	Coyote *Canis latrans*	Patógeno	EPEC	O55:H7	Ninguna	1994	Mapimí, Dgo.	México	Souza, V. Instituto de Ecología, UNAM

1639	4695	B2	B2	Ninguno	B22	2356	Animal silvestre	Mono aullador *Alouatta palliata*	Patógeno	EPEC	O127:H?	Ninguna	1994	Los Tuxtlas, Ver.	México	Souza, V. Instituto de Ecología, UNAM
1684	4863	A	A	ST10 Cplx	A1	10	Animal silvestre	Murciélago *Choeronycteris mexicana*	No patógeno	Ninguno	O148:H32	Ninguna	1994	Tepoztlán, Mor.	México	Souza, V. Instituto de Ecología, UNAM
1728	4700	A	A	ST10 Cplx	A1	10	Animal domesticado	Gallo *Galus gallus*	No patógeno	Ninguno	O140:H32	Ninguna	1995	México, DF.	México	Souza, V. Instituto de Ecología, UNAM
1735	5725	B1	B1	Ninguno	B1	1248	Animal silvestre	Manatí *Trichechus manatus*	No patógeno	Ninguno	O?:H?	Ninguna	1995	Chetumal, Q.Roo	México	Souza, V. Instituto de Ecología, UNAM
1743	4775	A-ST522	AxB1	ST522 Cplx	A1	2334	Animal en cautiverio	Oso hormiguero *Tamandua mexicana*	Patógeno	ETEC	O148:H-	Ninguna	1995	Chiapas	México	Souza, V. Instituto de Ecología, UNAM
1744	4608	A	A	ST10 Cplx	A1	48	Animal en cautiverio	Oso hormiguero *Tamandua mexicana*	No patógeno	Ninguno	O65:H11	Ninguna	1995	Chiapas	México	Souza, V. Instituto de Ecología, UNAM
1937	4702	B1	AxB1	Ninguno	B1	1308	Animal en cautiverio	Tapir *Tapirus bairdii*	No patógeno	Ninguno	O8:H7	Ninguna	1995	Chiapas	México	Souza, V. Instituto de Ecología, UNAM
1940	4633	A-ST522	AxB1	ST522 Cplx	A1	2334	Animal en cautiverio	Armadillo *Dasypus sp.*	No patógeno	Ninguno	O166:H-	Ninguna	1995	Chiapas	México	Souza, V. Instituto de Ecología, UNAM
1967	4722	D	D	ST349 Cplx	D2	2337	Animal domesticado	Oveja *Ovis aries*	Patógeno	ETEC	O166:H15	Ninguna	1995	Isla Socorro, Revillagigedo	México	Souza, V. Instituto de Ecología, UNAM
2055	5107	B1	AxB1	Ninguno	B1	1582	Animal silvestre	Pecarí *Tayassu tajacu*	No patógeno	Ninguno	O8:H11	Ninguna	1995	Chamela Cuixmala	México	Souza, V. Instituto de Ecología, UNAM
2064	5025	A	A	ST10 Cplx	A1	10	Animal silvestre	Mapache *Procyon lotor*	No patógeno	Ninguno	O70:H11	Ninguna	1995	Chamela Cuixmala	México	Souza, V. Instituto de Ecología, UNAM

2065	4601	A	A	ST10 Cplx	A1	2357	Animal silvestre	Mapache *Procyon lotor*	No patógeno	Ninguno	O70:H11	Ninguna	1995	Chamela Cuixmala	México	Souza, V. Instituto de Ecología, UNAM
3442	4602	A-ST522	AxB1	Ninguno	A1	716	Animal silvestre	Murciélago *Artibeus sp.*	No patógeno	Ninguno	O142:H25	Ninguna	1995	Cueva Salitre Tetecalitla, Mor.	México	Souza, V. Instituto de Ecología, UNAM
3456	4693	B1-ST86	ADB	ST86 Cplx	A1	2358	Agua	Agua	Unknown	Unknown	O135:H25	Ninguna	1995	Cueva Salitre Tetecalitla, Mor.	México	Souza, V. Instituto de Ecología, UNAM
3463	4634	A	A	Ninguno	A0	1434	Agua	Agua	Unknown	Unknown	O18ab:H14	Ninguna	1995	Cueva Salitre Tetecalitla, Mor.	México	Souza, V. Instituto de Ecología, UNAM
3470	4928	B2	B2xD	ST28 Cplx	B23	2360	Animal silvestre	Murciélago *Artibeus sp.*	No patógeno	Ninguno	O?:H6	Ninguna	1995	Cueva Salitre Tetecalitla, Mor.	México	Souza, V. Instituto de Ecología, UNAM
3524	4851	A	A	ST10 Cplx	A1	34	Humano	Humano	No patógeno	Ninguno	O30:H10	Ninguna	1995	Morelos	México	Faculty of Medicine, UNAM
3528	4731	A	A	Ninguno	A0	495	Humano	Humano	Patógeno	EPEC	O126:H27	Ninguna	1995	Morelos	México	Faculty of Medicine, UNAM
3535	4737	B1	AxB1	ST155 Cplx	D2	58	Humano	Humano	No patógeno	Ninguno	O8:H30	Ninguna	1995	Morelos	México	Faculty of Medicine, UNAM
3566	4602	A	A	Ninguno	A0	1434	Humano	Humano	No patógeno	Ninguno	O86:H12	Ninguna	1995	Morelos	México	Faculty of Medicine, UNAM
3607	5488	B1	AxB1	ST40 Cplx	B1	200	Humano	Humano	Patógeno	EAEC	O11:H-	Ninguna	1995	Morelos	México	Faculty of Medicine, UNAM
3609	5427	B2	B2	ST14 Cplx	B23	550	Humano	Humano	Patógeno	EAEC	O75:H5	Ninguna	1995	Morelos	México	Faculty of Medicine, UNAM
3614	4907	B1-ST86	ABD	ST86 Cplx	B1	453	Humano	Humano	Patógeno	EAEC	OR:H16	Diarrea	1995	Morelos	México	Faculty of Medicine, UNAM
3617	4892	A	A	ST10 Cplx	A1	34	Humano	Humano	Patógeno	EAEC	O136:H33	Diarrea	1995	Morelos	México	Faculty of Medicine, UNAM

3620	4863	A	A	ST10 Cplx	A1	10	Humano	Humano	Patógeno	EAEC	O92:H-	Diarrea	1995	Morelos	México	Faculty of Medicine, UNAM
3621	4252	D	D	Ninguno	D1	414	Humano	Humano	Patógeno	EAEC	O7:H3	Diarrea	1995	Morelos	México	Faculty of Medicine, UNAM
3622	5646	D	D	ST31 Clpx	D1	31	Humano	Humano	Patógeno	EAEC	O15:H7	Diarrea	1995	Morelos	México	Faculty of Medicine, UNAM
3623	5224	A	A	ST10 Cplx	A1	10	Humano	Humano	Patógeno	EAEC	O111:H12	Diarrea	1995	Morelos	México	Faculty of Medicine, UNAM
3625	4948	B1	AxB1	ST40 Cplx	B1	200	Humano	Humano	Patógeno	EAEC	O7:H?	Diarrea	1995	Morelos	México	Faculty of Medicine, UNAM
3627	4962	A	A	ST165 Cplx	A0	2342	Humano	Humano	Patógeno	EAEC	O130:H27	Diarrea	1995	Morelos	México	Faculty of Medicine, UNAM
3628	5147	A	A	ST10 Cplx	A1	2339	Humano	Humano	Patógeno	EAEC	O33:H19	Diarrea	1995	Morelos	México	Faculty of Medicine, UNAM
3629	4863	B1-ST86	ABD	ST86 Cplx	A0	453	Humano	Humano	Patógeno	EAEC	O7:H10	Diarrea	1995	Morelos	México	Faculty of Medicine, UNAM
3630	4182	D	D	Ninguno	D1	414	Humano	Humano	Patógeno	EAEC	O44:H18	Diarrea	1995	Morelos	México	Faculty of Medicine, UNAM
3631	5284	A	A	ST10 Cplx	A1	10	Humano	Humano	Patógeno	EAEC	O15:H7	Diarrea	1995	Morelos	México	Faculty of Medicine, UNAM
3635	5177	B1	B1	ST20 Cplx	B1	3	Humano	Humano	Patógeno	EPEC	O114:H2	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3637	4757	B1	AxB1	Ninguno	B1	173	Humano	Humano	Patógeno	ETEC	O78:H12	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3641	4898	B1	B1	ST205 Cplx	B1	443	Humano	Humano	Patógeno	ETEC	O167:H5	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3646	4686	B1	B1	ST23 Cplx	A1	410	Humano	Humano	Patógeno	ETEC	O8:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM

3648	4761	B1	B1	ST23 Cplx	A1	410	Humano	Humano	Patógeno	ETEC	O8:H9	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3650	4776	B1	B1	ST29 Cplx	B1	29	Humano	Humano	Patógeno	EHEC	O26:H11	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3651	5696	B1	B1	ST20 Cplx	B1	3	Humano	Humano	Patógeno	EPEC	O111:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3652	4608	B1	B1	ST278 Cplx	B1	316	Humano	Humano	Patógeno	ETEC	O27:H7	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3653	4802	B2	B2	Ninguno	B23	2362	Humano	Humano	Patógeno	EHEC	O125ac:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3655	5369	B1	AxB1	Ninguno	B1	734	Humano	Humano	Patógeno	ETEC	O78:H12	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3657	4697	A	A	ST10 Cplx	A1	10	Suelo	Suelo	Patógeno	EIEC	O135:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3658	5698	B1	AxB1	Ninguno	B1	2330	Suelo	Suelo	Unknown	Unknown	O34:H40	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3659	5606	B1	B1	ST20 Cplx	B1	3	Humano	Humano	Patógeno	EPEC	O111:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3662	4007	B2	B2	Ninguno	B23	2331	Humano	Humano	Patógeno	EIEC	O164:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3663	4927	B1	B1	ST205 Cplx	B1	443	Humano	Humano	Patógeno	EIEC	O167:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM

3664	4792	A	A	ST165 Cplx	A0	2343	Humano	Humano	Patógeno	EIEC	O152:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3665	4688	A	A	ST165 Cplx	A0	2355	Humano	Humano	Patógeno	EIEC	O112:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3673	4892	A	A	ST10 Cplx	A0	34	Humano	Humano	Patógeno	EIEC	O136:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3677	4648	A	A	ST165 Cplx	A1	2344	Humano	Humano	Patógeno	EIEC	O112:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3681	4533	B1	B1	ST278 Cplx	B1	316	Humano	Humano	Patógeno	EHEC	O157:H7	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3682	5316	E	B2xD	ST11 Clpx	D1	11	Humano	Humano	Patógeno	EHEC	O157:H7	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3683	4761	B1	AxB1	ST155 Cplx	A0	58	Humano	Humano	Patógeno	ETEC	O20:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3687	4397	A	A	ST10 Cplx	A1	2353	Humano	Humano	Patógeno	ETEC	O6:H16	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3691	6097	D	D	ST405 Cplx	D2	2348	Humano	Humano	Patógeno	EHEC	O4:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3693	5011	B2	B2	Ninguno	B23	2347	Humano	Humano	Patógeno	EPEC	O55:H6	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3694	4911	B1	B1	Ninguno	B1	2333	Suelo	Suelo	Unknown	Unknown	O170:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM

3697	4007	B2	B2	Ninguno	B23	2331	Humano	Humano	Patógeno	EHEC	O145:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3698	4158	D	D	Ninguno	D1	414	Humano	Humano	Patógeno	EPEC	O44:H18	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3700	4693	B2	B2	Ninguno	B23	2346	Humano	Humano	Patógeno	ETEC	O78:H12	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3702	4722	B2	B2	Ninguno	B23	2346	Humano	Humano	Patógeno	EHEC	O26:H11	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3707	5586	B1	B1	ST20 Cplx	B1	3	Humano	Humano	Patógeno	EPEC	O111ab:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3711	5063	A	A	Ninguno	A0	96	Humano	Humano	Patógeno	ETEC	O25:H42	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3712	5663	B1	AxB1	Ninguno	B1	2330	Suelo	Suelo	Unknown	Unknown	O34:H40	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3713	5348	D	D	ST405 Cplx	D2	405	Humano	Humano	Patógeno	ETEC	O6:H16	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3715	4714	A	A	Ninguno	A1	1312	Humano	Humano	Patógeno	ETEC	O25:H-	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3716	5136	B1	B1	ST205 Cplx	A0	443	Humano	Humano	Patógeno	ETEC	O167:H5	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3719	4736	E	B2xD	ST11 Clpx	D1	11	Humano	Humano	Patógeno	EHEC	O157:H7	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM

3720	4473	A	A	ST10 Cplx	A1	10	Drenaje	Drenaje	Unknown	Unknown	O?:H32	Ninguna	1995	DF	México	Souza, V. Instituto de Ecología, UNAM
3824	4693	B2	B2	Ninguno	B23	2335	Animal silvestre	Murciélago *Artibeus jamaiciencis*	No patógeno	Ninguno	O?:H6	Ninguna	1996	Yucatán	México	Souza, V. Instituto de Ecología, UNAM
3842	4706	B1	AxB1	ST155 Cplx	B1	58	Lodo	Lodo	Unknown	Unknown	O120:H27	Ninguna	1996	DF	México	Souza, V. Instituto de Ecología, UNAM
3859	4885	A	A	ST10 Cplx	A1	2349	Lodo	Lodo	Unknown	Unknown	O?:H30	Ninguna	1996	DF	México	Souza, V. Instituto de Ecología, UNAM
3873	4668	A	A	ST10 Cplx	B22	48	Lodo	Lodo	Unknown	Unknown	O8:H9	Ninguna	1996	DF	México	Souza, V. Instituto de Ecología, UNAM
3874	4668	A-ST522	AxB1	ST522 Cplx	A0	522	Lodo	Lodo	Unknown	Unknown	O148:H?	Ninguna	1996	DF	México	Souza, V. Instituto de Ecología, UNAM
3876	4247	A-ST522	AxB1	ST10 Cplx	A1	2345	Lodo	Lodo	Unknown	Unknown	O?:H-	Ninguna	1996	DF	México	Souza, V. Instituto de Ecología, UNAM
3885	4899	A	A	ST10 Cplx	A1	10	Lodo	Lodo	Unknown	Unknown	O71:H27	Ninguna	1996	DF	México	Souza, V. Instituto de Ecología, UNAM
4129	4742	D	D	ST394 Cplx	D1	394	Aire de cueva	Aire	Unknown	Unknown	O73:H18	Ninguna	1996	Yucatán	México	Souza, V. Instituto de Ecología, UNAM
4130	4719	D	D	ST394 Cplx	B23	394	Aire de cueva	Aire	Unknown	Unknown	O73:H18	Ninguna	1996	Yucatán	México	Souza, V. Instituto de Ecología, UNAM
4132	4742	D	D	ST394 Cplx	D1	394	Aire de cueva	Aire	Unknown	Unknown	O73:H18	Ninguna	1996	Yucatán	México	Souza, V. Instituto de Ecología, UNAM

4135	4642	B1	B1	ST205 Cplx	B1	2351	Aire de cueva	Aire	Unknown	Unknown	O139:H28	Ninguna	1996	Yucatán	México	Souza, V. Instituto de Ecología, UNAM
4136	4751	D	D	ST394 Cplx	D1	394	Aire de cueva	Aire	Unknown	Unknown	O73:H18	Ninguna	1996	Yucatán	México	Souza, V. Instituto de Ecología, UNAM
4952	4685	B1	B1	ST29 Cplx	B1	29	Animal domesticado	Cerdo *Sus scrofa domesticus*	Patógeno	EPEC	O26:H-	Ninguna	1998	Edo México	México	Souza, V. Instituto de Ecología, UNAM
4953	5277	B1	B1	ST29 Cplx	B1	29	Animal domesticado	Cerdo *Sus scrofa domesticus*	Patógeno	EPEC	O26:H-	Ninguna	1998	Edo México	México	Souza, V. Instituto de Ecología, UNAM
4957	4693	A	A	ST10 Cplx	A1	10	Aire	Aire	Patógeno	EPEC	O28:H-	Ninguna	1998	DF	México	Souza, V. Instituto de Ecología, UNAM
4958	4438	B1	ABD? B1	ST270 Cplx	B1	270	Humano	Humano	Patógeno	EPEC	O28:H-	Ninguna	1998	DF	México	Souza, V. Instituto de Ecología, UNAM
4959	5658	E	B2xD	ST11 Cplx	D1	335	Humano	Humano	Patógeno	EPEC	O55:H-	Ninguna	1998	DF	México	Souza, V. Instituto de Ecología, UNAM
4962	5646	E	B2xD	ST11 Cplx	D1	335	Humano	Humano	Patógeno	EPEC	O55:H-	Ninguna	1998	DF	México	Souza, V. Instituto de Ecología, UNAM
4964	5574	E	B2xD	ST11 Cplx	D1	335	Humano	Humano	Patógeno	EPEC	O55:H-	Ninguna	1998	DF	México	Souza, V. Instituto de Ecología, UNAM
4976	5370	B1	B1	ST20 Cplx	B1	3	Humano	Humano	Patógeno	EPEC	O111ab:H-	Ninguna	1998	DF	México	Souza, V. Instituto de Ecología, UNAM
4993	4617	B1-ST86	ABD	ST86 Cplx	B1	86	Aire	Aire	Patógeno	EPEC	O127:H9	Ninguna	1998	DF	México	Souza, V. Instituto de Ecología, UNAM

5014	5007	B1	B1	ST29 Cplx	B1	29	Animal domesticado	Cerdo *Sus scrofa domesticus*	Patógeno	EHEC	O26:H-	Ninguna	1998	Edo México	México	Souza, V. Instituto de Ecología, UNAM
5020	5908	B1	B1	ST29 Cplx	B1	16	Animal domesticado	Oveja *Ovis aries*	Patógeno	EHEC	O111ac:H-	Ninguna	1998	Edo México	México	Souza, V. Instituto de Ecología, UNAM
5021	4784	A-ST522	AxB1	Ninguno	A0	1303	Humano	Humano	Patógeno	EHEC	0128:H-	Ninguna	1998	Morelos	México	Souza, V. Instituto de Ecología, UNAM
5040	5122	A-ST522	A	Ninguno	A0	746	Aire	Aire	Unknown	Unknown	O181:H?	Ninguna	1998	DF	México	Souza, V. Instituto de Ecología, UNAM
5058	5614	B1-ST86	ABD	ST86 Cplx	B1	641	Aire	Aire	Unknown	Unknown	O?:H11	Ninguna	1998	DF	México	Souza, V. Instituto de Ecología, UNAM
6879	5387	B1	AxB1	Ninguno	A0	540	Humano	Humano	No patógeno	Ninguno	O11:H?	Ninguna	2003	Edo México	México	Souza, V. Instituto de Ecología, UNAM
6891	5073	D	D	ST394 Cplx	D1	2361	Humano	Humano	No patógeno	Ninguno	O8:H18	Ninguna	2003	Edo México	México	Souza, V. Instituto de Ecología, UNAM
6908	4156	B2	ABD	Ninguno	B22	2359	Humano	Humano	No patógeno	Ninguno	O?:H9	Ninguna	2003	Edo México	México	Souza, V. Instituto de Ecología, UNAM
6909	4608	B1	B1	Ninguno	B1	345	Humano	Humano	No patógeno	Ninguno	O9:H11	Ninguna	2003	DF	México	Souza, V. Instituto de Ecología, UNAM
6918	4826	A	A	ST10 Cplx	A1	2341	Humano	Humano	No patógeno	Ninguno	O35:H9	Ninguna	2003	DF	México	Souza, V. Instituto de Ecología, UNAM
41724	4766	B1	B1	Ninguno	A0	295	Humano	Humano	Patógeno	ETEC	O29:H21	Ninguna	1995	Morelos	México	Faculty of Medicine, UNAM
43221	5298	A	AxB1	Ninguno	B1	7	Humano	Humano	Patógeno	ETEC	O29:H21	Ninguna	1995	Morelos	México	Faculty of Medicine, UNAM

48339	4883	B1	AxB1	ST155 Cplx	B1	2340	Humano	Humano	Patógeno	ETEC	O6:H16	Ninguna	1995	Morelos	México	Faculty of Medicine, UNAM
50417	4706	B1	B1	Ninguno	B1	120	Humano	Humano	Patógeno	ETEC	O?:H-	Diarrea	1995	Morelos	México	Faculty of Medicine, UNAM
63880	4782	Sin grupo	ABD	Ninguno	A0	2332	Humano	Humano	Patógeno	ETEC	O6:H16	Diarrea	1995	Morelos	México	Faculty of Medicine, UNAM

[a] Grupo ancestral arrojado por el análisis de recombinación mediante el algoritmo STRUCTURE.
[b] Complejo Clonal
[c] Grupo Clermont: Asignación a grupo filogenético según el método de Clermont et al., 2000.
[d] ST: SecuencioTipo (Sequence Type)

5. Artículo 2

"El tamaño del genoma en *Escherichia coli* no se encuentra determinado por su historia filogenética"

5.1. Resumen

Las poblaciones bacterianas presentan variación en el tamaño del genoma tanto a nivel inter-específico como intra-específico. Esta variación es clasificada principalmente en dos partes: el genoma central (que se refiere a los genes compartidos por todas las cepas) y el genoma flexible (referente a los genes compartidos por algunas cepas). Se ha propuesto que ésta variación guarda una clara relación con el nicho ecológico donde habitan tales bacterias. *Escherichia coli* es un modelo apropiado para estudiar cómo el tamaño del genoma evoluciona debido a su versatilidad ecológica. Se ha propuesto que el genoma de *E. coli* presenta un rango de variación en el tamaño del cromosoma que va de los 4.6 a los 5.6 Mb y que esta variación correlaciona con la estructura filogenética de la especie. Lo anterior implica que las cepas pertenecientes a los grupos A y B1 cuentan con genomas pequeños con respecto a las cepas de los grupos B2, D y E. En un trabajo previo, encontramos que el genoma central de una muestra de *E. coli* aislada de diversos hospederos presenta una gran dinámica evolutiva. El objetivo de este trabajo es evaluar la influencia de la estructura filogenética y el nicho ecológico en el dinamismo del genoma flexible, reflejado en la variación en el tamaño del cromosoma de esta muestra de *E. coli* previamente estudiada. Utilizando la técnica de electroforesis en campos pulsados, encontramos que el genoma flexible de esta muestra es dinámico lo cual se encuentra representado por el amplio rango del tamaño del cromosoma (2 Mb) el cual no se encuentra determinado ni por la estructura filogenética y ni por el nicho ecológico de la muestra.

Chromosome size variation in *Escherichia coli* is not linked either to phylogeny nor ecological niche

González-González, A[1]., Delgado, G[2]., Eguiarte, L[1]., Souza, V[1,*]

[1] Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Apartado Postal 70-275, México, D.F. 04510, México

[2] Departamento de Microbiología y Parasitología, Facultad de Medicina, Universidad Nacional Autónoma de México, México, D.F., México.

[*] Corresponding author: Souza, V.

Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Apartado Postal 70-275, México, D.F. 04510, México.

Tel. (+52 55 56229006) Fax (+52 55 56228995)

E-mail: souza@servidor.unam.mx

Running title: Chromosome size variation in *E. coli*

Abstract

Bacterial populations show both inter-specific and intra-specific variation in genome size. The genome size variation within a bacterial species is mainly classified in two parts: the core genome (genes shared by all strains) and the flexible genome (the genes shared by some but not all the strains). It is proposed this variation has a clear relationship with the ecological niche where bacteria distribute. *Escherichia coli* is a suitable model to study how chromosome size evolve because its ecological versatility. It has been previously suggested that *E. coli* genomes vary in size from 4.6 to 5.6 Mb and that this variation correlates well with the phylogenetic structure of species having A and B1 strains smaller genomes than B2, D and E strains. In a previous work we described the core genome of a wide *E. coli* host sample as highly evolutionary dynamic. The objective of this study is to evaluate the influence of the phylogenetic structure and ecological niche in the dynamism of flexible genome reflected in the chromosome size variation of this *E. coli* sample. Using pulsed field gel electrophoresis technique, we found dynamism of flexible genome represented by a wide range of chromosome size not determined neither by the phylogeny nor ecology of the sample.

1. Introduction

It is well known that bacterial populations might show an intra-specific variation in chromosome size. This variability in gene content between isolates of the same species has been observed as the result of distinct molecular events such as point mutation, homologous recombination, deletions, duplications, inversions, and horizontal gene transfer (Mira et al., 2001; Gregory and DeSalle, 2005). Moreover, it has been suggested a clear relationship between the ecological niche and the bacterial genome size where, free-living bacteria has the largest genome size, the facultative pathogens the intermediate and the obligate symbionts or pathogens the smallest ones (Ochman and Dávalos, 2006).

Escherichia coli is a suitable model to study how chromosome size evolve because its ecological versatility. i.e. the majority of *E. coli* live as a harmless commensal in the gut and other *E. coli* strains have the capacity to cause disease in humans and many animals (Savageau, 1983; Kaper et al., 2004; Tenaillon et al., 2010). It has been previously suggested that *E. coli* genomes vary in size from 4.6 to 5.6 Mb (Bergthorsson and Ochman, 1998) suggesting the phenotypic diversity among pathogenic and commensal *E. coli* mirrors their genome content. This size diversity indicates the presence of different amounts of strain-specific genetic information, which may represent up to 30% of the complete genome content. Furthermore, an open pangenome it has been determined for this enteric species suggesting *E. coli* still evolves by gene acquisition and diversification processes (Rasko et al., 2008; Touchon et al., 2009). Thus, it has been suggested that evolution of the *E. coli* variants results from ordered gene acquisition events and loss of genetic information together with DNA rearrangements and point mutations (Lawrence et al., 1998; Ochman et al., 2000; Dobrindt et al., 2010). Furthermore, it is assumed that genome size in *E. coli* correlates well with their

phylogenetic structure having A and B1 strains smaller genomes than B2, D and E strains (Bergthorsson and Ochman, 1998).

To date, there are around four hundred *E. coli* genome sequences in databases. Nevertheless, most of the sequences come from human pathogenic isolates and only few are commensal or pathogenic animal and environmental associated. In a previous study where a non-outbreak related host-wide sample of 128 *E. coli* isolates were analyzed (González-González et al., 2013), we report high levels of genetic diversity driven mainly by homologous recombination and to a lesser extent by point mutation. Furthermore, we found both intra and inter lineage recombination differentiating the "core" genome of this sample. Moreover we proposed clonal complexes as evolutionary units since they display different genetic diversification processes coupled to particular lifestyles. Considering the evolutionary dynamics of the core genome above described, the objective of this study is to evaluate the magnitude of the flexible genome dynamism of *E. coli* under ecological and phylogenetic terms. To achieve partially this objective, we measure the chromosome size of the same sample of non-outbreak related *E. coli* strains isolated from a wide range of host to assesses if these traits are either associated with the phylogenetic groups where the strains belong or with the different ecological niches where this species lives.

2. Material and Methods

2.1. Bacterial strains

A total of 128 *E. coli* strains encompassing a wide range of hosts both pathogenic and

non-pathogenic isolates were analyzed in this study. Such sample was previously assessed by González-González et al. (2013) and considers strains exclusively isolated from Mexican hosts deposited in the Instituto de Ecología Collection of *E. coli* (IECOL) and in the Enteric Pathogen Reference Laboratory Collection, Public Health Department, Facultad de Medicina both at the Universidad Nacional Autónoma de México.

Three groups of *E. coli* strains were analyzed: a) 39 strains isolated from healthy animals from which, 25 were isolated from wild animals [5 enteropathogenic *E. coli* (EPEC), 4 enterotoxigenic *E. coli* (ETEC) and 16 non-pathogenic strains], 4 strains from captive animals [1 ETEC and 3 non-pathogenic strains] and 10 strains from domesticated animals [2 enterohemorragic *E. coli* (EHEC), 2 EPEC, 1 ETEC and 5 non-pathogenic strains]; b) 67 human strains from which, 14 strains were isolated from faeces of diarrheagenic babies [12 enteroaggregative *E. coli* (EAEC) and 2 ETEC strains], 10 strains from the faeces of healthy babies [1 EPEC, 2 EAEC, 1 EHEC, 3 ETEC and 3 non-pathogenic isolates] and 43 strains from healthy adults [8 EHEC, 6 enteroinvasive *E. coli* (EIEC), 11 EPEC, 13 ETEC and 5 non-pathogenic strains]; c) 22 environmental strains, from which 9 were isolated from air (2 EPEC, 7 unknown), 4 from soil (1 EIEC, 3 unknown), 1 from drainpipe, 6 from mud and 2 isolates from water (Supplementary Table 1).

2.2. Preparation of DNA

Genomic DNA preparation was carried out with modifications from the method described by Liu *et al.* (1993) and Matushek *et al.* (1996). Bacteria were grown overnight in 5ml of Luria-Bertani (LB) broth and harvested by centrifugation. The bacterial pellet was washed twice in 500 μl of PIV (10 mM Tris [pH 8], 1 M NaCl) and

adjusted to 5 OD at 600 nm. Then, 0.2 ml of this cell suspension were embedded in 0.2 ml of 1.5 % low melting point agarose (Seaplaque GTG agarose) temperated to 37ºC. This mixture was distributed into plug molds (BioRad) and waited to solidify. Agarose plugs were incubated twice at 37ºC in 2 ml of lysis solution (1 M Tris [pH 8], 1 M NaCl, 0.5 M EDTA [pH 8], 0.5% sodium deoxicolate, 12.5% N-lauroyl-sarcosine, 5 µg/ml RNAse, 10 µg/ml lysozyme). First incubation was for 3 h and the second one overnight in new lysis solution at 50ºC in ESP (10 mM Tris HCl [pH7.4], 1 mM EDTA, 0.25% N-lauroyl-sarcosine, 0.1 mg/ml proteinase K [Sigma]. Finally, plugs were washed seven times (1 h each wash) in TE (Tris-HCl 10 mM [pH 8], 1 mM EDTA [pH 8]) and store into fresh TE buffer at 4ºC.

2.3. I-CeuI endonuclease digestion

Agarose plugs were equilibrated at 37ºC in 1X corresponding buffer enzyme during 1 hour. Then, they were suspended in 0.1 ml of digestion mix (4 U of I-*Ceu*I [New England Biolabs], 1X enzyme buffer, 1X BSA [bovine serum albumin]) and incubated at 37ºC overnight.

