UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOMÉDICAS

INSTITUTO DE ECOLOGÍA

lución molecular y genética de poblaciones
la isla patogénica LEE en *Escherichia coli*

T E S I S

E PARA OBTENER EL GRADO ACADÉMICO DE

DOCTORA EN CIENCIAS

PRESENTA:

AMANDA CASTILLO COBIÁN

DIRECTORA DE TESIS: DRA. VALERIA SOUZA SALDÍVAR

MÉXICO D. F. OCTUBRE 2005

ÍNDICE

A mi madre Zoila Cobián Michel,

eres todas las razones por las que llegué hasta aquí.

A mi hijo Emiliano, que llena de luz mi vida y hace que todo sea posible.

Al niño que llevamos dentro, que un día sueña y al siguiente el sueño es una realidad, sigamos soñando...

"I seem to have been only like a boy playing on the seashore, and diverting myself in now and then finding a smooter pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me"

-Isaac Newton

"It is not enough to believe what you see. You must also understand what you see"

-Leonardo da Vinci

".....there are no victories in all our histories without love"

-Gordon Sumner

AGRADECIMIENTOS

A la UNAM porque como institución le debo mi formación completa, porque gracias a ella tuve la oportunidad de tener acceso a la educación formal lo que me permitió desarrollarme en todos los ámbitos de mi vida. Es un privilegio pertenecer a la máxima casa de estudios de nuestro país, doy gracias siempre por ello.

A la Doctora Valeria Souza, por ser una guía real en lo académico, una amiga en lo personal, pero sobre todo una gran mujer que con su ejemplo impacta más de lo que se imagina en la vida de sus estudiantes. Gracias por la paciencia, la fé, la disciplina y la presión que hizo que yo cumpliera con esta meta y extrajo de mi cosas que no me creía capaz de realizar.

Al Doctor Luis E. Eguiarte, por el rigor científico con que aborda los problemas, por los cuestionamientos agudos y precisos que fueron parte fundamental de mi formación como bióloga. Por ser un gran maestro en todo el sentido de la palabra.

Al Doctor David Romero ya que, al ser parte de mi comité tutoral del doctorado, estuvo presente durante la gestación y el desarrollo de todo el trabajo y siempre hizo observaciones relevantes, por su entusiasmo al escuchar y discutir los resultados. Sobre todo por su disposición a orientarme académicamente e impulsarme con sus comentarios.

Al jurado que reviso la presente tesis, Dra. Carmen Gómez, Dr. Daniel Piñero, Dr. David Romero, Dr. Guillermo Dávila, Dr. Pablo Vinuesa y Dr. Lorenzo Segovia, que se tomaron el tiempo de leerla, revisarla, hacer aportaciones relevantes para mejorarla y por las discusiones que le dieron otra perspectiva a mi trabajo.

A mis compañeros del Laboratorio de Evolución Molecular y Experimental, por compartir un espacio y un tiempo juntos, por compartir el sueño de la investigación, porque hay un gran equipo con un gran nivel de discusión que hace del laboratorio un lugar muy propicio para las ideas. Me siento privilegiada de haberlos conocido: Claudia, Lulú, Laura, Ana, Martha, Andrea, Ana Noguez, Erika, Luisa, Xitlali, Nuria, Doña Silvia, Salvador, René, Tobías, Germán, Toño, Aldo, Arturo y Eugenio.

Al Dr. Mike Clegg, miembro de la Academia Nacional de Ciencias de los Estados Unidos, que fue el angel guardián del trabajo, tuvo confianza en la calidad del mismo y cuya sencillez lo hace más excepcional de lo que ya es.

A mi familia, porque quien no sabe de donde viene no sabe a donde va....

A mi hermano Omar con amor, porque llenó de risas mi infancia, porque nunca falta cuando necesito ayuda, y porque nadie se parece más a mi.

A Juan Carlos, por el apoyo, por siempre estar ahí en las buenas y en las malas, por darme el mejor regalo que me pudo dar. (We are running in different sides of the river... but we are running together).

A mi abuelo que estará satisfecho y sonriendo desde allá arriba, porque ha sido continua inspiración en mi vida. A mi mamá Grande.

A mi Nina Guille por el amor incondicional, el apoyo en los peores momentos y por ser una segunda madre para mi.

A Guillita, por ser hermanas en muchos sentidos, por compartir la vida y porque cuento con ella para cualquier cosa. A su preciosa semilla que estoy viendo crecer.

A Zaisha que cuidó lo más valioso que tengo y que gracias a esto la tesis llego a buen término. A sus hermanos Dan y Luis porque ellos han llenado de atenciones a mi madre lo que también fue importante para mi concentración en el trabajo.

A mi Tía Nena porque me ha visto crecer y siempre tiene una sonrisa y apoyo en momentos difíciles.

A mi nina Hüera, se que estará feliz. A mi padrino Chema, siempre hay un lugar especial en mi corazón.

A mi tía Carmela, con muchísimo amor por creer en mi.

A mi tío Paco y mi tía Pilar, a mi tío Pepé y mi tía Lydia por el amor y la fé incondicionales.

A todos mis primos con amor.

A Yaayé, porque todos lo días me enseña el significado de la palabra amistad, y porque lo amigos son la familia que podemos escoger, ella es mi hermana. A su familia por la generosidad y amor.

Al Dr. Antonio Lazcano Araujo, por su pasión por la ciencia y su compromiso con la divulgación e inspiración, porque siempre será mi maestro, gracias a él mi vida cambio y se abrieron nuevas oportunidades.

A los compañeros y amigos que de manera indirecta tuvieron que ver con mi desarrollo y con este trabajo: Dorina, Arturo Becerra y León Martínez.

quiero seguir, ir más allá, y no puedo:
se despeñó el instante en otro y otro,
dormí sueños de piedra que no sueña
y al cabo de los años como piedras
oí cantar mi sangre encarcelada,
con un rumor de luz el mar cantaba,
una a una cedían las murallas,
todas las puertas se desmoronaban
y el sol entraba a saco por mi frente,
despegaba mis párpados cerrados,
desprendía mi ser de su envoltura,
me arrancaba de mí, me separaba
de mi bruto dormir siglos de piedra
y su magia de espejos revivía
un sauce de cristal, un chopo de agua,
un alto surtidor que el viento arquea,
un árbol bien plantado mas danzante,
un caminar de río que se curva,
avanza, retrocede, da un rodeo
y llega siempre:

-Octavio Paz

Resumen

El estudio de la genómica comparada ha resultado ser una herramienta poderosa para la comprensión de la evolución y la organización de los genomas. Las herramientas matemáticas y el marco teórico de la genética de poblaciones, en conjunto con el análisis genómico, nos brindan una poderosa aproximación al estudio de las heterogeneidades dentro de la evolución del genoma. En este trabajo se presenta, un análisis jerárquico de la isla LEE (Locus of enterocyte and effacement) (35kb), que se encuentra presente en las cepas enteropatogénica y enterohemorrágica de *Escherichia coli* y *Citrobacter rodentium* y, en la segunda parte, un estudio comparativo de genética de poblaciones de ocho marcadores localizados en distintos sitios dentro del genoma de *E. coli*. La isla LEE en *E. coli* se considera una unidad clonal dentro de un organismo clonal y se espera que evolucione como una unidad genética. El análisis del presente estudio prueba la hipótesis clonal mediante la determinación de la diversidad genética, el contenido de GC, así como, las tasas de substitución nucleotídica en varios niveles funcionales de organización: (i) la isla genómica, (ii) los cinco operones en los que la isla se encuentra organizada y (iii) cada uno de los 41 genes que comprenden la isla. Se encontró que existe un mosaico genético que parece estar moldeado de manera diferencial por la mutación y la selección. Nuestros resultados sugieren que tanto la recombinación como la selección pueden estar rompiendo la estructura clonal de la isla por lo que la mayoría de sus elementos se encontrarán débilmente ligados en su evolución. Estas observaciones sugieren que las unidades de selección no son las islas genómicas, sino unidades mucho más pequeñas dentro de los genes que la integran. Por otra parte, se llevo a acabo el estudio de genética de poblaciones utilizando cinco marcadores moleculares que cubren un amplio espectro de funciones celulares en *E. coli*, dichos marcadores se ubican en distintas regiones a todo lo largo del genoma de esta bacteria (*mdh*, *gapA*, *putP*, *mutS*, *fimA*). Además, se incluyeron los análisis de tres marcadores que se encuentran ubicados dentro de la isla LEE (*eae*, *tir*, *espB*) y que son considerados marcadores de virulencia. Los resultados preliminares del presente trabajo demuestran que existen diferencias significativas entre los procesos evolutivos que moldean a los distintos genes. En principio parece ser que existe gran heterogeneidad en las tasas de substitución molecular (rango: dS=0.097±0.024 para *mdh* a 0.618±0.102 para *tir* y dN=0.0068±0.0006 para *mdh* a 0.175±0.026 para *tir*) y de diversidad (rango de π=0.009±0.0003 para *mutS* a 0.203±0.0054 para *tir* y θ= 0.021±0.0009 para *putP* a 0.143±0.0128 para *tir*). Esto demuestra que los distintos tipos de genes se encuentran bajo diferentes presiones selectivas dentro de las poblaciones de *E. coli*. Este fenómeno tiene repercusiones directas en la descripción de la estructura poblacional de las cepas patógenas y no patógenas de esta bacteria. Los resultados obtenidos en conjunto sugieren que la patogénesis es un estado reciente en *E. coli*. Las pruebas de selección señalan que hemos subestimado el papel de la recombinación y la selección positiva Darwiniana en la evolución y estructura del genoma de *E. coli*. Además, nos señalan qué sitios dentro de los marcadores son propicios a ser sujetos como posibles blancos en el desarrollo de terapias génicas, mutación dirigida y diseño de vacunas lo que nos permitirá lidiar con *E. coli*, que ha demostrado ser un patógeno importante y difícil de controlar.

Abstract

Comparative genomic analysis is a powerful tool for understanding the history and organization of complete genomes. The mathematical tools of population genetics combined with genomic analysis provide a powerful approach to dissect heterogeneities in genome evolution. This study presents a hierarchical analysis of the enterocyte and effacement island (35kb), which is found in the enteropathogenic and enterohemorrhagic strains in *Escherichia coli* and in *Citrobacter rodentium*. In the second part of the study we made a comparative analysis of population genetics of eight molecular markers localized in different regions of the *E. coli* genome. LEE island in *E. coli* is considered a clonal unit inside a clonal organism and is expected to evolve as a single unit. This analysis examines the clonal assumption by determining genetic diversity, GC content, and the substitution rates at the different functional levels of i) the complete pathogenic island, ii) the five operons in which the island is organized, and iii) for each of the individual 41 genes that comprise the locus. We find a genetic mosaic that seems to be differentially affected by selection and mutation. Our results suggest that recombination and selection may be breaking this structure so that different elements are, at best, weakly coupled in their evolution. These observations suggest that the units of selection are not the complete island, but rather, much smaller units that comprise the island. On the other hand, the population genetics study was carried out using five molecular markers localized in different regions of the *E. coli* genome, these markers include a big spectrum of celullar activities in this bateria (*mdh*, *gapA*, *putp*, *mutS*, *fimA*).We also included the study of three virulence related genes localized at LEE island (*eae*, *tir*, *espB*). The preliminary results of the present study demonstrate that there are different selective pressures between and inside the genes. There seem to be heterogeneties in the molecular substitution rates (range: dS=0.097±0.024 for *mdh* to 0.618±0.102 for *tir* and dN=0.0068±0.0006 for *mdh* to 0.175±0.026 for *tir*) and genetic diversity (range: π=0.009±0.0003 for *mutS* to 0.203±0.0054 for *tir* and θ= 0.021±0.0009 for *putP* to 0.143±0.0128 for *tir*). These results suggest that there are different evolutionary processes acting in different genes at different populations of this bacteria which will have direct impact on the ideas about the genetic structure of the populations of *E. coli*. In conclusion this study suggest that pathogenesis in *E. coli* is a derived state, and selection analysis demonstrate that we have underestimated the role of recombination and positive Darwinian selection in the evolution and structure of *E. coli* genome. The selection study results highlight the relevant sites within the genes for future gene therapy analysis, directed mutation experiments, and the development of new vaccines to deal with *E. coli* that is an important pathogen.

PRÓLOGO

"Nothing in biology makes sense except in the light of evolution"

-T. Dobzhansky

Los procariontes a lo largo de su historia natural han desarrollado diversas estrategias ecológicas que han llevado a la colonización de prácticamente todos los hábitats conocidos en el planeta. Desde los ambientes templados hasta los más extremos, las bacterias marcan claramente los límites donde encontramos vida. Dentro de estas estrategias para colonizar nuevos nichos se encuentra el desarrollo de la patogénesis resultado de una depredación hacia un grupo de organismos en específico, generalmente eucariontes. Sin embargo, los mecanismos que subyacen a la evolución y emergencia de nuevas bacterias patógenas no son bien comprendidos, por lo consiguiente los estudios evolutivos acerca del desarrollo de la patogénesis utilizando diversos modelos son de gran importancia.

Este trabajo es una pequeña aproximación al estudio de la evolución de *Escherichia coli*, organismo fascinante debido a la plasticidad ecológica que posee, que de igual forma, se ve reflejada en su plasticidad genómica. El continuo análisis que se ha hecho con organismos modelo (como es el caso de *E. coli*) y el gran conocimiento generado a partir de éstos, nos permite hoy en día hacernos preguntas conceptuales acerca del proceso evolutivo. No pretendo abarcar todo lo que se conoce de *E. coli*, pues no existe tesis de doctorado que pueda hacerlo. Simplemente, intentaré dar un bosquejo acerca de algunos temas importantes relacionados directamente con el objetivo principal del presente trabajo. Todo lo que se presenta a continuación fue desarrollado siempre bajo la perspectiva evolutiva que, en mi caso personal, es el motor que sigue moviendo mi pasión por la ciencia y que no me permite ver los procesos biológicos de otra manera; también, fue lo que permitió que un modelo tan conocido como el trabajado en esta tesis, arrojara nuevos resultados a los ojos de los que nunca pensaron estudiarlo de esta forma.

La tesis fue gestada y desarrollada en el laboratorio de Evolución Molecular y Experimental del Instituto de Ecología de la UNAM, bajo la tutela de la Dra. Valeria Souza. Durante su desarrollo conté con el apoyo económico del proyecto Genómico CONACyT 0028.

Objetivo general y presentación

El objetivo general del presente trabajo es el estudio de la evolución molecular y estructura genética de la isla patogénica LEE en *Escherichia coli*. Para cumplir con este objetivo se utilizaron en conjunto las herramientas de la genómica comparada y la teoría de la genética de poblaciones. El trabajo se divide de manera general en dos partes, la primera comprende el estudio de genómica comparada de seis islas LEE. La segunda parte se integra por el estudio comparativo de genética de poblaciones de una submuestra de tres genes pertenecientes a LEE y cinco genes localizados en el cromosoma de *E. coli*.

La tesis está compuesta por seis capítulos. El primero corresponde a la introducción general. El capítulo dos está conformado por un artículo de divulgación acerca del desarrollo de la patogénesis *"The evolutionary ecology of Escherichia coli"* publicado en la revista **American Scientist** (2002) **90** (4): 332-341. En el tercer capítulo se desarrolla la teoría de la selección natural a nivel molecular, este capítulo a su vez formará parte de un libro de diversos temas acerca de Ecología Molecular cuya publicación se encuentra a cargo del Dr. Luis E. Eguiarte Fruns. El cuarto capítulo está integrado por el artículo *"A genomic population genetics analysis of the pathogenic enterocyte effacement island in Escherichia coli: the search for the unit of selection"* publicado en la revista **Proceedings of the National Academy of Sciences of the USA** (2005) **102** (5): 1542-1547. En el quinto capítulo se desarrolla una breve introducción a las principales ideas de la genética de poblaciones haciendo énfasis en la estructura poblacional de *E. coli* y algunos marcadores moleculares utilizados. Para el estudio de genética de poblaciones se presentan algunos resultados preliminares para una próxima publicación. El sexto capítulo corresponde a la discusión y conclusiones generales, haciendo una integración de la temática que se desarrolló durante el presente trabajo.

"Evolución molecular y genética de poblaciones de la isla patogénica LEE en *Escherichia coli*"

Capítulo 1. Introducción General

"Once we understand the biology of Escherichia coli we will understand the biology of an elephant"
-Jacques Monod

1.1 Escherichia coli

Escherichia coli es probablemente el organismo mejor estudiado. Ha sido el modelo para la descripción del metabolismo bacteriano, los procesos de división celular, la biosíntesis de la pared celular, la quimiotaxis, la mayor parte de la genética bacteriana y de la biología molecular en general. Pertenece al grupo de las Enterobacterias y a la subclase de las proteobacterias gamma asimismo, este grupo incluye a varios patógenos importantes de humanos como *Salmonella* y *Vibrio*. Metabólicamente se caracterizan por ser anaerobias facultativas, es decir, en ciertas condiciones ambientales pueden respirar y en condiciones de estrés oxigénico utilizan aceptores de electrones alternativos. Debido a esta plasticidad metabólica podemos encontrarlas como organismos de vida libre así como comensales y patógenas (Selander y col., 1987). Además, *E. coli* es una de las primeras especies bacterianas en colonizar al mamífero recién nacido, a partir del canal de parto y de las heces de su madre (Bettelheim, 1994).

1.2 El genoma de Escherichia coli

El primer genoma de *E. coli* secuenciado en su totalidad fue el de la cepa no patógena K12 MG1655 con un tamaño de 4.6 Megabases (Mb) (Blattner y col., 1997). A pesar de que no fue el primer genoma bacteriano secuenciado, si fue el primero en iniciar su completa secuenciación y es el que mejor descrito se encuentra. Debido a la técnica con que fue secuenciado (ya que se realizó en un inicio su mapa genético que sirvió posteriormente de base para la secuenciación del genoma), no pudo ser terminado en el tiempo estipulado y el primer genoma bacteriano corresponde a *Haemophilus influenzae* (Fleischmann y col., 1995).

Posteriormente, y debido a su importancia médica han sido secuenciados al menos cuatro genomas más de cepas de esta bacteria patógenas para el humano. Además, se encuentran en proceso de secuenciación varios genomas que abarcan un amplio espectro de la patología que esta especie posee: EPEC (Enteropatógena) E2348/69 (5.06 Mb), EAEC (Enteroagregativa) (5.35 Mb) y NMEC K1 (causante de la meningitis) (5.2 Mb) (www.genome.wisc.edu/sequencing.htm). Con el estudio comparado de distintos genomas de *E. coli* se pretende caracterizar la "patosfera" que correspondería a la poza genética que comparten las cepas patógenas de esta especie, y de esta forma, describir algunos patrones y rasgos característicos del desarrollo de la patogénesis con el fin de entender este fenómeno desde una perspectiva comparativa y evolutiva. Por otra parte, también se pretende la caracterización de genomas pertenecientes a otras enterobacterias, como es el caso de *Shigella flexneri*, *Shigella sonnei*, y *Shigella dysenteriae*, *Yersinia pestis* y varias especies del género *Salmonella* (www.genome.wisc.edu/resources/internet.htm#other).

Con la secuenciación de *E. coli* K12 se descubrió que existía una gran cantidad de genes extranjeros no pertenecientes a ésta especie y se determinó que alrededor del 18% de sus genes eran producto de transferencia horizontal (Lawrence y Ochman, 1998). Además, que el promedio de genes introducidos por este proceso fue de aproximadamente 16kb (kilobases) por millón de años desde la divergencia entre *E. coli* y *Salmonella* durante 234 eventos de transferencia horizontal (Lawrence y Ochman, 1998). Los resultados de este trabajo, entre otros, demuestran que existe un gran flujo de información genética entre los genomas bacterianos, no necesariamente entre la misma especie o entre especies cercanas; y que al parecer, la transferencia horizontal es un mecanismo muy importante dentro de la evolución y dinámica de los genomas procariontes (Gogarten y col., 2002), más de lo que se había sugerido con anterioridad (Cohan, 1996).

Por otra parte, el estudio comparativo entre los genomas completos de cuatro cepas de *E. coli* , las patógenas enterohemorrágicas (EHEC) (5.5 Mb) EDL933 (Perna y col., 2001) y O157 Sakai (5.5 Mb) (Hayashi y col., 2001), y la uropatógena CFT073 (5.2 Mb) (Welch y col., 2002), con la no patógena K12 MG1655 (4.6 Mb) (Blattner y col., 1997), ha mostrado la existencia de un esqueleto común de aproximadamente 4Mb que puede ser considerado como el núcleo genético de las cepas de *E. coli*. También mostró que las principales diferencias residen en la presencia de un mayor número de genes nuevos y

factores de virulencia en las cepas patógenas (Perna y col., 2001; Hayashi y col., 2001; Welch y col., 2002) y que la estructura general del genoma es un mosaico. Este mosaico está compuesto por el núcleo conservado de genes y numerosas islas genómicas intercaladas a lo largo del genoma. Uno de los descubrimientos más importantes derivado de la comparación múltiple de los genomas de *E. coli*, además del mosaicismo observado, es la gran diversidad presente entre las distintas cepas pertenecientes a la misma especie donde alrededor del 25% del genoma es único a las cepas patógenas (Welch y col., 2002). Además, es claro a partir de estos estudios que el genoma se encuentra organizado en varios niveles, las islas genómicas, los operones y los genes.

1.3 Islas Genómicas

Las islas genómicas fueron descritas por primera vez en *E. coli* uropatógena y posteriormente identificadas en otros grupos bacterianos (Knapp y col., 1986; Blum y col., 1994; Hentschel y col., 2000). Son elementos genéticos compuestos por grupos de genes, generalmente organizados en operones, y que en su conjunto codifican para realizar una actividad celular específica. Esta actividad puede ser el desarrollo de una estructura (fimbria, cilios), la producción de una toxina, la producción de un sistema de transducción de señales, el desarrollo de una determinada capacidad metabólica, resistencia a antibióticos, etc, (Hentschel y Hacker, 2001). Las islas genómicas comparten ciertas características genéticas como es su inserción en sitios específicos dentro del genoma como los tRNA (RNA de transferencia), un contenido de GC y uso de codones diferente al resto del genoma y se encuentran flanqueadas por elementos móviles como las secuencias de inserción (IS), lo que hace suponer que se mueven por transferencia horizontal (Hentschel y Hacker, 2001). Se considera que tanto las islas genómicas como los operones en que se encuentran organizadas son unidades genéticas compuestas por grupos de genes que se transcriben en conjunto y cuyos productos contribuyen a una función determinada, además de que se encuentran ligados en su historia evolutiva (Lawrence y Roth, 1996). Dentro de los numerosos tipos de islas genómicas identificadas encontramos las islas patogénicas.

1.4 Islas patogénicas

Las islas patogénicas (PAIs, por sus siglas en inglés) tienen un tamaño de 10 a 200 kb, codifican para funciones asociadas directamente a la patogénesis, así como habilidades metabólicas adicionales y producción de toxinas. Se encuentran presentes en las cepas patógenas y ausentes en las no patógenas, ya sea de la misma especie o de especies relacionadas. Se encontró que el contenido de GC y el uso de codones en las PAIs difieren del resto del genoma, lo que sugiere que dichas regiones fueron adquiridas por transferencia horizontal (Hacker y col., 1997; Hacker y Kaper, 2000). La mayor parte de los elementos que integran las PAIs se localizan dentro del genoma, aunque existen algunos factores que se pueden encontrar dentro de plásmidos y bacteriófagos. Las PAIs se insertan normalmente en sitios donde encontramos tRNA, siendo los más comunes el *selC*, *pheU* y *pheV* (Groisman y Ochman, 1996; Wieler y col., 1997; Hacker y Kaper, 2000). Se ha propuesto que la versatilidad del genoma de las cepas patógenas de *E. coli* es conferida por estos dos elementos genéticos, las islas patogénicas asociadas al cromosoma y los plásmidos que contienen ciertos factores de virulencia (Nataro y Kaper, 1998). Se debe tomar en cuenta que algunos de estos factores también pueden ser localizados en cepas no patógenas en donde funcionan en otro tipo de procesos no relacionados a la patogénesis (Burland y col., 1998). Una de las islas patogénicas más estudias es la isla LEE (Locus of enterocyte and effacement) (Jerse y col.,1990; McDaniel y col., 1995).

1.5 Cepas patógenas de Escherichia coli

Aunque el hábitat natural de *E. coli* es el tracto intestinal de animales de sangre caliente, donde es un componente normal de la microbiota, también existen cepas patógenas que pueden causar diferentes enfermedades en animales y humanos. Estas cepas expresan factores de virulencia y poseen otros elementos genéticos asociados a la patogénesis. Muchos de estos elementos permiten a los patógenos adaptarse y persistir dentro de las comunidades bacterianas (Donnenberg y Whittam, 2001). Las cepas patógenas se caracterizan por la producción de factores de virulencia como adhesinas, toxinas, proteínas invasivas, sistemas de secreción y de obtención de hierro, muchos de ellos codificados en las islas patogénicas (Hacker y col., 1990). Es la combinación de los factores de virulencia lo que determina el tipo de cepa patogénica y como resultado la patología que se va a

desarrollar. No se ha determinado con seguridad si la patogénesis es un estado ancestral en *E. coli* o es derivado, es decir, si *E. coli* es patógena desde su origen o fue una capacidad que adquirió de manera tardía dentro de su historia evolutiva. Algunos estudios sugieren que es un estado derivado (Lecointre y col., 1998; Reid y col., 2000; Castillo y col., 2005).

Cuatro síndromes clínicos generales en humanos resultan de la infección con alguna cepa patógena de *E. coli*: infecciones del tracto urinario, septicemia, meningitis y diarrea (Nataro y Kaper, 1998). Las cepas patógenas que han sido asociadas a las enfermedades diarreicas son las siguientes: *E.coli* EHEC, EPEC, enterotoxigénica (ETEC), enteroinvasiva (EIEC) y enteroagregativa (EAGEC) (Nataro y Kaper, 1998).

1.6 Modelo de estudio

En el presente trabajo, para llevar a cabo el estudio de la evolución de LEE se han tomado como modelo las cepas *E. coli* EPEC y EHEC y la especie hermana *Citrobacter rodentium*. Este grupo de organismos son patógenos que desarrollan una lesión específica en el intestino del hospedero. En el caso de *C. rodentium* este produce la lesión en ratones. La infección con EPEC y EHEC genera en el enfermo diarrea aguda y gran mortandad, sobre todo en niños menores de 2 años y adultos mayores. Las EPEC han sido caracterizadas por el serotipo que presentan, la histopatologia que producen y la ausencia de producción de la toxina tipo Shiga (las EHEC si producen la toxina). Estos patógenos causan un tipo de lesión localizada conocida como A/E (lesión de adherencia y esfacelamiento) (Cravioto y col., 1979; Moon y col., 1983), donde las bacterias se adhieren al intestino y borran las vellosidades de la mucosa del intestino delgado. También se caracteriza por producir cambios evidentes en el citoesqueleto, que incluyen la polimerización de actina y la formación de una estructura tipo pedestal característica de la lesión.

Los diferentes patotipos de *E. coli* EPEC y EHEC se encuentran concentrados en grupos clonales que comparten serotipos entre sí. Dentro de estos se han reconocido dos grupos de cepas EPEC (EPEC 1 y 2) y dos grupos de cepas EHEC (EHEC 1 y 2). Los grupos EHEC-1 y EPEC-1 se encuentran relacionados entre sí aunque sus secuencias son más divergentes, mientras que los grupos EPEC-2 y EHEC-2 son más cercanos y conservados (Donnenberg y Whittam, 2001). Además, se ha sugerido que las cepas EHEC

col., 1995), múltiples proteínas secretadas (*espA, espB, espD*) (Creasey y col., 2003), una adhesina denominada intimina (*eae*)(McDaniel y col., 1995; Nataro y Kaper, 1998), su receptor *tir* (Nataro y Kaper, 1998; Hartland y col., 1999) y algunas chaperonas (*cesD, cesT*) (Creasey y col., 2003). El SSTT es un sistema de proteínas encargadas de dirigir la transferencia de proteínas específicas a través de la membrana bacteriana y transferirlas como efectores hacia el hospedero (Jarvis y col., 1995). El receptor de la adhesina *tir* es transferido hacia las células del hospedero donde es modificado y posteriormente insertado en la membrana plasmática del hospedero, de tal forma que es presentado a la adhesina *eae*, promoviendo la adherencia íntima entre la célula bacteriana y la del hospedero (Kenny y col., 1997). Posteriormente, ocurren modificaciones que culminan con la alteración del citoesqueleto de la célula intestinal del hospedero y la formación de la estructura tipo pedestal característica de los patógenos A/E. Asimismo, la proteína *espA* forma un conducto en forma de filamento a través del cual otras proteínas secretoras viajan y llegan al poro (conformado por *espB* y *espD*) en la membrana plasmática del hospedero (Creasey y col., 2003; Clarke y col., 2003). Muchas de las proteínas antes de ser secretadas permanecen en el citoplasma asociadas a sus chaperonas, como *cesD*, *cesF* y *cesT* chaperonas de *espD*, *espF* y *tir*, respectivamente.

Los genes de LEE se encuentran organizados en cinco operones denominados LEE1, LEE2, LEE3, TIR y LEE4, todos ellos regulados positivamente por el gen *ler* localizado en LEE1. Recientemente, se han descrito dos nuevos posibles reguladores dentro de la isla: *orf10*, que al parecer codifica para un regulador negativo y *orf11* que parece ser un regulador positivo (Deng y col., 2004). Sin embargo, a pesar de que se encuentran bien reconocidos los genes pertenecientes a LEE, casi la mitad de ellos (a excepción del SSTT y la adhesina) no se les ha asignado una función específica y/o no poseen homólogos en otras especies bacterianas.

1.8 Origen y evolución de la isla LEE

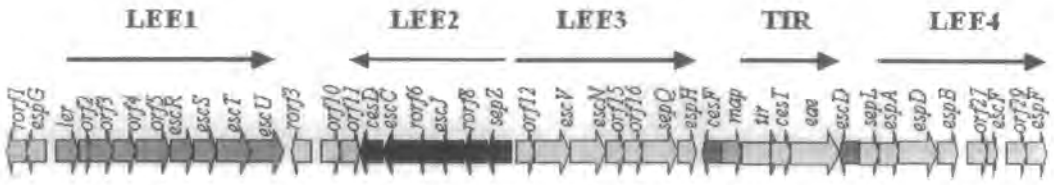
El conocimiento que se tiene de *E. coli* y de los mecanismos del desarrollo de la patogénesis y origen de LEE es fragmentario. Por una parte tenemos trabajos donde se abordan el análisis de genes específicos, así como, el mecanismo del desarrollo de la lesión en diferentes cepas como la EHEC (Perna y col., 1998), la ubicación y contenido de GC, la

propuesta de árboles filogenéticos usando como referencia la colección ECOR (Ochman y Selander, 1984; Reid y col., 2000), y algunos estudios comparativos generales (Lecointre y col., 1998). Por otra parte, en estudios realizados a partir de datos de isoenzimas obtenidos de cepas mexicanas de *E. coli*, se construyó un fenograma de parentesco utilizando el método de UPGMA (Rocha, 1999; Souza y col., 1999; Sandner y col., 2001). Se usó este análisis como punto de referencia para registrar la presencia o ausencia de los genes relacionados con la lesión A/E (Rocha, 1999). Los genes *eae*, *espB*, *per* y *bfp* fueron amplificados por PCR (Reacción en cadena de la polimerasa) y localizados en el árbol de isoenzimas. Los resultados obtenidos indican que los genes *eae* y *espB* pueden encontrarse a todo lo largo de la genealogía de las distintas cepas aisladas a partir de diferentes hospederos (Souza y col., 1999). Lo que nos hace suponer que la historia de la evolución de LEE en las poblaciones naturales de *E. coli* es más compleja de lo que se suponía con anterioridad a partir de los estudios en cepas clínicas. Por otra parte, en una submuestra de 54 cepas (36 de animales y 18 de humanos) se amplificaron los genes *eae*, *tir* y *espB* (Sandner y col., 2001). Estos genes fueron secuenciados y se encontró en un análisis preliminar que los genes *eae* y *tir* son sumamente variables y que al parecer las muestras poseen más alelos que los descritos con anterioridad, mientras que el gen *espB* es menos variable. En cuanto a los genes plasmídicos *per*, *bfp* y *eaf*, sólo se encontraron en las cepas EPEC y EHEC clínicas, lo cual hace pensar que es posible que existieran eventos de transferencia horizontal que dieron lugar al plásmido patogénico (Rocha, 1999).

Estudios con cepas que producen infecciones extraintestinales sugieren que la adquisición de las islas patogénicas es un fenómeno ancestral y que posteriormente se dispersó a otros linajes por transferencia horizontal (Boyd y Hartl, 1998). Pupo y colaboradores (1997) proponen que la adquisición de los factores de virulencia ocurrió polifiléticamente en numerosas ocasiones (Pupo y col., 1997). En general todos los trabajos sugieren que LEE es una unidad (Reid y col., 2000) y se ha considerado que existe diversidad en cada gen, no por los orígenes sino por las diferentes funciones y tasas de selección a las que están sujetos cada uno. Por otra parte, se ha propuesto que la adquisición de los plásmidos que codifican para factores de virulencia podría haber ocurrido de forma paralela y progresiva una vez que la isla se encontraba inserta en las cepas EPEC (Reid y col., 2000).

A pesar de que se han reconocido los factores genéticos involucrados en el desarrollo de la lesión A/E se sigue considerando que la evolución y la dinámica de la transferencia del locus LEE es una gran pregunta con varios puntos por resolver. Para tener una mayor comprensión acerca de la evolución de esta isla patogénica consideramos que es necesario conocer los parámetros evolutivos distintivos de cada gen que participa en su desarrollo y hacer un análisis comparativo con genes no relacionados a la patogénesis. Para esto, es necesario comparar las tasas de substitución, tipo de selección, mutación y recombinación. A partir de estos resultados podremos tener una idea más clara de la dinámica de la isla y esto nos llevará a una mejor comprensión de la transferencia de los factores de virulencia entre las distintas cepas de *E. coli,* que es uno de los principales eventos generadores de la plasticidad genómica que se ha registrado en esta bacteria.

Fig. 1.1. Estructura de la isla LEE en *E. coli*. La isla comprende 41 genes organizados en 5 operones, se señalan los genes, los operones (delimitados por los genes del mismo color) y la dirección de transcripción de los mismos.



Capítulo 2. Ecología evolutiva de *Escherichia coli.*

Resumen

Escherichia coli ha sido uno de los organismos probablemente mejor estudiados y temidos debido al amplio espectro de enfermedades que produce. Desde septicemia y meningitis hasta diversos tipos de diarreas. Los avances más recientes acerca de la biología de esta bacteria han arrojado resultados sorprendentes. La comparación entre genomas de cepas patógenas como lo es la enterohemorrágica EHEC y no patógena como la cepa K12, han mostrado la gran plasticidad genómica que esta bacteria posee, mucho más que la que se había supuesto con anterioridad. Además, demostraron que existía más de un solo camino para que un patógeno evolucione.

The Evolutionary Ecology of *Escherichia coli*

Abundantly studied and much feared, E. coli *has more genomic plasticity than once believed and may have followed various routes to become a pathogen*

Valeria Souza, Amanda Castillo and Luis E. Eguiarte

Bacteria are full of surprises. Consider the most familiar and most studied of all cellular life forms: *Escherichia coli*. Although it has been subjected to scientific scrutiny for more than a century and has occupied center stage in the development of genetic-engineering technologies in the laboratory, *E. coli* continues to confound our ideas of how bacteria reproduce, adapt and colonize new niches.

In 1994, for example, Stephen Jay Gould wrote that "the most salient feature of life has been the stability of its bacterial mode from the beginning of the fossil record until today." Just eight years later, new insights into the nature of *E. coli* and its close relatives have made such a view of the bacterium as a stable organism seem superficial. Having now sequenced certain *E. coli* genomes and studied the population genetics of numerous bacterial species, we know that although the bacteria have undergone little change in morphology, their genome is a small but dynamic and changing entity that has not stopped evolving.

Recent advances in our understanding of the genetics and physiology of *E. coli* have in fact been spectacular. We know the entire genome sequences of three strains of this species, and the *E. coli* genome is undoubtedly the best understood of any genome (approximately 70 percent of its genes being "annotated," in the terminology of genomics). It has also been used as a model organism in evolutionary studies, both in natural populations and in the laboratory in so-called "experimental evolution" studies. These investigations have allowed us to understand better the action of two evolutionary forces, selection and mutation, over a very long time. These studies were based on the prevailing notion that these bacteria are clonal, passing genes from generation to generation with little scrambling or swapping—a notion that was entirely upset as the 20th century came to a close.