2.4. Pulsed-field gel electrophoresis (PFGE) and chromosome size estimation

Electrophoresis was performed in Bio-Rad CHEF Mapper electrophoresis system in a 1% agarose (Seakem Gold agarose) gel and 0.5X TBE (90 mM Tris, 90 mM boric acid, 2 mM EDTA, pH8) buffer at 12°C. To resolve fragments ranging from 4000 kb's to 1400 kb's a two program blocks was used with the following conditions: block 1 with a pulse time ramped from 20 min to 29 min 45 s over 60 h followed by a block 2 with a pulse time ramped from 2.31 s to 2 min 25 s over 7 h. The device was at 2 V/cm with 53º angle for the first block and at 6 v/cm with 60º angle for the second block. For

125 fragments in the range of 1400 kb's to 40 kb's three program blocks was performed as
126 follows: in block 1 a pulse time ramped from 50 s to 2 min over 21 h, in block 2 a
127 pulsed time ramped from 20 s to 1 min 20 s over 10 h and in block 3 a pulse time
128 ramped from 3 s to 12 s over 7 h. The device for three blocks was at 4 V/cm with 60º
129 angle.
130
131 *2.5. Visualization of bands and chromosome size measurements*
132 Gels were stained with ethidium bromide, photographed under UV light and then
133 analyzed in Bionumerics software (Bionumerics version 2.5; Applied Maths, Kortrijk,
134 Belgium). *Escherichia coli* K12 MG1665 and *Pseudomonas aeruginosa* PAO1
135 chromosomes were used as molecular size markers in order to obtain the chromosome
136 size of bacterial isolates. Size of each molecular marker band was obtained *in silico*
137 using the program MapDraw implemented in the DNASTAR Lasergene package.
138 Chromosome size was estimated by adding up the sizes of the restriction fragments
139 produced in these digests.
140
141 *2.6. Statistical analysis*
142 A factorial ANOVA analysis implemented in SPSS 15.0 software (SPSS, Inc, Chicago,
143 IL) was performed in order to know which strain trait reveals differences in average
144 chromosome size.
145
146 **3. Results**
147
148 *3.1. Variation in chromosome size among natural isolates of* E. coli *in Mexico*

Based on the cumulative sizes of the *I-Ceu*-I digested fragments for each strain, natural isolates of Mexican *E. coli* can differ by over 2 Mbp in the lengths of their chromosomes, with sizes ranging from 4,007 to 6,097 kb (Supplementary Table S1, Supplementary Table S2, Figure 1).

3.2.Chromosome size differences according to E. coli *traits in Mexico*

Considering isolate traits such as phylogenetic group/lineage belonging, host and life style together, the factorial analysis suggests that only phylogenetic group/lineage trait yields significant differences in chromosome size (F = 2.7, P = 0.019) (Figure 1 and Figure 2). It has been previously reported that Mexican *E. coli* strains are clustered into A, B1, B2, D, and E phylogenetic groups in addition to A-ST522 and B1-ST86 lineages (González-González et al., 2013). However, as phylogenetic groups and lineages display heterogeneous variances (Levene's Test, F = 3.83, P = 0.006) this significant difference (Brown-Forsythe's Test, F = 3.657, P = 0.005) is due to the fact that A chromosomes are significantly smaller than B1 (Table 1), while the other groups are equal in their chromosomes sizes (*post hoc* Games-Howell test, P = 0.023) (Figure 1).

4. Discussion

Genome size is a bacterial trait which evolution is also modulated by natural selection and genetic drift. Genetic mechanisms such as horizontal gene transfer (HGT) gene duplication and genetic erosion, promote an intra specific genome size variation which is reflected in the pangenome concept in order to describe the total genetic diversity harboured by a species (Tettelin et al, 2008).

4.1. Similar chromosome size variation of Mexican E. coli to another samples

Regarding *E. coli* species, genome size variation has been previously reported in a subset of the ECOR collection (Bergthorsson and Ochman, 1998) as well as in whole genomic surveys of commensal and pathogenic type strains (Rasko et al., 2008, Touchon et al., 2009). Contrasting our results with these previous works we found that Mexican *E. coli* sample displays a little more wide genome size variation range (0.5 Mbp more bigger). It is important to note that the pulse field gel electrophoresis (PFGE) method applied in this study yields consistent results with other PFGE analysis (Bergthorsson and Ochman, 1998) and with the genome sequencing technique (Supplementary Table S3) (Rasko et al., 2008, Touchon et al., 2009; Kuo et al., 2011). Unlike the studies mentioned above we found strains (IE-5020 and IE-3691) with a genome size around 6 Mbp in our sample. This observation is supported by the existence of strains in other collections presenting this similar genome size (Supplementary Table S3). Remarkably, some of these isolates belong to O111 serogroup, which together with O26 and O103 serogroups are considered atypical enteropathogenic *E. coli* strains (aEPEC) (Bielaszewska et al., 2008), frequently associated to food-borne outbreaks and emergent diarrhoea among children worldwide (Brooks et al., 2005; Trabulsi et al., 2002, Robins-Browne et al., 2004; Nguyen et al., 2006). Furthermore, these serotypes are also associated with Non-O157 EHECs, the highest clinical importance isolates in many countries. Remarkably, it has been previously suggested that these serotypes contain surprisingly large numbers of prophages and integrative elements associated to virulence factors than the typical O157:H7 EHEC strains (Ogura et a., 2009) thus explaining their largest genome sizes.

4.2. *Chromosome size of Mexican Escherichia coli not present a phylogenetic component*

It has been proposed that genome size in *E. coli* correlates well with their phylogenetic structure having A and B1 strains smaller genomes than B2, D and E strains (Bergthorsson and Ochman, 1998). In contrast, our study suggests that B1group has larger chromosomes than A group being the other phylogenetic groups not statistically different. This discrepancy could be explained by the way strains cluster considering the molecular marker and phylogenetic reconstruction used in each study. Thus in the Bergthorsson study, the phylogenetic relationships among the 35 ECOR strains analyzed were inferred from variation at 38 polymorphic loci as detected by enzyme electrophoresis and by Neighborgh Joining (NJ) approach (Herzer et al., 1990). In contrast, the phylogenetic structure of the Mexican strains was assessed by MLST approach and by a clustering algorithm that considers the extent of homologous recombination present in the nucleotide data (González-González et al., 2013). Thus, we detected the same groups than Herzer et al. (1990) plus two new lineages: A-ST522 and B1-ST86, which cluster within the A+B1 group. In addition when we perform the statistical analysis not considering the new lineages, B1 phylogenetic group also clusters larger chromosomes than A group *(post hoc* Games-Howell test, $P = 0.012$) and E strains have bigger chromosomes than B2 strains (*post hoc* Games-Howell test, $P = 0.042$). Furthermore, considering the Clermont group assignation method, B1 group has larger chromosomes than A *(post hoc* Games-Howell test, $P = 0.005$) and B2 phylogenetic groups *(post hoc* Games-Howell test, $P = 0.006$) Supplementary Table S4.

223 It has been proposed that group B2 is the most phylogenetically distinct and
224 homogeneous *E. coli* phylogenetic group (Le Gall et al., 2007). Most extraintestinal
225 pathogenic *E. coli* (ExPEC) strains belong to group B2, and most group B2 strains are
226 ExPEC with subdivisions such as uropathogenic *E. coli* (UPEC), neonatal meningitis *E.*
227 *coli* (NEMEC), sepsis associated *E. coli* (SEPEC), and avian pathogenic *E. coli*
228 (APEC). Genome sequencing suggests chromosome sizes of these *E. coli* subtypes
229 around 4.6 and 5.6 Mbp (Welch et al., 2002; Rasko et al., 2008; Toh et al., 2010). In our
230 study we found that B2 group has a chromosome size mean of 4.7 Mbp clustering
231 commensal and pathogenic strains isolated from human and wild animal faeces. This
232 finding agrees with the chromosome size determined for a B2 human commensal
233 bacterium (SE15, O150:H5) containing fewer virulence-related genes than the typical
234 extra intestinal strains (Toh et al., 2010). It has been pointed out that ExPEC virulence
235 factors identified in commensal B2 strains may facilitate colonization of the human gut
236 acting as fitness factors for commensal *E. coli* strains (Le Gall et al., 2007). Thus,
237 genome sequencing of the wild animal and human strains here studied could shed light
238 on this niche colonization processes.
239
240 On the other hand, it has been suggested that D and E strains have larger chromosomes
241 sizes than B1 and A groups (Bergthorsson and Ochman, 1998), however in this study
242 we did not find statistically differences between D strains and the other phylogenetic
243 groups. Regarding E strains, we found genome sizes consistently larger than those
244 present in the other groups, however significant differences could not be detected
245 because the reduced number of E strains recovered in this sample. Remarkably, this
246 sample bias is the same inconvenient in the Bergthorsson and Ochman study.
247 Furthermore, when we consider the genome size of the strains conforming the clonal

248 complexes (CC) detected in this sample (González-González, et al., 2013), we found
249 that ST11-Cplx, ST20-Cplx and ST29-Cplx (associated to highly pathogenic *E. coli*
250 strains with different homologous recombination rates and represented in this sample by
251 O55, O157 O111 and O26 serogroups) (Figure 3) harbour similar genome sizes. In
252 contrast, CCs such as ST86-Cplx, ST394-Cplx, ST69-Cplx, ST28-Cplx and ST10-Cplx
253 are more diverse in their genome content. Interestingly, these latter CCs represent
254 environmental, wild animal and generalist ecotypes thus suggesting not epidemic clones
255 undergo a more dynamic genome diversification than others.
256
257 4.3. *Chromosome size of Mexican Escherichia coli not present an ecological component*
258 It has been previously suggested that in bacteria, there is a correlation between the
259 genome size and their lifestyle being largest genome sizes commonly considered an
260 adaptation to changing environments (Moran, 2002; Konstantidinis et al., 2004;
261 Ochman and Davalos, 2006). Likewise, overwhelming evidence indicates that these
262 larger chromosomes are the product of horizontal gene transfer, which entails the
263 incorporation of genetic elements transferred from another organism directly into the
264 recipient genome. These genetic elements called genomic islands conform the flexible
265 genome and have an important role in the evolution of bacteria, influencing traits such
266 as antibiotic resistance, symbiosis and fitness, pathogenesis, and adaptation in general
267 (Hacker and Carniel, 2001; Dobrindt et al., 2004).
268
269 In our study, we did not find statistical differences in chromosome size either among the
270 different hosts (animal, human and environment) nor life styles (non-pathogenic and
271 pathogenic). Our genome size estimates of some non-pathogenic wild animal and
272 human strains (5 to 5.4 Mbp) agree with those present in non-pathogenic wild animals

of Australia (Supplementary Table S5). Recently, the genome size of a commensal strain isolated from porcine faeces belonging to phylogenetic group B1 (AI27) was determinated in 4.9 Mbp. This strain has complete genetic regions coding for ExPEC type virulence factors which are absent in most B1 strains that (Lee et a., 2012). Likewise, other commensal members of B1 group have genome sizes around 5.1 Mbp and harbour several genetic elements absent in pathogenic strains (Kim et al., 2012). All the above suggests the existence of several genome diversification processes associated to the adaptation to different niches.

4.4. Ecological strategies and copy number of rRNA operons in E. coli and

As previously mentioned, we estimated the genome size of each isolated by pulsed field gel electrophoresis cutting with the I-CeuI endonuclease. Due the specific restriction site of I-Ceu I (Marshall and Lemieux, 1992), the total number of bands visualized in each fingerprint was considered as an estimator of the total number of rRNA operons for each strain (Liu et al., 1993; Bergthorsson and Ochman, 1998). As genes encoding the 5S, 16S, and 23S ribosomal RNAs are clustered into a rRNA operon, it has been previously suggested that the number of ribosomal RNA operons is an ecological strategy for responding to environmental variation and fluctuations in resource availability (oligotrophs have few rRNA genes than copiotrophs) (Klappenbach et al., 2000; Stevenson and Schmidt, 2004).

The copy number of rRNA operons per bacterial genome varies from 1 to as many as 15. For example, species as *Mycoplasma pneumonie* and *Roseobacter denitrificans* have one rRNA operon, while *Bacillus subtilis* and *Clostridium paradoxum* have fifteen copies (Klappenbach et al., 2000). Although variation in operon numbers between different bacterial species haven been well documented, variations between strains of

298 the same species are considered less often. The variation in operon numbers does not
299 appear to be large, but the phenomenon is not restricted to a specific phylogenetic group
300 since operon variation between strains of the same species was detected for diverse
301 species as *Vibrio cholerae*, *Bacillus cereus* and *Yersinia pestis* (Acinas et al., 2004).
302 Regarding *E. coli,* it has been proposed seven copies per genome being the number of
303 *rrn* operons conserved among natural strains of this species (Bergthorsson and Ochman,
304 1998). In this work, the majority of the 128 Mexican isolates produced seven fragments
305 as the result of convergent evolution although the different ecological niches where *E.*
306 *coli* inhabits. We suggest this convergence of the *rrn* operon number is explained by the
307 flux between the primary and secondary habitats where *E. coli* inhabits. Nevertheless
308 we found three Mexican *E. coli* isolates having eight *rrn* operons (Supplementary
309 information). Two strains, one isolated from drainpipe (IE-3720) and the other from
310 mud (IE-3885), belong to phylogenetic group A and to ST10-Cplx, the most widespread
311 clonal complex. The third is a B1 strain isolated from a healthy human (IE-6879).
312 Remarkably, the environmental strains have a "standard" chromosome size (4.5 and 4.9
313 Mbp) in contrast to the human strain, which has a bigger chromosome size (5.4 Mbp).
314 This suggests that there is not a tight relationship between the chromosome architecture
315 and its size. Interestingly, no relationship between chromosome size and *rrn* operon
316 number was found, suggesting that is not necessary to have more *rrn* operons to
317 replicate more faster larger chromosomes as previously suggested (Fegatella et al.,
318 1998; Strehl et al., 1999; Lauro et al., 2009). Moreover, further studies have to be
319 performed in order to elucidate the genetic mechanisms that origin a new *rrn* operon
320 copy, and their physiological and ecological implications in terms of the capacity to
321 respond to environmental fluctuation and resource availability. Regarding strains with
322 six *rrn* operons, we have to perform more studies in order to validate this result.

Although it has been suggested that at least one rRNA operon can be deleted from *E. coli* without obvious deleterious effects (Ellwood and Nomura, 1980; Stevenson and Schmidt, 2004), it is clear that some of the seven operons has distinct responses to particular physiological and genetic perturbations (Condon et al., 1992; Stevenson and Schmidt, 2004).

5. Conclusion and perspectives

In this study we investigated the chromosome size of Mexican non-outbreak related *E. coli* strains isolated from a host wide range. Overall, we found natural isolates of Mexican *E. coli* can differ by over 2 Mbp in the lengths of their chromosomes. Likewise, our analysis suggest that phylogenetic group B1 contains the largest chromosomes of the sample meanwhile B2 strains have similar chromosome sizes than A strains in contrast to the previously suggested. Moreover, we did not find a significant association with the ecological niche and chromosome size. All the above suggests a dynamic flexible genome represented by the chromosome size range variation present in this *E. coli* sample. However, in order to shed light on the molecular processes and gene flow dynamics governing the evolution of *E. coli* genome size, we propose the isolates analyzed in this study should be sequenced because the wide range of ecological niches displayed.

6. Acknowledgements

We thank José Luis Méndez for technical support during the development of the project. We also thank Jorge Valdivia for the fruitful discussion regarding the ecological strategies and copy number of rRNA operons. This paper is part of the doctoral research of the first author, who thanks the Doctorado en Ciencias Biomédicas (Universidad

Nacional Autónoma de México) and CONACYT (Grant: 169917). The project was supported by grant DGPA-UNAM PAPIIT IN219109.

7. References

Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., Polz, M. 2004. Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. Journal of Bacteriology. 186, 2629-2635.

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

Bielaszewska, M., Middendorf, B., Köck, R., Friedrich, A.W., Fruth, A., Karch, H., Schmidt, M.A., Mellmann, A., 2008. Shiga toxin-negative attaching and effacing *Escherichia coli*: distinct clinical associations with bacterial phylogeny and virulence traits and inferred in-host pathogen evolution. Clin. Infect. Dis. 47, 208-217.

Brooks, J.T., Sowers, E., Wells, J., Greene, K., Griffin, P., Hoekstra, R., Strockbine, N. 2005. Non-O157 Shiga toxin-producing *Escherichia coli* infections in the United States, 1983-2002. J. Infect. Dis. 192, 1422-1429.

Condon, C., Philips, J., Fu, Z.H., Squires, C., Squires, C.L. 1992. Comparison of the expression of the seven ribosomal RNA operons in *Escherichia coli*. The EMBO Journal. 11, 4175-4185.

Dobrindt, U., Agerer, F., Michaelis, K., Janka, A. Buchrieser, C., et al. 2003. Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays. Proc. Natl. Acad. Sci. USA 185, 1831-1840.

Dobrindt, U., Hochhut, B., Hentschel, U., Hacker, J. 2004. Genomic islands in pathogenic and environmental microorganisms. Nature Review Microbiology 2, 414-424.

Dobrindt, U., Geddam, M., Krumbholz, G., Hacker, J. 2010. Genome dynamics and its impact on evolution of *Escherichia coli*. Med. Microbiol. Immunol. 199, 145-154.

Ellwood, M., Nomura, M. 1980. Deletion of a ribosomal ribonucleic acid operon in *Escherichia coli*. Journal of Bacteriology. 143, 1077-1080.

Fegatella, F., Lim, J., Kjelleberg, S., Cavicchioli, R. 1998. Implications of rRNA operon copy number and ribosome content in the marine oligotrophic ultramicrobacterium *Sphingomonas sp*. strain RB2256. Appl. Environ. Microbiol. 64, 4433-4438.

Gregory, T.R., DeSalle, R. 2005. The evolution of genome size in prokaryotes. In: The evolution of the genome (Ed. T. R. Gregory). Elsevier, San Diego. 631-640.

González-González, A., Sánchez-Reyes, L.L., Delgado, G., Eguiarte, L.E., Souza, V. 2013. Hierarchical clustering of genetic diversity associated to different levels of mutation and recombination in *Escherichia coli*: A study based on Mexican isolates. Infection, Genetics and Evolution. 13, 187-197.

Hacker, J., Carniel, E. 2001. Ecological fitness, genomic islands and bacterial pathogenicity: A Darwinian view of the evolution of microbes. EMBO reports. 2, 376-381.

Herzer, P., Inouye, S., Inouye, M., Whittam, T.S., 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. J. Bacteriol. 172, 6175-6181.

Kaper, J.B., Nataro, J.P., Mobley, H.L., 2004. Pathogenic *Escherichia coli*. Nat. Rev. Microbiol. 2, 123-140.

Klappenbach, J.A., Dunbar, J.M., Schmidt, Th.M. 2000. rRNA operon copy number reflects ecological strategies of bacteria. Applied and Environmental Microbiology. 66, 1328-1333.

Kim, M., Yi, H., Cho, Y., Jang, J., Hur, H., Chun, J. 2012. Draft genome sequences of Escherichia coli W26, an enteric strain isolated from cow faeces. Journal of Bacteriology. 194, 5149-5150.

Konstantinidis, K.T., Tiedje, J.M. 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc. Natl. Acad. Sci. USA. 101, 3160-3165.

Kuo, C.H., Ochman, H., Raghavan, R. 2011. The genomics of Escherichia coli and beyond. In: Population Genetics of Bacteria: a tribute to Thomas Whittam. Walk, T., Feng, P.C.H. (eds). ASM Press, Washington, DC. Pp. 31-42.

Lane, D.J. 1991. 16S/23S rRNA sequencing. In: Stackebrandt, E., and Goodfellow, M. (Eds.), Nucleid Acid Techniques in Bacterial Systematics. Chichester, UK: Wiley, pp. 115-175.

Lauro, et al. 2009. The genomic basis of trophic strategy in marine bacteria. Proc. Natl. Acad. Sci. USA. 106, 5527-5533.

Lawrence, J.G., Ochman, H. 1998. Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci. USA. 95, 9413-9417.

Lee, K., Cho, Y., Jang, J., Hur, H., Chun, J. 2012. Draft genome sequence of Escherichia coli AI27, a porcine isolate belonging to phylogenetic group B1. Journal of Bacteriology. 194, 6640-6641.

Le Gall, T., Clermont, O., Gouriou, S., Picard, B., Nassif, X., et al., 2007. Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains. Mol. Biol. Evol. 24, 2373-2384.

Liu, S. L., A. Hessel and K. E. Sanderson. 1993. Genomic mapping with I-Ceu I, an intron-encoded endonuclease specific for genes for ribosomal RNA, in *Salmonella spp.*, *Escherichia coli*, and other bacteria. Proc. Natl. Acad. Sci. USA. 90:6874–6878.

Marshall, P., Lemieux, C. 1992. The *I-Ceu* I, endonuclease recognizes a sequence of 19 base pairs and preferentially cleaves the coding strand of the Chlamydomonas moewusii chloroplast large subunit rRNA gene. Nucleic Acids Research 20, 6401-6407.

Matushek, M. G., M. J. Bonten and M. K. Hayden. 1996. Rapid preparation of bacterial DNA for pulsed-field gel electrophoresis. J. Clin. Microbiol. 34:2598–2600.

Mira, A., Ochman, H., Moran, N. 2001. Deletional bias and the evolution of bacterial genomes. Trends in Genetics. 17, 589-596.

454 Mira, A., Ochman, H., Moran, N.A. 2001. Deletional bias and the evolution of bacterial
455 genomes. Trends in Genetics. 17, 589-596.
456
457 Moran, N.A. 2002. Microbial minimalism: genome reduction in bacterial pathogens. Cell. 108,
458 583-586.
459
460 Nguyen R.N., Taylor, L.S., Tauschek, M., Robins-Browne, R.M. 2006. Atypical
461 enteropathogenic *Escherichia coli* infection and prolonged diarrhea in children. Emerg. Infect.
462 Dis. 12, 597-603.
463
464 Ochman, H., Lawrence, J., Groisman, E. 2000. Lateral gene transfer and the nature of bacterial
465 innovation.
466
467 Ochman, H., Davalos, L.M. 2006. The nature and dynamics of bacterial genomes. Science 311,
468 1730-1733.
469
470 Ogura, Y., Tadasuke, O., Iguchi, A., Toh, H., Asadulghani, Md., Oshima, K., et al. 2009.
471 Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157
472 enteroherrhagic *Escherichia coli*. Proc. Natl. Acad. Sci. USA 106, 17939-17944.
473
474 Rasko, D.A, Rosovitz, M.J., Myers, G.S., Mongodin, E.F., Fricke, W.F. Gajer, P., Crabtree, J.,
475 et al. 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E.*
476 *coli* commensal and pathogenic isolates. J. Bacteriol. 190, 6881-6893.
477
478 Robins-Browne, R.M., Bordun, A., Tuschek, M., et al. 2004. *Escherichia coli* and community-
479 acquired gastroenteritis, Melbourne, Australia. Emerg. Infect. Dis. 10, 1797-805.
480
481 Sambrook J, Fritsch, E.F., Maniatis, T. 1989. Molecular cloning: a laboratory manual. Cold
482 Spring Harbor Laboratory Press.

483 Savageau, M.A., 1983. *Escherichia coli* habitats, cell types, and molecular mechanisms of gene
484 control. Am. Nat. 122, 732-744.
485
486 SPSS for Windows (Computer program). Version 15.0. Chicago: SPSS Inc; 2007.
487
488 Stevenson, B.S., Schmidt, Th.M. 2004. Life history implications of rRNA gene copy number in
489 *Escherichia coli*. Applied and Environmental Microbiology. 70, 6670-6677.
490
491 Strehl, B., Holtzendorff, J., Partensky, F., Hess, W.R. 1999. A small and compact genome in the
492 marine cyanobacterium *Prochlorococcus marinus* CCMP 1375: lack of an intron in the gene for
493 tRNA (Leu) UAA and a single copy of the rRNA operon. FEMS Microbiology Letters 181,
494 261-266.
495
496 Tettlelin, H., Riley, D., Cattuto, C., Medini, D. 2008. Comparative genomics: the bacterial pan-
497 genome. Current Opinion in Microbiology. 12, 472-477.
498
499 Tenaillon, O., Skurnik, D., Picard, B., Denamur, E., 2010. The population genetics of
500 commensal *Escherichia coli*. Nat. Rev. Microbiol. 8, 207-217.
501
502 Toh, H., Oshima, K., Toyoda, A., Ogura, Y., et al. 2010. Complete genome sequence of the
503 wild-type commensal *Escherichia coli* strain SE15, belonging to phylogenetic group B2. J.
504 Bacteriol. 192, 1165-1166.
505

506 Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., et al., 2009. Organised
507 genome dynamics in *Escherichia coli* species results in highly diverse adaptive paths. PLoS
508 Genetics 5, e1000344.
509
510 Trabulsi, L.R., Keller, R., Gomes T.A.T. 2002. Typical and atypical enteropathogenic
511 *Escherichia coli*. Emerg. Infec. Dis. 175, 508-513.
512
513 Welch, R.A., Burland, V., Plunkett III, P.R., Roesch, P., Rasko, D., et al. 2002. Extensive
514 mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*.
515 Proc. Natl. Acad. Sci. USA 99, 17020-17024.
516
517 Zhang, W., QI, W.H., et al. 2006. Probing genomic diversity and evolution of Escherichia coli
518 O157 by single nucleotide polymorphisms. Genome Research 16, 757-767.

Supplementary information

Determination of the rRNA operons number

Due the specific restriction site of *I-Ceu* I endonuclease (Marshall and Lemieux, 1992), the total number of bands visualized in each fingerprint was considered as an estimator of the total number of rRNA operons for strain (Liu et al., 1993; Bergthorsson and Ochman, 1998). In order to corroborate the number of *rRNA* operons in cases where more than seven *I-Ceu I* fragments were yielded, a hybridization analysis of the *E. coli* 16S and 23S rRNA genes was carried out. Thus, PFGE gels with the digested atypical strains were transferred overnight to positively charged nylon membrane (Hybond ^{TM}-N^{+}, Amersham Pharmacia Biotech UK) by Southern Blotting method (Sambrook et al 1989). Hybridization analysis of the *E. coli* 16S and 23S rRNA genes was carried out using a DIG High Prime DNA labeling and detection starter kit II (Roche Diagnostics). PCR probes were obtained as follows: 16S rRNA gene PCR amplification was performed as described previously in Lane (1991). For 23S rRNA gene PCR mixture consisted of 10 ng of DNA template, 20 pmol of each PCR primer (Invitrogen), 1 x PCR buffer (Invitrogen), 1.5 mM $MgCl_2$ (Invitrogen), 1 U of Taq DNA polymerase (Invitrogen), and 0.2 mM of each deoxynucleoside triphosphates (Invitrogen). PCR amplification consisted of 10 min of denaturation at 94°C followed by 30 cycles of 1 min of denaturation at 94°C, 1 min of annealing at 58ºC and 1 min of extension at 72°C and finally a cycle of 10 min of extension at 72ºC on a MJ Research PTC-200 Peltier Thermal Cycler (MJ Research Inc, Watertown, Mass., USA). *E. coli* K12 MG1655 was used as template to both genes. Primers for the PCR amplification and fragment size of each gene are listed in Table 1. DNA probes (1 μg) were labeled using the DIG High Prime DNA according to the manufacturer's protocol. Membranes were pre-hybridized in

hybridization solution at 58 °C for 30min. Hybridization was carried out in a hybridization solution containing labeled DNA (25 ng ml^{-1}) at 58 °C overnight to both genes. After hybridization, membranes were washed twice in 2x SSC and 0·1 % SDS solution at 15-25 ºC by 5 min, and washed again, twice in 0·5x SSC and 0·1 % SDS solution at 68 °C by 12 min. Chemiluminescence detection with CSPD ready-to-use was done by exposure of membranes to Lumi-Film Chemiluminescent Detection Film (Roche Diagnostics, GMBH, Mannheim, Germany) for 1 hr.

Table 1. *rRNA* genes selected to hybridization analysis

Gene	Primer	Sequence 5'-3'	Fragment size	Reference
16S	27F 1492R	AGAGTTTGATCCCTCAG ACCTTGTTACGACTT	≈ 1450 pb	Lane, 1991
23S	23S4F 23S4R	ACTGGGGCGGTCTCCTC CATCGCTGCGCTTACACA	≈ 616 bp	Delgados et al., en revision

References:

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

Lane, D.J. 1991. 16S/23S rRNA sequencing. In: Stackebrandt, E., and Goodfellow, M. (Eds.), Nucleid Acid Techniques in Bacterial Systematics. Chichester, UK: Wiley, pp. 115-175.

Liu, S. L., A. Hessel and K. E. Sanderson. 1993. Genomic mapping with I-Ceu I, an intron-encoded endonuclease specific for genes for ribosomal RNA, in *Salmonella spp.*, *Escherichia coli*, and other bacteria. Proc. Natl. Acad. Sci. USA. 90:6874–6878.

Marshall, P., Lemieux, C. 1992. The *I-Ceu* I, endonuclease recognizes a sequence of 19 base pairs and preferentially cleaves the coding strand of the Chlamydomonas moewusii chloroplast large subunit rRNA gene. Nucleic Acids Research 20, 6401-6407.

Sambrook J, Fritsch, E.F., Maniatis, T. 1989. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press.

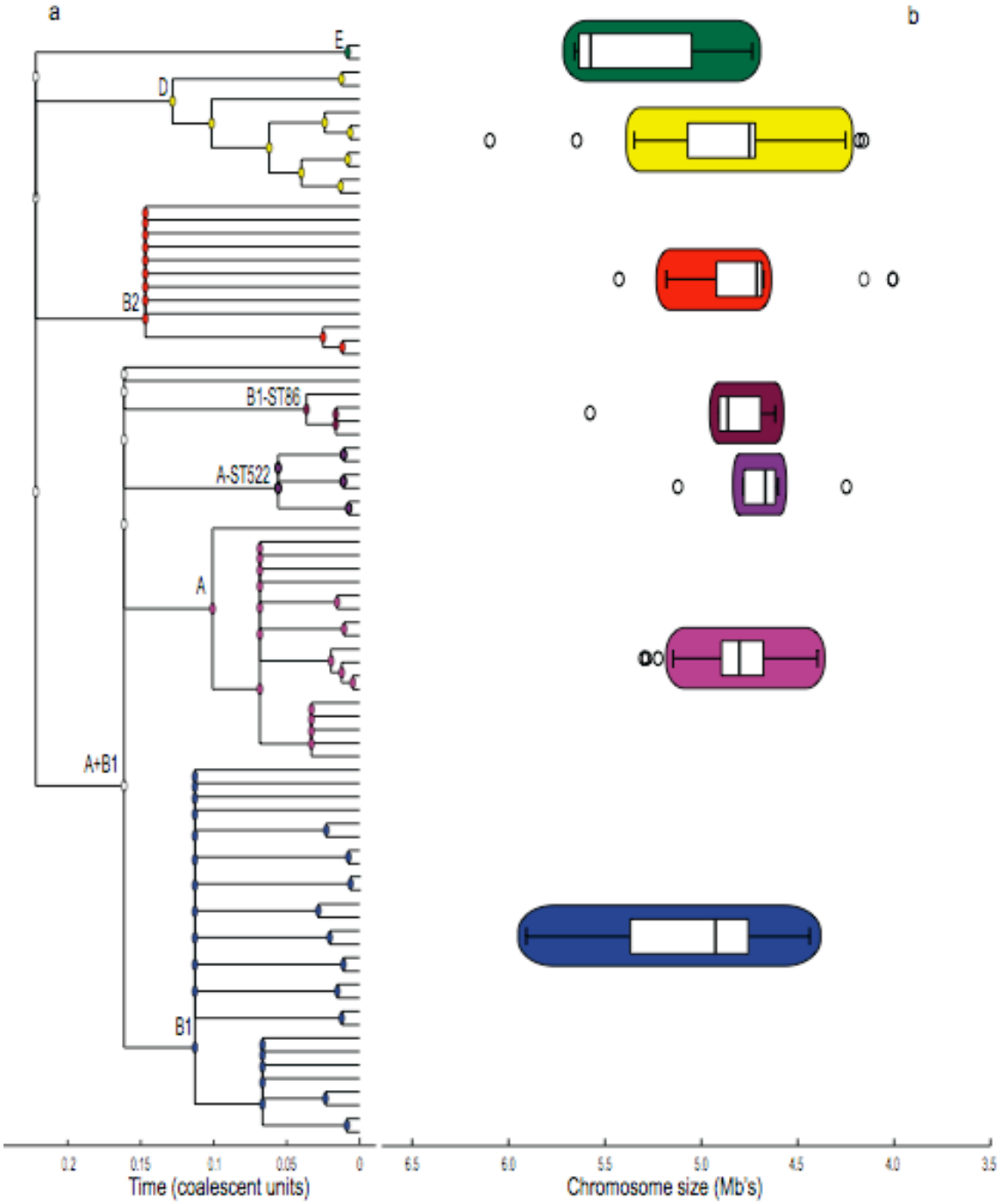

Figure 1 (a) Unrooted consensus tree displaying the relationship between the 82 STs of the Mexican *E. coli* sample at seven concatenated housekeeping genes (3423 nucleotides in total) (González-González et al., 2013). **(b)** Chromosome size of the Mexican *E. coli* phylogenetic groups in mega bases (Mbp).