However, we have just begun to investigate bacterial ecology and evolutionary biology in natural populations. Such studies have gained urgency in connection with recent outbreaks of some pathogenic, foodborne strains of *E. coli*. These strains have virulence factors and genetic "pathogenicity islands" that have made *E. coli*, long a killer of infants in poor countries, a growing threat everywhere in the world. It is important to understand how this happens.

Valeria Souza is a professor in the Faculty of Sciences at the National Autonomous University of Mexico (UNAM) and an investigator in the Laboratory of Molecular and Experimental Evolution, Department of Evolutionary Ecology in UNAM's Institute of Ecology. Amanda Castillo is a doctoral candidate in biomedical sciences at UNAM. Luis Eguiarte chairs the Department of Evolutionary Ecology. This article was adapted from a review in the October 2001 issue of Interciencia. *Address for Souza: Laboratorio de Evolución Molecular y Experimental, Depto. de Ecología Evolutiva, Instituto de Ecología, UNAM, Apto. Postal 70-275, C.U., C.P. 04510, México D.F., México. Internet: souza@servidor.unam.mx*

Clues in the Genome

Our familiarity with *E. coli* comes from our intimate experience with it as well as its widespread use as a workhorse in the genetics laboratory. *E. coli* lives in the gut of human beings and of many other mammals, domestic and wild; it also lives elsewhere in nature. *E. coli* makes headlines when a pathogenic strain contaminates food or drink, but in fact there are many benign strains living unnoticed among our normal intestinal fauna and in our environment.

E. coli is a member of the Enterobacteriaceae family. Molecular phylogeny studies indicate that it is closely related to some other pathogens of vertebrates, including *Shigella* and *Salmonella*, the *Vibrio* cholera bacteria and *Haemophilus*, which can cause pneumonia and meningitis. The enterobacteria are characterized by their capacity for facultative respiration: They are aerobic in the open air but live anaerobically inside the gut. Thanks to this versatility, many members of this family are free-living, whereas others live in commensal relationships with animals or plants.

A harmless strain of *E. coli* called K-12, widely used in genetic-engineering experiments, has been well studied, and the genome of one variant has been sequenced by Frederick R. Blattner and his colleagues at the *E. coli* Genome Project at the University of Wisconsin-Madison. This strain contains 4,639,221 base pairs, or 4.6 megabases, of double-stranded circular DNA (compared with the 3 *billion* base pairs of chromosomal DNA in the human genome). Of this genome 87.8 percent codes for proteins, and 0.8 percent codes for RNAs, or ribonucleic acids, key workers in protein synthesis. Another 0.7 percent consists of DNA without any known function. It is estimated that around 11 percent of the chromosome has regulatory functions. Some 28 percent of the 4,288 open reading frames (arrays coding for proteins) have no known function.

Other strains of *E. coli* may have differences in their genomic structure; it is now suspected that the maps are not always colinear and that the size of the genome varies from 4.4 to 5.5 megabases. The other strain that has been sequenced, a dangerous enterohemorrhagic *E. coli* (EHEC) known as O157:H7, appears to have acquired many of its genes by *horizontal transfer* since it diverged from K-12

Andrew Wallace/Reuters/Corbis

Figure 1. Pathogenic strains of the bacterium *Escherichia coli* cause dramatic illness that can be fatal. Here a child infected by the pathogenic strain O157:H7 in Canada's biggest *E. coli* epidemic is evacuated for hospital treatment. *E. coli* and *Campylobacter jejuni* from manure contaminated the water supply of Walkerton, Ontario, in May 2000, infecting at least 2,300 people and killing 7. The authors have found clues to the evolution of pathogenicity in *E. coli* and related bacteria by studying a wide range of harmful and harmless strains from around the world. Their studies support an emerging view of *E. coli* as a diverse organism possessing the genomic plasticity to successfully invade new ecological niches and assemble an array of tools for causing illness.

about 4 million years ago. Horizontal or lateral gene transfer is a special talent of bacteria, which can exchange DNA within or across species lines. Such gene-swapping takes place through bacterial *conjugation*, when two bacteria join and share DNA; through the intervention of *bacteriophages*, or bacteria-infecting viruses; or through *transformation*, wherein they take up "loose" DNA from their environment.

O157:H7, the culprit in several recent fatal outbreaks of foodborne disease in the U.S. and Europe, turns out to have 1,387 genes that K-12 lacks; these include virulence factors, certain metabolic pathways and *prophages* (DNA acquired from viruses), as well as genes enabling DNA elements to move around on a chromosome. It is remarkable that a strain can have so many novel genes; these comprise some 25 percent of the O157:H7 genome. By comparison, all human beings are thought to be genetically about 99 percent identical.

When they released the O157:H7 sequence last year, Nicole T. Perna of Blattner's group at the University of Wisconsin and her colleagues noted that a comparison of the K-12 and O157:H7 genomes shows that the enterobacteria are the subjects of a great deal more genetic *recombination*, or scrambling of genes, than had been suspected. Lateral transfer creates bacterial genomes that are mosaics of genes with different evolutionary histories. For example, geneticists can find markers of inheritance by looking for patterns in the nucleotide bases that link to form double-stranded DNA. One common statistic is the proportion of base pairs that have the arrangement G-C, in which the bases guanine and cytosine are linked. The *E. coli* genome on average consists of 50.8 percent G-C pairs. But a number of important genes (15 percent of the K-12 genome and 26 percent of the O157:H7 genome) contain different G-C proportions from the rest of the genome and also use *codons* (triplets of bases coding for a single amino acid) in a very different way. For this reason it has been suggested that these genes came from other bacterial lines and were acquired by *E. coli* recently via horizontal transfer.

Jeffrey G. Lawrence of the University of Pittsburgh and Howard Ochman of the University of Arizona in Tucson in 1998 estimated the rate of these

transfers to be at least in the range of 16 kilobases every million years; "pathogenic islands" (regions where genes that confer pathogenicity are found) dominate this activity.

Plasmidic Plasticity

Bacteria carry some of their genetic information in the form of extrachromosomal elements known as *plasmids*. These circular DNA molecules are the most dynamic component of the bacterial genome, since they move easily between strains. Plasmids are common in *E. coli*, although the bacterium can survive without a plasmid or, at the other extreme, store a good percentage of its genome in such elements. Around 300 kinds of plasmids have been described in the species. Within them one can find information for assimilating rare sugars and for producing *colicins* (substances that kill possible competitors of the same species); resistance to antibiotics and heavy metals; immunity against bacteria-targeting viruses and colicins; genes that code for genetic exchange; and filaments related to pathogenesis and the production of toxins.

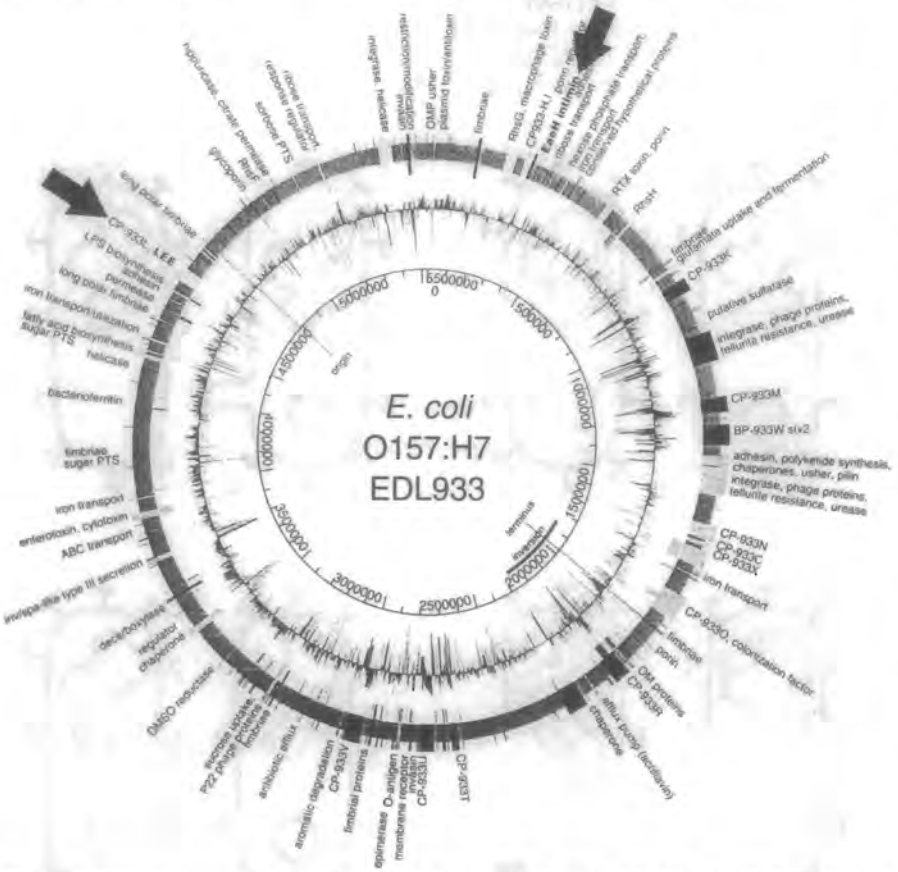


Figure 2. Circular genomic map of the *E. coli* strain O157:H7, published by Nicole T. Perna and colleagues at the University of Wisconsin in January 2001, showed characteristics that confounded scientists' understanding of the genomic stability of bacteria. O157:H7, an enterohemorrhagic pathogen, had 1,387 genes lacking in the previously sequenced *E. coli* strain K-12 genome. The new genes included virulence factors and DNA acquired from viruses. A number of important genes have a very different proportion of G-C (guanine-cytosine) nucleotide base pairs from the rest of the genome; this variation from the mean is shown in the second circle. Examples marked by arrows include a group of genes called LEE, or locus of enterocyte effacement, and another group coding for the attachment protein intimin. (Illustration adapted from Perna *et al.* 2001, courtesy of Nicole T. Perna; used by permission of *Nature*.)

Many plasmids are capable of transferring themselves among different species. Different models of plasmid distribution are possible in bacteria. Based on studies in 1997 of *Salmonella* and *E. coli* by E. Fidelma Boyd and

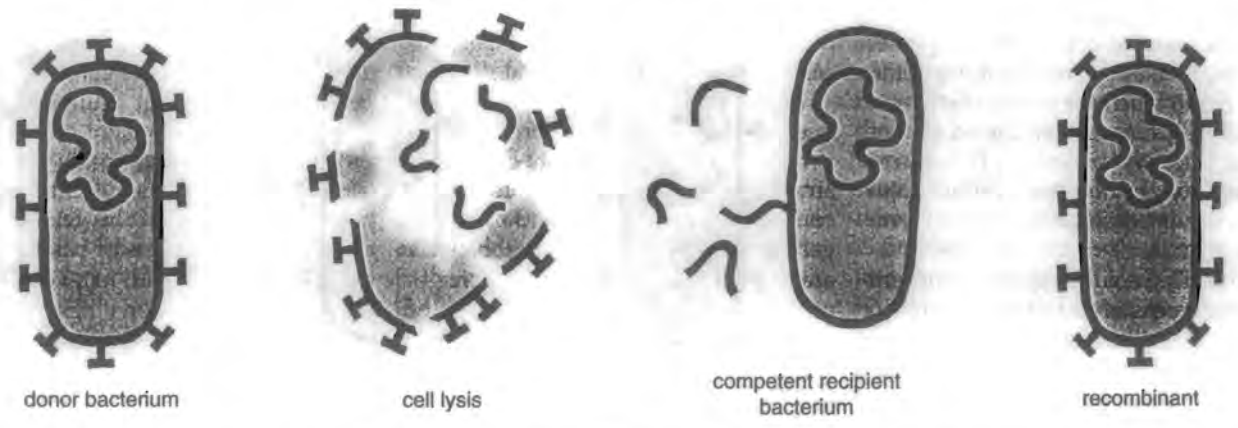


Figure 3. Bacteria acquire new genes through a variety of kinds of lateral transfer or gene-swapping. One method is transformation, shown here. A "competent" bacterium—one capable of acquiring DNA from the surrounding environment—acquires genes released from other bacteria that have broken up and incorporates the new DNA into its chromosome. Bacteria can also acquire genes from viruses called bacteriophages.

Daniel Hartl of Harvard University, our laboratory has advanced a *panmictic*, or random-mating, model of plasmid distribution. Some plasmids are extremely successful and, being promiscuous as well, are overrepresented in bacterial populations. These epidemic plasmids allow bacteria to acquire virulence factors or resistance to antibiotics by horizontal transfer. Promiscuous plasmids can contribute to coevolution; as they move between bacterial species, the extrachromosomal genomes of those species may evolve in parallel. However, there are also clonal plasmids, plasmids that are only transferred from parent to child in asexual reproduction, as well as plasmids whose transfer is limited to specific genomes within the same bacterial species.

The great genomic plasticity of *E. coli* has conferred on it an extraordinary ecological plasticity. *E. coli* can adapt rapidly to different environments and is capable of existing as a free-living organism or in commensal mutualism in the colons of mammals and birds. Additionally, in the interior of host organisms, it can invade other niches successfully. In this way it can become a dangerous pathogen that successfully colonizes people and animals.

E. coli in Its Environment

In spite of the abundance of bacteria, the study of their ecology is extraordinarily difficult and generally relies on indirect measurement. We believe the best way to understand bacterial ecology is through the use of genetic markers and the techniques of two fields: population genetics and molecular evolution.

Traditionally the colons of mammals and birds have been regarded as the natural habitat of *E. coli*. When *E. coli* causes illness, fecal contamination of food or water is commonly suspected; that is, health workers look for a way that *E. coli* from one mammalian colon could have gotten introduced into another's digestive system. It was long believed that the bacteria cannot reproduce on external media. However, recent results indicate that there are strains of *E. coli* that occupy niches other than the colon. Prominent among these are the pathogenic *E. coli* that can live in other parts of the digestive tract, in the blood, in the urogenital tract and in secondary environments. Strains found in drains and aquatic environments are in general more diverse than strains obtained directly from hosts.

A number of studies have found that aquatic and soil bacterial populations can increase their density over time, indicating that they grow and survive in these external environments. These studies suggest that *E. coli* infections can come from sources other than fecal contamination. But life for *E. coli* in a nutrient-poor environment such as water or mud is not the same as life in the rich environment of the mammalian gut. Bacteria in nutrient-poor environments divide at around 10 percent of the rate achieved in the laboratory.

Biologists know a number of things about the population ecology of *E. coli* that live in commensal relations with host animals. Generally there is one

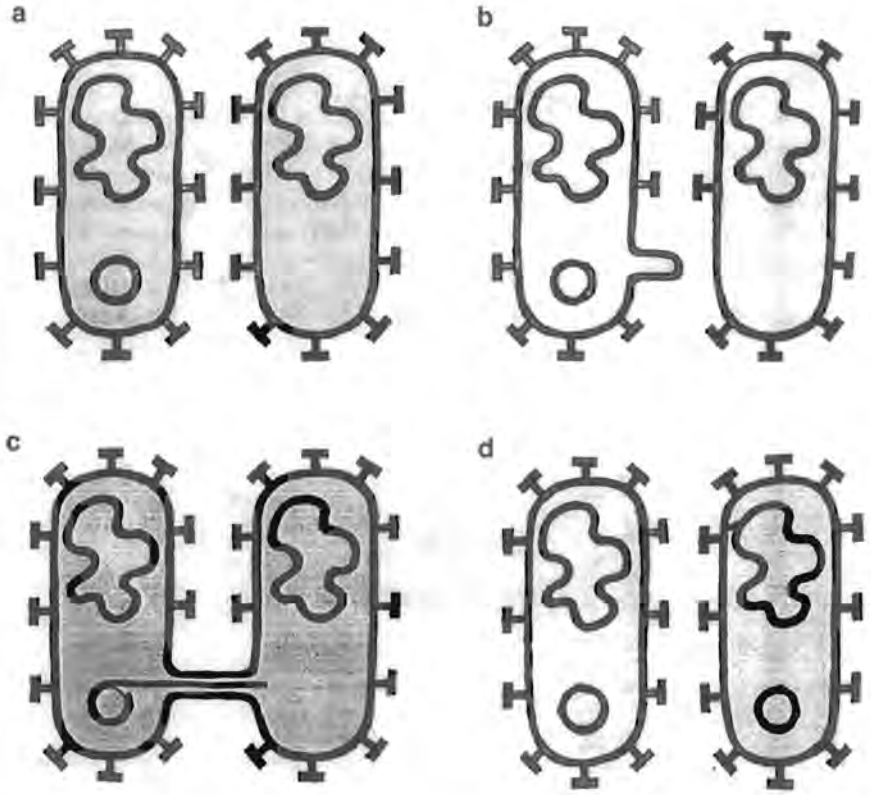


Figure 4. Bacterial conjugation is a second method of lateral transfer. In this example, one bacterium has an extrachromosomal genetic element, or plasmid, that the other lacks *(a)*. The plasmid includes genes coding for proteins that allow the genes to recognize each other and join. The bacterium with the plasmid extends a projection *(b)* to join with the second bacterium and makes a copy of the plasmid for its neighbor *(c)*. The neighbor thus acquires the traits of the original bacterium, which may include virulence factors or resistance to antibiotics.

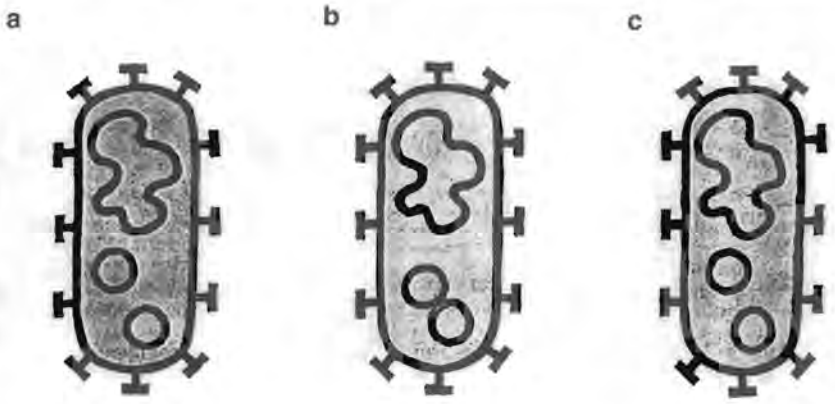


Figure 5. Genes can also be exchanged between plasmids and can move around on a chromosome or plasmid; in fact, the genes exchanged may be genes that allow recombination and transposition of genetic elements. Here a pair of plasmids with different traits *(a)* interact within the cell *(b)* so that an element on one plasmid is replicated and added to the second plasmid.