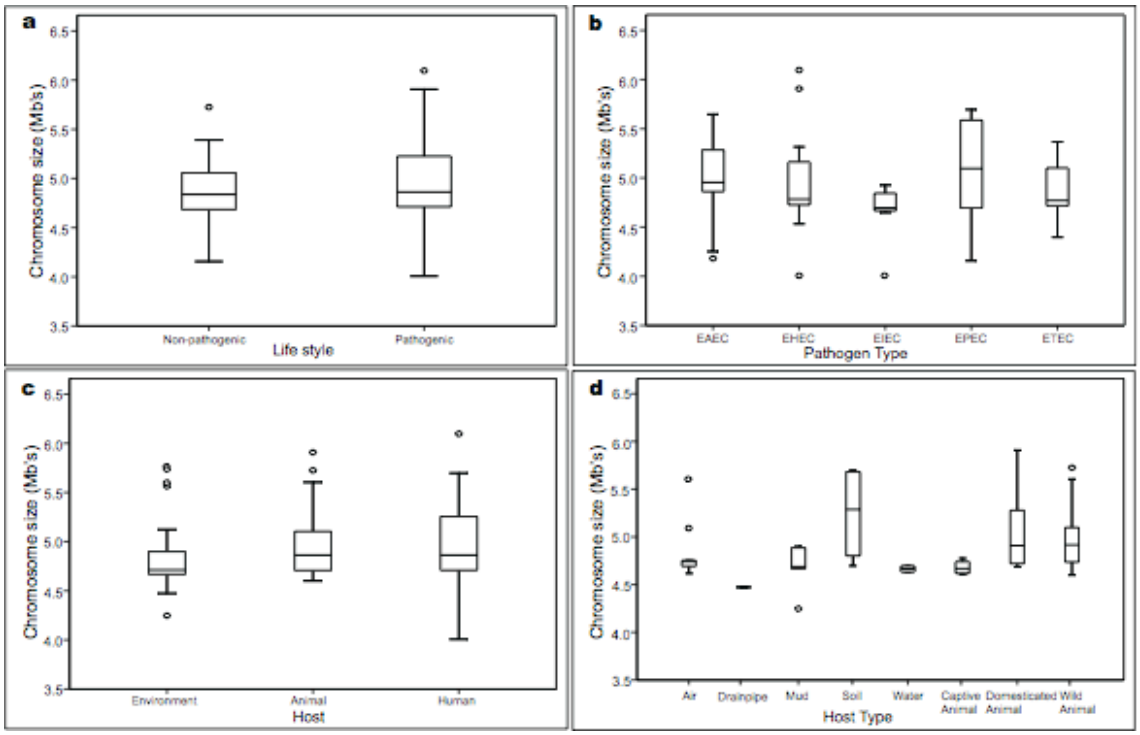

Figure 2. Chromosome size in mega bases (Mbp) associated to different *E. coli* **(a)** life style, **(b)** pathogenic type, **(c)** host and **(d)** host type. No significant differences were found between different lifestyles (Brown-Forsythe Test, $P = 0.211$) neither between host type (Brown-Forsythe Test, $P = 0.085$).

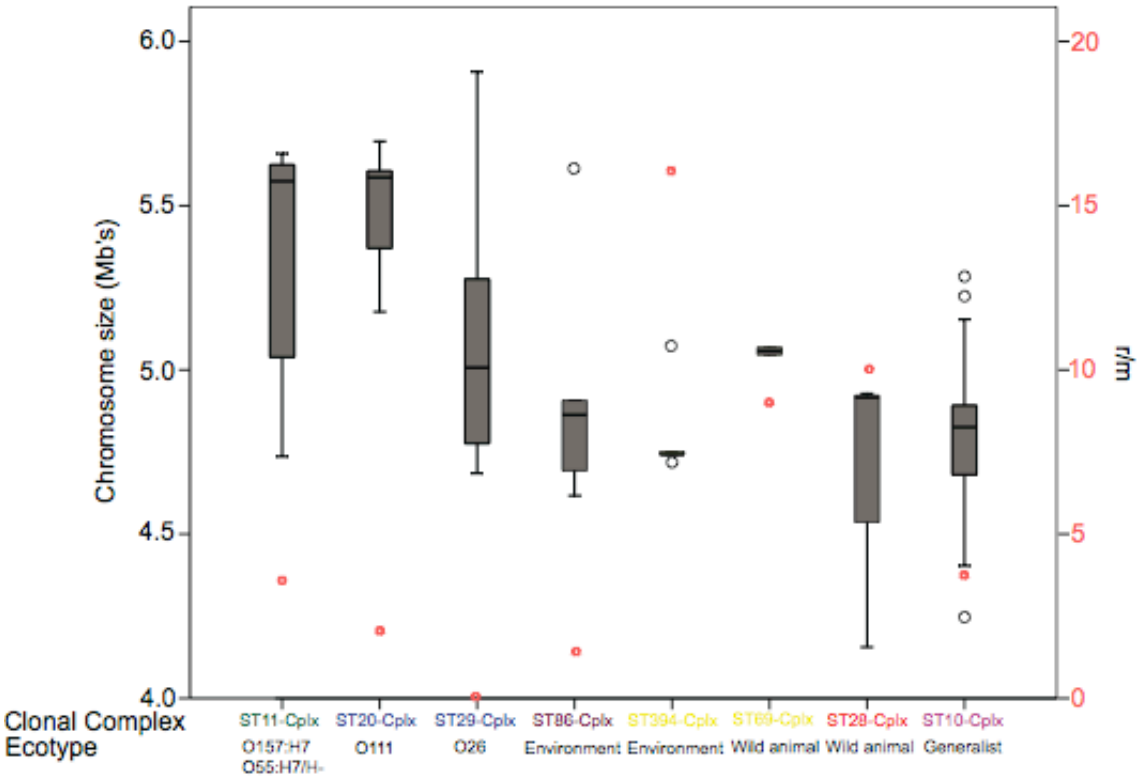

Figure 3. Chromosome size, recombination rate and ecotype associated to Mexican *E. coli* clonal complexes. *r/m:* Relative impact of recombination as compared to point mutation on the genetic differentiation of population. Each colour of clonal complex represents its phylogenetic group belonging as follows: phylogenetic group E is indicated by green, phylogenetic group B1 by blue, lineage B1-ST86 by magenta, phylogenetic group D by yellow, phylogenetic group B2 by red and phylogenetic group A by lilac colour.

Table 1
Chromosome size (Kb's) of the different Mexican *E.coli* phylogenetic groups and lineages

	N	Mean	Standard deviation	Standard error	95% Confidence Interval		Mínimum	Maximum
					Lower limit	Upper limit		
A-ST522	7	4.690,1	261,4684	98,8257	4.448,268	4.931,904	4.247,3	5.122,0
A	32	4.817,9	218,3320	38,5960	4.739,155	4.896,589	4.397,0	5.297,9
B1-ST86	5	4.938,8	395,7982	177,0063	4.447,352	5.430,248	4.617,0	5.614,0
B1	45	5.044,5	379,0253	56,5018	4.930,675	5.158,418	4.438,1	5.907,5
B2	15	4.708,5	398,4121	102,8695	4.487,833	4.929,100	4.007,0	5.427,0
D	15	4.895,3	527,3001	136,1483	4.603,271	5.187,289	4.158,0	6.097,1
E	7	5.327,8	412,0161	155,7274	4.946,734	5.708,837	4.736,0	5.658,4
Total	126	4.921,0	392,0905	34,9302	4.851,915	4.990,177	4.007,0	6.097,1

Supplementary Table S1

Mexican *Escherichia coli* isolates analyzed. Traits of isolates: Clonal Complex, Clermont group, ancestral group (Structure), MLST allele profiles, ST, host, pathotype, serotype

This supplementary table is the same of that present in paper 1

Supplementary Table S2. Chromosome size of Mexican *Escherichia coli* isolates.

Strain	Locus size (Kb's) [a]								*rrn* [b]	Total size (Kb's)
1	42	115	144	610	730	1100	2650,8		7	5392
3	42	95	135	521	610	900	2463		7	4766
19	42	95	160	521	670	730	2622,9		7	4841
33	42	115	144	550	730	1100	2500		7	5181
53	42	95	144	550	670	730	2581,9		7	4813
55	42	95	144	550	730	900	2607,6		7	5069
68	42	115	135	550	730	900	2670,7		7	5143
75	42	95	135	521	730	1100	2572,3		7	5195
88	42	95	135	521	705	730	2510,8		7	4739
90	42	95	144	580	705	730	2750		7	5046
95	42	95	310	348	521	1100	2500		7	4916
270	42	95	144	521	670	1100	2523,7		7	5096
271	42	95	144	521	670	1100	2500		7	5072
272	42	95	115	580	705	900	2446		7	4883
288	42	95	135	521	670	900	2500		7	4863
807	42	95	135	180	610	1700	2369,1		7	5131
814	42	115	160	550	670	705	2434,6		7	4677
815	42	115	160	521	670	705	2500		7	4713
825	42	95	144	580	670	730	2500		7	4761
830	42	95	144	580	705	900	3137,2		7	5603
1639	42	95	160	550	670	705	2472,6		7	4695
1684	42	95	135	521	670	900	2500		7	4863
1728	42	95	160	521	670	705	2506,8		7	4700
1735	42	95	144	521	705	1700	2518,2		7	5725
1743	42	95	144	521	670	730	2573,1		7	4775
1744	42	95	135	521	610	705	2500		7	4608
1937	42	95	144	521	670	730	2500		7	4702
1940	42	95	135	521	610	730	2500		7	4633
1967	42	95	135	550	670	730	2500		7	4722
2055	42	95	144	521	705	1100	2500		7	5107
2064	42	95	144	521	670	705	2848,3		7	5025
2065	42	95	128	521	610	705	2500		7	4601
3442	42	95	135	487	610	670	2562,5		7	4602
3456	42	95	160	521	670	705	2500		7	4693
3463	42	95	160	521	670	705	2441,1		7	4634
3470	42	115	135	550	705	730	2650,8		7	4928
3524	42	95	135	521	670	730	2657,7		7	4851
3528	42	95	144	550	670	730	2500		7	4731
3535	42	95	144	521	705	730	2500		7	4737
3566	42	95	135	550	610	670	2500		7	4602
3607	42	95	160	521	730	1100	2839,7		7	5488
3609	42	115	160	610	900	1100	2500		7	5427
3614	42	95	144	521	705	900	2500		7	4907
3617	42	95	135	550	670	900	2500		7	4892
3620	42	95	135	521	670	900	2500		7	4863
3621	**42**	**95**	**135**	**580**	**900**	**2500**			**6**	**4252**
3622	42	95	160	580	900	1100	2769,2		7	5646

3623	42	95	144	610	670	900	2763,2		7	5224
3625	42	95	160	521	730	900	2500		7	4948
3627	42	95	115	580	730	900	2500		7	4962
3628	42	95	160	580	670	1100	2500		7	5147
3629	42	95	135	521	670	900	2500		7	4863
3630	**42**	**95**	**135**	**610**	**900**	**2400**			**6**	**4182**
3631	42	95	135	550	730	900	2832,3		7	5284
3635	42	95	160	550	730	1100	2500		7	5177
3637	42	95	135	550	705	730	2500		7	4757
3641	42	95	135	521	705	900	2500		7	4898
3646	42	84	135	550	670	705	2500		7	4686
3648	42	95	144	580	670	730	2500		7	4761
3650	**42**	**95**	**255**	**357**	**705**	**3321,9**			**6**	**4776**
3651	42	95	310	341	900	1100	2907,9		7	5696
3652	42	95	135	521	610	705	2500		7	4608
3653	42	115	160	550	705	730	2500		7	4802
3655	42	95	144	580	705	1100	2702,5		7	5369
3657	42	95	135	550	670	705	2500		7	4697
3658	42	95	135	521	705	1700	2500		7	5698
3659	42	95	310	341	900	1100	2817,7		7	5606
3662	**42**	**95**	**144**	**521**	**705**	**2500**			**6**	**4007**
3663	42	95	135	550	705	900	2500		7	4927
3664	42	95	115	580	670	705	2585,1		7	4792
3665	42	115	135	521	670	705	2500		7	4688
3673	42	95	135	550	670	900	2500		7	4892
3677	42	95	115	521	670	705	2500		7	4648
3681	42	95	135	521	610	730	2400		7	4533
3682	42	95	144	580	670	705	3080		7	5316
3683	42	95	144	580	670	730	2500		7	4761
3687	42	95	135	475	550	730	2370		7	4397
3691	42	95	180	550	900	1700	2630,1		7	6097
3693	42	115	144	580	730	900	2500		7	5011
3694	42	95	135	521	610	730	2778,4		7	4911
3697	**42**	**95**	**144**	**521**	**705**	**2500**			**6**	**4007**
3698	**42**	**95**	**115**	**610**	**900**	**2396**			**6**	**4158**
3700	42	95	135	521	670	730	2500		7	4693
3702	42	95	135	550	670	730	2500		7	4722
3707	42	95	135	550	900	1100	2764		7	5586
3711	42	95	135	521	670	1100	2500		7	5063
3712	42	95	135	521	670	1700	2500		7	5663
3713	42	95	144	580	705	1100	2681,7		7	5348
3715	50	95	144	550	670	705	2500		7	4714
3716	42	95	144	550	705	1100	2500		7	5136
3719	42	95	144	580	670	705	2500		7	4736
3720	**84**	**115**	**135**	**186**	**265**	**459**	**705**	**2524**	**8**	**4473**
3824	42	95	135	521	670	730	2500		7	4693
3842	42	95	144	550	670	705	2500		7	4706
3859	42	95	135	521	610	705	2777,4		7	4885
3873	42	95	135	521	670	705	2500		7	4668
3874	42	95	135	521	670	705	2500		7	4668
3876	**42**	**95**	**135**	**550**	**670**	**2755,3**			**6**	**4247**
3885	**84**	**115**	**135**	**200**	**255**	**459**	**900**	**2750,8**	**8**	**4899**
4129	42	115	135	550	670	730	2500		7	4742

4130	42	115	135	550	670	730	2476,6		7	4719
4132	42	115	135	550	670	730	2500		7	4742
4135	42	95	144	521	610	730	2500		7	4642
4136	42	115	144	550	670	730	2500		7	4751
4952	42	95	265	348	705	730	2500		7	4685
4953	42	95	255	370	705	730	3079,7		7	5277
4957	42	95	160	521	670	705	2500		7	4693
4958	42	95	135	200	487	670	2809,1		7	4438
4959	42	95	160	580	705	900	3176,4		7	5658
4962	42	95	144	580	705	900	3179,7		7	5646
4964	42	95	144	580	670	1100	2943,2		7	5574
4976	42	95	310	341	730	1100	2752		7	5370
4993	42	95	144	521	610	705	2500		7	4617
5014	42	95	255	370	705	730	2809,8		7	5007
5020	42	95	144	521	730	1700	2675,5		7	5908
5021	42	95	144	521	670	705	2606,7		7	4784
5040	42	95	135	580	670	1100	2500		7	5122
5058	42	95	357	370	550	1700	2500		7	5614
6879	**42**	**95**	**144**	**521**	**610**	**670**	**705**	**2599,9**	**8**	**5387**
6891	42	115	144	550	670	900	2652		7	5073
6908	42	95	128	255	550	580	2506		7	4156
6909	42	95	135	521	610	705	2500		7	4608
6918	42	95	160	521	670	705	2632,6		7	4826
41724	42	95	144	550	705	730	2500		7	4766
43221	42	95	160	550	670	1100	2680,9		7	5298
48334	42	95	144	521	670	730	2680,9		7	4883
50417	42	95	144	550	670	705	2500		7	4706
63880	42	95	160	550	705	730	2500		7	4782

[a] Tamaño de cada uno de los fragmentos arrojados por la enzima de restricción *I-CeuI*.
[b] Número de operones ribosomales por aislado (*rrn*). Este número corresponde al número de cortes que arroja la enzima de restricción *I-CeuI*.

Supplementary Table S3. Chromosome size of sequenced *Escherichia coli* isolates

Organism	BioProject	Plasmids	Size (Mb)	Gene	Pathogenic/ NonPathogenic	Host
Escherichia coli O157:H7 str. Sakai	PRJNA57781, PRJNA226	2	5,59	5.460	p	
Escherichia coli SE11	PRJNA59425, PRJNA18057	6	5,16	5.105	n	
Escherichia coli str. K-12 substr. MG1655	PRJNA57779, PRJNA225	-	4,64	4.496	n	
Escherichia coli O26:H11 str. 11368	PRJNA41021, PRJDA32509	4	5,86	5.989	p	
Escherichia coli 'BL21-Gold(DE3)pLysS AG'	PRJNA59245, PRJNA30681	-	4,57	4.425		
Escherichia coli 042	PRJEA161985, PRJEA40647	1	5,36	5.036	p	
Escherichia coli 536	PRJNA58531, PRJNA16235	-	4,94	4.779	p	
Escherichia coli 55989	PRJNA59383, PRJNA33413	-	5,15	5.140	p	
Escherichia coli ABU 83972	PRJNA161975, PRJNA38725	1	5,13	4.906	n	
Escherichia coli APEC O1	PRJNA58623, PRJNA16718	2	5,5	4.968	p	
Escherichia coli ATCC 8739	PRJNA58783, PRJNA18083	-	4,75	4.408	n	
Escherichia coli B str. REL606	PRJNA58803, PRJNA18281	-	4,63	4.365		
Escherichia coli BL21(DE3)	PRJNA161947, PRJNA20713	-	4,56	4.334		
Escherichia coli BL21(DE3)	PRJEA161949, PRJEA28965	-	4,56	4.422		
Escherichia coli BW2952	PRJNA59391, PRJNA33775	-	4,58	4.267		
Escherichia coli CFT073	PRJNA57915, PRJNA313	-	5,23	5.579	p	
Escherichia coli DH1	PRJNA161951, PRJNA30031	-	4,63	4.375		
Escherichia coli DH1	PRJDA162051, PRJDA52077	-	4,62	4.450		
Escherichia coli E24377A	PRJNA58395, PRJNA13960	6	5,25	5.258	p	
Escherichia coli ED1a	PRJNA59379, PRJNA33409	-	5,21	5.325	n	
Escherichia coli ETEC H10407	PRJEA161993, PRJEA42749	4	5,33	5.086	p	
Escherichia coli HS	PRJNA58393, PRJNA13959	-	4,64	4.629	n	
Escherichia coli IAI1	PRJNA59377, PRJNA33373	-	4,7	4.633	p	
Escherichia coli IAI39	PRJNA59381, PRJNA33411	-	5,13	5.097	p	
Escherichia coli IHE3034	PRJNA162007, PRJNA43693	-	5,11	4.970	p	
Escherichia coli KO11FL	PRJNA52593, PRJNA33875	2	5,03	4.885		
Escherichia coli KO11FL	PRJNA162099, PRJNA62299	1	5,03	4.825		
Escherichia coli LF82	PRJNA161965, PRJNA33825	-	4,77	4.545	p	
Escherichia coli NA114	PRJNA162139, PRJNA66975	-	4,97	4.975	p	upec
Escherichia coli O103:H2 str. 12009	PRJNA41013, PRJDA32511	1	5,52	5.545	p	ehec
Escherichia coli O111:H- str. 11128	PRJNA41023, PRJDA32513	5	5,77	5.935	p	ehec
Escherichia coli O127:H6 str. E2348/69	PRJNA59343, PRJEA32571	2	5,07	5.015	p	epec
Escherichia coli O157:H7 str. EC4115	PRJNA59091, PRJNA27739	2	5,7	6.066	p	ehec
Escherichia coli O157:H7 str. EDL933	PRJNA57831, PRJNA259	1	5,62	5.528	p	ehec (isolated from USA food)
Escherichia coli O157:H7 str. TW14359	PRJNA59235, PRJNA30045	1	5,62	5.588	p	ehec
Escherichia coli O55:H7 str. CB9615	PRJNA46655, PRJNA42729	1	5,45	5.371	p	epec
Escherichia coli O55:H7 str. RM12579	PRJNA162153, PRJNA68245	5	5,45	5.255	p	epec

Escherichia coli O7:K1 str. CE10	PRJNA162115, PRJNA63597	4	5,38	5.269	p	nmec
Escherichia coli O83:H1 str. NRG 857C	PRJNA161987, PRJNA41221	1	4,89	4.690	p	aiec
Escherichia coli P12b	PRJNA162061, PRJNA59455	-	4,94	4.581	n	
Escherichia coli S88	PRJNA62979, PRJNA33375	1	5,17	5.191	p	nmec
Escherichia coli SE15	PRJDA161939, PRJDA19053	1	4,84	4.594	n	comensal
Escherichia coli SMS-3-5	PRJNA58919, PRJNA19469	4	5,22	5.127	n	free-living
Escherichia coli UM146	PRJNA162043, PRJNA50883	1	5,11	4.891	p	aiec
Escherichia coli UMN026	PRJNA62981, PRJNA33415	2	5,36	5.298	p	upec
Escherichia coli UMNK88	PRJNA161991, PRJNA42137	5	5,67	5.754	p	etec (porcine/post-weaning diarrhea in pigs)
Escherichia coli UTI89	PRJNA58541, PRJNA16259	1	5,18	5.272	p	upec
Escherichia coli W	PRJNA162011, PRJNA48011	2	5,01	4.880		
Escherichia coli W	PRJNA162101, PRJNA62301	2	5,01	4.848		
Escherichia coli Xuzhou21	PRJNA163995, PRJNA45823	2	5,52	5.298	p	ehec
Escherichia coli str. 'clone D i14'	PRJNA162049, PRJNA52023	-	5,04	5.049	p	commensal/upec (dog)
Escherichia coli str. 'clone D i2'	PRJNA162047, PRJNA52021	-	5,04	5.049	p	commensal/upec (dog)
Escherichia coli str. K-12 substr. DH10B	PRJNA58979, PRJNA20079	-	4,69	4.356	n	
Escherichia coli str. K-12 substr. W3110	PRJNA161931, PRJNA16351	-	4,65	4.440	n	
Escherichia coli O157:H7 str. TW14588	PRJNA55087, PRJNA28847	1	5,67	6.011	p	ehec
Escherichia coli UMNF18	PRJNA48455	5	-	-	p	ETEC (porcine/post-weaning diarrhea in pigs)
Escherichia coli str. K-12 substr. MG1655	PRJNA40075	-	4,64	-	n	
Escherichia coli str. K-12 substr. MG1655star	PRJNA51747	-	4,64	-	n	
Escherichia coli 1.2264	PRJNA51097	-	5,5	5.661		isolated from a goat *Capra hircus* (O76:H-)
Escherichia coli 1.2741	PRJNA51085	-	5,69	5.914		isolated from a cow (O2:H4
Escherichia coli 101-1	PRJNA54363, PRJNA16193	-	4,98	5.155	p	eaec
Escherichia coli 1827-70	PRJNA60615, PRJNA40257	-	4,8	4.822	?	
Escherichia coli 2.3916	PRJNA51123	-	5,64	6.179		isolated from a pig
Escherichia coli 2.4168	PRJNA51127	-	4,74	4.804		isolated from water
Escherichia coli 2362-75	PRJNA60613, PRJNA40275	-	5,17	5.343	p	epec
Escherichia coli 2534-86	PRJNA48251	-	5,29	5.642		etec (O8:K87)
Escherichia coli 3.2303	PRJNA51129	-	4,95	5.149		isolated from water
Escherichia coli 3.2608	PRJNA51105	-	5,45	5.670		isolated from a horse
Escherichia coli 3.3884	PRJNA51125	-	5,23	5.311		isolated from a cow
Escherichia coli 3003	PRJNA51131	-	4,92	4.982		isolated from water
Escherichia coli 3030-1	PRJNA48253	-	5,25	5.456	?	
Escherichia coli 3431	PRJNA40265	-	5,22	5.322		da-epec
Escherichia coli 4.0522	PRJNA51109	-	5,83	6.308		isolated from a cow
Escherichia coli 4.0967	PRJNA51121	-	5,87	6.266		isolated from a rabbit
Escherichia coli 4_1_47FAA	PRJNA39385	-	5,26	5.389	?	
Escherichia coli 5.0588	PRJNA51089	-	5	5.137		isolated from a cow
Escherichia coli 5.0959	PRJNA51115	-	5,42	5.710		isolated from an unknown host
Escherichia coli 53638	PRJNA54321, PRJNA15639	2	5,37	5.654		eiec
Escherichia coli 541-1	PRJNA47101	-	5	4.984		Homo sapiens ileum
Escherichia coli 541-15	PRJNA47099	-	5,03	5.065		Homo sapiens ileum
Escherichia coli 576-1	PRJNA47103	-	5,19	5.292	p	invasive E. coli
Escherichia coli 75	PRJNA47105	-	4,6	4.484		
Escherichia coli 83972	PRJNA55485, PRJNA31467	-	5,07	5.380		isolated from a clinical ABU episode
Escherichia coli 9.0111	PRJNA51119	-	5,73	6.149		O128 isolated from human, stx1 stx2 genes