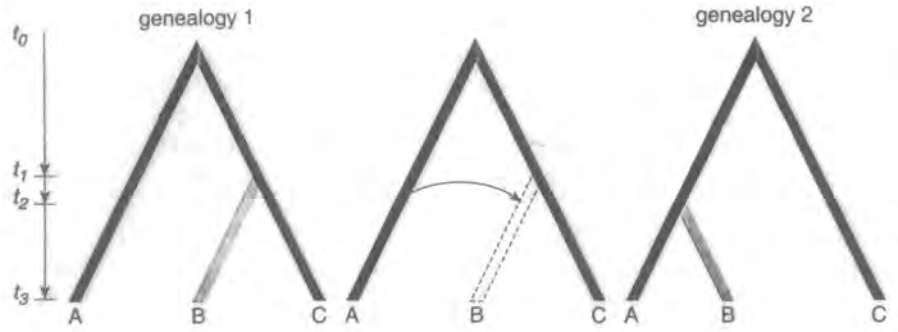


Figure 6. Bacterial lineages reflect the genetic contributions of random mutations and genetic drift—patterns that dominate in clonal species—as well as recombination through lateral transfer and other processes. Evolution is deciphered by statistical reconstruction of gene phylogenies, but the statistics may give confusing signals. This diagram shows how recombination would affect the genealogy seen by a biologist. Genealogy 1 represents a clonal state: Two strains diverged from a common ancestor at time t_0. One survives as strain A; strains B and C diverged from their most recent common ancestor at time t_1. But if genetic material from strain A is introduced into strain B by recombination at time t_2 *(arrow)*, strains A and B appear more closely related than either is to strain C, so that they seem to be branches sharing a recent common ancestor *(genealogy 2)*. (Illustration adapted from Guttman 1997.)

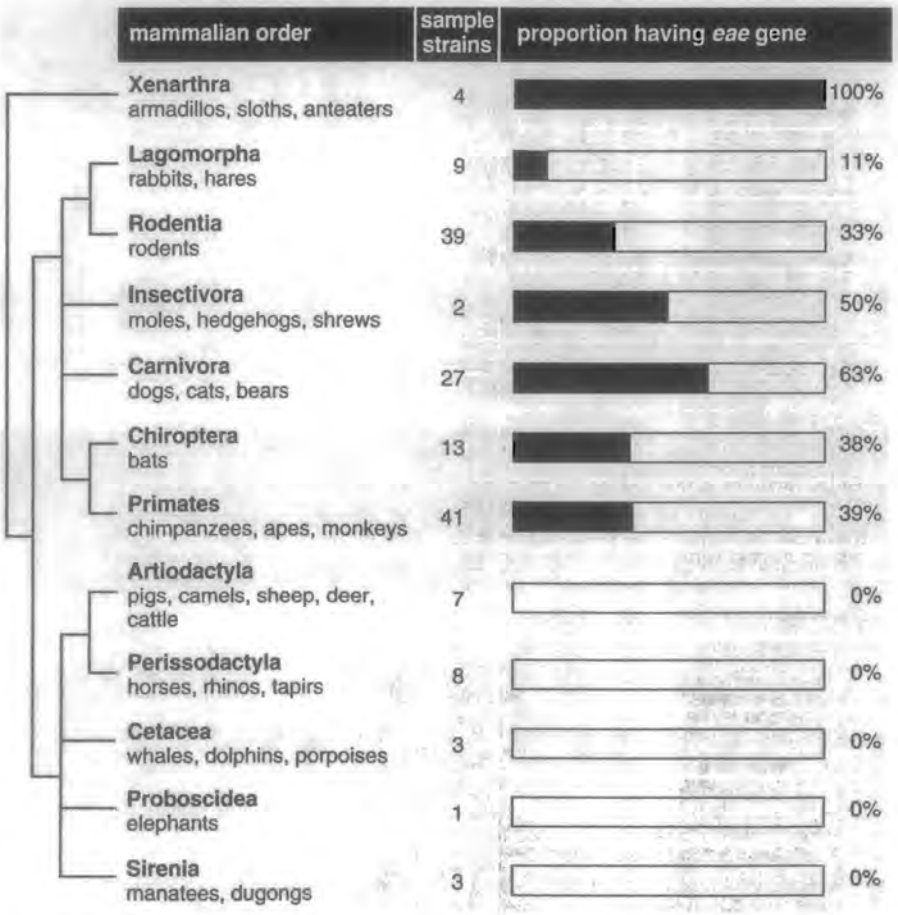


Figure 7. Authors' collection of *E. coli* strains from various hosts and environments shows intriguing patterns in the distribution of the genes associated with pathogenicity. The *eae* gene codes for a protein called intimin, which forges a tight attachment between the pathogen and a host epithelial cell in the intestine. The gene was found in strains associated with the most ancient mammals—anteaters, armadillos and rabbits, for example—and did not turn up in the small number of samples taken from mammalian taxonomic groups more distantly related to human beings. It appears possible that there may be an association between species phylogenies and the evolution of some pathogenic genes.

dominant strain of *E. coli* per host, but the appearance of new genotypes indicates that this dominance is temporary. Traditional evolution is at work here: The population can change through adaptive processes, in which a better strain displaces a less competitive one, or via the random processes known as genetic drift. The primary weapons of intraspecific competition are the colicins, which can destroy strains of the same species that do not display a plasmid coding for the same colicin. Colicins act by disrupting critical cellular functions such as the production of adenosine triphosphate, or ATP, the energy molecule central to cell metabolism.

E. coli is one of the first bacterial species to colonize mammals after birth; it is acquired from the birth canal and from the mother's feces. It has been calculated that the density of *E. coli* in the large intestine of mammals and birds is from 1 million to 10 million cells per gram of colon. This makes *E. coli* a minor component of the microbiota of this part of the intestine, which is primarily anaerobic, and has a total bacterial density calculated at some 100 billion cells per gram of colon. It is believed that in the intestine there is one cell division daily, whereas in the middle of a rich culture medium in the laboratory, *E. coli* K-12 can be seen to double six times a day or more.

Population Genetics

In bacteria, reproduction is not tied to sexuality. Bacteria divide by binary fission to produce clones. Genetic variation comes about primarily by way of mutations passed along to clones. Horizontal transfer, which can be regarded as a parasexual process, serves as an additional source of variability in populations. In a population or species, the balance between these two processes is called the *degree of clonality*.

A highly clonal species is distinguished by a collection of independently evolved lineages. In these cases, it is difficult to speak of a species, since there is no pool of shared genes. It is much more difficult to apply classical population-genetics theory and concepts. Evolution is accomplished by substitutions of complete lineages, whether by selection or by genetic drift. If, however, bacterial species exhibit high levels of recombination, one obtains panmictic populations, and it is possible to apply approximately the same ideas about populations and species that we use

with diploid organisms (organisms with double sets of chromosomes). Until the mid-1990s most evidence led scientists to describe *E. coli* as a clonal species, although its capability for gene transfer was well known.

The fundamental problem is that most bacteria exhibit a great number of possible mechanisms of recombination, but these are not used in every generation (that is, reproduction is uncoupled from sexuality). Also, it is hard to get direct estimates of the degree of sexuality of bacterial populations—the relative importance of processes other than fission. In order to study them one needs to use indirect methods derived from the theory of population genetics.

Classical studies of the genetic structure and clonality of *E. coli* revealed high levels of genetic variation within its populations; values of one common measure of variation reach from 0.47 to 0.52 on a scale of 0 (no variation) to 1 (each individual unique). But the number of multiloci genotypes, or complete genetic patterns, is small. In fact, initially it was estimated that the "linkage disequilibrium," a patterning of gene arrangements different from what random mixing of the gene pool (sexual reproduction) would produce, was around the maximum. This would indicate that few of all the possible genotypes were found, or that there was no recombination between strains. Studies completed from 1980 through 1992 almost without exception concluded that recombination is a rare phenomenon in *E. coli* asssociated with humans or domesticated animals and suggested (using population-genetics theory) that the effective size of the reproducing *E. coli* population was about 10^7 or 10 million genetically distinct organisms, a number relatively small in comparison to the expected total population of the organism, which has been estimated at 10^{20} or 100,000,000,000,000,000,000 cells. The effective population was sufficiently low to assure that random processes of the birth and death of strains were the dominant evolutionary force.

If there is little recombination, how does one explain the high genetic diversity we observe? Periodic selection could be one answer; a genotype might selectively displace others present in the population. In an asexual population, once a favored mutation spreads by natural selection, it replaces not just the gene involved, but a complete genotype. Such a story might make sense as a mechanism of adaptation to particular niches. In other words, *E. coli* would be, according to these ideas, a collection of very different strains, each adapted to a different environment.

A New Evolutionary Paradigm

The studies described above examined mostly strains associated with humans. Indeed, with *E. coli* implicated in a large portion of the more than 2 million annual deaths from diarrheal diseases, the human strains are understandably the focus of attention. However, this tight focus may have limited our understanding of evolutionary issues. For this reason our group initiated a long-term study of the evolutionary ecology of *E. coli* with a particular emphasis on populations in wild hosts. The first step was to organize a collection of strains from various continents (mainly from America, Australia and Antarctica) associated with wild mammals and birds, as well as environmental strains, including samples from the air, water and soil. This collection amounts to more than 5,000 strains. We call it IECOL, the Institute of Ecology Collection of *E. coli*.

In the first study we used 201 strains associated with mammals, in which we completed a population-genetics analysis using 12 *polymorphic* genes—genes whose alleles code for a number of different enzymes. In comparing these strains we studied the use of 12 sugars, resistance to five antibiotics, their serotypes, and the number and size of their plasmids. We found that the diversity is even higher than reported from human strains of *E. coli* (0.68 on the 0-to-1 variation scale). The genetic diversity is greater in Mexican than in

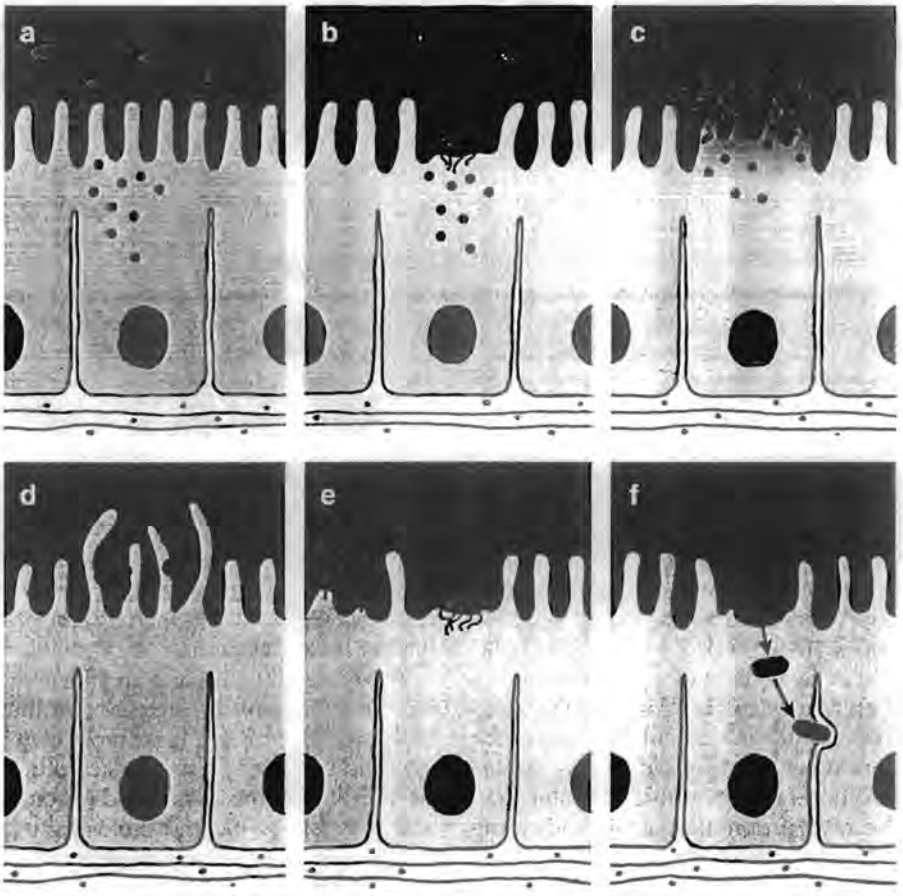


Figure 8. ***E. coli*** **actually has several forms capable of inducing diarrhea in a mammalian host; these share some genes but can vary widely. Enterotoxic** ***E. coli*** **inject toxins into the epithelial cells of the intestine (*a*). Enterohemorrhagic** ***E. coli*****, including O157:H7, cause bloody diarrhea (*b*). They rub off the intestinal microvilli and induce the polymerization of actin fragments in the host cell's cytoskeleton to form an intimate attachment with the host cell, and also inject toxins. Enteroaggregative** ***E. coli*****, a cause of persistent diarrhea among children in developing countries, form a biofilm and secrete cytotoxins into the epithelium (*c*). Diffusely adherent** ***E. coli*** **may cause elongation of the microvilli (*d*). Enteropathogenic** ***E. coli*****, a worldwide cause of infant diarrhea, exhibit actions similar to enterohemorrhagic bacteria, inducing the host cells to form pedestals to which they attach (*e*). Finally, enteroinvasive** ***E. coli*** **can invade epithelial cells and move laterally within the epithelium (*f*).**

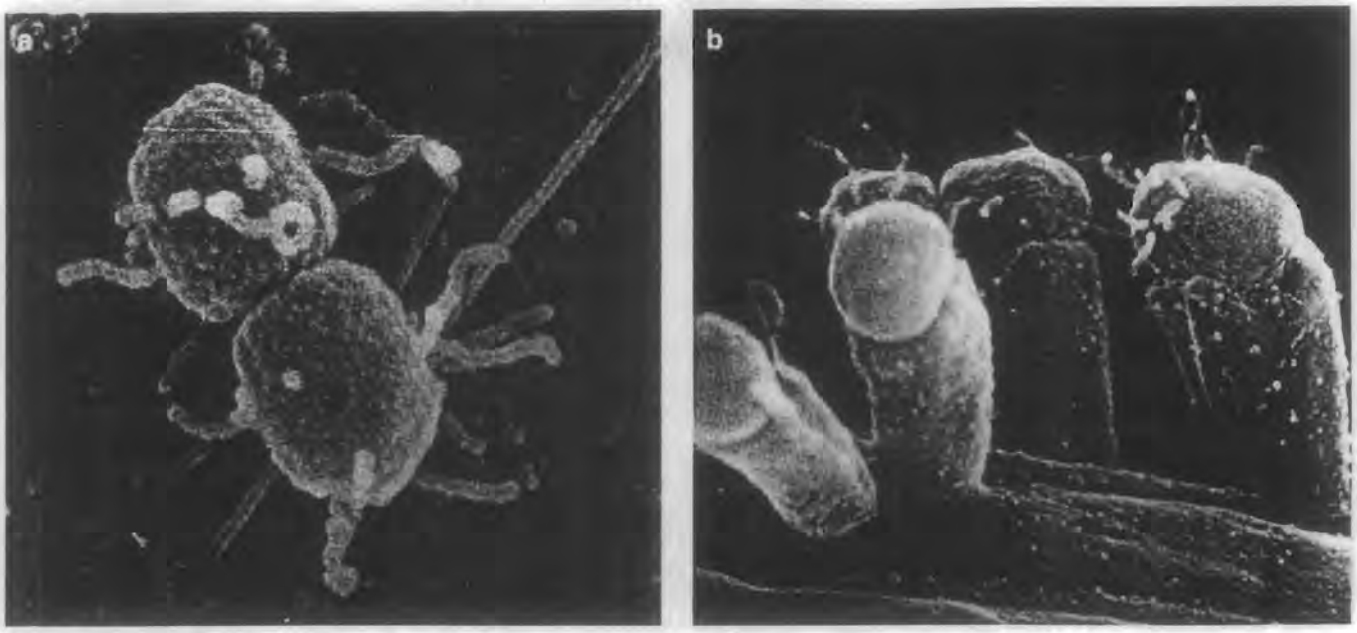


Figure 9. Enteropathogenic *E. coli* extend filaments to attach to epithelial host cells and form a bridge through which toxins and other substances are passed (connecting filaments can be seen at lower left in *a*). The bridge shortly becomes a pipeline funneling chemical instructions to the host cell, causing the host cell to build a pedestal for the invader *(b)* and form an intimate attachment. (Images courtesy of Stuart Knutton, University of Birmingham, U.K.; used by permission of Oxford University Press.)

Australian strains, and each group of host organisms displays a particular group of related strains. The Mexican mammalian strains are those that exhibit statistically the lowest linkage disequilibrium. Such a low level in a bacterial strain signals recombination and lateral transfer and suggests that the *E. coli* associated with mammals in general are not, on average, as clonal as was suggested by data from strains solely derived from humans.

It is interesting that in certain groups of strains, such as those associated with carnivores, rodents and primates, recombination is more frequent than it is in hosts with very specific diets. We have now estimated that the overall effective population size of *E. coli* is 5.3×10^9—that is, two orders of magnitude higher than that calculated for *E. coli* isolated from human hosts. Our estimate of the parameter of recombination is almost two orders of magnitude greater than the rate of mutation, again indicating that lateral transfer happens far more frequently than had been thought. Additionally, we found that the worldwide population of *E. coli* may not be homogeneous, but rather there may be isolated subpopulations. We draw this conclusion from the fact that the estimated global migration rate is less than the estimated migration within Mexico by roughly an order of magnitude. We have estimated the rate of recombination by comparing the congruence between the genealogies of different genes and those derived from analysis of various genetic loci, and we conclude that intragene and intergene recombination in *E. coli* is more important than mutation.

We have also conducted an analysis of the genetic sequences of four metabolic genes using 50 strains in the IECOL collection. These results were consistent with our emerging hypothesis: The genetic diversity from intergene and intragene recombination was greater in strains associated with animals than in those associated with humans. This difference is especially clear in the gene *gapA*, which appears to have suffered a reduction in its diversity after invading humanity.

Thus, our best sampling and the new, more detailed studies suggest that the basic biology of *E. coli* is far more complex and interesting than classical studies indicated. *E. coli* is endowed not only with a great genetic and ecological diversity, but also with a high level of genetic recombination and exchange. These phenomena permit the generation of a large quantity of genotypes, even though they may not occur in each generation. It is not only genes that move; there is considerable recombination of plasmids and fragments of genes. Some of these combinations turn out to be successful and invade many environmental niches and new hosts, and so continues the spread of new variants of *E. coli* able to become highly competitive—for example, O157:H7, first identified as a human pathogen in 1982. This generates a structure populated with ecotypes and at the same time produces strains that can live in a large number of environments that before were believed secondary or atypical for the species.