Escherichia coli 9.1649	PRJNA51117	-	5,1	-		O2:H- isolated from pig, stx1 stx2
Escherichia coli 900105 (10e)	PRJNA51137	-	5,56	5.983		O26:H11 isolated from calf in 1990, stx1
Escherichia coli 93-001	PRJNA65775	-	5,38	5.805		isolated from *Bos taurus* (O157:H7 related)
Escherichia coli 93.0624	PRJNA51107	-	5,28	5.459		O103:H6 isolated from human, stx1
Escherichia coli 95.0941	PRJNA51095	-	4,79	-		O45:H2 isolated from human, stx1 and stx2
Escherichia coli 96.0497	PRJNA51101	-	5,01	5.026		O91:H21 isolated from human, stx1
Escherichia coli 96.154	PRJNA51113	-	4,95	4.981		O113:H12 isolated from human, stx2
Escherichia coli 97.0246	PRJNA51087	-	5,5	5.924		O5:H- isolated from cow, stx1, stx2
Escherichia coli 97.0259	PRJNA51091	-	5,21	5.269		O11:H- isolated from cow, stx2
Escherichia coli 97.0264	PRJNA51099	-	5,23	-		O88:H25 isolated from cow, stx1, stx2
Escherichia coli 99.0741	PRJNA51103	-	5,45	5.724		O91:H- isolated from food, stx1, stx2
Escherichia coli AA86	PRJNA65321	1	4,98	4.819	n	isolated from healthy cow feces
Escherichia coli AI27	PRJNA89369	-	4,9	4.750		isolated from porcine feces in South Korea
Escherichia coli B088	PRJNA47003, PRJNA38905	-	4,94	4.671		isolated from bird (*Circus assimilis*)
Escherichia coli B093	PRJNA38907	-	5,19	4.990		isolated from bird (*Strepera graculina*)
Escherichia coli B171	PRJNA54319, PRJNA15630	1	5,5	5.932	p	epec
Escherichia coli B185	PRJNA47001, PRJNA38915	-	5,11	4.797		isolated from bird (*Turdus merula*)
Escherichia coli B354	PRJNA46999, PRJNA38917	-	4,83	4.724		isolated from bird (*Monarca trivirgatus*)
Escherichia coli B41	PRJNA51135	-	5,01	5.075		O101:HNM isolated from pig, does not encode for stx1 nor stx2
Escherichia coli B799	PRJNA38929	-	5,3	5.203		isolated from bird (*Sericornis frontalis*)
Escherichia coli B7A	PRJNA54297, PRJNA15572	-	5,3	5.529	p	etec (O148:H28)
Escherichia coli CUMT8	PRJNA47109	-	4,71	4.635	p	adherent and invasive pathotype (O8:H21)
Escherichia coli DEC10A	PRJNA50999	-	5,37	5.838	p	
Escherichia coli DEC10B	PRJNA51001	-	5,58	6.104	p	ehec (O26:H11 isolated from australian human)
Escherichia coli DEC10C	PRJNA51003	-	5,53	6.026	p	ehec (O26:H11 isolated from USA infant human with acute gastroenteritis)
Escherichia coli DEC10D	PRJNA51005	-	5,4	5.878	p	
Escherichia coli DEC10E	PRJNA51007	-	5,17	5.361	p	ehec (O26:H11 isolated from USA cow-calf)
Escherichia coli DEC10F	PRJNA51009	-	5,76	6.202	p	
Escherichia coli DEC11A	PRJNA51011	-	5,22	5.365	p	epec (O128a:H2 isolated from USA human with diarrhea)
Escherichia coli DEC11B	PRJNA51013	-	5,23	5.390	p	diarrheagenic E. coli (O128a:H2 isolated from USA infant human with diarrhea)
Escherichia coli DEC11C	PRJNA51015	-	5,55	5.998	p	
Escherichia coli DEC11D	PRJNA51017	-	5,3	5.498	p	diarrheagenic E. coli (O128:H2 isolated from UK infant human, source Dr. Cravioto)
Escherichia coli DEC11E	PRJNA51019	-	5,13	5.294	p	diarrheagenic E. coli (O128:H2 isolated from Brazil infant human with diarrhea)
Escherichia coli DEC12A	PRJNA51021	-	5,36	5.769	p	epec (O111:H2 isolated from UK infant human with diarrhea)
Escherichia coli DEC12B	PRJNA51023	-	5,49	5.920	p	epec (O111:H2 isolated from USA human with diarrhea outbreak)
Escherichia coli DEC12C	PRJNA51025	-	5,45	5.922	p	epec (O111:HNM isolated from Panama human)
Escherichia coli DEC12D	PRJNA51027	-	5,4	5.775	p	
Escherichia coli DEC12E	PRJNA51029	-	5,3	5.556	p	
Escherichia coli DEC13A	PRJNA51031	-	4,69	4.758	p	diarrheagenic E. coli (O128:H7 isolated from USA human with diarrhea)
Escherichia coli DEC13B	PRJNA51033	-	4,71	4.801	p	diarrheagenic E. coli (O128:H7 isolated from USA human with diarrhea)
Escherichia coli DEC13C	PRJNA51035	-	5,29	5.522	p	diarrheagenic E. coli (O128:H7 isolated from Tanzania infant human with diarrhea)
Escherichia coli DEC13D	PRJNA51037	-	5,01	5.123	p	diarrheagenic E. coli (O128:H7 isolated from Rwanda infant human with diarrhea)
Escherichia coli DEC13E	PRJNA51039	-	4,98	5.049	p	diarrheagenic E. coli (O128:H47 isolated from USA human with diarrhea)
Escherichia coli DEC14A	PRJNA51043	-	4,95	5.022	p	diarrheagenic E. coli (O128:H21 isolated from Peru infant humanwith diarrhea)
Escherichia coli DEC14B	PRJNA51045	-	5,27	5.519	p	diarrheagenic E. coli (O128:H21 isolated from India infant human with diarrhea)
Escherichia coli DEC14C	PRJNA51047	-	5,15	5.407	p	diarrheagenic E. coli (O128a:H21 isolated from Peru human with diarrhea)
Escherichia coli DEC14D	PRJNA51049	-	5,18	5.368	p	diarrheagenic E. coli (O128:HNM isolated from USA human with diarrhea)
Escherichia coli DEC15A	PRJNA51053	-	5,25	5.414	p	epec (O111:H21 isolated from USA human with diarrhea from outbreak)
Escherichia coli DEC15B	PRJNA51055	-	5,26	5.464	p	diarrheagenic E. coli (O111:H21 isolated from USA human with diarrhea)
Escherichia coli DEC15C	PRJNA51057	-	5,2	5.308	p	diarrheagenic E. coli (O111:H21 isolated from USA

						human with diarrhea)
Escherichia coli DEC15D	PRJNA51059	-	5,21	5.317	p	diarrheagenic E. coli (O111:H21 isolated from USA human with diarrhea)
Escherichia coli DEC15E	PRJNA51061	-	5,23	5.418	p	diarrheagenic E. coli (O111:H21 isolated from USA human with diarrhea)
Escherichia coli DEC1A	PRJNA50903	-	5,11	5.322	p	epec (O55:H6 isolated from USA infant human with diarrhea outbreak)
Escherichia coli DEC1B	PRJNA50905	-	5,2	5.472	p	epec (O55:H6 isolated from Dutch infant human with diarrhea)
Escherichia coli DEC1C	PRJNA50907	-	5,28	5.492	p	epec (O55:H6 isolated from Germany infant human with diarrhea)
Escherichia coli DEC1D	PRJNA50909	-	5,16	5.420	p	
Escherichia coli DEC1E	PRJNA50911	-	5,15	5.424	p	
Escherichia coli DEC2A	PRJNA50913	-	5,09	5.362	p	epec (O55:H6 isolated from Congo infant human with diarrhea)
Escherichia coli DEC2B	PRJNA50915	-	5,13	5.337	p	epec (O55:HNM isolated from USA infant human with diarrhea outbreak)
Escherichia coli DEC2C	PRJNA50917	-	5,3	5.599	p	epec (O55:H6 isolated from USA infant human with diarrhea outbreak)
Escherichia coli DEC2D	PRJNA50919	-	5,09	5.258	p	
Escherichia coli DEC2E	PRJNA50921	-	5,21	5.425	p	epec (O55:H6 isolated from USA human)
Escherichia coli DEC3A	PRJNA50923	-	5,45	5.692	p	
Escherichia coli DEC3B	PRJNA50925	-	5,5	5.824	p	ehec (O157:H7 isolated from USA human)
Escherichia coli DEC3C	PRJNA50927	-	5,48	5.850	p	ehec (O157:H7 isolated from USA human with gastroenteritis)
Escherichia coli DEC3D	PRJNA50929	-	5,44	5.751	p	ehec (O157:H7 isolated from USA human)
Escherichia coli DEC3E	PRJNA50931	-	5,52	5.900	p	ehec (O157:H7 isolated from Canada human)
Escherichia coli DEC3F	PRJNA50933	-	5,41	5.692	p	
Escherichia coli DEC4A	PRJNA50935	-	5,38	5.720	p	ehec (O157:H7 isolated from Argentinca cow-calf)
Escherichia coli DEC4B	PRJNA50937	-	5,63	6.097	p	ehec (O157:H7 isolated from Denmark human woth diarrhea)
Escherichia coli DEC4C	PRJNA50939	-	5,53	5.919	p	
Escherichia coli DEC4D	PRJNA50941	-	5,39	5.676	p	ehec (O157:H7 isolated from Japan cow-calf)
Escherichia coli DEC4E	PRJNA50943	-	5,41	5.683	p	ehec (O157:H7 isolated from Denmark human with diarrhea)
Escherichia coli DEC4F	PRJNA50945	-	5,4	5.727	p	
Escherichia coli DEC5A	PRJNA50947	-	5,37	5.425	p	ehec (O55:H7 isolated from USA infant human with diarrhea for more than 7 weeks)
Escherichia coli DEC5B	PRJNA50949	-	5,48	5.700	p	ehec (O55:H7 isolated from USA infant human with diarrhea)
Escherichia coli DEC5C	PRJNA50951	-	5,33	5.480	p	ehec (O55:H7 isolated from USA human with diarrhea)
Escherichia coli DEC5D	PRJNA50953	-	5,23	5.320	p	ehec (O55:H7 isolated from Sri Lanka human)
Escherichia coli DEC5E	PRJNA50955	-	5,68	5.936	p	ehec (O55:H7 isolated from Iran human)
Escherichia coli DEC6A	PRJNA50957	-	5,39	5.679	p	diarrheagenic E. coli (O111:H21 isolated from USA infant human with diarrhea from an outbreak)
Escherichia coli DEC6B	PRJNA50959	-	5,32	5.661	p	
Escherichia coli DEC6C	PRJNA50961	-	5,12	5.503	p	diarrheagenic E. coli (O111:H12 isolated from Guatemala human with diarrhea)
Escherichia coli DEC6D	PRJNA50963	-	4,98	5.312	p	
Escherichia coli DEC6E	PRJNA50965	-	5,08	5.323	p	
Escherichia coli DEC7A	PRJNA50967	-	5,25	5.399	p	diarrheagenic E. coli (O157:H43 isolated from USA pig with septicemia)
Escherichia coli DEC7B	PRJNA50969	-	5,24	5.414	p	diarrheagenic E. Coli (O149:HNM isolated from small intestine of USA cow)
Escherichia coli DEC7C	PRJNA50973	-	5,31	5.436	p	diarrheagenic E. coli (O157:H43 isolated from USA pig with colibacilosis)
Escherichia coli DEC7D	PRJNA50975	-	5,36	5.517	p	
Escherichia coli DEC7E	PRJNA50977	-	5,14	5.232	p	
Escherichia coli DEC8A	PRJNA50979	-	5,65	6.141	p/n?	ehec (O111a:HNM isolated from stool of USA human)
Escherichia coli DEC8B	PRJNA50981	-	5,37	5.759	p	
Escherichia coli DEC8C	PRJNA50983	-	5,91	6.505	p	ehec (O111:HNM isolated from USA cow-calf with scours)
Escherichia coli DEC8D	PRJNA50985	-	5,46	5.763	p	ehec (O111:H11 isolated from Cuba infant human with diarrhea)
Escherichia coli DEC8E	PRJNA50987	-	5,32	5.711	p	ehec (O111:H8 isolated from Denmark human with diarrhea)
Escherichia coli DEC9A	PRJNA50989	-	5,41	5.735	p	ehec (O26:H11 isolated from USA human with diarrhea and fever)
Escherichia coli DEC9B	PRJNA50991	-	5,36	5.616	p	ehec (O26:HN isolated from USA human with diarrhea)
Escherichia coli DEC9C	PRJNA50993	-	5,19	5.369	p	
Escherichia coli DEC9D	PRJNA50995	-	5,49	5.750	p	ehec (O26:H11 isolated from Denmark infant human with diarrhea)
Escherichia coli DEC9E	PRJNA50997	-	5,43	5.768	p	

Escherichia coli E101	PRJNA38935	-	5,18	5.015		isolated from the environment (Lake Ginnderra, ACT, Australia)
Escherichia coli E110019	PRJNA54303, PRJNA15578	-	5,38	5.681	p	atypical EPEC (O111:H9 isolated from a diarrhea outbreak in Finland, without EAF plasmid
Escherichia coli E1167	PRJNA38941	-	4,92	4.724	p/n?	This strain will be used for comparative analysis
Escherichia coli E128010	PRJNA40269	-	5,22	5.500	p	aEPEC (O114:H2) diarrhea, human, Bangladesh
Escherichia coli E1520	PRJNA38945	-	4,9	4.777	p/n?	
Escherichia coli E22	PRJNA54301, PRJNA15577	-	5,53	5.831	p	epec O103:H2 model strain, with Locus LEE, Kaper strain, human, diarrhea, intestinal
Escherichia coli E482	PRJNA38949	-	4,83	4.637	p/n?	
Escherichia coli EC1734	PRJNA65941	-	5,43	5.685	p/n?	Human
Escherichia coli EC1738	PRJNA65939	-	5,48	5.710	p	ehec (O157:H7 isolated from USA human)
Escherichia coli EC1845	PRJNA65949	-	5,36	5.749	p	ehec (O157:H7 isolated from USA human)
Escherichia coli EC1863	PRJNA65965	-	5,38	5.693	p	ehec (O157:H7 isolated from USA human)
Escherichia coli EC4013	PRJNA65927	-	5,35	5.714	p	ehec (O157:H7 isolated from USA human)
Escherichia coli EC4100B	PRJNA61479	-	5,11	5.117	p	ehec (O157:H7 isolated from USA human)
Escherichia coli EC4196	PRJNA65917	-	5,36	5.830	p	ehec (O157:H7 isolated from USA human)
Escherichia coli EC4203	PRJNA65915	-	5,39	5.824	p	ehec (O157:H7 isolated from USA human)
Escherichia coli EC4402	PRJNA65929	-	5,45	5.953	p	ehec (O157:H7 isolated from USA human)
Escherichia coli EC4421	PRJNA65923	-	5,32	5.664	p	ehec (O157:H7 isolated from USA human)
Escherichia coli EC4422	PRJNA65925	-	5,32	5.625	p	ehec (O157:H7 isolated from USA human)
Escherichia coli EC4436	PRJNA65933	-	5,42	5.780	p	ehec (O157:H7 isolated from USA human)
Escherichia coli EC4437	PRJNA65935	-	5,41	5.820	p	ehec (O157:H7 isolated from USA human)
Escherichia coli EC4439	PRJNA65931	-	5,4	5.862	p	ehec (O157:H7 isolated from USA human)
Escherichia coli EC4448	PRJNA65937	-	5,49	5.969	p	ehec (O157:H7 isolated from USA human)
Escherichia coli EPEC C342-62	PRJNA49113	-	5,24	5.485	p	epec (O126:H2 from human USA
Escherichia coli EPECa12	PRJNA40273	-	5,23	5.479	p	epec from human stool Brazil
Escherichia coli EPECa14	PRJNA40267	-	5,44	5.994	p	epec
Escherichia coli F11	PRJNA54299, PRJNA15576	-	5,22	5.307	p	ExPEC (O6:H31, cystitis, infection of the bladder, woman
Escherichia coli FDA505	PRJNA65755	-	5,35	5.756	p	ehec(O157:H7 isolated from USA human
Escherichia coli FDA517	PRJNA65759	-	5,52	5.990	p	ehec(O157:H7 isolated from USA human
Escherichia coli FRIK1985	PRJNA65765	-	5,54	6.159	p	ehec (O157:H7 isolated from *Bos taurus*, USA
Escherichia coli FRIK1990	PRJNA65769	-	5,52	6.058	p	ehec (O157:H7 isolated from *Bos taurus*, USA
Escherichia coli FRIK1996	PRJNA65763	-	5,43	5.936	p	ehec (O157:H7 isolated from *Bos taurus*, USA
Escherichia coli FVEC1302	PRJNA49707, PRJNA39915	-	5,29	5.199		human FVEC
Escherichia coli FVEC1412	PRJNA46997, PRJNA39917	-	5,19	5.055		human FVEC
Escherichia coli G58-1	PRJNA48249	-	4,87	5.047	p	etec (O101:K28:NM, isolated from a terrestial animal) Emerging Diarrhea pathogens project
Escherichia coli H120	PRJNA38971	-	5,04	4.955	p/n?	This strain will be used for comparative analysis PATRIC
Escherichia coli H252	PRJNA38981	-	5,26	5.126	p/n?	This strain will be used for comparative analysis PATRIC
Escherichia coli H263	PRJNA38985	-	5,19	5.126	p/n?	This strain will be used for comparative analysis PATRIC
Escherichia coli H299	PRJNA52527, PRJNA38991	-	5,25	5.714		human
Escherichia coli H30	PRJNA51093	-	5,49	-	p	diarrheagenic E. coli (O26:H11 isolated from UK human with diarrhea)
Escherichia coli H397	PRJNA39001	-	5,21	5.121		human
Escherichia coli H489	PRJNA39013	-	4,83	4.739	p/n?	This strain will be used for comparative analysis PATRIC
Escherichia coli H494	PRJNA39015	-	5,09	4.960		human
Escherichia coli H591	PRJNA52525, PRJNA39021	-	4,92	5.143		human
Escherichia coli H730	PRJNA39031	-	5,05	4.960		human
Escherichia coli H736	PRJNA52485, PRJNA39033	-	4,65	4.781		human
Escherichia coli HM605	PRJNA47107	-	5,12	5.163	p	adherent and invasive pathotype (O8:H1), human patient suffering with Crohn's disease, not diarrhea associated
Escherichia coli HM605	PRJEA66487	-	5,2	-	p	adherent and invasive pathotype (O8:H1), human patient suffering with Crohn's disease, not diarrhea associated
Escherichia coli J53	PRJNA82837	-	4,65	4.567		
Escherichia coli JB1-95	PRJNA51111	-	5,41	5.768		
Escherichia coli KD1	PRJNA47095	-	4,84	4.694		
Escherichia coli KD2	PRJNA47097	-	4,97	4.828		
Escherichia coli LCT-EC106	PRJNA89725	-	5,12	-		
Escherichia coli LT-68	PRJNA40261	-	5,19	5.697		

Escherichia coli M605	PRJNA52483, PRJNA39039	-	5,45	5.750		isolated from mammal (*Nyctophilus geoffroyi*- bat)
Escherichia coli M718	PRJNA52481, PRJNA39043	-	5,37	5.514		isolated from mammal (*Sarcophilus harrisii* - Tasmanian devil)
Escherichia coli M863	PRJNA39045	-	5,25	5.057		
Escherichia coli M919	PRJNA39047	-	5,25	5.136		isolated from mammal (*Sarcophilus harrisii* - Tasmanian devil)
Escherichia coli MS 107-1	PRJNA50575, PRJNA40713	-	4,91	5.294		
Escherichia coli MS 110-3	PRJNA47225	-	5,06	5.536		
Escherichia coli MS 115-1	PRJNA50627, PRJNA47227	-	4,8	5.133		
Escherichia coli MS 116-1	PRJNA50635, PRJNA47229	-	4,84	5.276		
Escherichia coli MS 117-3	PRJNA47231	-	5,01	5.491		
Escherichia coli MS 119-7	PRJNA50577, PRJNA40709	-	4,97	5.477		
Escherichia coli MS 124-1	PRJNA50763, PRJNA40707	-	5,39	6.114		
Escherichia coli MS 145-7	PRJNA59467, PRJNA40703	-	5,12	5.642		
Escherichia coli MS 146-1	PRJNA50897, PRJNA47241	-	4,74	5.071		
Escherichia coli MS 153-1	PRJNA47257	-	5,09	5.578		
Escherichia coli MS 16-3	PRJNA47259	-	4,95	5.394		
Escherichia coli MS 175-1	PRJNA50639, PRJNA47263	-	4,67	4.984		
Escherichia coli MS 182-1	PRJNA50641, PRJNA47265	-	4,98	5.433		
Escherichia coli MS 185-1	PRJNA50657, PRJNA47267	-	4,94	5.296		
Escherichia coli MS 187-1	PRJNA50629, PRJNA47269	-	4,38	4.605		
Escherichia coli MS 196-1	PRJNA50655, PRJNA47271	-	5,21	5.713		
Escherichia coli MS 198-1	PRJNA50625, PRJNA47273	-	5,24	5.724		
Escherichia coli MS 200-1	PRJNA50645, PRJNA47275	-	5,04	5.445		
Escherichia coli MS 21-1	PRJNA50631, PRJNA47205	-	5,28	5.860		
Escherichia coli MS 45-1	PRJNA50643, PRJNA47207	-	4,99	5.346		
Escherichia coli MS 57-2	PRJNA47209	-	4,94	5.334		
Escherichia coli MS 60-1	PRJNA47211	-	5,2	5.759		
Escherichia coli MS 69-1	PRJNA50653, PRJNA47213	-	5,21	5.592		
Escherichia coli MS 78-1	PRJNA50771, PRJNA47217	-	4,75	5.104		
Escherichia coli MS 79-10	PRJNA40701	-	4,9	5.397		
Escherichia coli MS 84-1	PRJNA50623, PRJNA47219	-	5,29	5.828		
Escherichia coli MS 85-1	PRJNA40699	-	5,46	6.149		
Escherichia coli NC101	PRJNA52343, PRJNA47121	-	5,02	4.805		
Escherichia coli NCCP15647	PRJNA157365	-	5,15	-		
Escherichia coli NCCP15657	PRJNA157429	-	5,02	-		
Escherichia coli NCCP15658	PRJNA157427	-	5,46	-		
Escherichia coli O103:H2 str. CVM9450	PRJNA129421	-	5,39	5.477		
Escherichia coli O103:H25 str. CVM9340	PRJNA129419	-	5,26	5.340		
Escherichia coli O103:H25 str. NIPH-11060424	PRJNA74417	-	5,16	-		
Escherichia coli O104:H4 str. 01-09591	PRJNA67931	-	5,49	5.630		
Escherichia coli O104:H4 str. 04-8351	PRJNA68211	-	5,39	5.241		
Escherichia coli O104:H4 str. 09-7901	PRJNA68213	-	5,29	5.138		
Escherichia coli O104:H4 str. 11-3677	PRJNA68215	-	5,4	5.229		
Escherichia coli O104:H4 str. 11-4404	PRJNA70733	-	5,39	5.245		
Escherichia coli O104:H4 str. 11-4522	PRJNA70735	-	5,38	5.214		
Escherichia coli O104:H4 str. 11-4623	PRJNA70737	-	5,4	5.245		

Escherichia coli O104:H4 str. 11-4632 C1	PRJNA70739	-	5,38	5.226		
Escherichia coli O104:H4 str. 11-4632 C2	PRJNA70741	-	5,39	5.225		
Escherichia coli O104:H4 str. 11-4632 C3	PRJNA70743	-	5,37	5.201		
Escherichia coli O104:H4 str. 11-4632 C4	PRJNA70745	-	5,39	5.234		
Escherichia coli O104:H4 str. 11-4632 C5	PRJNA70747	-	5,39	5.246		
Escherichia coli O104:H4 str. C227-11	PRJNA68153	-	5,39	5.225		
Escherichia coli O104:H4 str. C227-11	PRJNA68253	-	5,54	5.536		
Escherichia coli O104:H4 str. C236-11	PRJNA68155	-	5,4	5.231		
Escherichia coli O104:H4 str. Ec11-5536	PRJNA70753	-	-	-		
Escherichia coli O104:H4 str. Ec11-5537	PRJNA70751	-	-	-		
Escherichia coli O104:H4 str. Ec11-5538	PRJNA70749	-	-	-		
Escherichia coli O104:H4 str. GOS1	PRJNA71263	-	5,31	-		
Escherichia coli O104:H4 str. GOS2	PRJNA71339	-	5,31	-		
Escherichia coli O104:H4 str. H112180280	PRJNA67929	-	5,38	-		
Escherichia coli O104:H4 str. H112180282	PRJNA68661	-	5,37	-		
Escherichia coli O104:H4 str. H112180283	PRJNA70805	-	5,34	-		
Escherichia coli O104:H4 str. H112180540	PRJNA70809	-	5,28	-		
Escherichia coli O104:H4 str. H112180541	PRJNA68163	-	5,05	-		
Escherichia coli O104:H4 str. LB226692	PRJNA67613	-	5,46	5.585		
Escherichia coli O104:H4 str. ON2010	PRJNA81621	-	5,1	-		
Escherichia coli O104:H4 str. ON2011	PRJNA81623	-	5,23	-		
Escherichia coli O104:H4 str. TY-2482	PRJNA67657	-	5,29	-		
Escherichia coli O111 str. CVM9455	PRJNA129407	-	6	6.546		
Escherichia coli O111:H11 str. CVM9534	PRJNA129415	-	5,46	5.738		
Escherichia coli O111:H11 str. CVM9545	PRJNA89667	-	5,61	6.164		
Escherichia coli O111:H11 str. CVM9553	PRJNA129417	-	5,59	5.987		
Escherichia coli O111:H8 str. CVM9570	PRJNA129413	-	5,51	5.884		
Escherichia coli O111:H8 str. CVM9574	PRJNA129397	-	5,36	5.633		
Escherichia coli O111:H8 str. CVM9602	PRJNA129393	-	5,1	5.176		
Escherichia coli O111:H8 str. CVM9634	PRJNA129401	-	5,77	6.081		
Escherichia coli O113:H21 str. CL-3	PRJNA72243	-	5,05	-		
Escherichia coli O121:H19 str. MT#2	PRJNA72249	-	5,26	-		
Escherichia coli O145:H28 str. 4865/96	PRJNA72253	-	5,23	-		
Escherichia coli O157:H- str. 493-89	PRJNA60059	-	5,05	4.946	p	ehec (isolated from German infant human, from HUS outbreak 1989)
Escherichia coli O157:H- str. 493-89	PRJNA72247	-	5,36	-		
Escherichia coli O157:H- str. H 2687	PRJNA60061	-	5,05	4.938		
Escherichia coli O157:H43 str. T22	PRJNA81821	-	4,33	4.670		
Escherichia coli O157:H7 str. 1044	PRJEA61463	-	5,49	5.580		
Escherichia coli O157:H7 str. 1125	PRJNA61473	-	5,57	5.683		
Escherichia coli O157:H7 str. EC1212	PRJEA61465	-	5,51	5.593		

Escherichia coli O157:H7 str. EC4009	PRJNA42809	-	5,19	-		
Escherichia coli O157:H7 str. EC4024	PRJNA54969, PRJNA27747	-	6,2	6.540		
Escherichia coli O157:H7 str. EC4042	PRJNA54961, PRJNA27737	-	5,62	5.976		
Escherichia coli O157:H7 str. EC4045	PRJNA54957, PRJNA27733	-	5,63	5.930		
Escherichia coli O157:H7 str. EC4076	PRJNA54967, PRJNA27745	-	5,71	6.071		
Escherichia coli O157:H7 str. EC4084	PRJNA42813	-	5,3	-		
Escherichia coli O157:H7 str. EC4113	PRJNA54965, PRJNA27743	-	5,66	5.826		
Escherichia coli O157:H7 str. EC4127	PRJNA42815	-	5,3	-		
Escherichia coli O157:H7 str. EC4191	PRJNA42817	-	5,2	-		
Escherichia coli O157:H7 str. EC4192	PRJNA42811	-	5,35	-		
Escherichia coli O157:H7 str. EC4196	PRJNA54963, PRJNA27741	-	5,62	5.891		
Escherichia coli O157:H7 str. EC4205	PRJNA42819	-	5,29	-		
Escherichia coli O157:H7 str. EC4206	PRJNA54959, PRJNA27735	-	5,63	6.062		
Escherichia coli O157:H7 str. EC4401	PRJNA54971, PRJNA27749	-	5,73	6.001		
Escherichia coli O157:H7 str. EC4486	PRJNA54973, PRJNA27751	-	5,93	6.189		
Escherichia coli O157:H7 str. EC4501	PRJNA54975, PRJNA27753	-	5,68	5.947		
Escherichia coli O157:H7 str. EC508	PRJNA54977, PRJNA27755	-	5,66	5.853		
Escherichia coli O157:H7 str. EC536	PRJNA42821	-	5,26	-		
Escherichia coli O157:H7 str. EC869	PRJNA54979, PRJNA27757	-	5,73	5.926		
Escherichia coli O157:H7 str. FRIK2000	PRJNA55943, PRJNA36543	-	5,41	5.383		
Escherichia coli O157:H7 str. FRIK966	PRJNA55561, PRJNA32275	-	5,38	5.393		
Escherichia coli O157:H7 str. G5101	PRJNA60057	-	5,06	4.988		
Escherichia coli O157:H7 str. LSU-61	PRJNA60067	-	5,05	4.939		
Escherichia coli O25b:H4-ST131 str. EC958	PRJEA61443	-	5,16	-		
Escherichia coli O26:H11 str. CVM10021	PRJNA129425	-	5,49	5.752		
Escherichia coli O26:H11 str. CVM10026	PRJNA129427	-	5,57	5.869		
Escherichia coli O26:H11 str. CVM10030	PRJNA129429	-	5,49	5.743		
Escherichia coli O26:H11 str. CVM10224	PRJNA129431	-	5,34	5.581		
Escherichia coli O26:H11 str. CVM9942	PRJNA129423	-	5,62	5.991		
Escherichia coli O26:H11 str. CVM9952	PRJNA94343	-	5,5	5.701		
Escherichia coli O32:H37 str. P4	PRJNA91125	-	4,87	4.693		
Escherichia coli O45:H2 str. 03-EN-705	PRJNA72251	-	5,3	-		
Escherichia coli O55:H7 str. 3256-97	PRJNA60063	-	5,08	4.995		
Escherichia coli O55:H7 str. USDA 5905	PRJNA60065	-	5,11	4.980		
Escherichia coli O91:H21 str. B2F1	PRJNA72245	-	5,01	-		
Escherichia coli OK1180	PRJNA40289	-	5,55	5.910		
Escherichia coli OK1357	PRJNA40291	-	5,28	5.341		
Escherichia coli OP50	PRJNA49051, PRJNA41499	-	4,42	6.145		
Escherichia coli PA10	PRJNA65853	-	5,57	6.081		
Escherichia coli PA14	PRJNA65855	-	5,47	5.976		
Escherichia coli PA15	PRJNA65857	-	5,44	5.884		
Escherichia coli PA22	PRJNA65859	-	5,49	5.752		
Escherichia coli PA24	PRJNA65863	-	5,32	5.684		

Escherichia coli PA25	PRJNA65867	-	5,37	5.769		
Escherichia coli PA28	PRJNA65869	-	5,4	5.804		
Escherichia coli PA3	PRJNA65841	-	5,37	5.805		
Escherichia coli PA31	PRJNA65873	-	5,37	5.735		
Escherichia coli PA32	PRJNA65875	-	5,37	5.810		
Escherichia coli PA33	PRJNA65877	-	5,35	5.752		
Escherichia coli PA39	PRJNA65883	-	5,36	5.827		
Escherichia coli PA40	PRJNA65885	-	5,53	5.753		
Escherichia coli PA41	PRJNA65887	-	5,48	5.949		
Escherichia coli PA42	PRJNA65889	-	5,39	5.806		
Escherichia coli PA5	PRJNA65845	-	5,35	5.708		
Escherichia coli PA9	PRJNA65851	-	5,42	5.797		
Escherichia coli PCN033	PRJNA64999	-	5,06	4.966		
Escherichia coli RN587/1	PRJNA40279	-	5,06	5.108		
Escherichia coli SCI-07	PRJNA82801	-	4,97	4.758		
Escherichia coli STEC_7v	PRJNA48269	-	5,2	5.220		
Escherichia coli STEC_94C	PRJNA48271	-	5,03	5.244		
Escherichia coli STEC_B2F1	PRJNA48273	-	4,99	5.006		
Escherichia coli STEC_C165-02	PRJNA48275	-	5,01	5.019		
Escherichia coli STEC_DG131-3	PRJNA48277	-	5,3	5.609		
Escherichia coli STEC_EH250	PRJNA48279	-	5,21	5.353		
Escherichia coli STEC_H.1.8	PRJNA48281	-	5,39	5.815		
Escherichia coli STEC_MHI813	PRJNA48285	-	5,23	5.185		
Escherichia coli STEC_O31	PRJNA48267	-	5,2	5.250		
Escherichia coli STEC_S1191	PRJNA48287	-	5,15	5.260		
Escherichia coli TA007	PRJNA39063	-	5,25	5.217		
Escherichia coli TA124	PRJNA39075	-	4,92	4.714		isolated from mammal (*Dasyurus geoffroii* - western quoll/chudchit, marsupial)
Escherichia coli TA143	PRJNA52477, PRJNA39079	-	4,83	4.856		isolated from mammal (*Antechinus bellus* - fawn antechinus, a small carnivorous marsupial)
Escherichia coli TA206	PRJNA52479, PRJNA39085	-	5,06	5.177		isolated from mammal (*Perameles nasuta* - long-nosed bandicoot, marsupial)
Escherichia coli TA271	PRJNA52521, PRJNA39091	-	5,02	5.197		isolated from mammal (*Dasyurus viverrinus* - eastern quoll/eastern native cat, marsupial)
Escherichia coli TA280	PRJNA52523, PRJNA39093	-	5,26	5.477		isolated from mammal (*Dasyurus viverrinus* - eastern quoll/eastern native cat, marsupial)
Escherichia coli TW06591	PRJNA65899	-	5,48	5.650		
Escherichia coli TW07793	PRJNA51133	-	4,64	4.595		
Escherichia coli TW07945	PRJNA65901	-	5,36	5.701		
Escherichia coli TW09098	PRJNA65907	-	5,49	5.911		
Escherichia coli TW09109	PRJNA65909	-	5,58	5.881		
Escherichia coli TW09195	PRJNA65911	-	5,47	6.084		
Escherichia coli TW10119	PRJNA65913	-	5,55	5.873		
Escherichia coli TW10246	PRJNA65903	-	5,45	5.681		
Escherichia coli TW10509	PRJNA39103	-	5,35	5.149		
Escherichia coli TW10598	PRJNA59743	-	5,24	-		
Escherichia coli TW10722	PRJNA59745	-	5,69	-		
Escherichia coli TW10828	PRJNA59747	-	5,28	-		
Escherichia coli TW11039	PRJNA65905	-	5,6	5.955		
Escherichia coli TW11681	PRJNA59749	-	5,31	-		
Escherichia coli TW14301	PRJNA65921	-	5,3	5.649		
Escherichia coli TW14313	PRJNA65919	-	5,33	5.722		
Escherichia coli TW14425	PRJNA59751	-	5,21	-		
Escherichia coli TX1999	PRJNA40287	-	5,25	5.465		
Escherichia coli W	PRJNA52603, PRJNA42709	-	4,94	4.772		
Escherichia coli W26	PRJNA88641	-	5,12	4.920		
Escherichia coli WV_060327	PRJNA61477	-	4,68	4.618		
Escherichia coli XH001	PRJNA70787	1	4,84	4.716		
Escherichia coli XH140A	PRJNA70725	1	4,41	4.246		
Escherichia coli cloneA_i1	PRJNA63489	-	4,76	4.727		