Mutation and Pathogenesis

Recently, the study of mutation in microorganisms has taken on an interesting character. First there was the controversy over "directed mutation." Patterns of mutation were seen in bacteria that seemed less than random; were certain types of mutations favored over others? This debate, which began in 1989, was put to rest by studies whose results strongly contradicted this notion. But in 1996, a different question of mutation pattern surfaced when J. Eugene Leclerc and his colleagues at the U.S. Food and Drug Administration found that pathogenic bacteria such as O157:H7 are "hypermutable": They exhibit a much higher mutation rate than nonpathogens. According to Leclerc's proposition, errors in the DNA repair system may constitute a lifestyle adaptation that helps the bacteria escape the immune system of their hosts. The challenges of life as a pathogen may select for "mutators," variants that possess the ability to mutate rapidly during invasion, coloniza-

tion and immune warfare. A year later, however, another study found the frequency of mutator strains to be similar between commensal and pathogenic *E. coli* in human beings.

The findings came in the midst of intense interest in the rise of antibiotic resistance. Resistance to antibiotic drugs appears to arise mainly through the sharing of resistance plasmids via conjugation. But high mutation rates might also play a role, and the idea continues to be hotly debated. Some studies of experimental evolution indicate that only under certain conditions can bacteria reach anomalously high mutation rates, especially in changing environments and small populations. Travis Kibota of Clark College and Michael Lynch of the University of Oregon presented in 1996 a model where hypermutability itself is always unstable, since there are more deleterious mutations than there are those that either carry pathogenicity, or are adaptive and provide escape from the host's attack.

Recently Antonio Oliver and his colleagues at INSALUD, Spain's National Institute of Health, studied patients with cystic fibrosis who were subjected at various ages to high doses of antibiotics to combat the bacterium *Pseudomonas aeruginosa*. The *P. aeruginosa* of these patients displayed a significant number of hypermutations associated with gravely ill patients with other diseases. From this they proposed in 2000 that hypermutation is an adaptation that permits sufficient resistance to antibiotics for survival in the altered lung mucosa of a patient with this disease.

In *E. coli* we can compare evolutionary patterns between pathogenic and nonpathogenic strains by looking at pathogenic islands in the genome. Much work is being done to understand the evolution of a pathogenic island called the locus of enterocyte effacement, or LEE, which contains a group of genes that enable the bacterium to create lesions in the intestinal lining. We are currently conducting such comparisons by studying a part of the IECOL collection. There seems not to be any correlation between mutation rate and presence of the LEE or elements of this island, but we have begun to learn about other aspects of how bacteria acquire pathogenic abilities.

Where Pathogenesis Comes From

Two of the most dangerous types of *E. coli* are the EHEC, or enterohemorrhagic, strains such as O157:H7 often implicated in food poisoning, and the EPEC, or enteropathogenic, strains, which are an important cause of infantile diarrhea in the developing world. These strains can adhere to the wall of the intestine and rub off the epithelial villi, creating a lesion that disrupts the function of the intestinal wall and causes severe diarrhea that is often fatal, particularly in children and the elderly. The lesion is called an attachment-and-effacement lesion, abbreviated A/E.

Genetic and genomic studies have identified the genetic underpinnings

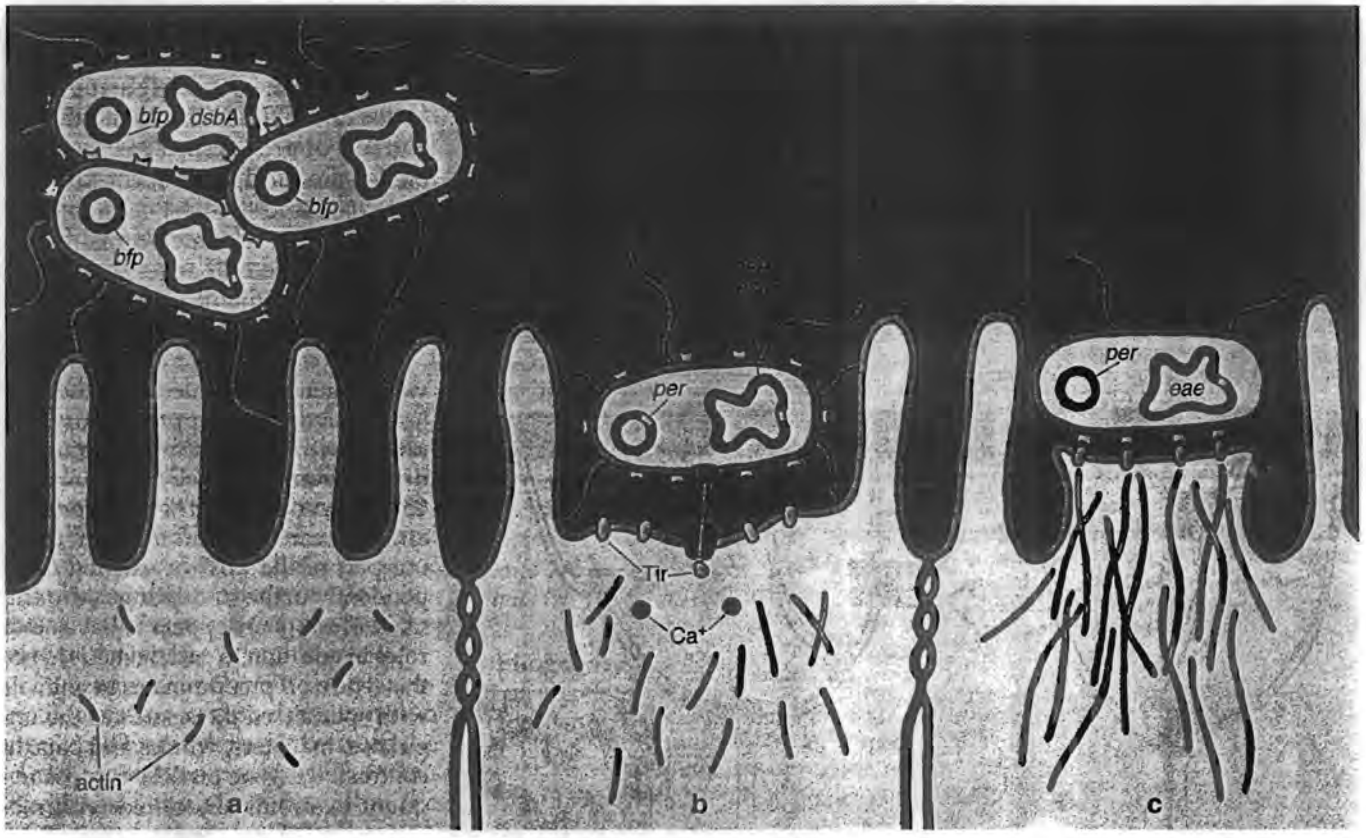


Figure 10. Attachment-effacement lesion formed by enteropathogenic *E. coli* results from the action of genes on the bacterium's chromosome as well as extrachromosomal genes carried on plasmids. The plasmid gene cluster *bfp* (for bundle-forming pilus, an attachment factor) and the chromosomal *dsbA* gene (for disulphide bond) initiate colonization of the mucosal surface of the intestine. The remainder of the work is done primarily by a group of genes on the chromosome (including *sep*, *espB*, *tir* and *eae*) that together form the LEE, or locus of enterocyte effacement. Some of these genes are involved in assembling a type III secretion system, a mechanism many pathogens use to penetrate the host-cell membrane and transmit signal-transduction molecules that cause the movement of calcium ions (Ca^+), the effacement of the epithelium, the polymerization of actin filaments in the cytoskeleton and the translocation of the Tir (translocated intimin receptor) protein to the cell surface (*b*). These activities are regulated by plasmidic genes such as *per*. The chromosomal *eae* gene codes for the protein intimin, which binds with Tir to form an intimate attachment that allows the bacterium to freely absorb nutrients from the host cell.

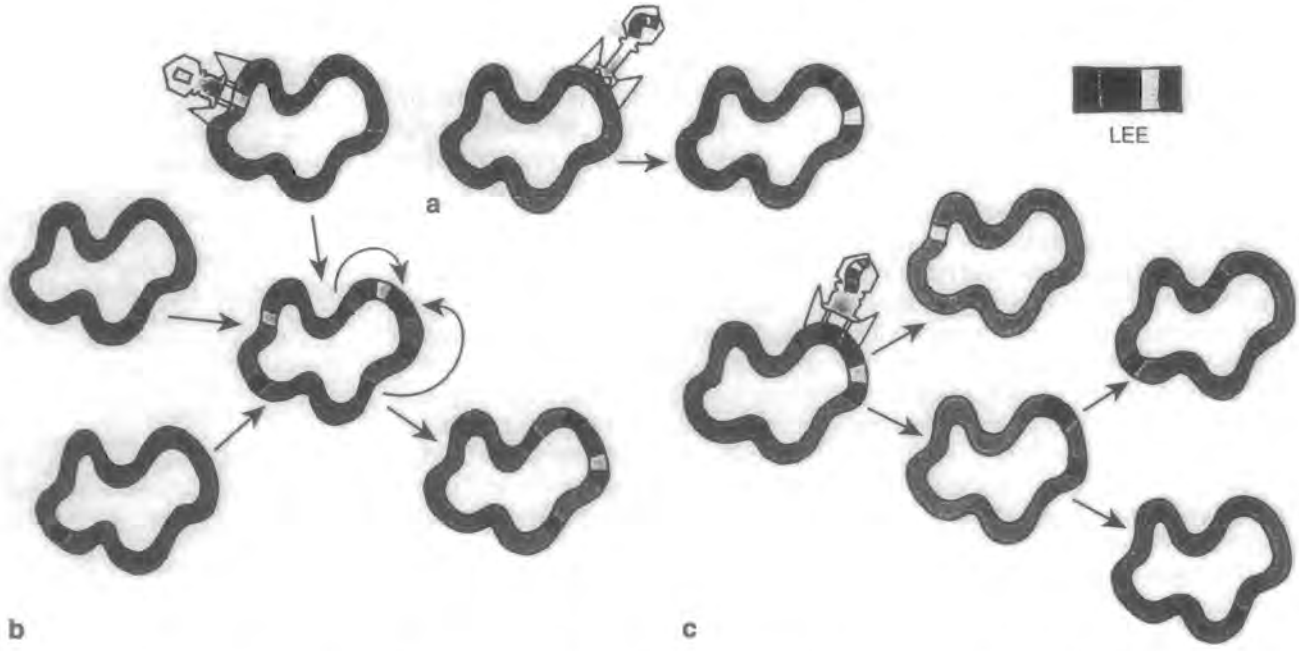


Figure 11. Cassettes of pathogenic genes, or "pathogenic islands," can be assembled in the bacterial genome in various ways. A virus called a bacteriophage can insert individual genes or packages of genes such as the LEE, or locus of enterocyte effacement. (The virus actually attaches to the cell membrane.) A bacterium therefore can acquire the entire LEE from a phage virus by transduction (*a*). But elements of the LEE are found independently in many strains. This suggests that the LEE may be constructed—its elements acquired in pieces and then moved around in the genome in a sequence of events, leaving copies of genes in various arrangements (*b*). Alternatively, an unstable island might break up, disarming the pathogen but again leaving fragments (*c*).

of this process. The chromosomes of these *E. coli* (and of some other enteric pathogens, such as *Shigella* and *Salmonella*) include the LEE. In addition, they have a plasmid called EAF (for "EPEC adherence factor"). Within the LEE we find genes for a type III secretion system, a system used to deliver toxins and other proteins to an infected host cell. The cassette also includes signal-transduction proteins and a gene called *eae*, which codes for intimin, a protein that tightly binds the bacterium to the host cell. The LEE package also includes *tir*, a gene coding for the intimin receptor Tir; genes for some secreted proteins (*espA*, *espD*, *espB*); and the regulator *Ler*. In the EAF plasmid there is a group of genes called *bfp*, responsible for producing a filament to attach the bacteria to the host cells, and the regulator *Per*, which controls the expression of the LEE genes through interaction with the chromosomal regulator *Ler*.

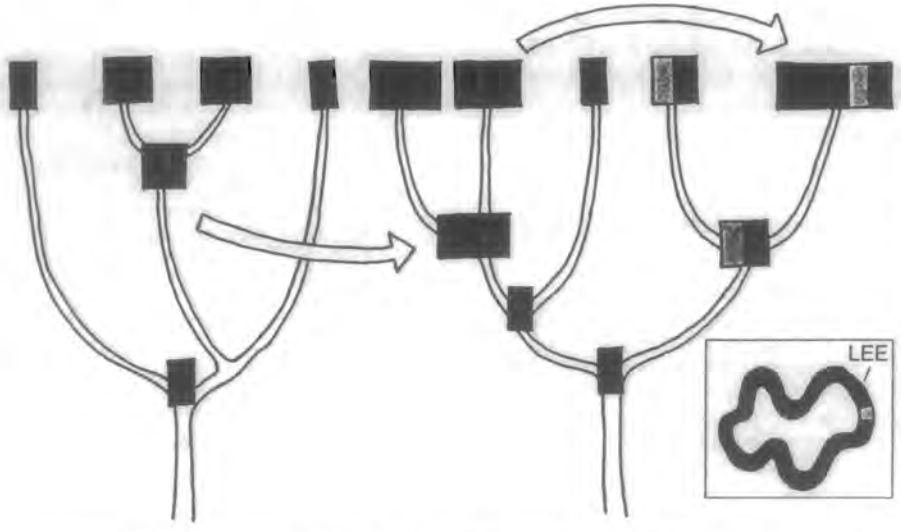


Figure 12. Evolutionary microbiologists attempt to reconstruct evolutionary events. Part of the LEE may have evolved through clonal inheritance and mutation (*left*), but genes may have been transferred to another lineage by recombination, followed by additional recombination within the second lineage (*arrows*). Competing explanations for the evolutionary patterns revealed in *E. coli* genomes include hypermutation—rapid opportunistic mutation—as well as recombination. The LEE, which is found in other diarrhea-causing pathogens in addition to *E. coli*, appears to be a highly dynamic assemblage of diverse and widely dispersed genes.

Within the past year the relations of various genes associated with the A/E lesion have been mapped in our laboratory with a branching diagram, or dendrogram, obtained by examining 155 Mexican strains. The genes *cesT/eae* and *espB* can be found together, forming part of the LEE, or existing independently in the strains of healthy animals. These genes appear to have other roles in addition to pathogenesis, since the gene *espB* predominates in animals with specialized diets, such as the ungulates, manatees, whales and bats. In contrast, the gene *cesT/eae* is most frequent in strains associated with rabbits, armadillos and anteaters. Both independent genes are found distributed throughout the dendrogram.

This suggests that *cesT/eae* and *espB* are ancient genes in *E. coli*, and that they have a role in the normal, nonpathogenic association with their host. In a sequence analysis of 50 strains that do and do not possess the LEE, we

found that the genes *eae* and *tir* are the most variable in their sequences, and that the gene *espB* is less variable. There is an association between the types of sequences in *eae* and *tir*, although this is not clear with *espB*. From the same sample of the collection we also sequenced five genes outside the island (*gapA, fimA, mdh, nuiS* and *putP*). Additional statistical analysis suggests that pathogenic and nonpathogenic genes have different types of selection acting on them. Some genes show diversifying selection (an increase of diversity to escape the host response), whereas others show low levels of diversity owing to purifying selection (low diversity being selected in order to conserve a specific function). More interesting, LEE genes are almost twice as diverse as nonpathogenic genes, and they have very different substitution rates.

In 2000 Sean D. Reid and his colleagues at Pennsylvania State University proposed that pathogenic islands including the LEE are evolutionary units whose genes have integrated at the same time and evolved coordinately. If this is true, the origin of new pathogens could be explained by successive events of horizontal transfer, for example from EPEC to EHEC. But when we compared the genealogies for individual genes, we were surprised to see that each has a very different history to tell. In evolutionary terms, the LEE appears to be not a unit but rather a dynamic assemblage of genes that act together to make a lesion.

On the other hand, the plasmid genes *per* and *bfp* have never been detected in the strains associated with wild animals and are found only in EPEC and EHEC strains associated with human patients. This is evidence that horizontal-transfer events are important in the history of plasmid-carried pathogenesis.

It seems that the origin of new pathogens is a complex phenomenon produced by the dynamic between the chromosomal component (LEE, for example) and the plasmids within the bacterial populations. From these results we conclude that the evolution and emergence of new pathogens is still not well understood. We need to combine more work on population genetics, population ecology and molecular evolution in order to have a wider view and a better idea of the dynamics of *E. coli* natural populations; this is the only way to understand the evolutionary biology of this organism.

Perspectives

In spite of the fact that *E. coli* is the best-known bacterium in the world, we are just starting to understand its ecology and evolutionary biology. It is clear that *E. coli* is a very diverse bacterium and that its genome is highly dynamic. It is not the strictly clonal organism suggested in the first population-genetics studies. It is a bacterium with a generous and complex sexuality, which has played a role in, among other things, its success as a pathogen. Successful combinations can be dispersed in epidemic fashion in human or animal populations, giving a false signal of clonality.

The diverse tools of molecular genetics and population genetics offer the possibility of completing adequate ecological and evolutionary studies of bacterial populations. These studies should be based on a solid knowledge of natural populations, their ecology and their biology. In our studies with the IECOL collection we have tried to make advances in this direction in *E. coli*. We are pleased to place our collection at the disposal of persons interested in working with it, and we look forward to the continuing evolution of our understanding of *E. coli*.

Acknowledgments

The authors thank Aldo Valera, Laura Espinosa and Andy Peek for technical assistance; Becky Gaut for lots of ideas; Martha Rocha, Luisa Sandner and Claudia Silva for interesting data and endless discussions; and Alejandro Cravioto for showing them the fascinating world of the enteropathogens. This project was supported by CONACyT grants 27557-M, and 0028 and DGAPA grants IN-218698 and IN-208601.

Bibliography

Elliott, S. J., L. A. Wainwright, T. K. McDaniel, K. G. Jarvis, Y. K. Deng, L.-C. Lai, B. P. McNamara, M. S. Donnenberg and J. B. Kaper. 1999. The complete sequence of the locus of enterocyte effacement (LEE) from enteropathogenic *Escherichia coli* E2348/69. *Molecular Microbiology* 28:1–4.

Guttman, D. S. 1997. Recombination and clonality in natural populations of *Escherichia coli*. *Trends in Ecology & Evolution* 12(127):15–22.