Supplementary Table S4. Chromosome size (Kb's) of the Mexican *E.coli* phylogenetic groups defined by the Clermont assignation method

Clermont group assignation	N	Mean	Standard deviation	Standard error	95% Confidence Interval		Minimum	Maximum
					Lower limit	Upper limit		
A	47	4.822,602	241,3156	35,1995	4.751,749	4.893,455	4.247,3	5.391,8
B1	41	5.068,654	393,0560	61,3850	4.944,590	5.192,717	4.438,1	5.907,5
B2	16	4.677,038	363,7672	90,9418	4.483,200	4.870,875	4.007,0	5.427,0
D	22	5.033,732	525,2087	111,9749	4.800,867	5.266,596	4.158,0	6.097,1
Total	126	4.921,046	392,0905	34,9302	4.851,915	4.990,177	4.007,0	6.097,1

Supplementary Table S5. Chromosome size of sequenced *Escherichia coli* strains isolated from animals

Organism	BioProject	Plasmids	Size (Mb)	Gene	Host
Escherichia coli 1.2741	PRJNA51085	-	5,69	5.914	Cow (O2:H4)
Escherichia coli 2.3916	PRJNA51123	-	5,64	6.179	Pig
Escherichia coli 2.4168	PRJNA51127	-	4,74	4.804	Water
Escherichia coli 3.2303	PRJNA51129	-	4,95	5.149	Water
Escherichia coli 3.2608	PRJNA51105	-	5,45	5.670	Horse
Escherichia coli 3.3884	PRJNA51125	-	5,23	5.311	Cow
Escherichia coli 3003	PRJNA51131	-	4,92	4.982	Water
Escherichia coli 4.0522	PRJNA51109	-	5,83	6.308	Cow
Escherichia coli 4.0967	PRJNA51121	-	5,87	6.266	Rabbit
Escherichia coli 5.0588	PRJNA51089	-	5	5.137	Cow
Escherichia coli AA86	PRJNA65321	1	4,98	4.819	Healthy cow feces
Escherichia coli AI27	PRJNA89369	-	4,9	4.750	Porcine feces in South Korea
Escherichia coli B088	PRJNA47003, PRJNA38905	-	4,94	4.671	Bird (*Circus assimilis*)
Escherichia coli B093	PRJNA38907	-	5,19	4.990	Bird (*Strepera graculina*)
Escherichia coli B185	PRJNA47001, PRJNA38915	-	5,11	4.797	Bird (*Turdus merula*)
Escherichia coli B354	PRJNA46999, PRJNA38917	-	4,83	4.724	Bird (*Monarca trivirgatus*)
Escherichia coli B799	PRJNA38929	-	5,3	5.203	Bird (*Sericornis frontalis*)
Escherichia coli E101	PRJNA38935	-	5,18	5.015	Environment (Lake Ginnderra, ACT, Australia)
Escherichia coli M605	PRJNA52483, PRJNA39039	-	5,45	5.750	Mammal (*Nyctophilus geoffroyi*-bat)
Escherichia coli M718	PRJNA52481, PRJNA39043	-	5,37	5.514	Mammal (*Sarcophilus harrisii* - Tasmanian devil)
Escherichia coli M919	PRJNA39047	-	5,25	5.136	Mammal (*Sarcophilus harrisii* - Tasmanian devil)
Escherichia coli TA124	PRJNA39075	-	4,92	4.714	Mammal (*Dasyurus geoffroii* - western quoll/chudchit, marsupial)
Escherichia coli TA143	PRJNA52477, PRJNA39079	-	4,83	4.856	Mammal (*Antechinus bellus* - fawn antechinus, a small carnivorous marsupial)

Escherichia coli TA206	PRJNA52479, PRJNA39085	-	5,06	5.177	Mammal (*Perameles nasuta* - long-nosed bandicoot, marsupial)
Escherichia coli TA271	PRJNA52521, PRJNA39091	-	5,02	5.197	Mammal (*Dasyurus viverrinus* - eastern quoll/eastern native cat, marsupial)
Escherichia coli TA280	PRJNA52523, PRJNA39093	-	5,26	5.477	Mammal (*Dasyurus viverrinus* - eastern quoll/eastern native cat, marsupial)

6. Artículo 3

"Dinámica evolutiva del pangenoma de *Escherichia coli* y adaptación a diferentes nichos ecológicos"

6.1. Resumen

El pangenoma de una especie bacteriana consiste en el genoma "central" (genes compartidos por todos los aislados) y el genoma flexible (genes compartidos por sub-poblaciones y genes específicos a determinadas cepas). Este último se encuentra en su mayoría compuesto por genes adquiridos horizontalmente, los cuales codifican nuevas funciones metabólicas. Debido a lo anterior, se asume que el genoma flexible es el que permite a las bacterias adaptarse a nuevos nichos, ignorando así el papel del genoma central en el origen y evolución de adaptaciones tales como la patogénesis y el comensalismo. Para examinar este supuesto, llevamos a cabo un estudio de genómica de poblaciones en *Escherichia coli*, bacteria con diferentes estilos de vida. Al comparar la dinámica evolutiva tanto del genoma central como del flexible de los dos "eco-grupos" en las cuales se dividió la muestra, encontramos que las cepas patógenas presentan mayor diversidad genética que las cepas no patógenas, no solamente a nivel del genoma flexible sino también a nivel del genoma central. Asimismo, encontramos pocos loci en desequilibrio de ligamiento lo que sugiere a *E. coli* como especie sexual. Con respecto a los patrones de la selección natural, las cepas patógenas mostraron señales de selección positiva en genes con diversas funciones (genes de transporte/unión, metabolismo energético y transcripción), al contrario de las cepas no patógenas en las cuales, predominó la selección negativa. Finalmente, encontramos que los niveles de diversidad genética, clonalidad y selección son heterogéneos a lo largo del cromosoma. Solamente pocas regiones del cromosoma mostraron patrones similares en ambos ecogrupos. Así, concluimos que la adaptación de *E. coli* a diferentes nichos ocurre no solamente por la adquisición horizontal de genes sino también por la evolución del genoma central.

Dinámica evolutiva del pangenoma de *Escherichia coli* y adaptación a diferentes nichos ecológicos

Sánchez-Reyes, L[1,a]., González-González, A[1]., Eguiarte, L[1]., Souza, V[1]

[1] Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Apartado Postal 70-275, México, D.F. 04510, México

[a] Dirección actual: Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, D.F., México.

Running title: Population genomics of *E. coli*

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). Given that flexible genome is mainly composed by horizontally acquired genes encoding new metabolic functions, the idea that the flexible genome is what allows bacterial adaptation to new niches is pervasive, thus ignoring the role of the core genome in the origin and evolution of adaptations, such as pathogenesis and commensalism. To evaluate this assumption, we conducted a population genomics study in *Escherichia coli*, a bacterial species with different lifestyles. 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. 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 syntenic 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 eco-groups. We conclude that adaptation of *E. coli* to different niches occurs not only by the horizontal acquisition of genes, but also by the evolution of the core genome and its regulation.

1. Introducción

La comparación de genomas entre diferentes especies y dentro de individuos de la misma especie, ha permitido la identificación de mecanismos genéticos tales como la duplicación, pérdida y ganancia de genes así como de re-arreglos cromosómicos los cuales en su conjunto promueven la variabilidad tanto en el contenido y estructura de los genomas (Mira et al. 2002). Es así que para describir el total de la información genética de una especie se utiliza el concepto de pangenoma el cual se encuentra conformado principalmente por un conjunto de genes compartidos por todas las cepas a la que se le llama genoma central y por un conjunto de genes o regiones genómicas presentes solamente en ciertos aislados al que se le llama genoma flexible (Tettelin et al. 2005).

Generalmente se asume que el genoma central es aquel que codifica para las funciones metabólicas básicas comunes al nicho ecológico y que el genoma flexible producto de transferencia horizontal de genes principalmente, codifica las funciones especializadas las cuales en determinados casos aumentan la adecuación en ambientes particulares constituyendo así un mecanismo de adaptación bacteriana (Dobrindt et al. 2004). Es así que 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, sobre todo para explicar la existencia de bacterias que son generalmente comensales del humano, pero que también pueden llegar a ser patógenas al adquirir genes de virulencia, como es el caso de *Escherichia coli* (Nataro y Kaper 1998).

Asimismo, la comparación de genomas de diferentes patotipos ha confirmado que cepas comensales comparten gran parte de los factores de virulencia (Rasko et al. 2008) lo que sugiere 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 y adecuación. Entonces la adquisición de ciertos factores de virulencia no implican forzosamente la transformación inmediata de una cepa a-virulenta 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?

Para responder a esta pregunta, realizamos un estudio de genómica de poblaciones en donde analizamos la dinámica evolutiva de los diferentes componentes del pangenoma arrojado por 12 genomas completos de *Escherichia coli* representativos de los diferentes estilos de vida: comensales, de vida libre, patógena extra-intestinal y patógena intestinal.

Al comparar el genoma de diferentes aislados de una misma especie, la genómica de poblaciones permite estudiar a fondo el papel que juegan tanto la selección natural como la deriva génica, mutación y recombinación en la estructuración de las poblaciones. Permitiendo así la identificación y separación de los efectos que tienen en la divergencia de las poblaciones bacterianas, los procesos históricos promovidos por la deriva génica, de los contemporáneos o ecológicos, moldeados por la selección natural (Luikart et al., 2003). Gracias a lo cual, se vuelve factible el identificar los patrones evolutivos que subyacen a la adecuación y adaptación a nuevos nichos (Black et al, 2001; Luikart et al., 2003; Nadeau y Jiggins, 2010).

2. Material y Métodos

2.1. Organismo de estudio

Escherichia coli es un organismo que se ha usado como modelo en muchos ámbitos de la biología. Las cepas asociadas a hospedero son las que se han estudiado con mayor profundidad y se conocen mejor debido a que pueden ser patógenas tanto a nivel intra- como extra-intestinal. 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 por lo que tradicionalmente son consideradas como simbiontes comensales (Kaper et al. 2004).

Desde el punto de vista médico, se han identificado principalmente 6 grupos de cepas de *E. coli* que comparten un proceso de patogénesis similar y que comparten un conjunto de características fenotípicas y de factores genéticos (de virulencia) a los que se les llama patotipos (Kaper et al. 2004). A pesar de que cada uno de estos patotipos despliega un patrón de adherencia y de toxicidad particular, existe cierta superposición de los factores de virulencia entre los mismos (Rasko et al. 2008). Además, se han descrito

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 adecuado para estudiar los mecanismos evolutivos a nivel genómico, que dan origen a los diferentes estilos de vida y que permiten la adaptación a distintos nichos ecológicos.

2.2. Muestra

En la Tabla 1 se enlistan los números de acceso de las 12 cepas de *Escherichia coli* que se utilizaron en este análisis. La muestra se encuentra conformada por aislados representativos de diferentes estilos de vida: comensales (4 cepas), de vida libre (1 cepa), patógena extra-intestinal (4 cepas) y patógena intestinal (3 cepas) cuyos genomas se encuentran secuenciados con una cobertura $\geq$ 5x, el mínimo necesario para obtener un ensamblado correcto y una mayor seguridad sobre la información de secuencia (Blattner et al. 1997; Rasko et al. 2008).

2.3. Alineación

Los cromosomas se alinearon con el algoritmo implementado en el programa MAUVE versión 2.1.1 para Linux (Darling et al. 2004). Este método de alineación múltiple utiliza comparaciones locales y globales para identificar regiones ortólogas conservadas y delimitar puntos de quiebre exactos de re-arreglos de secuencia a las que se les llama sub-alineación o bloques localmente colineares (LCBs; Locally Colinear Blocks). Gracias a lo cual, este método de alineación permite identificar eventos de re-arreglo cromosómico, inversiones y translocaciones, al detectar bloques de colinearidad local y su posición relativa en cada genoma (Darling et al. 2004). Para detectar errores en el alineamiento, se realizó una revisión "a mano" de cada sub-alineación, con ayuda del programa BIOEDIT (Hall 1999) versión 7.0.5.

2.4. 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 de regiones 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; 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.

2.5. Reconstrucción filogenética

La reconstrucción filogenética se realizó a partir de los datos del genoma central. Se infirió un árbol por cada uno de los LCBs con el programa PHYML versión 3.0 para Linux (Guindon y Gascuel 2003), el cual permite un análisis rápido de una gran cantidad de datos de secuencia, tanto en número de individuos como en la longitud de sus secuencias. En cada caso, se usó el modelo de sustitución nucleotídica GTR y se hicieron 1,000 réplicas de bootstrap.

2.6. Genómica de poblaciones

Se estimó diversidad con el parámetro de diversidad nucleotídica π (pi) y el parámetro de mutación poblacional θ (theta), se buscó evidencia de selección natural con 3 pruebas de neutralidad diferentes (D de Tajima; Tajima, 1983; D* y F* de Fu-Li; Fu & Li, 1993) y se estimaron niveles de desequilibrio de ligamiento con el estimador Zns (Kelly, 1997). Para no restringir el análisis a unidades funcionales, como son los genes, se tomó como unidad de análisis una longitud de 1000 pares de bases, a lo que se consideró como loci de análisis. Todos los estimados de genética de poblaciones se calcularon con el programa VARISCAN versión 2.0.2 (Vilella et al. 2005).

2.7. Análisis funcional y estadístico

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 1). Bajo este criterio, dividimos la muestra en dos grupos funcionales, a los que llamamos eco-grupos. 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 eco-grupo 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 estimados para el total de la muestra de *E. coli* y para cada eco-grupo por separado. Las comparaciones de parámetros estimados, entre componentes del pangenoma y entre eco-grupos 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 detectaran 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 eco-grupos se acumulaban en algún tipo particular de genes.

3. Resultados y Discusión

3.1. Variación en la estructura cromosómica de E. coli

El número de bloques de colinearidad local (LCBs por sus siglas en inglés: Locally Colinear Blocks) nos da un aproximado del número de eventos de re-arreglo cromosómico que han ocurrido en una muestra de genomas. En nuestra muestra de *E. coli* encontramos evidencia de al menos 21 eventos de re-arreglo cromosómico, lo cual es ligeramente menor a lo que se ha reportado para muestras de *E. coli* y *Shigella* (34 LCBs en 6 cepas de *Shigella*; Mau et al. 2006) (Figura 1; Tabla S1). Esta diferencia se puede explicar por la presencia de un número elevado de secuencias de inserción en cepas de *Shigella*, cuya presencia favorece la ocurrencia de re-arreglos cromosómicos (Touchon et al. 2009).

El número de re-arreglos que arroja la muestra de *E. coli* aquí estudiada sugiere que el cromosoma de *E. coli* es tan plástico como el del patógeno oportunista *Neisseria meningitidis* (10 LCBs en 20 cepas; Budroni et al., 2011). Pero más recombinante 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 oportunistas como *Legionella pneumophila* (16 LCBs en 5 cepas; D'Auria et al. 2010), el género *Yersinia* (98 LCBs en 11 cepas; Chen et al. 2010; Darling et al. 2008), así como en la bacteria de vida libre *Rhodobacter sphaeroides* (382 LCBs en 3 cepas; Choudhary et al. 2007).

Estas observaciones nos sugieren 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, lo que se puede explicar por presiones selectivas que mantienen órdenes particulares de las regiones genómicas, y que eliminan aquellas cepas con re-arreglos que no son adaptativos (Mira et al., 2002).

La ausencia de diferencias en la estructura cromosómica de bacterias patógenas de *E. coli* como es el caso de los aislados enterohemorrágicos, y en particular el serotipo O157:H7 (Tabla 1) puede ser explicado por la estructura epidémica de esta especie (Maynard-Smith et al. 1993), en donde determinados linajes clonales se expanden al interior de las poblaciones. A la larga, ésta situación aumenta la probabilidad de que el intercambio genético ocurra entre aislados idénticos generando así poca o nula variación en la estructura del cromosoma.

Al comparar el número de re-arreglos entre individuos de un mismo eco-grupo, observamos que resulta ser parecido en ambos eco-grupos (Tabla 2). Esto sugiere que la estructura del cromosoma se conserva al interior de los eco-grupos y que la frecuencia de los eventos de re-arreglos cromosómicos no se encuentra asociado al estilo de vida. Sin embargo, el número de re-arreglos es mayor cuando se comparan cepas de diferente eco-grupo los cuales llegan a ser hasta 9 entre las cepas K12 8739ATCC y UTI89 (Tabla 2). La única excepción notable es el número de re-arreglos (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.

En general, los re-arreglos 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), las cuales fueron detectadas alrededor del origen y término de replicación, hecho que se considera como 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 para conservar procesos celulares básicos como la replicación (Darling et al. 2008; Touchon et al. 2009).

Entre las cepas patógenas, los re-arreglos 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 1). Entonces la estructura del cromosoma es cohesiva dentro de los eco-grupos, pero diferente entre ellos. Es decir que la naturaleza de los eventos de re-arreglo cromosómico es similar entre individuos con estilos de vida semejantes. Sin embargo, aparentemente la frecuencia con la que ocurren los eventos de re-arreglo es parecida en ambos eco-grupos. 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 re-arreglos entre diferentes especies (Mira et al. 2002). Dado que el presente estudio describe la estructura cromosómica dentro de una misma especie, la cohesividad estructural refleja la cohesividad de la especie, como ha sido sugerido por Touchon et al. (2009).

Finalmente, a pesar de la aparente homogeneidad en la estructura del cromosoma entre individuos del mismo eco-grupo, los re-arreglos observados entre aislados 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 (Tabla 2), así como en el trabajo de Ferenci et al. (2009), nos indica que el cromosoma de estas bacterias posee un alto potencial de diversidad estructural.

3.2. 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

279 más divergentes sean dos individuos ó grupos de individuos, la proporción de genes que
280 compartirán será menor (Mushegian y Koonin 1996). Dado que las cepas de *Shigella*
281 constituyen un subgrupo parafilético de *E. coli* y altamente divergente del resto de las cepas
282 de esta especie (Pupo et al. 2000), se espera que la comparación de contenido genómico
283 resulte en una menor cantidad de genes compartidos entre *E. coli* y *Shigella* que entre
284 individuos de *E. coli* solamente.
285
286 Por otro lado, encontramos que las regiones del genoma flexible, producto de
287 eventos de transferencia horizontal ó de deleción genómica, son abundantes y representan
288 la mayor parte del pangenoma de esta muestra de *E. coli* (Figura S1). Esto, aunado a los
289 altos niveles de identidad nucleotídica encontrados en el genoma central (≥ 0.96; Tabla S2)
290 apoyarían la idea de que la transferencia horizontal y el flujo génico son los principales
291 procesos que generan la diversidad en la especie (Darling et al. 2004).
292
293 Al comparar los elementos del genoma flexible compartidos por los individuos de
294 cada eco-grupo, encontramos que las cepas no-patógenas tienen mayor número de regiones
295 del genoma flexible que las patógenas (3.3% contra 1.3%). Este resultado pareciera
296 contradecir a la teoría clásica de evolución de la patogénesis, pues se ha sugerido que son
297 los elementos del genoma flexible los que han permitido a las bacterias la adaptación al
298 nicho patógeno. En ese caso las cepas patógenas deberían poseer una poza de genes
299 flexibles más amplia que aquellas cepas que carecen de potencial patogénico. Sin embargo,
300 no encontramos tal patrón. Esto pudiera ser explicado por dos motivos. Por una parte los
301 elementos genéticos que permiten la patogénesis en *E. coli* son muy diversos y las
302 combinaciones de éstos elementos que permiten la colonización de diferentes regiones de
303 un hospedero de manera exitosa son muy grandes (Kuhnert et al. 2000; Tenaillon et al.
304 2010), por lo que en teoría no se necesita un repertorio genético particular para que las
305 bacterias posean capacidad patogénica. Asimismo, se ha propuesto que las cepas patógenas
306 pueden usar diferentes genes con funciones similares para el proceso de infección (Mokady
307 et al. 2005).
308

Por otra parte, hay numerosos estudios de genómica comparada y de evolución experimental han encontrado que las cepas comensales comparten también 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 de elementos genéticos 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 (Rasko et al. 2008), debemos preguntarnos si existen otras regiones del genoma que promuevan la patogénesis en *E. coli*.

3.3. 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* ($\pi = 0.00010$; $\pi = 0.00847$; Nübel et al. 2008 y Takuno et al., 2012 respectivamente), *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, ya 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 ($V(\pi)=0.00032863$; Tabla S3). Los valores abarcaron un rango muy amplio, desde diversidad nula ($\pi = 0$) en varios loci tanto codificantes como no-codificantes, hasta una diversidad de $\pi = 0.224242$ 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 $\pi = 0.27$ (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 $\pi = 0.24$ (Castillo et al. 2005).

3.4. Patrones de desequilibrio de ligamiento: estructura clonal ó panmíctica?

La prueba de desequilibrio de ligamiento Zns se utiliza como una medida del grado de clonalidad que puede existir en una población (Maynard- Smith 1993) ya que se ha visto que es inversamente proporcional a la presencia de recombinación (Dawson et al. 2002; Hedrick 2005). Cuando los valores son significativamente diferentes de 0 indican presencia de desequilibrio de ligamiento, i.e. clonalidad. Valores de 0 indican ausencia de desequilibrio de ligamiento i.e. recombinación. En nuestra muestra de *E. coli*, de los 4,122

loci del genoma central analizados, solamente 202 loci fueron significativamente diferentes de 0 en la prueba Zns (promedio Zns=0.3642279 ± 0.0020288). Estos 202 loci representan el 4.9% de loci del genoma central. Es decir que solamente una porción muy pequeña del genoma central está en desequilibrio de ligamiento (i.e., es clonal) (Figura 2 d; Tabla S3). Utilizando otra aproximación estadística, Didelot y colaboradores (2012) sugieren que la tasa de recombinación es casi homogénea a lo largo del genoma de *E. coli* a excepción de tres sitios donde la tasa de recombinación es mucho mayor (un sitio alrededor del operón *rfb*, un sitio alrededor de *fimA* y un sitio alrededor de *aroC*). Asimismo, 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).

3.5. Selección natural a nivel molecular

Bajo el modelo de evolución neutral de Kimura (1989), 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 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* (Figura S2).

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* (Charlesworth y Eyre-Walker 2006; Chen et al. 2006; Petersen et al. 2007; Chattopadhay et al. 2009), en la bacteria patógena *Listeria monocytogenes* (Orsi et al. 2008) y en los endosimbiontes de insectos *Buchnera sp.* y *Blochmannia sp.* (Toft et al. 2009).

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 (Tabla S3). 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*, 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).

Asimismo, 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 S2, Figura S3A y S3B).

3.6. Diversidad ecológica a nivel molecular en E. coli

Los dos eco-grupos aquí definidos presentaron una estructura cromosómica diferente entre sí, aunque conservada al interior de cada grupo (Tabla 2). A un nivel más fino, la diversidad genética también fue diferente (Tabla 3). El eco-grupo de cepas patógenas registró mayor diversidad tanto en el genoma central como en el genoma flexible con respecto al eco-grupo de las cepas no-patógenas. El rango de diversidad nucleotídica π en los loci del genoma central fue similar en ambos eco-grupos, 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 a 0.214285 en las no-patógenas, y de 0 a 0.36 en las patógenas. Los valores máximos, tanto de genoma central como de genoma flexible de ambos eco-grupos, 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. Estas regiones altamente variables funcionarían como anclas a la recombinación, tanto homóloga como ilegítima, promoviendo la acumulación de nuevos alelos y de diversidad a comparación de otras regiones del genoma (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 amplio 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 cromosómicos. 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, es necesario más adelante, 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; 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 eco grupo de *E. coli* patógenas fue significativamente menos clonal que el eco grupo de *E. coli* no-patógenas (Tabla 3; Figura 2d). 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 (Hendrickson y Lawrence 2006). Esto se puede explicar desde un punto de vista 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 la selección lo que está generando diferentes patrones en la diversidad del genoma central en donde la selección natural promueve la diversificación de este componente del genoma en las cepas patógenas a comparación de las cepas no patógenas (Figura 2a y 2b, Figura 3A y 3B; Figura 4A y 4B).

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

555 la disponibilidad elevada de recursos en el ambiente (Wedekind et al. 2010), sobre todo si
556 el tamaño efectivo es grande (Stevens et al. 2007). En el caso de *E. coli* la capacidad de
557 colonizar y vivir en ambientes nuevos distintos al colon como el sistema urinario y sistema
558 nervioso, pulmones, en donde hay menor competencia por los recursos y por lo tanto mayor
559 disponibilidad de ellos, es lo que probablemente favoreció el surgimiento de las cepas
560 patógenas y lo que ha permitido que permanezcan evolutivamente, como grupos diferentes
561 de las comensales.
562
563 Adicionalmente, dentro del eco-grupo de las cepas patógenas analizamos la
564 diversidad del subgrupo de cepas patógenas intestinales y del subgrupo de patógenas extra-
565 intestinales por separado. Encontramos que la diversidad fue mayor en el grupo de *E. coli*
566 patógenas extra- intestinales (π = 0.0125736 ± 0.0002437) que en las cepas patógenas
567 intestinales (π = 0.0090948 ± 0.0001952). Chattopadhay et al. (2009) reportan valores de
568 diversidad promedio de genes en los genomas de cepas de UPEC (CFT, 536, UTI89) y *E.*
569 *coli* patógena de ave (APEC01) de 0.004 ± 0.001, menor a lo que nosotros encontramos en
570 todo el genoma central, que incluye no solamente genes sino también regiones no-
571 codificantes, solamente en las cepas CFT, 536 y UTI89. De acuerdo a la clasificación
572 filogenética tradicional de *E. coli* (Selander et al. 1986), las cepas extra-intestinales se
573 encuentran en el grupo B2 (Johnson et al. 2006a, 2006b; Touchon et al. 2009), y las 3 cepas
574 que usamos en el presente estudio están asignadas a dicho grupo (Chen et al. 2006). De
575 acuerdo a estudios recientes (Jaureguy et al. 2008; Touchon et al. 2009), este linaje B2 sería
576 el más antiguo, lo que puede explicar por qué las cepas de este grupo de *E. coli* patógena
577 son las que tienen la mayor diversidad acumulada. Además de esta explicación "histórica",
578 cabe señalar que las cepas urinarias poseen un gran número y diversidad de islas de
579 patogénesis, factores de virulencia y adecuación provenientes hasta de otras especies a lo
580 largo de su genoma (Le Gall et al. 2007; Rasko et al. 2008). Lo anterior pareciera indicar
581 que en estas cepas hay una mayor frecuencia de eventos de transferencia horizontal,
582 sugiriendo que son más recombinantes que las cepas intestinales. Sería interesante saber si
583 la recombinación es algo que ya estaba presente en estas cepas antes de que comenzaran a
584 invadir el nicho de patogénesis extra-intestinal, ó sí es una característica que surgió en
585 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 (EHEC y 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 conjunto particular de cepas, cuyo genotipo, al originarse y expandirse clonalmente en el tiempo y en el espacio (clonas epidémicas) reflejan un comportamiento diferente con respecto a *E. coli* 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*.

4. Conclusiones

Con este estudio proponemos que la adaptación al nicho patógeno de *E. coli* no está determinada únicamente por los elementos del genoma flexible sino que el genoma central juega un papel importante en este proceso debido a las huellas de la acción de selección natural y de la recombinación detectadas lo que sugiere un proceso de adaptación ecológica a la patogénesis a nivel de los elementos del genoma central. Asimismo, encontramos que los patrones de diversidad del genoma central de *E. coli* no concuerdan en su totalidad con el modelo de evolución neutral del genoma sino que existe un dinámica diferencial entre la acción de la selección natural y la recombinación homóloga a lo largo del genoma. Sin embargo, es necesario en futuros análisis evaluar también el papel de la deriva génica -en términos de la variación en el tamaño efectivo poblacional- en la evolución de la patogénesis.

5. Agradecimientos

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. Al Dr. José Luis Puente García, Dr. Arturo Carlos II Becerra Bracho y Dr. León Patricio Martínez Castilla por los constructivos y fructíferos comentarios al presente estudio. Este trabajo fue financiado el proyecto PAPIIT/DGAPA IN219109. Durante la elaboración de este trabajo Sánchez-Reyes, L recibió la beca de ayudante SNI de Luis E. Eguiarte Fruns dentro del proyecto**:** "Transferencia horizontal en bacterias entéricas diarréicas y la evolución de la patogénesis".

6. Referencias

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.

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

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.

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

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.

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

Black, W.C. Baer, C.F., Antolin, M.F., DuTeau, N.M. 2001. Population genomics: genome-wide sampling of insect populations. Annu. Rev. Entomol. 46, 441-469.

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.

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

651 Budroni, S, E Siena, JC Dunning Hotopp, KL Seib, D Serruto, Ch Nofroni et al. 2011. *Neisseria meningitidis*
652 is structured in clades associated with restriction modification systems that modulate homologous
653 recombination. Proc. Natl. Sci. Acad. 108(11): 4494-4499.

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

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

658 Castillo A, LE Eguiarte, V Souza. 2005. A genomic population genetics analysis of the pathogenic enterocyte
659 effacement island in Escherichia coli: the search for the unit of selection. Proc Natl Acad Sci U S A.
660 102(5):1542-7.

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

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

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

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

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

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

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

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

679 Dawson E, GR Abecasis, S Bumpstead, Y Chen, S Hunt, DM Beare, J Pabial, T Dibling, E Tinsley, S Kirby,
680 D Carter, M Papaspyridonos, S Livingstone, R Ganske, E Lõhmussaar, J Zernant, N Tõnisson, M Remm, R
681 Mägi, T Puurand, J Vilo, A Kurg, K Rice, P Deloukas, R Mott, A Metspalu, DR Bentley, LR Cardon, I

682 Dunham. 2002. A first-generation linkage disequilibrium map of human chromosome 22. Nature.
683 418(6897):544-8.

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

685 Didelot X, Méric G, Falush D, Dariling AE. Impact of homologous and non-homologous recombination in the
686 genomic evolution of *Escherichia coli*. BMC Genomics, 13:256.

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

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

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

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

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

696 Dos Vultos T, O Mestre, J Rauzier, M Golec, N Rastogi, V Rasolofo, T Tonjum, C Sola, I Matic, B Gicquel.
697 2008. Evolution and Diversity of Clonal Bacteria: The Paradigm of Mycobacterium tuberculosis. PLoS ONE
698 3(2): e1538.

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

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

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

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

707 González-González A, Sánchez-Reyes LL, Delgado G, Eguiarte LE, Souza V. 2013. Hierarchical clustering
708 of genetic diversity associated to different levels of mutation and recombination in *Escherichia coli*: A study
709 based on Mexican isolates. Infection, Genetics and Evolution. 13, 187-197.

710 Groisman EA, H Ochman. 1996. Pathogenicity islands: bacterial evolution in quantum leaps. Cell 87: 791–
711 794.

712 Hacker J, JB Kaper. 2000. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol. 54:641–
713 679.

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

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

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

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

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

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

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

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

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

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

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

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

739 Kimura M. 1989. The neutral theory of molecular evolution and the world view of the neutralists. Genome.
740 31(1):24-31.

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

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

744 Lawrence JG. 2001. Catalyzing bacterial speciation: correlating lateral transfer with genetic headroom. Syst.
745 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.

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.

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.

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.

Maynard-Smith J, CG Dowson, BG Spratt. 1991. Localized sex in bacteria. Nature. 349(6304):29-31.

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.

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.

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.

Nadeau, N.J., Jiggins, C.D. 2010. A golden age for evolutionary genetics? Genomic studies of adaptation in natural populations. Trends Genet. 26, 484-492.

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

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

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

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

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

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

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

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

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

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

799 Roos V, GC Ulett, MA Schembri, P Klemm. 2006. The asymptomatic bacteriuria Escherichia coli strain
800 83972 outcompetes uropathogenic E. coli strains in human urine. Infect Immun. 74:615–624.

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

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

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

807 Savageau MA. 1983. Escherichia coli habitats, cell types, and molecular mechanisms of gene control. Am Nat
808 122:732–744.

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

811 Selander RK, BR Levin. 1980. Genetic diversity and structure in Escherichia coli populations. Science
812 210:545-547.

813 Selander RK, Caugant DA, Ochman H, Musser JM, Gilmour MN, Whittam TS. 1986. Methods of multilocus
814 enzyme electrophoresis for bacterial population genetics and systematics. Appl Environ Microbiol.
815 51(5):873–884.

816 Shapiro JA, Huang W, Zhang C, Hubisz MJ, Lu J, Turissini DA, Fang S, Wang HY, Hudson RR, Nielsen R,
817 Chen Z, Wu CI. 2007. Adaptive genic evolution in the Drosophila genomes. Proc Natl Acad Sci U S A.
818 104(7):2271-6.

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

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

824 Sokurenko EV, HS Courtney, J Maslow, A Siitonen, DL Hasty. 1995. Quantitative differences in
825 adhesiveness of type 1 fimbriated Escherichia coli due to structural differences in fimH genes. J. Bacteriol.
826 177:3680–3686.

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

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

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

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

836 Takuno S, Kado T, Sugino RP, Nakhleh L, Innan H. 2012. Population genomics in bacteria: a case study of
837 *Staphylococcus aureus*. Mol. Biol. Evol. 29 (2): 797-809.

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

840 Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL,
841 Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR,
842 Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC,
843 Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor

844 KJ, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels
845 MR, Rappuoli R, Fraser CM. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus
846 agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 102(39):13950-5.
847 Erratum in: Proc Natl Acad Sci U S A. 2005 Nov 8;102(45):16530.

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

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

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

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

862 Weissman SJ, Beskhlebnaya V, Chesnokova V, Chattopadhyay S, Stamm WE, Hooton TM, Sokurenko EV.
863 2007. Differential stability and trade-off effects of pathoadaptive mutations in the Escherichia coli FimH
864 adhesin. Infect Immun. 75(7):3548-55. Erratum in: Infect Immun. 2009 Apr;77(4):1720.

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

869 Whittam TS. 1989. Clonal dynamics of Escherichia coli in its natural habitat. Antonie Leeuwenhoek 55:23–
870 32.

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

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

875 Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H,
876 Achtman M. 2006. Sex and virulence in Escherichia coli: an evolutionary perspective. Mol Microbiol.
877 60(5):1136-51.

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

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

883 Yan F, DB Polk. 2004. Commensal bacteria in the gut: learning who our friends are. Curr Opin Gastroenterol.
884 20:565–571.

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

Información suplementaria

2. Material y Métodos

Alineación de los cromosomas de E. coli

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, las cuales usualmente fungen como anclas par el alineamiento. 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).

Para detectar 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 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 una $E \leq 10\text{-}70$ 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 (*sensu* Didelot et al. 2009). Las regiones alineadas incorrectamente se realinearon por separado con el programa CLUSTALW (versión 1.8.4) con los parámetros dados por omisión.

Referencias:

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

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

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

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.

Tabla 1. Genomas completos analizados en este estudio

Hábitat	Características ecológicas	Cepa/ plásmido	Grupo filogenético (de acuerdo a Wirth et al. 2006)	Serotipo	Tamaño del replicón (Mb)	Contenido GC	Número de plásmidos	Número de acceso NCBI	Referencia
Intestinal	Comensal Adaptada al laboratorio	K12 MG1655	A	N.E. O16	4.639675	50%	-	NC_000913	Blattner y col. 1997
		K12 W3110	A	N.E. O16	4.646332	50%	-	AC_000091	Hayashi y col. 2006
		K12 ATCC8739	A	N.E. O16	4.746218	50%	-	NC_010468	Joint Genome Institute 2008
	Comensal silvestre	HS	A	O9	4.643538	50%	-	NC_009800	Rasko y col. 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	50% 50% 46% 49% 43%	4	NC_010498	Fricke y col. 2008
Intestinal/ Extra-intestinal	Patógena APEC	APEC-O1 pAPEC-O1-R pAPEC-O1-ColBM	B2	O1:K1:H7	5.082025 0.241387 0.174241	50% 46% 49%	2	NC_008563	Johnson y col. 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	50% 47% 51% 50% 52% 49% 49%	6	NC_009801	Rasko y col. 2008
	Patógena EHEC	EDL933 pO157	-	O157:H7	5.528445 0.092077	50% 47%	1	NC_002655	Perna y col. 2001
	Patógena EHEC	Sakai pO157 pOSAK1	-	O157:H7	5.498450 0.092721 0.003306	50% 47% 43%	2	NC_002695	Hayashi y col. 2001
Extra-intestinal	Patógena, UPEC	UTI 536	B2	O6:K15:H31	4.938920	50%	-	NC_008253	Hochhut y col. 2006
	Patógena, UPEC	UTI CFT073	B2	O6:K2:H1	5.231428	50%	-	NC_004431	Welch y col. 2002
	Patógena UPEC	UTI89 pUTI89	B2	O18:K1:H7	5.065741 0.114230	50% 51%	1	NC_007946	Chen y col. 2006

N.E. No expresado (Stevenson y col. 1994).

Tabla 2. 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 eco-grupo.

Tabla 3. 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 el 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
	D de Tajima [2]	**-0.454711****	**-0.467283****
	D* de Fu-Li [2]	**-0.498445****	**-0.5052****
	F* de Fu-Li [2]	**-0.522249****	**-0.523444****
	Hd [3]	0.7501213	0.6338858
	Zns [3]	0.7282182*	0.7732863*
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
	D de Tajima [2]	**0.7365571****	**0.6695001****
	D* de Fu-Li [2]	**0.5570546****	**0.4874307****
	F* de Fu-Li [2]	**0.6593397****	**0.5790057****
	Hd [3]	0.8732007	0.6882265
	Zns [3]	0.6347493*	0.6694768*
TOTAL	N	*4,122*	-
	N_P	*4,096*	-
	N_{DL}	*4,037*	-
	π [1]	0.0217819	-
	θ [1]	0.0195899	-
	D de Tajima [2]	**0.4484568****	-
	D* de Fu-Li [2]	**0.3270199****	-
	F* de Fu-Li [2]	**0.4089824****	-
	Hd [3]	0.9079772	-
	Zns [3]	0.3642279*	

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

* loci significativos positivos en la prueba de neutralidad; * loci significativos negativos en la prueba de neutralidad; * loci significativos en la prueba de desequilibrio de ligamiento Zns.

Figura 1. Alineamiento múltiple del cromosoma de *E. coli*. 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 esta 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.

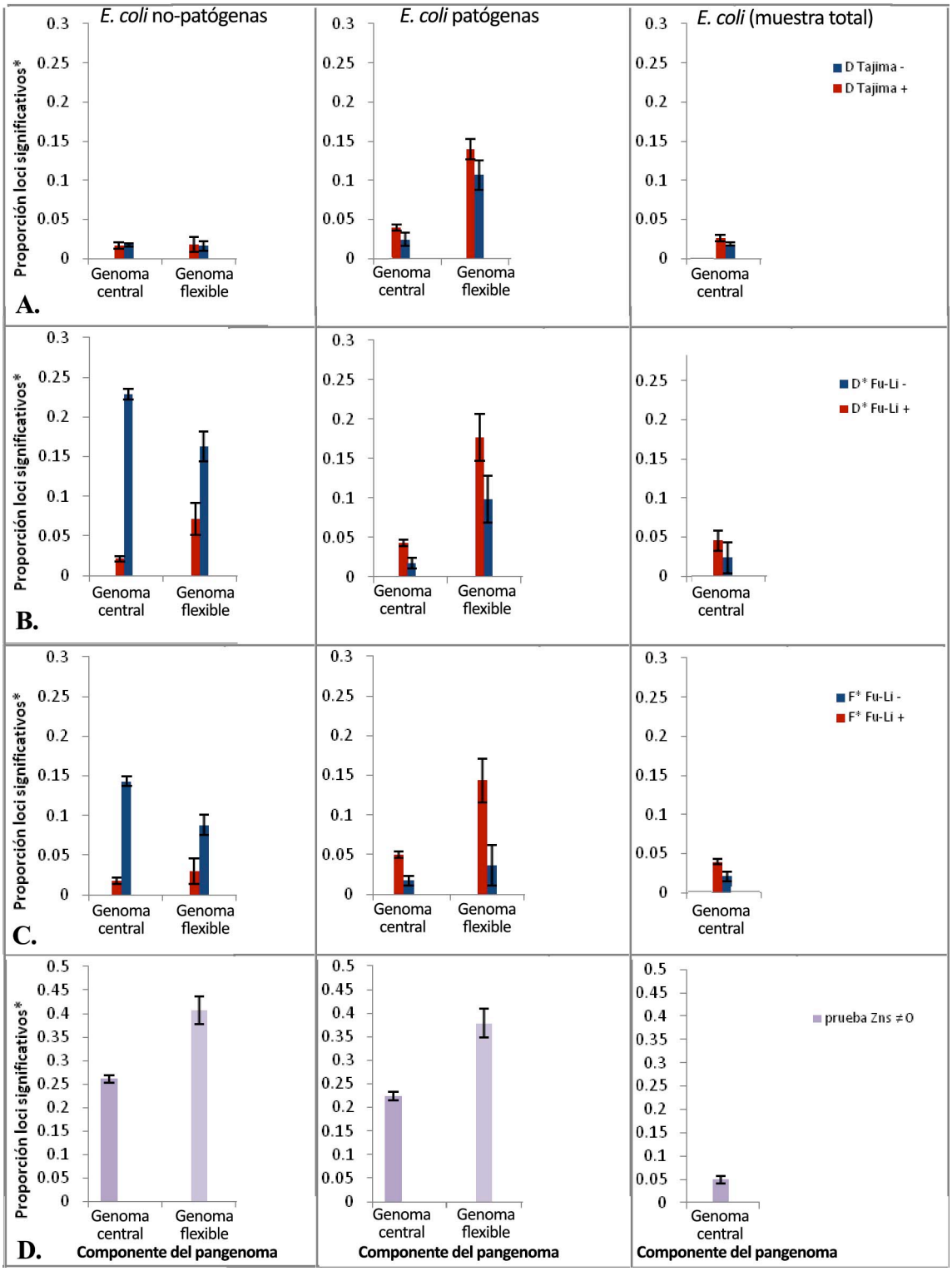

Figura 2. Proporción de loci significativos* en las pruebas de A) D de Tajima, B) D* de Fu-Li, C) F* de Fu-Li y D) desequilibrio de ligamiento Zns, en los diferentes ecogrupos de *E. coli*.
* Significativos a $p<0.05$ en las pruebas de D de Tajima, D* y F* de Fu-Li, y a $p<0.025$ en la prueba de desequilibrio de ligamiento Zns.

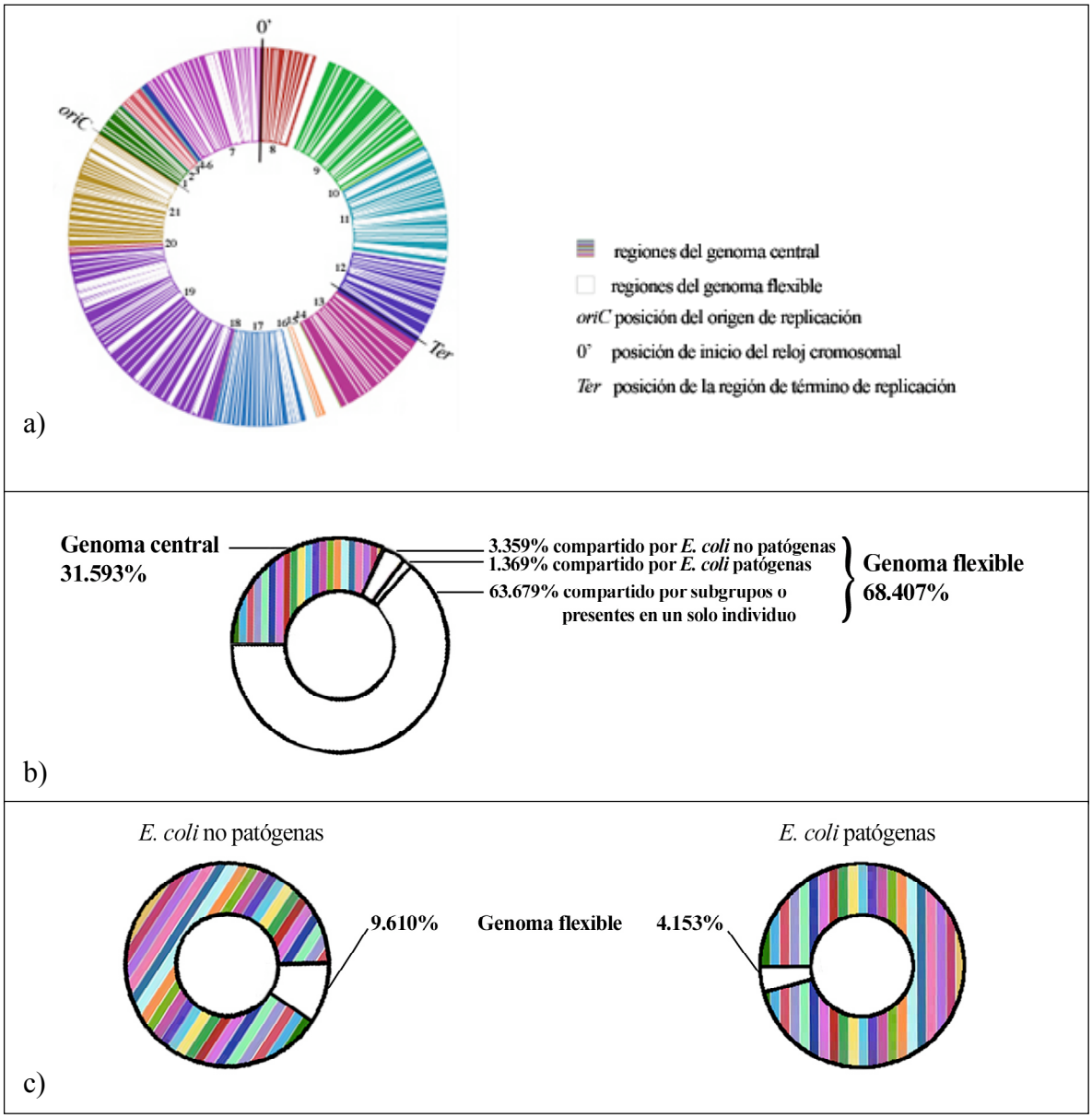

Figura S1 a) Distribución de regiones del genoma central y del genoma flexible en el alineamiento consenso de cromosoma de *E. coli*. Los LCBs se ordenaron de acuerdo al cromosoma de la cepa K12 MG1655. Los colores que representan cada LCB son los mismos que asignó el programa MAUVE, como se ve en la 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*. b) Proporción de sitios del alineamiento correspondientes al genoma central y al genoma flexible compartido por los individuos de un mismo ecogrupo. c) Proporción de sitios del alineamiento correspondientes al genoma flexible, relativa al número de sitios analizados en cada ecogrupo.

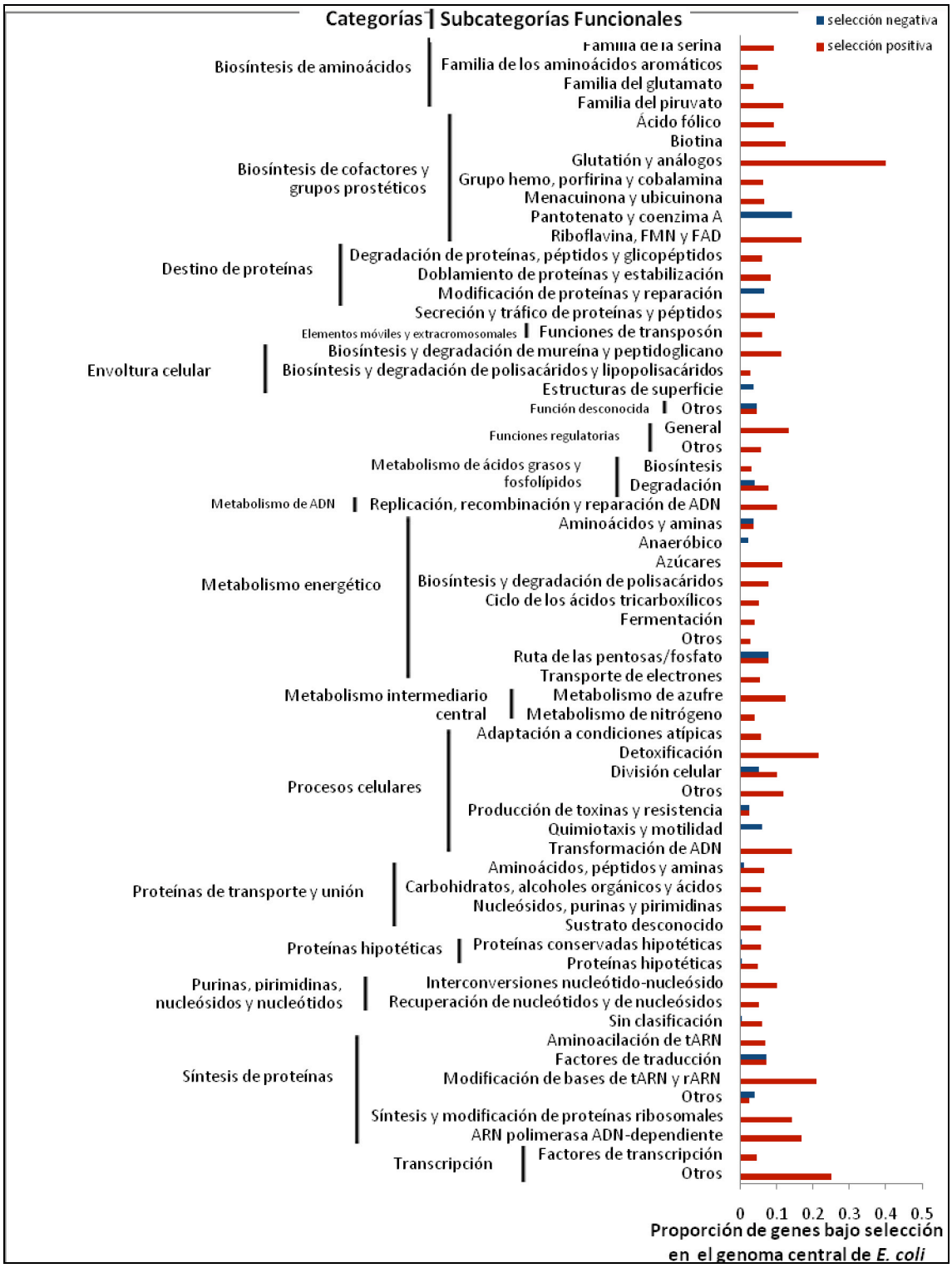

Figura S2. Proporción de genes con evidencia de selección positiva y selección negativa en el genoma central de la muestra total de *E. coli*, dentro de las diferentes categorías funcionales. Se incluyeron genes presentes en loci que fueran significativos (a $p<0.05$) en al menos una de las tres pruebas de neutralidad.

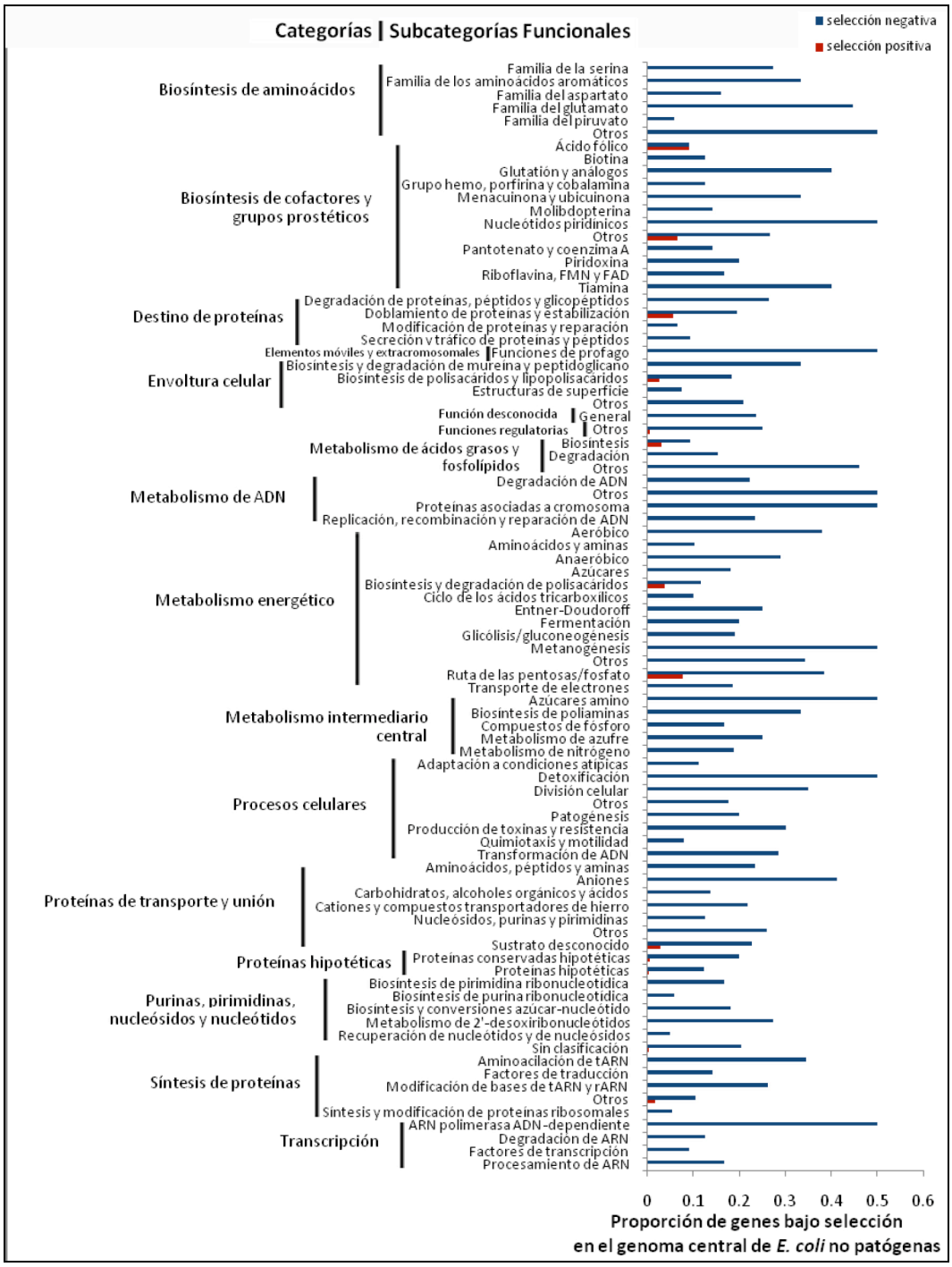

Figura S3 A. Proporción de genes con evidencia de selección positiva y selección negativa dentro de las diferentes categorías funcionales, en el ecogrupo de *E. coli* no-patógenas, A. en el genoma central y B. en el genoma flexible. Se incluyeron genes que tuvieran evidencia de selección (significativos a $p<0.05$) en al menos una de las tres pruebas de neutralidad.

B.

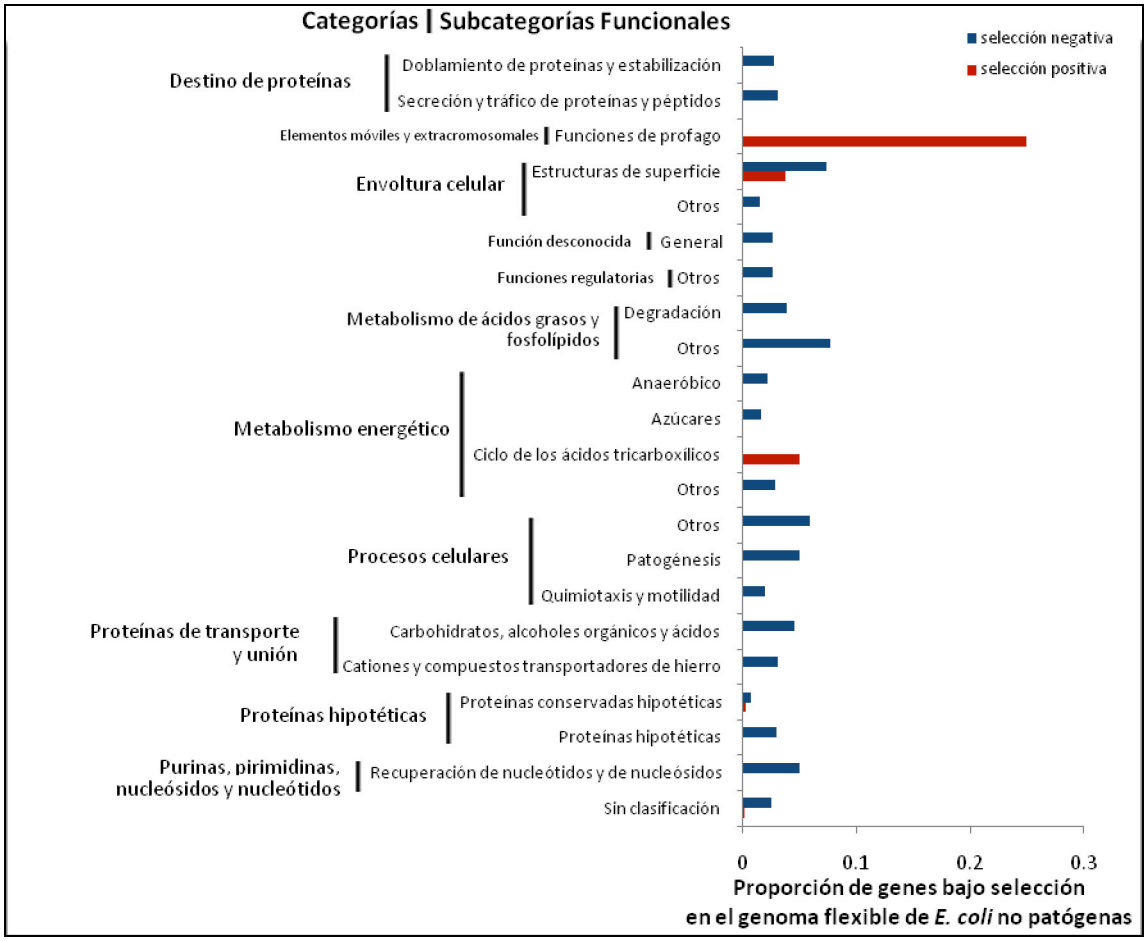

Figura S3B. *Continuación.*

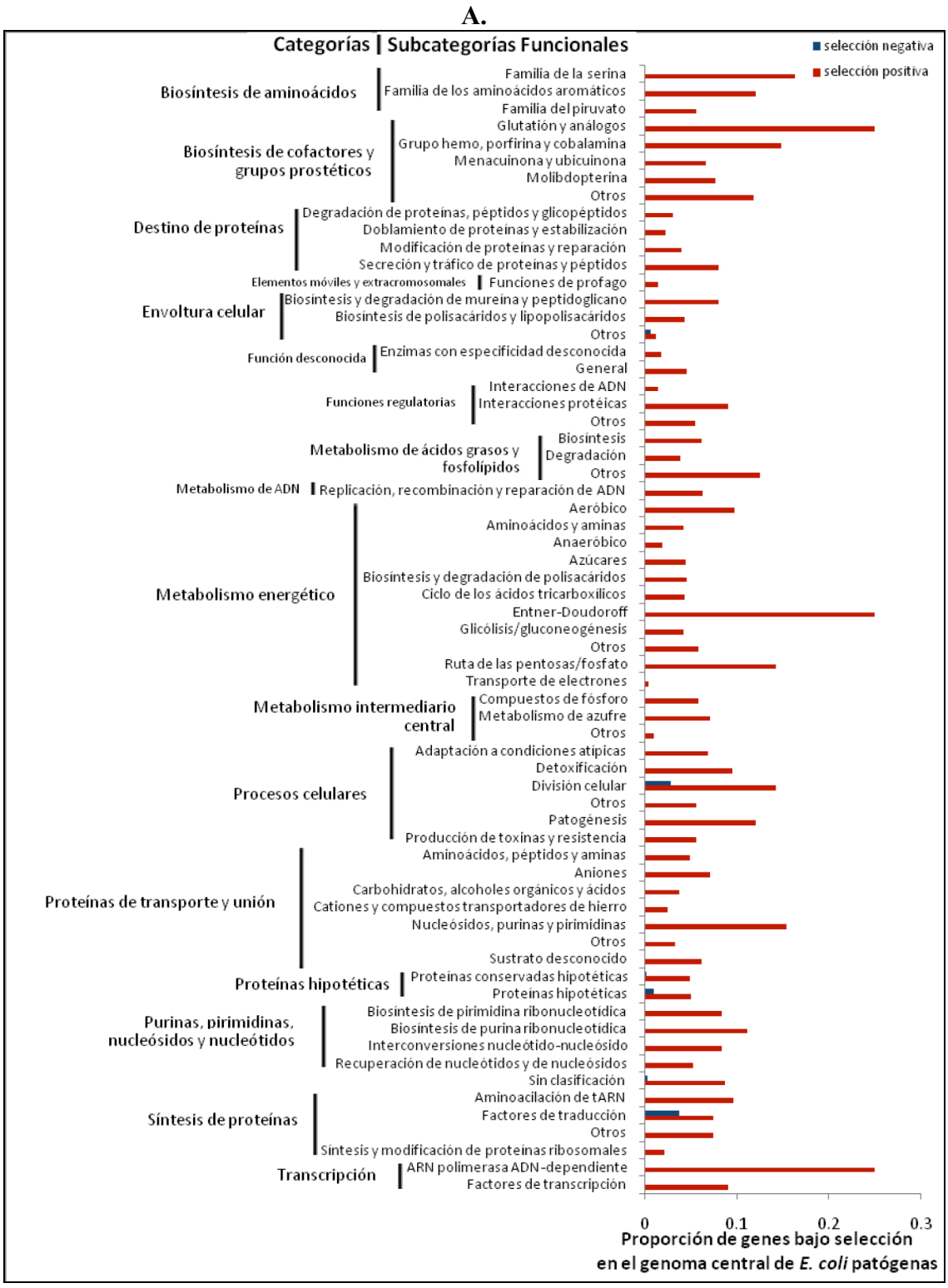

Figura S4 A. Proporción de genes con evidencia de selección positiva y selección negativa dentro de las diferentes categorías funcionales, en el ecogrupo de *E. coli* patógenas, A. en el genoma central y B. en el genoma flexible. Se incluyeron genes que tuvieran evidencia de selección (significativos a $p<0.05$) en al menos una de las tres pruebas de neutralidad.