Levin, B. R., and C. T. Bergstrom. 2000. Bacteria are different: Observations, interpretations, speculations and opinions about the mechanisms of adaptive evolution in bacteria. *Proceedings of the National Academy of Sciences of the U.S.A.* 97:6981–6985.

Maynard-Smith, J. 1991. The population genetics of bacteria. *Proceedings of the Royal Society of London B* 245:37–41.

Nataro, J. P., and J. B. Kaper. 1998. Diarrheagenic *Escherichia coli*. *Clinical Microbiology Reviews* 11:142–201.

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

Perna, N. T., G. Plunkett, III, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor, H. A. Kirkpatrick, G. Posfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E. J. Grotbeck, N. W. Davis, A. Lim, E. T. Dimalanta, K. D. Potamousis, J. Apodaca, T. S. Anantharaman, J. Lin, G. Yen, D. C. Schwartz, R. A. Welch and F. R. Blattner. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 409:529–533.

Reeves, P. R. 1992. Variation in O-antigens, niche-specific selection and bacterial populations. *FEMS Microbiology Letters* 100:509–516.

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

Sandner, L., L. E. Eguiarte, A. Navarro, A. Cravioto and V. Souza. 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.

Souza, V., and L. E. Eguiarte. 1997. Bacteria gone native vs. bacteria gone awry?: Plasmidic transfer and bacterial evolution. *Proceedings of the National Academy of Sciences of the U.S.A.* 94:5501–5503.

Souza, V., P. E. Turner and R. E. Lenski. 1997. Long-term experimental evolution in *Escherichia coli*. V. Effects of recombination with immigrant genotypes on the rate of bacterial evolution. *Journal of Evolutionary Biology* 10:743–769.

Souza, V., A. Castillo, M. Rocha, L. Sandner, C. Silva and L. E. Eguiarte. 2001. Ecología evolutiva de *Escherichia coli*. *Interciencia* 26: 513–517.

Whittam, T. S. 1996. Genetics variation and evolutionary processes in natural populations of *Escherichia coli*. In Escherichia coli *and* Salmonella *Cellular and Molecular Biology*, ed. F. C. Neidhart. Washington DC: ASM Press, pp. 2708–2720.

Links to Internet resources for further exploration of "The Evolutionary Ecology of *Escherichia coli*" are available on the *American Scientist* Web site:

http://www.americanscientist.org/articles/02articles/souza.html

Capítulo 3. La selección natural a nivel molecular.

" DNA: that registry of chance, that tone-deaf conservatory where the noise is preserved with the music"
-Jaques Monod

3.1 Introducción

Las dos ideas principales derivadas de la teoría Darwin-Wallace sobre la evolución de las especies señalan que los organismos son producto de una historia evolutiva, es decir, de ancestría-descendencia a partir de modificaciones de un ancestro en común. Además, que el mecanismo principal de la evolución es la selección natural de estas variaciones hereditarias (Futuyma, 1986). A partir de estas ideas se generaron dos campos de estudio dentro de la biología. Por un lado el estudio de la historia evolutiva de los organismos, y por el otro, la elucidación de las fuerzas evolutivas que moldean la biodiversidad y las adaptaciones que producen (Futuyma, 1986). El papel de la selección natural es ampliamente aceptado dentro de la comunidad científica, sobre todo en cuanto a caracteres morfológicos se refiere, es decir, a las variaciones fenotípicas, de las cuales se tienen abundantes ejemplos (Lack, 1947; Wiesenfeld, 1967; Bishop y Cook, 1975; Jope, 1976; Grant, 1986; Sato y col., 1999; Zhang y col., 2002). Sin embargo, la importancia de la selección natural a nivel molecular es materia de un continuo debate.

La selección se refiere a la reproducción diferencial de algunos fenotipos/genotipos sobre otros bajo ciertas condiciones ambientales que prevalecen en un momento determinado (Futuyma, 1986; Li, 1997). En otras palabras, es un mecanismo evolutivo que genera un cambio en las frecuencias relativas de los fenotipos/genotipos, de acuerdo a su adaptación relativa dentro de la población. Durante mucho tiempo se creyó que la selección era la principal fuerza generadora de la variación fenotípica y genotípica. Sin embargo, a partir del desarrollo de nuevas técnicas para el estudio de la variación a nivel DNA y proteínas se inicio de manera formal el estudio de la evolución a nivel molecular (Zuckerkandl y Pauling, 1965). Los resultados de toda una serie de trabajos de genética de poblaciones con diferentes organismos arrojaron información acerca de la naturaleza de la variación a nivel molecular y trajeron como resultado la evidencia de que existía una gran variación genética contenida en las proteínas y genes, misma que no podía ser justificada

del todo por efectos de la selección natural. Estas ideas culminan con la propuesta de la teoría Neutral de la evolución (Kimura, 1968 y 1983).

De acuerdo a la teoria neutral de la evolución molecular, la mayor parte de la variación al nivel molecular no es producto de la selección tipo Darwiniano (selección adaptativa o positiva), sino producto de la deriva génica y de mutaciones neutras a casi neutras (Kimura, 1983; Ohta, 1992). Es decir, la mayor parte de la variación a nivel molecular ocurría de manera aleatoria y no tenía importancia adaptativa. En su conjunto, los postulados de la teoría neutral parecían regir a la evolución molecular.

Con el advenimiento y desarrollo de la era de la genómica y la existencia de grandes bases de datos moleculares, surge una gran curiosidad en la comunidad científica de reevaluar la participación de la selección natural a nivel molecular, lo que ha generado el desarrollo de diferentes metodologías para su detección. Existe un gran debate acerca de cual método es más adecuado para cuantificar la selección y cómo se ven afectados por fenómenos demográficos que hacen que resulte difícil su detección inequívoca, así como la distinción del tipo de selección de que se trata. A lo largo del presente capítulo discutiremos los más utilizados desde las pruebas clásicas hasta los métodos basados en máxima verosimilitud y estadística bayesiana, señalando sus cualidades y desventajas.

3.2 Tipos de selección

Básicamente, la selección natural a nivel molecular puede ser considerada de dos tipos: selección positiva de tipo Darwiniano (que incluye a la selección balanceadora, direccional, diversificadora o adaptativa) y selección negativa (purificadora) (Li, 1997).

La selección positiva es el mecanismo evolutivo mediante el cual nuevos mutantes poseen adecuaciones mayores que el promedio de la población, y las frecuencias de dichos mutantes se incrementan en la siguiente generación (Li, 1997). Por otra parte, en la selección negativa los nuevos mutantes poseen adecuaciones menores que el promedio de la población, y la frecuencia de estos mutantes disminuye en las siguientes generaciones (Li, 1997). La selección positiva promueve la diversidad genética y la selección negativa disminuye o purga esta diversidad eliminando a las variantes de la población.

El estudio de la adaptación a nivel molecular se ha abordado sólo de manera reciente como un intento de evaluar cuál es el papel de la selección a nivel molecular, ya que inicialmente la mayor parte de los estudios parecían indicar que la selección de tipo

Darwiniano tendría un papel secundario (Suzuki y Gojobori, 1999). Pero ¿qué tal si los métodos desarrollados no eran lo suficientemente sensibles para detectar la selección de manera inequívoca y directa? ¿Cómo podemos aprovechar toda la información contenida en las bases de datos moleculares para inferir el papel que la selección positiva juega dentro de la evolución molecular?

3.3 "The neutral expectation"

De manera histórica la selección positiva es generalmente inferida de manera indirecta cuando un grupo de genes parecen desviarse del modelo neutral, es decir, al no encajar en las predicciones neutrales se supone que estas desviaciones son producto de la selección que las está modificando. Es entonces que para poder detectar la participación de la selección en una región del genoma o en un gen en particular, lo primero que se tiene que probar es el rechazo de la hipótesis nula de que la selección no ha actuado en las secuencias en cuestión. Los postulados de la teoría de Kimura (Kimura 1968 y 1983) nos permiten establecer qué se debe esperar si existen desviaciones del modelo neutro. Nos describen los patrones de evolución de las secuencias bajo las fuerzas de la mutación y la deriva génica sin la participación de la selección. Un resultado particularmente importante derivado del modelo neutro es que la tasa de cambio a la cual un nucleótido es remplazado (substituido) por otro dentro de una población es igual a la tasa de mutación o cambio de ese sitio sin importar el tamaño poblacional. El número de mutaciones se encuentra en balance con la baja probabilidad de que cada mutación se fije en una población grande. Además, la teoría neutra hace predicciones acerca de los patrones de polimorfismo esperados dentro de las secuencias de las mismas especies. Por ejemplo, la probabilidad de que dos muestras tomadas al azar difieran en una posición particular del genoma se determina por la siguiente relación: $\theta = 4Ne\mu$ (θ = diversidad genética Ne = tamaño efectivo poblacional μ = tasa de mutación) o para el caso de un organismo haploide $\theta = 2Ne\mu$. Esta ecuación tiene sentido pues dos individuos poseen mayor probabilidad de ser diferentes si la población es grande y/o la tasa de mutación es alta. Este tipo de predicciones hacen que la teoría neutra sea tan importante y central a la biología evolutiva moderna.

3.4 Adaptación a nivel molecular

El estudio de la adaptación molecular se ha realizado a partir de dos aproximaciones estadísticas diferentes; la primera, es la distribución del polimorfismo en las secuencias de DNA. La segunda, en la determinación de las substituciones sinónimas y no sinónimas en secuencias codificantes de DNA, utilizando el codón como unidad evolutiva mínima. Ambas aproximaciones hacen uso de las predicciones del modelo neutral para determinar la participación de la selección positiva.

3.4.1 Métodos basados en la distribución del polimorfismo

Los primeros trabajos desarrollados para el estudio de la selección a nivel molecular fueron los basados en la distribución del polimorfismo en la secuencia de DNA. Estos trabajos clásicos son la prueba de Tajima (Tajima, 1983 y 1989), el método de Fu-Li (Fu y Li, 1993) que es una modificación a la prueba de Tajima, la prueba de Hudson-Kreitman-Aguadè (HKA) (Hudson y col., 1987) y el método de McDonald-Kreitman (McDonald y Kreitman, 1991).

Para entender con mayor facilidad los fundamentos de las pruebas basadas en la distribución del polimorfismo en el DNA es necesario definir algunos parámetros básicos en genética de poblaciones que además son fundamentales en las predicciones de la teoría neutra. Uno de estos parámetros es el de la diversidad genética que puede ser descrita por dos estimadores: π (pi) y θ (teta). El primero (π) se refiere a la diversidad nucleotídica la cual es el número de nucleótidos diferentes por sitio entre dos secuencias tomadas al azar (Nei y Li, 1979). El segundo (θ), es calculado a partir de la expresión $\theta = 4Ne\mu$ (Kimura, 1968). Sin embargo, es difícil tener la determinación exacta de los parámetros Ne (tamaño efectivo poblacional) y μ (tasa de mutación), de esta forma una manera indirecta de calcular θ es utilizando el número total de sitios segregativos en un grupo de secuencias (un sitio segregativo es un sitio en donde las secuencias difieren), $\theta = K/a$, donde K es igual al número total de sitios segregativos en una muestra de secuencias dada y $a = 1 + 1/2 + \ldots + 1/n-1$ (n = número de secuencias en la muestra) (Watterson, 1975 y Tajima, 1983). Se deduce de estas expresiones que π se ve afectada mayormente por los alelos que poseen mayor frecuencia y es independiente del tamaño de la muestra, mientras que θ sí se ve afectada por el tamaño de la muestra y por la deriva génica, es decir, por los alelos poco

frecuentes (Tajima, 1983 y 1989). La relación que existe entre ambos estimadores nos permite determinar si nuestras secuencias se encuentran bajo el modelo neutral o se desvían del mismo. Si ambos estimadores nos dan el mismo resultado en cuanto a variación genética quiere decir que el polimorfismo observado es neutro y se encuentra distribuido aleatoriamente (Tajima, 1983 y 1989). En cambio, si existen diferencias ente ambos, indica que la selección está afectando alguno de ellos, promoviendo su incremento o decremento. Es decir, si existe algún tipo de selección positiva este incrementará las frecuencias alélicas y eso se reflejará en el incremento de π. Por otra parte, si existe un mayor número de alelos deletéreos en la muestra, θ se verá incrementada (Watterson, 1975; Tajima, 1983 y 1989). De esta forma, si determinamos estadísticamente las diferencias entre ambas podremos detectar de manera indirecta la participación de la selección positiva o negativa en el mantenimiento de cierto tipo de polimorfismo dentro las poblaciones.

3.4.1.1 Prueba de Tajima

La prueba de Tajima (Tajima, 1983) está basada precisamente en la detección de las diferencias entre los dos estimadores de la diversidad genética π y θ. La prueba se basa en la determinación de la D de Tajima calculada a partir de la siguiente expresión:

$$D = \frac{\pi - K/a}{\sqrt{V(\pi - K/a)}}$$

V - varianza

Si la D resulta negativa quiere decir que θ posee un valor mayor que π, lo que indica la presencia de mutaciones deletéreas. En cambio, si D resulta positiva quiere decir que π tiene un mayor valor que θ, indicación de que algunos alelos se encuentran bajo algún tipo de selección positiva (por ejemplo, balanceadora) incrementando sus frecuencias. Si D es igual a cero quiere decir que no existe diferencia alguna entre ambos estimadores y nos encontramos bajo equilibrio neutral. La relación ente ambos tipos de índices utilizada como base en esta prueba posee un razonamiento muy interesante ya que cada uno se ve afectado diferencialmente por procesos demográficos al interior de la poblaciones. El número de sitios segregativos, que es parte de la estimación de θ, se puede ver afectado seriamente por

la existencia de mutantes deletéreos dentro de la muestra, pues son tomados en cuenta no por su frecuencia sino por la aportación de sitios segregativos a la estimación de la diversidad. En cambio, para la estimación de π este hecho no la afectaría ya que se refiere al promedio de la muestra, y en este caso sí toma en cuenta la frecuencia de los mutantes, por lo cual no altera de manera significativa la diversidad genética. Por otra parte, un fenómeno demográfico que altera significativamente la prueba es la existencia de un cuello de botella reciente dentro de la población (Tajima, 1989). Es decir, un valor negativo de D se puede obtener debido a selección negativa o debido a que la población pasó por un cuello de botella reciente (Wright y Gaut, 2004). En este caso, la comparación entre regiones codificantes y no codificantes ayudaría a identificarlo ya que un cuello de botella debe de afectar a todos los tipos de polimorfismo presentes (Tajima, 1989). Otro caso que se tendría que examinar cuidadosamente es cuando se tienen sitios que se encuentran ligados a sitios que sufrieron un evento de selección, por lo que la D del sitio neutral se verá afectada por el sitio que se encuentra bajo selección ("hitchhiking"). Los supuestos bajo los cuales trabaja esta prueba son: a) se considera una población de individuos N tomados al azar, b) no existe recombinación ni selección, c) utiliza el modelo de sitios infinitios, d) el tamaño poblacional es constante (Tajima, 1989; Kimura, 1969).

3.4.1.2 Método Fu-Li

El método de Fu-Li (Fu y Li, 1993) fue desarrollado con la misma lógica que el de la prueba de Tajima, la diferencia reside en que los autores consideran la distribución de las mutaciones en una genealogía de muestras tomadas al azar a partir de una población dada. Suponen que las mutaciones antiguas tendrán que encontrarse con mayor probabilidad en las ramas más antiguas de la genealogía, mientras que las mutaciones originadas más recientemente en las ramas más nuevas. La parte más antigua de la genealogía consiste de las ramas internas y la parte más reciente de las ramas externas. Si existe selección de tipo negativa o purificadora se observará un exceso de las mutaciones en las ramas más externas. Por otra parte, si existe algún tipo de selección positiva se verá una disminución de las mutaciones en las ramas externas. Es entonces, que al comparar el número de mutaciones en ramas internas y externas con los valores esperados de acuerdo al modelo

neutro se tendrá una prueba estadística para la detección de la participación de la selección. La prueba consiste en evaluar la siguiente expresión:

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

Donde,

ne = número de mutaciones en ramas externas
ni = número de mutaciones en ramas internas

Ambos métodos parecen ser suficientemente efectivos para determinar si existe la participación de la selección, sin embargo, un estudio demostró que la prueba de Tajima es mucho más efectiva (Simonsen y col., 1995). Sin embargo, ambos asumen que las poblaciones se encuentran en equilibrio mutación-deriva génica por suficiente tiempo y toman como base el modelo de sitios infinitos, lo cual no siempre es válido sobre todo al estudiar secuencias codificantes de DNA donde las tres posiciones de los aminoácidos poseen distintas probabilidades de cambio. Además, ambas pruebas se ven afectadas por procesos demográficos que muchas veces pueden llevar a confusiones acerca del proceso evolutivo.

3.4.1.3 Hudson-Kreitman-Aguadè (HKA)

La prueba de Hudson-Kreitman-Aguadè (HKA) (Hudson y col., 1987) es uno de los primeros intentos por determinar la participación de la selección a partir de las desviaciones del modelo neutral. Esta prueba describe si existen diferencias entre el polimorfismo intrapoblacional y el polimorfismo entre diferentes poblaciones. De acuerdo a los postulados de la teoría Neutra, si todo el polimorfismo existente es neutro, no deberían de existir diferencias entre la variación genética dentro de los individuos de una especie de una población específica y entre poblaciones distintas de la especie, ya que todo el polimorfismo es originado al azar. En este caso se realiza una prueba donde se registra la correlación entre la diversidad genética intrapoblacional y entre poblaciones. El

razonamiento de la prueba es el siguiente, si se considera un loci específico (m) entre las especies 1 y 2 y se toma al azar una secuencia n_1 y n_2 de la especie 1 y 2, de las cuales se tienen un número determinado de sitios segregativos o polimórficos K_{1i} del locus *i* en la especie 1 y K_{2i} para la especie 2, donde el número de diferencias en el locus *i* entre las dos secuencias es D_i. Los parámetros K_{1i} y K_{2i} ($i = 1,\ldots, m$) son las medidas del polimorfismo intraespecífico y la D_i es la medida de la variación entre las especies (Hudson y col., 1987). Por otra parte, se realiza la determinación de los valores esperados de estos parámetros bajo el modelo neutro y el modelo HKA prueba si existen diferencias estadísticamente significativas entre los parámetros observados y los esperados mediante la siguiente expresión:

$$X^2 = \sum_{i=1}^{m}[K_{1i} - E(K_{1i})]^2 \Big/ V(K_{1i}) + \sum_{i=1}^{m}[K_{2i} - E(K_{2i})]^2 \Big/ V(K_{2i}) + \sum_{i=1}^{m}[D_i - E(D_i)]^2 \Big/ V(D_i)$$

La determinación de los parámetros esperados y de las varianzas de estos parámetros se obtiene usando las siguientes expresiones a partir del modelo neutral:

$$E(K_{1i}) = a_1\theta_1$$
$$V(K_{1i}) = a_1\theta_1 + (b_1\theta_1)^2$$

Donde,

$a_1 = 1 + 1/2 + \ldots + 1/n-1$

$b_1 = 1 + 1/2^2 + \ldots + 1/(n-1)^2$

Si la variación genética observada es muy diferente a la esperada por la teoría neutra, entonces se sugiere que la selección está participando en el mantenimiento del polimorfismo en las poblaciones. Los postulados del modelo HKA son los siguientes: 1) las generaciones son discretas; 2) todas las mutaciones son neutras; 3) el número de sitios en la secuencia es suficientemente grande para que cada mutación ocurra en un sitio que no ha sido mutado con anterioridad (modelo de sitios infinitos); 4) las mutaciones ocurren de manera independiente en cada locus; 5) el número de mutaciones por locus en cada

generación tiene una distribución de Poisson; 6) no existe recombinación entre los loci; 7) los loci no se encuentran ligados genéticamente; 8) las especies 1 y 2 poseen los tamaños poblacionales N y N*f*, respectivamente; 9) ambas especies son derivadas de un ancestro común en el tiempo T con una tamaño poblacional de $2N(1+f)/2$ (promedio entre las poblaciones de las especies A y B) (Hudson y col., 1987). Estos postulados hacen que esta prueba no sea tan efectiva cuando se tienen muestras poblacionales pequeñas, cuando las secuencias son muy divergentes entre sí y existe alta recombinación, y cuando se trabajan con muestras grandes de loci diferentes que se encuentran ligados genéticamente.

3.4.1.4 McDonald-Kreitman

Esta prueba está basada en principios muy sencillos, en donde se considera las relaciones entre las diferencias sinónimas y no sinónimas entre y dentro de las poblaciones. La idea es parecida a la descrita por la prueba HKA, pero en este caso se prueba si el cociente de las sustituciones no sinónimas a sinónimas entre las especies es el mismo que dentro de las poblaciones. Por ejemplo, si consideramos una región codificante de DNA y tomamos dos secuencias m1 y m2 a partir de dos especies cercanamente relacionadas, 1 y 2. Sólo tomamos en cuenta los sitios que son variables entre las secuencias y descartamos los invariables. Dentro de éstos sitios variables agrupamos dos categorías diferentes, los sitios que son variables en un nucleótido (por ejemplo, G) en todos los miembros de la especie 1, y que difieren en otro nucleótido (por ejemplo, A) en todos los miembros de la especie 2, a estos sitios los denominamos sitios fijos. Son sitios fijos, ya que son iguales para cada población en cada especie. Todos los demás sitios variables son considerados polimórficos. Estos sitios polimórficos serán diferentes dentro de las poblaciones de cada especie o entre las poblaciones de cada especie. Ambos tipos de sitios son clasificados en los que poseen sustituciones sinónimas y no sinónimas. La hipótesis nula en la prueba de McDonald y Kreitman es la siguiente:

$$E(n_f)/Es_f = E(n_p)/Es_p$$

Donde,

n_f = sitios fijos con sustituciones no sinónimas

s_f= sitios fijos con sustituciones sinónimas

n_p = sitios polimórficos con sustituciones no sinónimas

s_p= sitios polimórficos con sustituciones sinónimas

Para ilustrar mejor las diferencias entre los parámetros estimados se puede construir una tabla de contingencia donde se comparen ambos tipos de sitios con sus respectivas sustituciones como lo realizado por McDonald y Kreitman (1991), al realizar las comparaciones entre tres especies de la mosca de la fruta *Drosophila* (Tabla 3.1).

Tabla 3.1. Tabla de contingencia (2 X 2) para los sitios fijos y polimórficos y sus respectivos tipos de sustitución, resultado de la comparación del gen de la *Adh* (alcohol deshidrogenasa) entre *D. melanogaster* (12 secuencias), *D. simulans* (6) y *D. yakuba* (24).

Sustituciones	Sitios fijos	Sitios polimórficos	Total
Sinónimas	s_f (17)	s_p (42)	$s_f + s_p$ (59)
No Sinónimas	n_f (7)	n_p (2)	$n_f + n_p$ (9)
Sumatoria	$s_f + n_f$ (24)	$s_p + n_p$ (44)	68

A estos resultados se les aplicó una prueba de Fisher, que mostró que existían diferencias estadísticamente significativas entre los cocientes de los sitios fijos y los sitios polimórficos (n_f/s_f = 7/17 y n_p/s_p = 2/42), siendo mayor el cociente de los sitios fijos que el de los polimórficos, lo que sugiere que la selección positiva promovía un incremento en las sustituciones no sinónimas entre especies. Una debilidad de este método es que no contempla la existencia de sustituciones múltiples en un mismo sitio, algunos estudios posteriores sugieren que si esto se toma en cuenta no se puede rechazar la neutralidad para el gen de *Adh* en las tres especies de la mosca de la fruta (Whittam y Nei, 1991). Además, ya que la teoría neutra incluye la participación de la selección negativa y mutación, la prueba de McDonald y Kreitman no necesariamente descarta la neutralidad. Por lo mismo es necesario ser cuidadoso en la interpretación de los resultados obtenidos utilizando este método.

3.4.2 Métodos basados en sustituciones moleculares

El codón está formado por tripletes de bases que son traducidas durante la síntesis de proteínas, es decir, cada codón corresponde a un tipo de aminoácido, siendo estos últimos los pilares que conforman las proteínas. Dentro de cada codón se encuentran tres posiciones diferentes denominadas primera, segunda y tercera posición. Cada una de estas posiciones se ven afectadas por distintas tasas de mutación, por lo tanto, ni la mutación ni la selección es uniforme en cada uno de los codones. Por ejemplo, de acuerdo al código genético, si un cambio es introducido en la primera posición en su mayoría será un cambio sinónimo y un porcentaje bajo corresponderá a un cambio no sinónimo, es decir se mantendrá codificando para el mismo tipo de aminoácido. Si un cambio se introduce en la tercera posición en su mayoría corresponderá a un cambio sinónimo. En cambio, si un cambio se introduce en la segunda posición, éste repercutirá directamente en el tipo de aminoácido para el cual está codificando, originando un cambio. Por lo tanto, es la única posición que nos brinda la posibilidad de evaluar un cambio que potencialmente podría ser adaptativo y que será sujeto directamente a la selección natural.

3.4.3 Métodos para la estimación de substituciones sinónimas y no-sinónimas

Existen numerosos métodos para la estimación de las tasas de sustitución molecular y pueden ser clasificados en tres grupos: a) métodos evolutivos, b) métodos basados en el modelo de Kimura-2, c) métodos de máxima verosimilitud basados en un modelo de sustitución de codones.

3.4.3.1 Métodos Evolutivos

3.4.3.1.1 Detectando selección positiva con el método de Nei-Gojobori (1986)

Dentro de los métodos evolutivos, el más utilizado es el método de Nei-Gojobori (Nei y Gojobori, 1986) y su modificación (Ina, 1995). En el método de Nei-Gojobori original las sustituciones sinónimas y no sinónimas se determinan calculando el número de sustituciones sinónimas y no sinónimas y el número de sitios con sustituciones potencialmente sinónimas y no sinónimas. En el caso de los sitios potencialmente sinónimos y no sinónimos, estos son descritos a partir de cada codón asumiendo que existe

igual probabilidad de cambio para todos los nucleótidos. El número de sitios potencialmente sinónimos (s) y no sinónimos (n) para un codón en específico es calculado con la siguiente relación (Nei y Gojobori, 1986):

$$s = \sum_{i=1}^{3} fi$$

$$n = 3 - s$$

donde f es igual a la proporción de cambios sinónimos (se refiere a el cociente del número de cambios sinónimos y no sinónimos excluyendo mutaciones sin sentido)

Por ejemplo, para el caso del codón de la fenilalanina:

T T T

$$s = 0 + 0 + \frac{1}{3}$$

porque todos los cambios en la primera y segunda posición resultan en un cambio no sinónimo y en la tercera posición 1 de cada 3 posibles cambios resulta en una sustitución sinónima (TTC) (tomado de Nei y Kumar, 2000).

Como todos los cambios restantes originan una sustitución no sinónima,

$$n = 3 - \frac{1}{3} = \frac{8}{3}$$

En el caso de que cualquier cambio origine un codón de TÉRMINO, este no será tomado en cuenta para el cálculo.

Para obtener el número total de sitios sinónimos (S) y no sinónimos (N) en una secuencia entera usamos la fórmula:

$$S = \sum_{j=1}^{c} sj$$

$$N = 3C - S$$

donde sj= valor de s para el codon j

C= número total de codones

$S + N = 3C$ (igual al número total de nucleótidos comparados)

Aplicando esta fórmula en un ejemplo en concreto, tenemos la comparación de dos codones (TTG- Leu y AGA- Arg) (tomado de Nei y Kumar, 2000). Para la transición de Leu a Arg tenemos 6 escenarios evolutivos posibles:

1) TTG (Leu) - ATG (Met) - AGG (Arg)- AGA (Arg)
2) TTG (Leu) - ATG (Met) - ATA (Ile) - AGA (Arg)
3) TTG (Leu) - TGG (Trp) - AGG (Arg) - AGA (Arg)
4) TTG (Leu) - TGG (Trp) - TGA (Alto) - AGA (Arg)
5) TTG (Leu) - TTA (Leu) - ATA (Ile) - AGA (Arg)
6) TTG (Leu) - TTA (Leu) - TGA (Alto) - AGA (Arg)

Como los escenarios 4 y 6 involucran un codón de término, no son tomados en cuenta para el estudio. El número de sustituciones sinónimas en los escenarios 1, 2, 3 y 5 son 1, 0, y 1, respectivamente, mientras que para el caso de las sustituciones no sinónimas es 2, 3, 2, y 2, respectivamente. Ya que asumimos que los 4 escenarios son igualmente probables, tenemos que *sd*= 3/4 y *nd*= 9/4. Podemos obtener las sustituciones sinónimas y no sinónimas en la comparación de dos secuencias sumando los valores de todos los codones. Esto es,

$$Sd = \sum_{j=1}^{C} sdj$$

$$Nd = \sum_{j=1}^{C} ndj$$

donde *sdj* y *ndj* corresponden a los números de las diferencias sinónimas y no sinónimas para el codón *j* y *C* es el número de codones comparados.

Por lo tanto,

Sd + *Nd* es igual al número total de diferencias entre dos secuencias comparadas.

A partir de esto podemos estimar la proporción de cambios sinónimos (*Ps*) y no sinónimos (*Pn*) como:

$$PS = \frac{Sd}{S}$$

$$PN = \frac{Nd}{N}$$

donde S y N se refieren al número promedio de sitios sinónimos y no sinónimos para dos secuencias comparadas.

A partir del modelo de Jukes-Cantor (1969) para calcular el número de diferencias de nucleótidos (d) entre dos secuencias, tenemos la siguiente relación:

$$d = -\left(\frac{3}{4}\right)\ln\left(1-\left(\frac{4}{3}\right)p\right)$$

donde,

$p = 1 - q$, es la proporción de nucleótidos diferentes entre dos secuencias,

q, es el número de nucleótidos idénticos entre dos secuencias.

Podemos usar esta misma ecuación para determinar el número de sustituciones sinónimas y no sinónimas por sitio reemplazando p por PS o PN. Esta ecuación nos da buenos resultados, siempre y cuando las frecuencias de las bases A, T, G y C sean casi iguales y no exista una desviación en el cociente de transciciones y transversiones, es decir, que ambos cambios sucedan con la misma probabilidad (Ota y Nei, 1994). Para el cálculo de la varianza de dS y dN podemos usar la misma fórmula de la varianza basada en el método de Jukes-Cantor (Ota y Nei, 1994):

$$V(dS) = \frac{V(PS)}{\left(1-\frac{4}{3}PS\right)^2}$$

$$V(dN) = \frac{V(PN)}{\left(1-\frac{4}{3}PN\right)^2}$$

donde,

$$V(PS) = \sum_{i=1}^{c} (sdi - PSsi)^2 / S^2$$

$$V(PN) = \sum_{i=1}^{c} (ndi - PNni)^2 / N^2$$

Otra forma de calcular sus varianzas es el método de bootstrap, siempre y cuando las muestras de *Sd*, *Sn*, *S* y *N* sean suficientemente grandes (Nei y Kumar, 2000).

Como ya vimos con anterioridad, para poder detectar selección positiva a nivel molecular es necesario mostrar que *dN* es significativamente más grande que *dS*. Un método muy simple es ver las diferencias entre las tasa de sustitución sinónima y no sinónima $D = dN - Ds$ y su varianza *V(D)* (Nei y Kumar, 2000). Siempre y cuando nuestra muestra sea grande, la distribución aproximada de *D* es la normal y, en este caso en particular, la $V(D) = V(dN) + V(dS)$, por lo tanto tendríamos la siguiente relación a usar en una prueba estadística de *Z*:

$$Z = D / s(D)$$

donde,

$$s(D) = V(D)^{1/2}$$

Como estamos interesados en saber si $dN > dS$, se analiza como una prueba de una cola con un infinito número de grados de libertad. La varianza de *D* también puede ser calculada a partir de una prueba de bootstrap, en donde se toma como unidad de muestreo sucesivo los codones en lugar de los nucleótidos de la secuencia y se procede a realizar múltiples repeticiones de los parámetros *Psb*, *Pnb*, *dsb*, *dnb* de *PS*, *PN*, *dS*, *dN* y se obtienen las varianzas de estos valores. También podemos calcular el error estandar de *D* mediante el bootstrap y luego usar una prueba de *Z* (si la muestra es grande) (Efron y Tibshirani, 1993). Una forma más adecuada en el caso de que la secuencia de nucleótidos sea pequeña es calcular directamente *dS* y *dN* y sus varianzas y usar la hipótesis nula de que ambas son iguales $dN = dS$ con una prueba de *Z*. Otra forma es construir una tabla de contingencia para las sustituciones sinónimas y no sinónimas calculadas directamente de la

secuencia y realizar una prueba de Fisher (Zhang y col. 1997). Si el valor de P es menor a 0.05, entonces la hipótesis nula de evolución neutral es rechazada.

3.4.3.1.2 Método modificado de Nei-Gojobori

Al realizar el cálculo del número de sitios sinónimos y no sinónimos el método de Nei-Gojobori asume que la sustitución de los 4 nucleótidos sucede de forma azarosa, es decir con la misma probabilidad para cada uno (Nei y Gojobori, 1986). En la práctica esto no es necesariamente cierto, ya que pueden existir un mayor número de sustituciones de un tipo de nucleótido u otro. El cociente de transiciones (cambio de purina por purina o pirimidina por pirimidina) es usualmente mayor que el de transversiones (cambio de purina por pirimidina y viceversa). En este caso se espera que el número de sitios potenciales que producen sustituciones sinónimas sea mayor que el determinado por el método de Nei-Gojobori, ya que las transiciones en la tercera posición son en su gran mayoría sinónimas y serán más frecuentes. Para corregir este detalle se propuso modificarlo utilizando como base el modelo de dos parámetros de Kimura (Kimura-2) (Kimura, 1980; Ina, 1995), en este modelo la proporción de transiciones del total de cambios es dada por la siguiente relación:

$$\frac{\alpha}{\alpha + 2\beta} = \frac{R}{1 + R}$$

α = transiciones, cada nucleótido puede tener un cambio transicional, por ejemplo A (purina) a G (purina).

β = transversiones, 2β tomando en cuenta que cada nucleótido puede tener dos cambios transversionales, por ejemplo, A (purina) puede cambiar a T (pirimidina) ó C (pirimidina).

R = cociente de transición/transversión, que es igual a 0.5 cuando las transiciones y las tranversiones se producen con la misma probabilidad.

Teóricamente la modificación al método de Nei-Gojobori es mejor al ser un modelo que se apega más a la realidad de la naturaleza del proceso de sustituciones en una secuencia de DNA. Sin embargo, en la práctica la determinación del cociente de transiciones/transversiones (R) en los sitios sinónimos o no sinónimos es complicada y usualmente cae en sobreestimaciones de cambios sinónimos. Por otra parte, cuando se

sobreestima este cociente se puede concluir que las sustituciones no sinónimas son significativamente mayores a las sinónimas cuando no es el caso. Hay que señalar que la mayor parte de las veces los patrones de sustitución son más complicados que los descritos por el modelo de Kimura-2, y bajo ciertas condiciones tanto el método de Nei-Gojobori como su modificación pueden sobreestimar las sustituciones sinónimas sobre las no sinónimas, haciendo difícil la detección correcta de selección positiva. Lo recomendable es usar ambos métodos, si ambas versiones detectan la presencia de selección positiva se tiene una mayor seguridad acerca de las conclusiones derivadas de este análisis y un fuerte indicio de selección positiva.

3.4.3.2 Métodos basados en el modelo de dos parámetros de Kimura

Además de la modificación al modelo de Nei-Gojobori, existen otros métodos que intentan detectar la selección basados en el modelo de Kimura-2 (Kimura, 1980). Entre los más importantes encontramos el método de Li-Wu-Luo (Li y col., 1985), Pamilo-Bianchi-Li (Pamilo y Bianchi, 1993 y Li, 1993) y el de Comeron y Kumar (Nei y Kumar, 2000).