B.

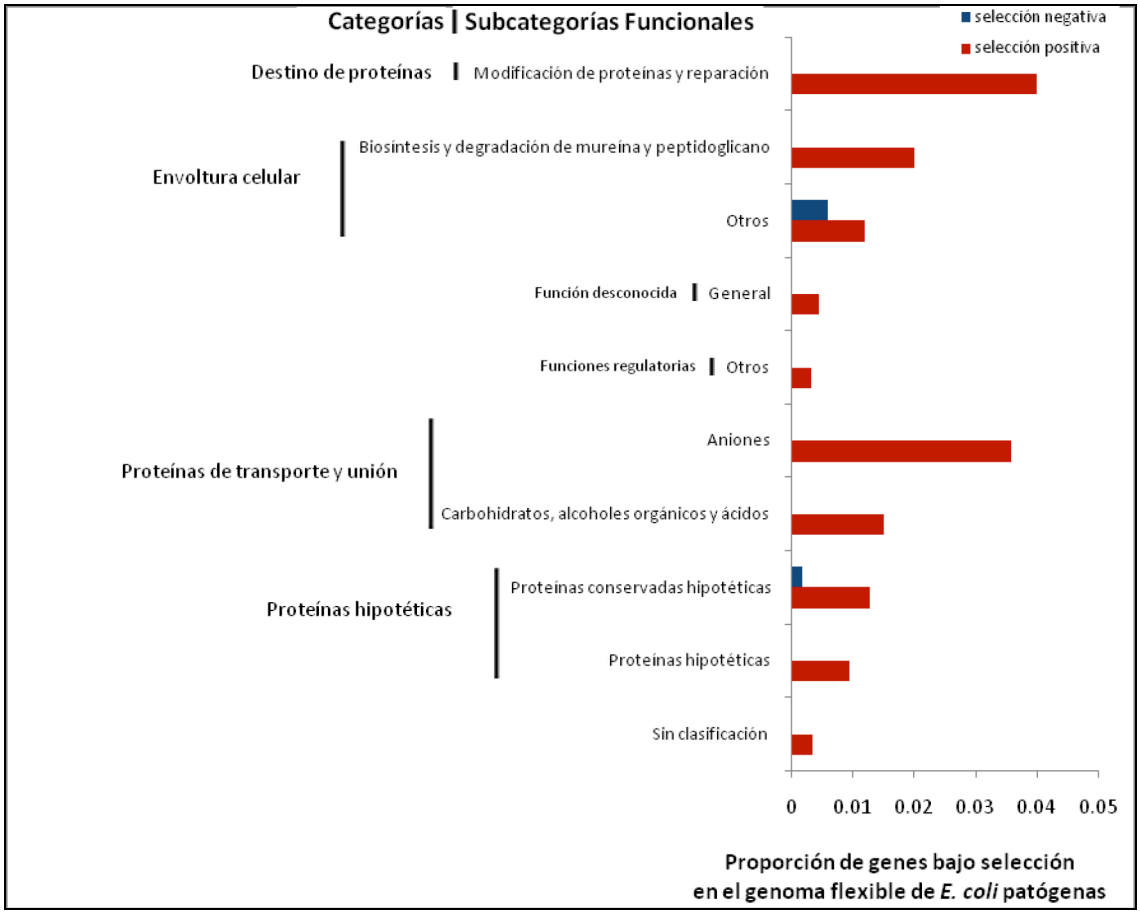

Figura S4 B. *Continuación.*

Tabla S1. 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 cadauno 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.

Tabla S2. 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*

Tabla S3. Diversidad, pruebas de selección natural y desequilibrio de ligamiento en el genoma central de losbloques 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 pruea ZnS (N_{DL}) y [3] número de loci polimórficos (N_P), de cada LCB.
[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 Tarima, D* y F* de Fu-Li y [c] en diversidad Hd y en los valores de desequilibrio de ligamiento Zns (p de Wilcoxon).
- 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.

7. Discusión y conclusiones

Los trabajos seminales de genética de poblaciones realizados para *E. coli* se realizaron utilizando principalmente aislados patogénicos, por razones obvias para el humano, aunque posteriormente se utilizaron aislados "naturales" (Caugant et al., 1981; Selander et al., 1987). Debido a estos trabajos, durante mucho tiempo se asumió que *E. coli* era una especie clonal y que evolucionaba por eventos de selección periódica cada que un alelo benéfico confería mayor adecuación a la especie (Levin, 1981). Sin embargo, esta conclusión de clonalidad fue obtenida utilizando la técnica de electroforesis de enzimas multilocus (MLEE: Multi Locus Enzyme Electrophoresis), la cual a pesar de arrojar ciertos niveles de diversidad genética no es el marcador molecular apropiado para explorar la presencia de recombinación.

Al mismo tiempo se fue extendiendo el estudio de genética de poblaciones a otras especies bacterianas principalmente patógenas. Debido a que generalmente los aislados patogénicos corresponden a clonas epidémicas con serotipos definidos y de origen reciente, una sobre representación de determinados genotipos en las poblaciones fue detectado y esto, aunado al hecho de que las bacterias se dividen por fisión binaria, alimentó la idea de que las especies bacterianas eran clonales. Este supuesto comienza a perder fuerza en la medida en la que comienzan a aparecer las secuencias de ADN de ciertos genes utilizados en los estudios de MLEE al igual que el desarrollo de algoritmos matemáticos para interpretar toda esta información. Es así, que se demuestra que no todas las especies bacterianas patógenas, a pesar de compartir un mismo estilo de vida, presentan el mismo tipo de estructura poblacional ni niveles de recombinación (Figura 7,) (Spratt et al., 2001; Pérez-Losada et al., 2006; Vos y Didelot 2009). Esta variación se debe a que tanto especies patógenas obligadas como oportunistas y comensales como es el caso de *E. coli*, presentan una amplia gama de especies hospederas, rangos de hospederos, virulencia, sitios de infección, mecanismos de evasión al sistema inmune así como de transmisión de hospedero a hospedero. Sin embargo, como se observa en la Figura 8, los patógenos obligados son los más clonales en comparación de los oportunistas, comensales y bacterias de vida libre los cuales presentan un espectro amplio de tasas de recombinación y mutación. Esto nos sugiere que la evolución de la patogénesis en diferentes especies bacterianas puede haber presentar diferentes procesos de adaptación y diversificación. Es así que en esta tesis se

considera a la patogénesis como una efectiva (aunque desafortunada para otros) estrategia más de supervivencia por parte de las bacterias, al igual que el comensalismo o la simbiosis. Por lo que estudiar a una especie que presente diferentes estilos de vida así como una versatilidad en los nichos ecológicos en los que habita, arrojará información más amplia concerniente a los procesos y mecanismos que promueven la diversificación de los linajes bacterianos.

Así, el objetivo de este trabajo fue describir los diferentes mecanismos genéticos y procesos evolutivos que promueven la diversificación y diferenciación genética de *E. coli* en términos de la historia filogenética y versatilidad ecológica de esta especie.

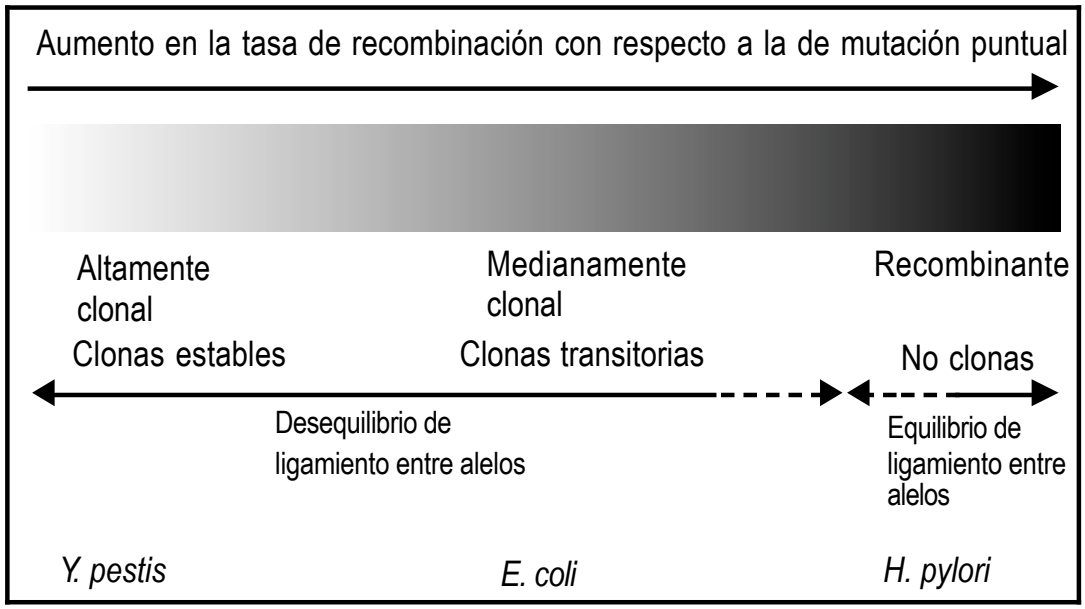

Figura 7. Impacto de la recombinación homóloga en la estructura poblacional bacteriana

En una población bacteriana en donde la recombinación es prácticamente inexistente, la diversificación de las clonas es lenta ya que depende completamente de la acumulación de mutaciones puntuales. Asimismo, existen niveles de desequilibrio de ligamiento altos y la población es altamente clonal consistiendo en linajes que evolucionan independientemente. En tanto aumenta la contribución de la recombinación en el cambio evolutivo de los loci neutrales, las clonas comienzan a ser transitorias hasta que las altas tasas de recombinación con respecto a las de mutación evitan la emergencia de clonas debido a que los genomas diversifican muy rápido. No obstante, puede existir desequilibrio de ligamiento significativo en clonas o complejos clonales aún cuando la recombinación es mayor que la mutación puntual (Figura modificada de Spratt et al., 2001 y complementada con información de Achtman et al., 1999; González-González et al., 2013; Vos y Didelot, 2009 para *Y. pestis*, *E. coli* y *H. pylori* respectivamente).

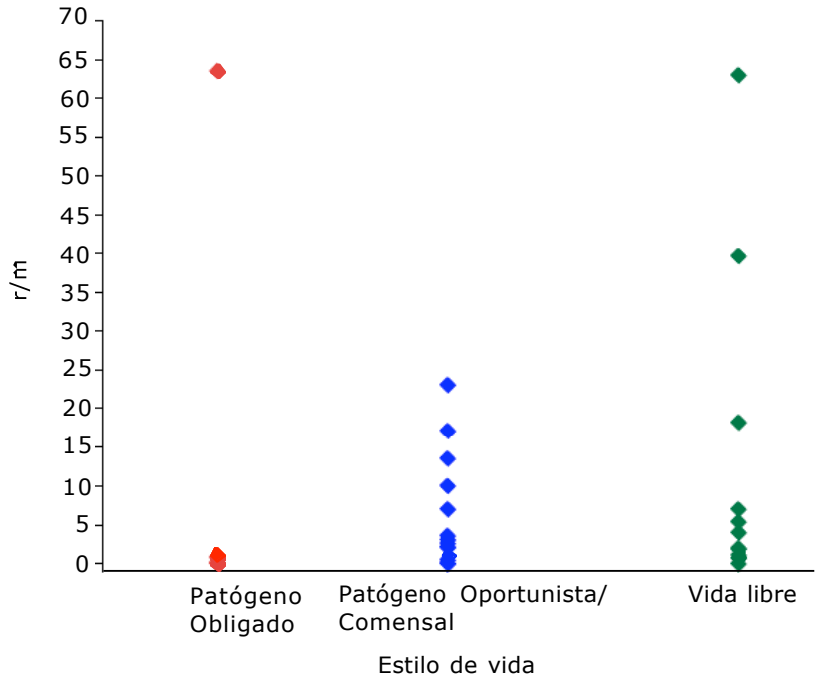

Figura 8. Valores de recombinación correspondientes a diferentes estilos de vida bacterianos

El estimado de recombinación *r/m* se refiere al impacto relativo de la recombinación comparado con la mutación puntual en la generación de diversidad genética de una población. **Patógenos obligados**: *Flavobacterium psychrophilum*, *Porphyromonas gingivalis*, *Chlamydia trachomatis*, *Bordetella pertussis*, *Bordetella henselae*. **Patógenos Oportunistas/Comensales**: *Escherichia coli* (3.24; este trabajo), *Streptococcus pneumoniae*, *Streptococcus pyogenes*, *Helicobacter pylori*, *Moraxella catarrhalis*, *Neisseria lactamica*, *Haemophilus influenzae*, *Campilobacter jejuni*, *Enterococcus faecium*, *Bacillus cereus*, *Enterococcus faec*alis, *Chlamydia trachomatis*, *Clostridium difficile*, *Staphylococcus aureus*, *Leptospira interrogans*. **Bacterias de vida libre**: *Pelagibacter ubique* (SAR 11), *Vibrio parahaemolyticus*, *Vibrio vulnificus*, *Microsystis aeruginosa*, *Myxococcus xanthus*, *Bacillus weihenstephanensis*, *Microcoleus chthonoplates*, *Oenococcus ovni*, *Rizobium gallicum*. (Figura armada con la información de Vos y Didelot, 2009).

7.1 Los niveles de diversidad genética del genoma central de *Escherichia coli* son producto principalmente de la recombinación.

El análisis de genética de poblaciones realizado en este estudio utilizando siete genes constitutivos sugiere que en promedio, la recombinación homóloga tiene más impacto que la mutación puntual en la diversificación genética de toda la muestra de *E. coli* (Capítulo 4). Lo mismo encontramos cuando analizamos el genoma central completo de 12 aislados de *E. coli* si consideramos los altos niveles de diversidad nucleotídica, rearreglos cromosómicos y regiones bajo equilibrio de ligamiento encontrados (Capítulo 6).

A pesar de los niveles de recombinación encontrados, el análisis de reconstrucción filogenética realizado en esta tesis y en otros trabajos donde se han estudiado diferentes tipos de muestras así como diversos marcadores moleculares (Wirth et al., 2006; Walk et al., 2007; Jaureguy et al., 2008; Okeke et al., 2010; Bergholz et al., 2011), recupera los mismos grupos filogenéticos que se describieron utilizando isoenzimas en un principio (Selander et al., 1986, Goullet y Picard, 1989). Esta aparente incongruencia también llamada paradigma clonal, ha sido previamente estudiada utilizando un enfoque filogenético utilizando la información proveniente de 20 genomas completos (Touchon et al, 2009). Lo que este estudio se sugiere es que la estructura clonal de *E. coli* se mantiene a pesar de los niveles de recombinación homóloga presentes debido a que la recombinación se lleva a cabo mediante el intercambio de pequeños fragmentos de ADN entre los diferentes genomas.

Una de las aportaciones más importantes de esta tesis es la explicación alternativa de tipo poblacional a este paradigma clonal. Nuestra propuesta considera que *E. coli* además de estar conformada por diferentes poblaciones (los grupos filogenéticos), también se encuentra dividida en unidades evolutivas menores como linajes y complejos clonales. Así encontramos que cada una de estas unidades de organización genética presenta diferentes mecanismos de diversificación los cuales pueden encontrarse asociados o no a determinados nichos ecológicos y estilos de vida. Gracias a esta estructura y dinámica evolutiva poblacional es que *E. coli* mantiene su estructura clonal a pesar de que ciertas unidades evolutivas presenten niveles de recombinación altos.

Nuestra explicación "poblacional" no invalida la propuesta filogenética de Touchon y colaboradores (2009), más bien implica una perspectiva complementaria (la de genética de poblaciones), la cual junto con la perspectiva filogenética nos muestra un panorama mucho más amplio de cómo sucede el proceso evolutivo en *E. coli*. Además de proponer que *E. coli* diversifica gracias a que tanto la recombinación como la mutación suceden a tasas diferentes al interior de cada población, linaje y grupo filogenético, esta tesis sugiere que la recombinación homóloga tienen un efecto tanto homogeneizante como heterogeneizante en términos de su historia filogenética, lo cual reafirma el papel importante que juega tanto en el mantenimiento de la poza génica de una especie así como en la divergencia de las poblaciones que la conforman.

Otro resultado importante de esta tesis consiste en que la estructura filogenética de *E. coli* no coincide con la estructura ecológica, es decir, no existe una asociación estricta entre los grupos filogenéticos que conforman a la especie y los diferentes nichos ecológicos (diferentes hospederos) en los que habita y estilos de vida que presenta (comensales y patógenos) como previamente se había sugerido. Esto es importante en términos de la identificación de las barreras que promueven la especiación en las poblaciones bacterianas. Como se mencionó en la introducción, en bacterias se han identificado barreras de tipo geográficas, ecológicas y sexuales como en los eucariontes. En el caso de *E. coli*, proponemos que este proceso de diferenciación genética acoplada a barreras ecológicas se lleva a cabo no a nivel de grupo filogenético (A, B1, B2, D, E) sino más bien, a un nivel de organización de la diversidad genética menor como son los complejos clonales, a los cuales consideramos como las unidades evolutivas mínimas mediante las cuales diversifica *E. coli*. Debido a que dichos complejos clonales se encuentran asociados a determinada dinámica de recombinación y mutación acopladas a cierto nicho ecológico y/o estilo de vida.

Considerando lo anterior, así como los resultados de los análisis de genómica de poblaciones (Capítulo 6), encontramos necesario explorar la asociación ecológica de cada uno de los complejos clonales que conforman a *E. coli* en términos más detallados como son por ejemplo, fenotipos metabólicos y sus niveles de expresión así como los niveles de regulación génica involucrados en determinados tipos de dieta del hospedero o en el hecho de ser animales silvestres o domésticos o asociados a ambientes secundarios como suelo, agua y aire. Asimismo, esta tesis sugiere además, realizar estudios profundos sobre las barreras sexuales, es decir, las restricciones en los mecanismos genéticos de recombinación homóloga que pudieran estar promoviendo diferenciación poblacional como sucede en diferentes complejos clonales de *Salmonella enterica*, otra especie facultativa (Didelot et al., 2011) y en *Neisseria meningitidis* en donde se han descrito 22 sistemas de restricción y modificación los cuales generan barreras al intercambio de ADN consistentes con la estructura poblacional de esta especie (Budroni et al., 2011). Con respecto a las barreras geográficas, a pesar de que *E. coli* es una especie que se encuentra distribuida en todo el mundo, proponemos que es necesario hacer estudios como los realizados en esta tesis pero enfocados a muestreos más localizados por región geográfica como por ejemplo país. Ya que gracias al estudio

detenido de la dinámica evolutiva de la poza génica de cada región se podrán establecer programas de vigilancia epidemiológica más robustos.

7.2. El tamaño del cromosoma en *E. coli* es variable y no se encuentra asociado a un nicho ecológico y estilo de vida en particular, ni tampoco a un grupo filogenético determinado.

En el presente estudio encontramos que el cromosoma de *E. coli* presenta un rango de variación de alrededor 2 Megabases (Capítulo 5). Pero lo más importante que arroja esta tesis es el hecho de que ésta variación no se encuentra estructurada ni filogenética y ni ecológicamente. Asimismo, si tomamos en cuenta que esta diferencia en el tamaño del cromosoma corresponde principalmente a la parte del genoma flexible, entonces, podemos considerar que la transferencia horizontal de genes (THG) junto con la pérdida de genes, juegan un papel muy importante en la diversificación del genoma de *E. coli* (Leimbach et al., 2013). Y aunque se ha considerado tradicionalmente que *E. coli* no es un transformante natural, trabajos recientes sugieren que esta especie presenta mecanismos de transformación no caracterizados bajo ciertas condiciones ambientales (Etchuuya et al, 2011). Asimismo, existe evidencia de THG de factores de virulencia y de adecuación entre las diferentes especies que conforman a la familia Enterobacteriaceae (Dobrindt et al., 2010; Paauw et al., 2010). Por otro lado, debido a que *E. coli* se encuentra en contacto con el microbioma del hospedero y por ende por una comunidad bacteriana diversa, existe una gran poza génica flexible disponible para transferirse horizontalmente (Tenaillon et al., 2010). Un trabajo reciente sugiere que la THG se da principalmente entre los grupos filogenéticos A con B1 y B2 con D. Sin embargo no se encontraron diferencias significativas en el flujo existente entre cepas con diferentes estilos de vida (comensal y patogénicas) y tampoco entre diferentes ambientes (intestinal y extraintestinal) (Skippington y Ragan, 2012).

Dado que mediante la transferencia horizontal de islas genómicas e islas de patogénesis, se adquieren genes que promueven la adaptación a nuevos nichos ecológicos así como la adquisición de diferentes estilos de vida (Hacker and Carniel, 2001; Dobrindt et al., 2004), se ha propuesto que este tipo de recombinación juega un papel muy importante en la especiación simpátrica. En el caso de *E. coli*, el genoma central -el cual presenta pequeños cambios entre diferentes aislados generados por

recombinación homóloga- define lo que es común a la especie y los genes adquiridos por THG definirán las adaptaciones específicas a microambientes específicos lo que a la larga puede constituir una diversificación de tipo simpátrica mediante ecotipos. Este tipo de diversificación simpátrica ha sido descrita en otras especies de microorganismos como las bacterias *Sinorhizobium medicae* y *Vibrio cyclitrophicus* y la arquea *Sulfolobus islandicus* (Bailly et al., 2011; Shapiro et al., 2012; Cadillo-Quiroz et al., 2012; respectivamente) en donde las condiciones ambientales en las que se distribuyen son ligeramente diferentes debido al tipo de hospedero como es el caso de *S. medicae* cuyos hospederos son diferentes especies de plantas del género *Medigaco*, o diferentes condiciones ambientales a lo largo de la columna de agua de mar, como en el caso de *V. cyclitrophicus* o el aprovechamiento diferencial de los nutrientes por parte de *S. islandicus*.

7.3. La adaptación a diferentes nichos ecológicos y estilos de vida va más allá de adquirir los genes nicho-específicos.

Como se mencionó en un principio, gracias a la genómica de poblaciones podemos estudiar los procesos históricos y contemporáneos que moldean la diversificación y evolución de las especies bacterianas. Con el estudio de genómica de poblaciones realizado en esta tesis analizamos una parte del componente contemporáneos o ecológicos de la evolución de *E. coli*. Encontramos que la evolución de la patogénesis no se da nada más gracias a la transferencia horizontal de los factores de virulencia, sino que el cromosoma central también juega un papel importante al representar un fondo genético que ha diversificado en función del estilo de vida patogénico. Es decir, encontramos evidencias del proceso de adaptación a la patogénesis a nivel del genoma central debido a la acción de la selección natural y de la recombinación en funciones metabólicas básicas como la trascripción, estrés oxidativo, biosíntesis de estructuras celulares y elementos regulatorios. Un estudio reciente en donde comparan los patrones de transcripción de una especie de listeria patógena y otra no patógena *Listeria monocytogenes* y *Listeria innocua* respectivamente, sugieren patrones diferenciales en la expresión de la región no codificante de ambos genomas (Wurtzel et al., 2012).

E. coli es una especie que explota diferentes nichos ecológicos al encontrarse asociada tanto a un hospedero así como al ambiente exterior. En los últimos años, se ha

volteado la mirada al estudio detallado de la fisiología de esta especie en los diferentes microhábitats en los que se le encuentra con el fin de demostrar la relación que existe entre los procesos metabólicos básicos y la patogénesis (Marteyn et al., 2010; Alteri et al., 2011). Se ha encontrado que el metabolismo energético pudiera ser una señal importante utilizada por las bacterias patógenas para identificar y cambiar entre microambiente específico como pueden ser las diferentes regiones del intestino, el ambiente urinario, el nervioso y el respiratorio (Alteri y Mobley, 2012). Interesantemente, nuestro estudio de genómica de poblaciones arroja fuertes señales de selección diversificadora en los diferentes genes involucrados en el metabolismo energético en las cepas patógenas en comparación con las cepas no patógenas en donde actúa la selección purificadora. Estos resultados nos llevan a proponer que estudiar cómo cambia la arquitectura transcripcional entre las cepas patógenas y no patógenas, así como entre los diferentes ambientes intestinal, extraintestinal en los que se encuentra esta especie arrojará mayor información concerniente al ecología y evolución de esta especie facultativa.

7.4. Estilos de vida, asociación a nicho, recombinación homóloga, recombinación no homóloga y diversificación simpátrica en *E. coli*

A diferencia de otras especies bacterianas como *Yersinia pestis* cuyo pangenoma es cerrado, su nicho ecológico restringido y cuya evolución es moldeada principalmente por la deriva génica dados los cambios demográficos sufridos con cada brote epidémico (Cui et al 2013), *E. coli* presenta una dinámica evolutiva compleja.

Esta tesis propone que *E. coli* consta de un genoma central que diversifica en mayor medida por recombinación homóloga que por mutación puntual. Esta gran poza génica se organiza en diferentes grupos genotípicos correspondientes a poblaciones o grupos filogenéticos, linajes y complejos clonales. Es a nivel de complejo clonal que se encuentra una coherencia ecológica entre los miembros que lo conforman asociada a procesos de diversificación genética (tasas de recombinación y mutación) característicos a cada uno. Por lo que cada uno de los complejos clonales es considerado un ecotipo o la unidad evolutiva mínima en *E. coli*.

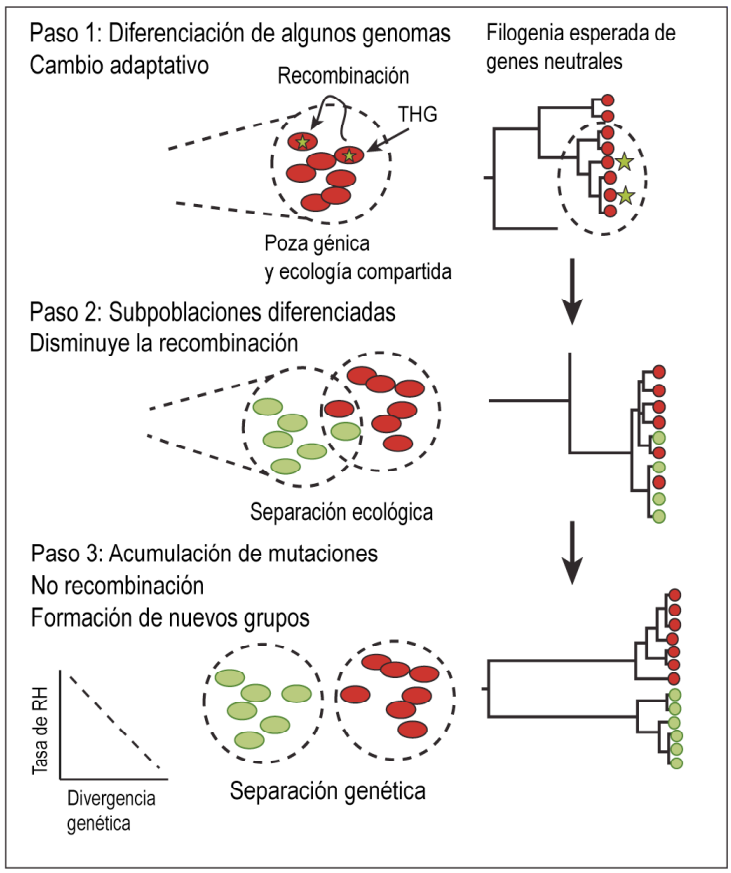

Figura 9. Modelo del proceso de divergencia genotípica vía especialización ecológica simpátrica.

Consideremos una población ancestral de *E. col*i asociada al ambiente intestinal por ejemplo. En el Paso 1 al menos un genoma adquiere variación (indicado por la estrella verde) vía recombinación homóloga, mutación puntual, pérdida o transferencia horizontal de genes. Esta variante puede dispersarse a los miembros de la población por recombinación homóloga. En un árbol filogenético elaborado con genes neutrales, la población permanece homogénea. En el Paso 2, la separación ecológica comienza a disminuir el flujo génico entre las dos poblaciones (la nueva población denotada por color verde, comienza a ocupar un nicho diferente, por ejemplo, una región diferente del colon o el ambiente extra-intestinal o el de vida libre o también un estilo de vida diferente como la patogénesis). En este paso, diferentes genes arrojarán diferentes árboles filogenéticos debido a los diferentes grados de recombinación y arrastre con los genes adaptativos. En el Paso 3, las mutaciones y todos los cambios genéticos se han acumulado al interior de cada población. Debido a que en las bacterias existe un decrecimiento exponencial de la tasa de recombinación homóloga en tanto aumenta la divergencia de entre las secuencias, este proceso rápidamente disminuye el flujo génico entre las poblaciones nacientes. En este punto, el árbol filogenético muestra grupos genéticos consistentes con la diferenciación ecológica es decir un ecotipo intestinal o un ecotipo extra-intestinal o un ecotipo de vida libre (Figura modificada de Polz, et al., 2013)

A pesar de que en esta tesis solamente estudiamos cepas aisladas de un ambiente intestinal, podemos generalizar que la asociación a un nicho ecológico ya sea intestinal, extra-intestinal o de vida libre, así como la adquisición de un estilo de vida determinado (patogénesis y comensalismo), está dado por dos mecanismos principales. Por un lado,

por la adquisición de genes mediante transferencia horizontal los cuales promueven la adaptación a dicho nicho y su posterior diseminación al interior del ecotipo mediante recombinación homóloga (Figura 9). Y por el otro, por la posterior afinación de las rutas fisiológicas dada por mutación puntual y recombinación homóloga lo que promueve el ajuste a las nuevas condiciones ecológicas del nicho.

Con el tiempo, si deja de haber flujo génico con miembros de diferentes ecotipos, se originan clonas estables y "exitosa" en términos adaptativos aunque sujetas en mayor medida a los efectos demográficos de la deriva génica lo que a la larga puede llevarlas a su extinción. Pero al mismo tiempo surgen en la población nuevos ecotipos debido a la versatilidad ecológica de la especie por lo que si tomamos una fracción en el tiempo evolutivo de *E. coli* vemos una serie de ecotipos que coexisten cada uno con su propia dinámica de recombinación homóloga, no-homóloga, mutación puntual y selección natural los cuales van cambiando a lo largo del tiempo.

8. Perspectivas

A partir del trabajo realizado en esta tesis surgieron diversas preguntas a explorar. En concreto nos enfocaremos a las siguientes:

En el capitulo 5 referente a la estimación del tamaño del cromosoma resta evaluar el papel que tienen tanto los re-arreglos como la duplicación en la diversificación del cromosoma de *E. coli*. Esta inquietud surge debido a la diversidad en los perfiles de restricción arrojados por diversas cepas estudiadas en esta tesis lo cual nos sugiere dinamismo y flexibilidad a nivel genómico. Al mismo tiempo, en el capítulo 6 de este trabajo, encontramos la presencia de re-arreglos al comparar el cromosoma completo de diversos aislados de *E. coli* previamente secuenciados. Gracias a estos indicios, proponemos que caracterizar los mecanismos de diversificación del genoma a gran escala de aislados provenientes de hospederos tan diversos como animales con diferentes niveles de domesticación, aportaría información valiosa para el entendimiento de la evolución del genoma de *E. coli*.

Para llevar a cabo este objetivo se determinará mediante hibridaciones, la identidad de cada fragmento de restricción arrojado por la enzima I-*Ceu*I utilizada previamente para estimar el tamaño del genoma. Como referencia se utilizará el genoma de *E. coli* K12. Los loci a identificar serán: *dif* en la posición 34.3' para el fragmento **A**; *rpoS* en la posición 61.8' para el fragmento **B**; *oriC* en la posición 84.6' para el fragmento **C**; *uvrD* para el fragmento 86.2' para el fragmento **D**; *glnLG* en la posición 87.3' para el fragmento **E** y *ileS* en la posición 0.5' para el fragmento **G** (Tomado de Bergthorsson y Ochman, 1998).

Por otro lado, del capítulo 6 de esta tesis queda pendiente el evaluar la influencia de la deriva génica en la diversificación y estructuración poblacional de *E. coli*. Así como también incorporar estimados de tasas de mutación y recombinación históricas con el fin de darle un componente temporal a la diversificación y divergencia de *E. coli*. Para cubrir este objetivo se realizará un estudio de genómica de poblaciones pero se tiene contemplado ampliar la muestra de genomas analizados en el capítulo 6 con algunos de los genomas enlistados en la Tabla suplementaria 3 y 5 del capítulo 5 de esta tesis. Los genomas extras que analizaremos cubrirán diversos tipos de hospederos como por ejemplo, más aislados ambientales así como cepas provenientes de animales silvestres. Al mismo tiempo, aumentaremos la muestra por cada nicho ecológico (intestinal, extra-intestinal, vida libre) lo cual nos permitirá evaluar la dinámica evolutiva de cada uno de estos ecotipos.

9. Referencias

Achtman, M., et al. 1999. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. Proc. Natl. Acad. Sci USA. 96, 14043-14048.

Alm, E.W., Walk, S.T., Gordon, D.M. 2011. The niche of *Escherichia coli*. En: Walk, S.T. y Feng, P.C.H. (Editores). Population genetics of bacteria: a tribute to Thomas S. Whittam., ASM Press, Washington, D.C. pp., 69-89.

Alteri, C.J., Lindner, J.R., Reiss, D.J., Smith, S.N., Mobley, H.L. 2011. The broadly conserved regulator PhoP links pathogen virulence and membrane potential in *Escherichia coli*. Mol. Microbiol. 82, 145-163.

Alteri, C.J., Mobley, H.L.T. 2012. *Escherichia coli* physiology and metabolism dictates adaptation to diverse host microenvironments. Curr. Opin. Microbiol. 15, 3-9.

Bailly, X., Giuntini, E., Sexton, M.C., Lower, R.P.J., Harrison, P.W., Kumar, N., Young, J.P. 2011. Population genomics of *Sinorhizobium medicae* based on low-coverage sequencing of sympatric isolates. ISME J. 5, 1722-1734.

Balloux, F. 2010. Demographic influences on bacterial population structure. En Robinson, D.A., Falush, D., Feil, D. (Eds). Bacterial population genetics in infectious disease. Wiley-Blackwell, USA. pp., 103-120.

Berg, R.D. 1996. The indigenous gastrointestinal microflora. Trends Microbiol. 4, 430-435.

Bergholz, P.W., Noar, J.D., Buckely, D.H., 2011. Environmental patterns are imposed on the population structure of *Escherichia coli* after fecal deposition. Appl. Environ. Microbiol. 77, 211-129.

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

Bisharat, N., Cohen, D.I., Maiden, M.C., Crook, D.W., Peto, T., Harding, R.M. 2007. The evolution of genetic structure in the marine pathogen, *Vibrio vulnificus*. Infect. Genet. Evol. 7, 685-693.

Black, W.C. Baer, C.F., Antolin, M.F., DuTeau, N.M. 2001. Population genomics: genome-wide sampling of insect populations. Annu. Rev. Entomol. 46, 441-469.

Budroni, S., Siena, E., Dunning-Hotopp, C.D., Seib, K.L., Serruto, D., Nofroni, C., Comanducci, M., Riley, D.R., Daugherty, S.C., Angiuoli, S.V., et. al. 2011. *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. Proc. Natl. Acad. Sci USA. 108, 4494-4499.

Bumbaugh, A.C. y Lacher, D.W. 2011. Gene acquisition and loss in the phylogenetic lineages of the invasive *Escherichia coli*. En Walk, S.T. y Feng, P.C.H. (Editores). Population genetics of bacteria: a tribute to Thomas S. Whittam. Editores. ASM Press, Washington, D.C. pp., 135-156.

Cadillo-Quiroz, H., Didelot, X., Held, N.L., Herrera, A., Darling, A., Reno, M.L., Krause, D., Whitaker, R.J. 2012. Patterns of gene flow define species of thermophilic archaea. PLos Biol 10, e1001265.

Caugant, D.A., Levin, B.R., Selander, R.K., 1981. Genetic diversity and temporal variation in the *E. coli* population of a human host. Genetics, 98, 467-490.

Caugant, D.A., Mocca, L.F., Frasch, C.E., Froholm, L.O., Zollinger, W.D., Selander, R.K. 1987. Genetic structure of *Neisseria meningitidis* populations in relation to serogroup, serotype, and outer membrane protein pattern. J. Bacteriol. 169, 2781-2792.

Clermont, O., Bonacorsi, S., Bingen, E. 2000. Rapad and simple determination of the *Escherichia coli* phylogenetic group. Appl. Environ. Microbiol. 66, 4555-4558.

Clermont, O., Christenson, J.K., Denamur, E., Gordon, D. 2013. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environ. Microb. Reports. 5, 58-65.

Cohan, F.M. 2001. Bacterial species and speciation. Syst. Biol. 50, 513-524.

Cui, Y., Yu, Ch., Yan, Y., Li, D., Li, Y., Jombart, Weinert, L.A., Wang, Z., Guo, Z., et al. 2013. Historical variations in mutation rate in an epidemia pathogen, *Yersinia pestis*. Proceedings of the National Academy of Sciences. 110, 577-582.

Chaudhuri, R.R., Henderson, I.R. 2012. The evolution of the *Escherichia coli* phylogeny. Infect. Genet. Evol. 12, 214-226.

Cho, J-Ch., Tiedje, J.M. 2000. Biogeography and degree of endemicity of fluorescent Pseudomonas strains in soil. Appl. Environ Microb. 66, 5448-5456.

Denamur, E., Picard, B., Tenaillon O., 2010. Population genetics of pathogenic *Escherichia coli*, En: Robinson, D.A., Falush, D., Feil, E.J. (Eds.), Bacterial population genetics in infectious diseases. Wiley-Blackwell, West Sussex, United Kingdom, pp. 269-286.

Didelot, X., Falush. D. 2007. Inference of bacterial microevolution using multilocus sequence data. Genetics. 175, 1251-1266.

Didelot, X., Maiden, M.C.J. 2010. Impact of recombination on bacterial evolution. Trends Microb. 8, 315-322.

Didelot, X., Bowden, R., Street, T., Golubchik, T., Spencer, Ch., McVean, G., Sangal, V., Anjum, M.F., Achtman, Falush, D., Donnelly, P. 2011. Recombination and population structure in *Salmonella enterica*. Plos Genet. 7, e1002191.

Didelot, X., Méric, G., Falush, D., Dariling, A.E. 2012. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. BMC Genomics, 13:256.

Dobrindt, U., Hochhut, B., Hentschel, U., Hacker, J. 2004. Genomic islands in pathogenic and environmental microorganisms. Nat. Rev. Microbiol. 2, 414-424.

Dobrindt, U., Hacker, J. 2008. Targeting virulence traits: potential strategies to combat extraintestinal pathogenic *Escherichia coli* infections. Curr. Opin. Microbiol. 11, 409-413.

Dobrindt, U., Chowdary, M.G., Krumbholz, Hacker, J. 2010. Genome dynamics and its impact on evolution of *Escherichia coli*. Med. Microbiol. Immunol. 199, 145-154

Dudley, E.G. y Rasko, D.A. 2011. Genomic and virulence heterogeneity of Enteroaggregative *Escherichia coli*. En: Walk, S.T. y Feng, P.C.H (Editores). Population genetics of bacteria: a tribute to Thomas S. Whittam. ASM Press, Washington, D.C. pp 181-198.

Dykuizen, D.E., Green, L. 1991. Recombination in *Escherichia coli* and the definition of biological species. J. Bacteriol. 173, 7257-7268.

Eguiarte, L.E. 1999. Una guía para principiantes a la genetica de poblaciones. En: La evolución biológica. Núñez, J. y Eguarte, L.E. (Eds.). México, D.F., UNAM. 35-50.

Escobar-Páramo, P., Clermont, O., Blanc-Potard, A.B., Bui, H., Le Bouguénec, C., Denamur, E., 2004. A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. Mol. Biol. Evol. 21, 1085-1094.

Escobar-Paramo, P., Le Menac'h A, Le Gall T., et al. 2006. Identification of forces shaping the commensal *Escherichia coli* genetic structure by comparing animal and human isolates. Environ. Microbiol. 8, 1975-1984.

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.

Etchuuya, R., Ito, M., Kitano, S., Shigi, F, Sobue, R., Maeda, S. 2011. Cell-to-cell transformation in Escherichia coli: A novel type of natural transformation involving cell-derived DNA and a putative promoting pheromone. Plos ONE 6, e16355.

Ettema, J.G., Andersson, S.G. 2009. The α-proteobacteria: the Darwin finches of the bacterial world. Biol. Lett. 5, 429-432.

Falkowski, P.G., Fenchel, T., Delong, E.F. 2008. The microbial engines that drive earth's biogeochemical cycles. Science. 320, 1034-1039.

Feil, E.J., Spratt, B.G. 2001. Recombination and the population structures of bacterial pathogens. Annu. Rev. Microbiol. 55, 561-590.

Feil, E.J. 2004. Small change: Keeping pace with microevolution. Nat. Rev. Microbiol. 2, 483-495.

Feil, N. 2010. Linkage, selection, and the clonal complex. In: Robinson, D.A., Falush, D., Feil, E.J. (Eds.), Bacterial Population Genetics in Infectious Diseases. Wiley-Blackwell, West Sussex, United Kingdom, pp. 19-35.

Fierer, N., Bradford, M.A., Jackson, R.B. 2007. Toward an ecological classification of soil bacteria. Ecology. 88, 1354-1364.

Fricke, W.F., Wright, M.S., Lindell, A.H., Harkins, D.M., Baker-Austin, C., Ravel, J., Stepanauskas, R. 2008. Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate *Escherichia coli* SMS-3-5. J. Bactriol. 190, 6779-6794.

Futuyma, D. 2005. Evolution. Sinauer Associates, USA. 603 pp.

Garrigues, C., Johansen, E., Crittenden, R. 2013. Pangenomics –an avenue to improved industrial starter cultures and probiotics. Curr. Opin. Biotechnol. 24, 187-191.

Gevers, D. Vandepoele, K. Simillion, C., Van de Peer, Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol. 12, 148-154.

González-González A., Sánchez-Reyes L.L., Delgado, G., Eguiarte, L.E., Souza, V. 2013. Hierarchical clustering of genetic diversity associated to different levels of mutation and recombination in *Escherichia coli*: A study based on Mexican isolates. Infect. Genet. Evol. 13, 187-197.

Gordon, D.M. 2010. Strain typing and the ecological structure of *Escherichia coli*. J AOAC Int. 93, 974-984.

Gordon, D.M., Clermont, O., Tolley, H., Denamur, E., 2008. Assigning *Escherichia coli* strains to phylogenetic groups: Multi-locus sequence typing versus the triplex method. Environ. Microbiol. 10, 2484-2496.

Gregory, T.R. y DeSalle, R. 2005. Comparative genomics in prokaryotes. En: Gregory, T.R. (Editor). The evolution of the genome. Elsevier Academic Press. USA. pp. 585-675.

Gogarten, J.P., Doolittle, W.F., Lawrence, J.G. 2002. Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. 19, 2226-2238.

Gordon, D.M., Cowling, A. 2003. The distribution and genetic structure of Escherichia coli Australian vertebrates: host and geographic effects. Microbiology. 12, 3575-3586.

Goullet, P., Picard, B., 1989. Comparative electrophoretic polymorphism of esterases and other enzymes in *Escherichia coli*. J. Gen. Microbiol. 135, 135-143.

Gregory, T.R., DeSalle, R. 2005. The evolution of genome size in prokaryotes. In: The evolution of the genome (Ed. T. R. Gregory). Elsevier, San Diego. 631-640.

Hacker, J., Carniel, E. 2001. Ecological fitness, genomic islands and bacterial pathogenicity – a Darwinian view of the evolution of microbes. EMBO Rep. 2, 376-381.

Hanage, W.P. Fraser, C., Spratt, B.G. 2006. The impact of homologous recombination on the generation of diversity in bacteria. J. Theor. Biol. 239, 210-219.

Hanson, C.A., Fuhrman, J.A., Horner-Devine, M.C., Martiny, J.B.H. 2012. Beyond biogeographic patterns: processes shaping the microbial landscape. Nat. Rev. Microbiol. 10, 497-506.

Hartl, D.L., Clark, A.G. 1989. Principles of Population Genetics. 2nd Edition. Sinauer Associates, Inc. Publishers, MA, EUA.

Hedrick, P.W. 2000. Genetics of Populations. 2nd Edition. Jones & Bartlett Publishers, Sudbury, Massachusetts.

Herczegh, A., Ghidan, A., Deseo, K., Kamotsay, K., Tarjan, I. 2008. Comparison of *Streptococcus mutants* strains from children with caries-active, caries-free and gingivitis clinical diagnosis by pulsed-field gel electrophoresis. Acta Microbiologica et Immunologica Hungarica. 55, 419-427.

Herzer, P., Inouye, S., Inouye, M., Whittam, T.S., 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. J. Bacteriol. 172, 6175-6181.

Hughes, D. 1999. Impact of homologous recombination on genome organization and stability. En: Charlebois R.L. (editor). Organization of the prokaryotic genome. ASM Press, Washington, D.C. pp 109-128.

Hughes, J.B. et al. 2006. Microbial biogeography: putting microorganisms on the map. Nat. Rev. Microbiol. 4, 102-112.

Jaureguy, F., Landraud, L., Passet, V., Diancourt, L., Frapy, E., Guigon, G., Carbonnelle, E., Lortholary, O., Clermont, O., Denamur, E., et al., 2008. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. BMC Genomics 9, 560-573.

Johnson, J.R., Stell, A.L. 2000. Extended virulence genotypes of *Escherichia coli* strains from patients with urosepsis in relation to phylogeny and host compromise. J. Infect. Dis. 181, 261-272.

Johnson, Z.I., Zinser, E.R., Coe, A., McNulty, N.P., Woodward, E.M.S., Chisholm, S.W. 2006. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science. 311, 1737-1740.

Johnson, T.J., Wannemuehle, Y., Johnson, S.J., Stell, A.L., Doetkott, C., Johnson, J.R., Kim, K.S., Spanjaard, L., Nolan, L.K. 2008. Comparison of extraintestinal pathogenic *Escherichia coli* strains from human and avian sources reveals a mixed subset representing potential zoootic pathogens. Appl. Environ. Microbiol. 74, 7043-7050.

Johnson, J. R. 2011. Molecular epidemiology and population genetics of extraintestinal pathogenic *Escherichia coli*. En: Walk, S.T. y Feng, P.C.H. (Editores). Population genetics of bacteria: a tribute to Thomas S. Whittam. ASM Press, Washington, D.C. pp 91-107.

Kaper, J.B., Nataro, J.P., Mobley, H.L. 2004. Pathogenic *Escherichia coli*. Nat. Rev. Microbiol. 2: 123-140.

Kettler, G.C., Martiny, A.C., Huang, K., Zucket, J., Coleman, M.L., Rodrigues, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., Steglich, C., Church, G.M., Richardson, P., Chisholm, S.W. 2007. Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLos Genetics. 3, e231.

Koonin, E.V. 2009. Darwinian evolution in the light of genomics. Nucl. Ac. Res. 37, 1011-1034.

Ksoll, W.B., Ishii, S., Sadowsky, M.J., Hicks, R.E. 2007. Presence and sources of fecal coliform bacteria in epilithic periphyton communities of Lake Superior. Appl. Environ. Microbiol. 73, 3771-3778.

Lawrence, J.G., Hendrickson, H. 2003. Lateral gene transfere: when will adolescence end? Mol. Microbiol. 50, 739-749.

Leimbach, A., Hacker, J., Dobrindt, U. 2013. E. coli as an all-rounder: the thin line between commensalism and pathogenicity. Curr. Top. Microbiol. Immun. DOI: 10.1007/82_2012_303.

Leopold, Sh., Sawyer, S.A., Whittam, T.S., Tarr, P.I. 2011. Obscured phylogeny and posible recombinational dormancy in *Escherichia coli*. BMC Evolutionary Biology, 11: 183.

Levin, B.R., 1981. Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. Genetics 99, 1-23.

Lloyd, A.L. y Mobley, H.L.T. 2011. Fitness islands in uropathogenic *Escherichia coli*. En: Robinson, D.A., Falush, D., Feil, D. (Editores). Bacterial population genetics in infectious disease. Wiley-Blackwell, USA. pp. 157-179.

Logan, N.A. 1994. Bacterial systematics. Blackwell Scientific Publications, Oxford. 263 pp.

Luikart, G., England, P.R., Tallmon, D., Jordan, S., Taberlet, P. 2003. The power and promise of population genomics: from genotyping to genome typing. Nat. Rev. Genet. 4, 981-994.

Luo, C., Walk, S.T., Gordon, D.M., Feldgarden, M., Tiedje, J.M., Konstantinidis, K.T., 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. Proc. Natl. Acad. Sci. USA 108, 7200-7205.

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B.G., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. USA 95, 3140-3145.

Marteyn, B., West, N.P., Browning, D.F., Cole, J.A., Shaw, J.G., Palm, F., Mounier, J., Prevost, M.C., Sansonetti, P., Tang, C.M. 2010. Modulation of *Shigella* virulence in response to available oxygen in vivo. Nature. 465, 355-358.

Maynard-Smith, J. Smith, N.H., O'Rourke, M., Spratt, B.G. 1993. How clonal are bacteria? Proc Natl Acad Sci USA. 90, 4384-4388.

Maynard-Smith, J., Feil, E.J., Smith, N.H. 2000. Population structure and evolutionary dynamics of pathogenic bacteria. BioEssays. 22, 1115-1122.

Medini, D., Donati, C., Tettelin, H., Masignani, V., Rappuoli. 2005. The microbial pan-genome. Curr. Opin. Gen. Dev. 15, 589-594.

Mira, A., Ochman, H., Moran, N.A. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17, 589-596.

Mira, A., Martin-Cuadrado, A.B., D'Auria, G., Rodriguez-Varela, F. 2010. The bacterial pan-genoma: a new paradignm in microbiology. Int. Microbiol. 13, 45-57.

Moran, N., Wernegreen, J.J. 2000. Lifestyle evolution in symbiotic bacteria: insights from genomics. TREE, 15, 321-326.

Morelli, G., Song, Y., Mazzoni, C., Eppinger, M., Roumagnac, Ph., Wagner, D., Feldkamp, M. et al. 2010. Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. Nat. Genet. 42, 1140-1143.

Nadeau, N.J., Jiggins, C.D. 2010. A golden age for evolutionary genetics? Genomic studies of adaptation in natural populations. Trends Genet. 26, 484-492.

Nakamura, Y. Itoh, T., Matsuda, H., Gojobori, T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat. Genet. 36, 760-766.

Narra, H., Ochman, H. 2006. Of what use is sex to bacteria? Curr. Biol. 16, R705-R710.

Nataro, J.P., Kaper, J.B. 1998. Diarrheagenic *Escherichia coli*. Clin. Microbiol. Rev. 11, 142-201.

Ochman, H., Dávalos, L. 2006. The nature and dynamics of bacterial genomes. Science. 311, 1730-1733.

Okeke, I.N., Wallace-Gadsden, F., Simons, H.R., Matthews, N., Labar, A.S., Hwang, J., Wain, J., 2010. Multi-locus sequence typing of enteroaggregative *Escherichia coli* isolates from Nigerian children uncovers multiple lineages. PLoS One 5, e14093.

OMS. 2008. Global tuberculosis control –Surveillance, planning, financing. World Health Organization, Geneva. http://www.who.int/tb/publications/global_report/2008/en/index.html.

O'Rourke, M., Stevens, E. 1993. Genetic structure of *Neisseria gonorrhoeae* populations: a non-clonal pathogen. J. Gen. Microbiol. 139, 2603-2611.

Orskov, F., Orskov, I. 1983. Summary of a workshop on the clone concept in the epidemiology, taxonomy, and evolution of the enterobacteriaceae and other bacteria. J. Infect. Dis. 148, 346-357.

Pace, N.R. 1997. A molecular view of microbial diversity and the biosphere. Science. 276, 734-740.

Palys, T., Berger, E., Mitrica, I., Nakamura, L.K., Cohan, F.M. 2000. Protein-coding genes as molecular markers for ecologically distinct populations: the case of two *Bacillus* species. Int. J. Syst. Evol. Microbiol. 50, 1021-1028.

Papke, R.Th., Ramsing, N.B., Bateson, M.M., Ward, D.M. 2003. Geographical isolation in hot spring cyanobacteria. Environ. Microb. 5, 650-659.

Paauw, A., Leverstein-van Hall, M.A., Verhoef, J., Fluit, A.C. 2010. Evolution in quantum leaps: Multiple combinatorial pf HPI and other genetic modules in *Enterobacteriaceae*. PLos ONE 5: e8662.

Pérez-Losada, M., Browne, E.B., Madsen, A., Wirth, T., Viscidin, R.P., Crandall, K.A. 2006. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect. Genet. Evol. 6, 97-112.

Philippot, L., Andersson, S.G.E., Battin, T.J., Prosser, J.I., Schimel, J.P., Whitman, W., Hallin, S. 2010. The ecological coherence of high bacterial taxonomic ranks. Nat. Rev. Microbiol. 8, 523-529.

Picard, B., Sevali Garcia, J., Gouriou, S., Duriez, P., Brahimi, N., Bingen, E., et al. The link between phylogeny and virulence in *Escherichia coli* extra-intestinal infection. Infect. Immun. 67, 546-553.

Polz, M., Alm, E.J., Hanage, W.P. 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet. 29, 170-175.

Power, M.L. Littlefield-Wyer, J., Gordon, D.M., Veal, D.A., Slade, M.B. 2005. Phenotypic and genotypic characterization of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes. Environ. Microbiol. 7, 631-640.

Rasko, D.A, Rosovitz, M.J., Myers, G.S., Mongodin, E.F., Fricke, W.F. Gajer, P., Crabtree, J., et al. 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J. Bacteriol. 190, 6881-6893.

Reid, S.D., Herbelin, C.J., Bumbaugh, A.C., Selander, R.K., Whittam, T.S. 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. Nature. 406, 64-67.

Rocap, G., et al. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects ocean niche differentiation. Nature. 424, 1042-1047.

Rocha, E.P.C. 2008. The organization of the bacterial genome. Annu. Rev. Genet. 42, 211-233.

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

Salas, R. 2007. La recombinación: relevancia evolutiva y métodos de estimación, con énfasis en microorganismos. En: Eguiarte, L.E., Souza, V., Aguirre, X. (Editores). Ecología Molecular SEMARNAT-CONABIO. México, D.F. pp: 281-311.

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

Savageau, M.A. 1983. *Escherichia coli* habitats, cell types, and molecular mechanisms of gene control. Am. Nat. 122, 732-744.

Selander, R.K., Levin, B.R. 1980. Genetic diversity and structure in *Escherichia coli* populations. Science. 210, 545-547.

Selander, R.K. Caugant, D.A., Ochman, H., Musser, J.M., Gilmour, M.N., Whittam, T.S. 1986. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl. Environ. Microbiol. 51, 873-884.

Selander, R.K., Caugant, D.A., Whittam, T.S. 1987. Genetic structure and variation in natural populations of *Escherichia coli*. En: Neidhardt, F.C. (Editor). *Escherichia coli* and *Salmonella Typhimurium*. Cellular and Molecular Biology. American Society for Microbiology, Washington, DC. pp., 1625-1648.

Schierup, M.H., Wiuf, C. 2010. The coalescent of bacterial populations. En: Robinson, D.A., Falush, D., Feil, D. (Editores). Bacterial population genetics in infectious disease. Wiley-Blackwell, USA. pp 3-18.

Schubert, S., Darlu, P., Clermont, O., Wieser, A., Magistro, G., Hoffman, Ch., Weinert, K., Tenaillon, O., Matic, Denamur, E. 2009. Role of the spread of pathogenicity islands within the *Escherichia coli* species. PLosPathogens, 5, e1000257.

Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabó, G., Polz, M.f., Alm, E.J. 2012. Events in the ecological differentiation of bacteria. Science. 336, 48-51.

Skippington, E., Ragan, M.A. 2012. Phylogeny rather than ecology or lifestyle biases the construction of *Escherichia coli-Shigella* genetic exchange communities. Open Biol. 2: 120112.

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, 1260-1268.

Souza, V., Rocha, M., Valera, A., Eguiarte, L.E. 1999. Genetic structure of natural populations of *Escherichia coli* in wild hosts on different continents. Appl. Environ. Microbiol. 65, 3373-3385.

Souza, V., Rocha, M., Valera, A., Eguiarte, L.E. 2002. The evolutionary ecology of *Escherichia coli*. Am. Sc. 90, 332-341.

Spratt, B.G., Hanage, W.P., Feil, E. 2001. The relative contributions of recombination and point mutation to the diversification of bacterial clones. Curr. Opin. Microbiol. 4, 602-606.

Suerbaum, S., Smith, J.M., Bapumia, K., Morelli, G., Smith, N.H., Kunstmann, E., Dyrek, I., Achtman, M. 1998. Free recombination within *Helicobacter pylori*. Proc. Natl. Acad. Sci. USA. 95, 12619-12624.

Tettelin, H., Masignani, V., Cieslewicz, M.J., et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. USA. 102, 13950-13955.

Thomas, C.M., Nielsen, K.M. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat. Rev. Microbiol. 3, 711-721.

Tenaillon, O., Skurnik, D., Picard, B., Denamur, E., 2010. The population genetics of commensal *Escherichia coli*. Nat. Rev. Microbiol. 8, 207-217.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., et al., 2009. Organised genome dynamics in *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genetics 5, e1000344.

Tyson, G.W. Chapman, J., Hugenholtz, Ph., Allen, E.E., Rachna, J., et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 428, 37-43.

Van Belkum, A., Melles, D.C., Nouwen, J., van Leeuwen, W.B., van Wamel., Vos, M.C., Wertheim, H.F.L. Verbrugh, H.A. 2009. Coevolutionary aspects of human colonisation and infection by *Staphylococcus aureus*. Infect. Genet. Evol. 9, 32-47.

Vos, M., Didelot, X. 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 3, 199-208.

Walk, S.T., Alm, E.W., Calhoun, L.M., Mladonicky, J.M., Whittam, T.S., 2007. Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. Environ. Microbiol. 9, 2274-2288.

Ward, B.B., O'Mullan, G.D. 2002. World distribution of Nitrosococcus oceani, a marine ammonia-oxidizing y-Proteobacterium, detected by PCR and sequencing of 16 rRNA and amoA genes. Appl. Environ. Microbiol. 68, 4153-4157.

Whitaker, R.J., Grogan, D.W., Taylor, J.W. 2003. Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science. 301, 976-978.

Whitaker, R.J. Grogan, D.W., Taylor, J.W. 2005. Recombination shapes the natural population structure of the hyperthermophilic Archaeon *Sulfolobus "islandicus"*. Mol. Biol. Evol. 22, 2354-2361.

Whitman, W.B., Coleman, D.C., Wiebe, W.J. 1998. Prokaryotes: the unseen majority. Proc. Natl. Acad. Sci. USA. 95: 6578-6583.

Whittam, T.S., Ochman, H., Selander, R.K. 1983. Geographic components of linkage disequilibrium in natural populations of *Escherichia coli*. Mol. Biol. Evol. 1, 67-83.

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L., Karch, H., Reeves, P., Maiden, M., Ochman, H., Achtman, M., 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol. Microbiol. 60, 1136-1151.

Wu, D., et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature. 462, 1056-1060.

Wurtzel, O., Sesto, N., Mellin, J.R., Karunker, I., Edelheit, S., Bécavin, C., Archambaud, Cossart, P., Sorek, R. 2012. Comparative transcriptomics of pathogenic and non-pathogenic *Listeria* species. Mol. Syst. Biol. 8:583.

Xu, J. 2006. Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. Mol. Ecol. 15, 1713-1731.