3.4.3.2.1 Li-Wu-Luo

Li y sus colaboradores (1985) desarrollaron otro método considerando la degeneración del código genético. Los sitios en los codones pueden ser clasificados en 4-veces degenerados, 2-veces degenerados, y 0- veces degenerados. Un sitio es reconocido como 4- veces degenerado si todos los cambios posibles en ese sitio son sinónimos, 2-veces degenerado si uno de tres cambios o dos de tres cambios es sinónimo y 0-veces degenerado si todos los cambios son no sinónimos o se refieren a mutaciones sin sentido que generan codones de término. Por ejemplo, la mayoría de los cambios en la tercera posición son 4-veces degenerados, mientras que todos los cambios en la segunda posición son 0- veces degenerados. Bajo esta regla podemos calcular los tres tipos de cambios L0, L2, y L4 (0-veces, 2-veces y 4-veces degenerados, respectivamente) para dos secuencias que quieran ser comparadas. Se comparan las dos secuencias codón por codón y se clasifica cada diferente nucleótido según sea una transición ó transversión. Entonces se calcula la proporción de transiciones (Pi) y transversiones (Qi) en cada clase (i) de nucleótido (i= 0, 2 ó 4) tomando en cuenta la probabilidad de verosimilitud de ocurrencia de cada aminoácido (se denomina

en inglés "likelihood occurrence", es un parámetro de probabilidad de que se encuentre presente un determinado aminoácido). Ellos estiman la proporción de cambios transicionales y de transversiones con la siguiente relación:

$$A_i = \frac{1}{2}\ln(a_i) - \frac{1}{4}\ln(b_i)$$

$$B_i = \frac{1}{2}\ln(b_i)$$

Donde,

Pi = número de transiciones tomando en cuenta la clase de sitio (0, 2 ó 4)

Qi = número de transversiones tomando en cuenta la clase de sitio (0, 2 ó 4)

$a_i = 1/(1-2p_i - Q_i)$

$b_i = 1/(1-2Q_i)$

Sabemos que la mayoría de las sustituciones en los sitios 4-veces degenerados son sinónimas y todas las sustituciones en los sitios 0-veces degenerados son no sinónimas. Sin embargo, para los sitios 2-veces degenerados las transiciones (A2) son en su mayoría sinónimos, mientras que las transversiones (B2) son en su mayoría no sinónimas. Si suponemos que la probabilidad de sustitución de cada uno de los 4 nucleótidos (A, G, C y T) se lleva a cabo con la misma frecuencia, los autores de este método sugirieron que 1/3 de los cambios en los sitios 2-veces degenerados son potencialmente sinónimos y 2/3 son potencialmente no sinónimos. Bajo estas ideas, propusieron que *dN* y *dS* pueden ser estimados con la siguiente relación:

$$dS = \frac{3\left[L_2 A_2 + L_4(A_4 + B_4)\right]}{L_2 + 3L_4}$$

$$dN = \frac{3\left[L_0(A_0 + B_0) + L_2 B_2\right]}{3L_0 + 2L_2}$$

A su vez, estas fórmulas se basan en ciertos supuestos, el primero y más evidente es que el tipo de sitio en una determinada secuencia puede no corresponder al mismo sitio o sitio homólogo en la secuencia contra la cual se está comparando. Es decir, en una secuencia puede tratarse de un sitio 4-veces degenerado, mientras que en su equivalente este sitio puede ser un sitio 2-veces degenerado. Esto último sucede con frecuencia, sobre todo si trabajamos con secuencias muy divergentes entre sí. En este caso en particular la mitad de los sitios son considerados 2-veces degenerados y la otra mitad 4-veces degenerados. El segundo supuesto es que los cambios sin sentido son considerados cambios no sinónimos (lo que es totalmente incorrecto). Se sabe que este tipo de cambios suceden con una probabilidad de 4%. Por lo tanto se espera que el método de Li-Wu-Luo sobreestime las sustituciones no sinónimas (*dN*). Aunado a esto tenemos casos específicos de ciertos aminoácidos, donde surgen problemas en las determinaciones. Tal es el caso de la leucina donde en la tercera posición de sus tres codones (2-veces degenerados), algunas transversiones son sinónimas. Otro caso es el de la arginina, en donde las transiciones en las primeras posiciones de 4 de sus codones (2-veces degenerados) son no sinónimas con excepción de un codón en donde el cambio genera una mutación sin sentido.

A pesar de estos detalles, éste método proporciona estimados parecidos al de Nei-Gojobori con precisión, siempre y cuando se trate del análisis de muchos codones (>100) y secuencias poco divergentes entre sí.

3.4.3.2.2 Pamilo-Bianchi-Li

Este método es en realidad una extensión del anterior. Los autores notaron que las transiciones que resultan en cambios sinónimos suceden sólo en los sitios 2-veces y 4-veces degenerados, como las transversiones en los sitios 4-veces degenerados son también sinónimas, el número total de sustituciones sinónimas (por sitios sinónimos) se estima de la siguiente manera:

$$dS = (L_2 A_2 + L_4 A_4)/(L_2 + L_4) + B_4$$

Por lo tanto las sustituciones no sinónimas,

$$dN = A_0 + (L_0 B_0 + L_2 B_2)/(L_0 + L_2)$$

3.4.3.2.3 Comeron y Kumar

Ninguno de los métodos anteriores toma en cuenta casos específicos como la arginina, aunque en primera instancia esto no parece tener importancia pues se trata de sólo aminoácido, para algunos genes de los mamíferos (como la protamina) es un problema real, ya que sus proteínas son ricas en arginina precisamente. Utilizando las mismas fórmulas del método de Pamilo-Bianchi-Li (Pamilo y Bianchi, 1993 y Li, 1993), Cameron (1995) y Kumar y col. (2000) dividieron los sitios 2-veces degenerados en sitios 2-veces degenerados simples y complejos. Los sitios simples se refieren a sitios en que las transiciones dan lugar a un cambio sinónimo y dos transversiones generan sustituciones no sinónimas o sin sentido. Todos los casos restantes se incluyen en los sitios 2-veces degenerados complejos.

3.4.3.3 Métodos de verosimilitud con modelos de sustitución de codones

Como alternativa a los métodos descritos anteriormente, surgen los modelos basados en la aproximación de máxima verosimilitud. En estos modelos se pretende incluir algunos detalles que los demás no incluían, como es el caso de que una clase de sitio (0-veces degenerado, por ejemplo) haya cambiado en el tiempo, además de la inclusión de múltiples tasas de sustitución para un sitio determinado (Muse y Gaut, 1994 y Goldman y Yang, 1994).

Dos modelos surgieron al mismo tiempo utilizando esta aproximación. Uno de ellos es el trabajo llevado a cabo por Muse y Gaut (1994), utilizando como modelo el genoma del cloroplasto, y el segundo fue desarrollado por Goldman y Yang (1994). Ambas propuestas se basan en un modelo específico de sustitución de codones y la determinación de la mayoría de los parámetros del modelo por el método de Máxima Verosimilitud descrito por Felsenstein (1981). En estos trabajos se supone que cada sitio en la secuencia evoluciona de manera independiente de sus sitios vecinos y de los sitios homólogos en las otras secuencias. De esta forma el parámetro de máxima verosimilitud de cada sitio (algo así como la probabilidad de observar dicho nucleótido en un lugar específico de la secuencia) es el producto de la máxima verosimilitud de cada uno de los sitios.

Comenzaremos por detallar el funcionamiento de estos modelos analizando los trabajos desarrollados por Goldman y Yang (1994), Nielsen y Yang (1998) y Yang y

Nielsen (2000), entre algunos, ya que son los más desarrollados y sobre los cuales se basan en la actualidad la mayoría de los trabajos de estudio de adaptación molecular con el método de máxima verosimilitud. Los autores definen el parámetro ω como el cociente de las tasas de substituciones no sinónimas y sinónimas (*dN/dS*), que al igual que en los demás modelos mide la relación entre los dos tipos de tasas de substituciones. Si un cambio en un aminoácido es neutral se fijará con la misma probabilidad que una mutación sinónima y ω = *1*. Si el cambio en el aminoácido es deletéreo la selección negativa o purificadora reducirá su probabilidad de fijación y ω < *1*. Sólo cuando el cambio de un aminoácido tiene una ventaja selectiva se fija con una mayor probabilidad que una mutación sinónima y ω > *1*. Por lo tanto, un valor de ω significativamente mayor a uno es una evidencia de selección positiva. Desarrollan el modelo de sustitución de codones con el cual determinan la tasa de sustitución del codon *i* al *j (i≠j)* y que está dado por las siguientes condiciones:

$$qij \begin{cases} 0, \text{ si } i \text{ y } j \text{ difieren en una posición} \\ \pi j, \text{ para transversiones sinónimas} \\ \kappa\pi j, \text{ para transiciones sinónimas} \\ \omega\pi j, \text{ para transversiones no sinónimas} \\ \omega\kappa\pi j, \text{ para transiciones no sinónimas} \end{cases}$$

πj – frecuencia del codon *j* en equilibrio (tomado de la frecuencia del codón en la muestra en estudio)

κ - cociente de transiciones y transversiones determinado a partir de la muestra

ω - *dN/dS*

Si se quiere considerar la sustitución a través de un tiempo *t* la probabilidad de transición está dada por:

$$P(t) = \{p_{ij}(t)\} = e^{Qt}$$

Donde $p_{ij}(t)$ es la probabilidad de que el codón *i* se convierta en el *j* en un tiempo *t*, y *Q* es igual a p_{ij}.

Después de que se realizan las determinaciones de los parámetros correspondientes se procede a realizar unas pruebas estadísticas que nos permitirán determinar si la tasa de sustitución no sinónima es significativamente mayor que la sinónima. En el caso de los métodos basados en máxima verosimilitud, se utiliza una prueba de máxima verosimilitud con modelos alternativos donde se fijan varios valores a ω y se comparan con los obtenidos en la muestra (Goldman y Yang, 1994, Nielsen y Yang, 1998, Yang y Nielsen, 2000 y Yang y Bielawski, 2000). La función de verosimilitud obtenida para cada modelo se compara con una distribución de χ^2 con un grado de libertad y se prueba la hipótesis nula de que ω es igual o mayor a uno (Goldman y Yang, 1994, Nielsen y Yang, 1998, Yang y Nielsen, 2000 y Yang y Bielawski, 2000). Si éste es el caso, entonces se puede decir con una alta probabilidad que existe selección positiva actuando en el gen en estudio.

Como podemos darnos cuenta, hemos trabajado definiendo el cociente de *dN/dS* para toda la secuencia del gen y lo que obtenemos siempre es un valor promedio de todos los sitios de ésta secuencia. En este caso, sólo detectamos la selección si ésta es en promedio mayor a uno, lo cual no sucede en la mayoría de los casos (Goldman y Yang, 1994, Nielsen y Yang, 1998, Yang y Nielsen, 2000 y Yang y Bielawski, 2000). Sin embargo, ésto no quiere decir que no existan sitios dentro de la secuencia que se encuentren seleccionados positivamente. Si lo que pretendemos es detectar cuáles aminoácidos dentro de la secuencia se encuentran bajo selección positiva con el modelo de máxima verosimilitud, entonces la prueba de selección involucra dos pasos. El primero es probar si existen sitios dentro de la secuencia donde ω sea mayor que 1, lo que se lleva a cabo con la determinación del cociente de verosimilitud comparando un modelo en donde se tengan sitios que posean una $\omega > 1$, con otro modelo donde no existan esos sitios; el segundo paso implica la aplicación del teorema de Bayes para identificar los sitios donde realmente existe selección positiva (Goldman y Yang, 1994, Nielsen y Yang, 1998, Yang y Nielsen, 2000 y Yang y Bielawski, 2000).

Para poder tener una mayor comprensión del método, haremos un paréntesis para explicar con un modelo cómo funcionan estas pruebas (tomado de Yang y Bielawski, 2000). Supongamos que tenemos una población hipotética dividida en dos grupos I y II que corresponden al 60% y 40% de la población total, respectivamente. En el grupo I ocurre un tipo de enfermedad en una proporción de 1% y en el grupo II de 0.01%. Supongamos ahora

que hacemos un muestreo al azar de 100 individuos de la población total (incluyendo grupo I y II), ¿cuál es la probabilidad de que 3 de ellos estén enfermos? Esta probabilidad (D) es un promedio entre los dos grupos G1 y G2, y está definida por:

$$p = P(D) = P(G1) \times P(D \mid G1) + P(G2) \times P(D \mid G2) = 0.6 \times 0.01 + 0.4 \times 0.001 = 0.0064$$

De igual manera la probabilidad de que el individuo no esté enfermo es:

$P(D_2) = P(G1) \times P(D_2 \mid G1) + P(G2) \times P(D_2 \mid G2) = 0.6 \times 0.99 + 0.4 \times 0.999 = 0.9936$

lo que es lo mismo que $1 - p$ (de la primera ecuación).

La probabilidad de que 3 de cada 100 individuos posean la enfermedad está dada por la probabilidad:

$$P = \frac{100! p^3 (1-p)^{97}}{3! \times 97!} = 0.0227$$

Si ésta última ecuación involucra un parámetro desconocido tal como la probabilidad de aparición de la enfermedad en el grupo I, por ejemplo, éste parámetro puede ser estimado maximizando esta fórmula. En este caso, la ecuación nos da la probabilidad de observar el dato de la muestra y se le denomina la función de verosimilitud.

La segunda cuestión involucra realizar el razonamiento de forma inversa, es decir, calcular la probabilidad de que un individuo de la muestra tomada al azar que se encuentre enfermo pertenezca al grupo I. Es aquí que usamos el teorema de Bayes:

$$P(G1 \mid D) = P(G1) \times P(D \mid G1)/P(D) = 0.6 \times 0.01/0.0064 = 0.94$$

Por lo tanto, la probabilidad de que un individuo enfermo pertenezca al grupo I es mayor, de la misma forma la probabilidad de que un individuo sano pertenezca a este grupo es mayor.

Para el caso del estudio de la selección positiva, la *D* del ejemplo corresponde a un sitio de la secuencia y *Gi* es la clase de sitio *i* a la que pertenece con un determinado valor de *Wi*. La probabilidad de observar un dato determinado en un sitio es un promedio de las clases de sitios. Entonces, el producto de las probabilidades de clases de esos sitios constituye la verosimilitud a partir de las cual se estiman los parámetros desconocidos, tal

como la distribución de ω sobre los sitios y la longitud de las ramas de los árboles filogenéticos o genealogías. Después de que éstos parámetros son estimados, se utiliza el teorema de Bayes para el cálculo de la probabilidad posterior de que cualquier sitio pertenece a una clase determinada de valores de ω, y dentro de ellos se encuentre el modelo de selección positiva cuyos valores serán mayores a uno.

Una de las ventajas de los modelos basados en máxima verosimilitud es que no necesitan de las reconstrucciones filogenéticas, además de que pueden incluir ciertos parámetros característicos del proceso evolutivo del DNA, tales como el cociente de transiciones/transversiones y el uso de codones.

3.4.3.4 Métodos basados en reconstrucciones filogenéticas

Dentro de los métodos que intentan determinar la selección a nivel intragénico llaman la atención los trabajos que incluyen una reconstrucción de los posibles ancestros basándose en una filogenia o genealogía determinada. Uno de los trabajos más sistemáticos es realizado por Suzuki y Gojobori (1999). Los autores calculan el número de sitios sinónimos y no sinónimos a lo largo de un árbol filogenético construido a partir del modelo de distancia del vecino más cercano (Neighbor joining), y luego prueban si la proporción de sustituciones no sinónimas difieren del modelo neutral ($\omega = 1$). Las tasas de sustitución las calculan a partir del modelo de Nei-Gojobori y para cada sitio en el codón se infiere su secuencia ancestral a partir de cada nodo del árbol. Otro trabajo basado en reconstrucciones genealógicas es el realizado por Fitch y colaboradores (1997), donde básicamente usan el modelo de máxima parsimonia para la reconstrucción de una genealogía y a partir de ella calculan los cambios en cada sitio de cada codón a lo largo de las ramas de las genealogías obtenidas asumiendo que las tasas de sustitución son constantes en todos los sitios de los codones (lo que en general no es cierto). Estos métodos necesitan de muestras grandes para ser eficientes y no tener varianzas muy grandes, recordando que a partir de las reconstrucciones filogenéticas ya existen errores dependiendo del modelo evolutivo usado, así es que hay que tener cuidado en su implementación.

En conclusión, la mayoría de los métodos presentados aquí son vigentes en la literatura de investigación de los procesos evolutivos moleculares. Todos son efectivos dependiendo del tipo de datos y las características de la muestra a usar, sobre todo si

trabajamos con secuencias muy conservadas y poco divergentes entre sí a través del tiempo. Sin embargo, lo que sí es claro es que cuando trabajamos con secuencias poco conservadas y divergentes entre sí, los métodos de máxima verosimilitud y las reconstrucciones basadas en modelos de distancia resultan ser más efectivas que los demás (Zhang y Nei, 1997). Los métodos de verosimilitud y con uso de un modelo de codones no se ven afectados por los procesos demográficos y son los únicos que determinan de manera directa la participación de la selección (Nielsen, 2001). Por otra parte, se ha visto que la probabilidad de reconstrucción correcta de las secuencias ancestrales, utilizando como modelo la secuencia de la lisozima C, es de 98.7% para el método de máxima verosimilitud contra 91.3% para los métodos de máxima parsimonia (Yang y col., 1995).

A continuación en la tabla 2 se comparan las tasas de sustitución obtenidas por varios métodos descritos en este trabajo determinados a partir del estudio de la subunidad α_2 de las globinas entre humanos y orangutanes (142 codones) (tomado de Yang y Bielawski, 2000) y el gen mitocondrial para la NADH deshidrogenasa de humanos y chimpancés (603 codones) (Modificado de Nei y Kumar, 2000).

Tabla 3.2. Tasas de sustitución de nucleótidos de la subunidad α_2 de las globinas y de la NADH deshidrogenasa determinadas por los distintos métodos descritos en el presente capítulo.

Método	**dN**	**dS**	**dN/dS**	**Referencia**
globinas α_2				
Nei-Gojobori	0.0095	0.0569	0.168	(Nei y Gojobori, 1986)
Li	0.0104	0.0517	0.201	(Li, 1993)
Ina	0.0101	0.0523	0.193	(Ina,1995)
Yang y Nielsen	0.0083	0.1065	0.078	(Yang y Nielsen, 2000)
NADH deshidrogensasa				
Nei-Gojobori	0.0379	0.4151	---	(Nei y Kumar, 2000)
Li-Wu-Luo	0.0378	0.4277	---	(Nei y Kumar, 2000)
Modificación a Nei-Gojobori	0.0438	0.2730	---	(Nei y Kumar, 2000)

Ina	0.0438	0.3031	---	(Nei y Kumar, 2000)
Pamilo-Bianchi-Li	0.0438	0.3018	---	(Nei y Kumar, 2000)
Comeron-Kumar	0.0438	0.3018	---	(Nei y Kumar, 2000)
Goldman-Yang	0.0442	0.2872	---	(Nei y Kumar, 2000)

Como podemos observar no existe una diferencia importante entre los distintos modelos para estos genes.

3.5 Genes donde se ha detectado selección positiva

Hemos intentado dar un panorama de los diferentes métodos utilizados en la actualidad para el estudio de la selección positiva a nivel molecular. No existe un consenso dentro de la comunidad científica acerca de cuál es el mejor método a usar para la determinación de la participación de la selección. En general existe una amplia gama de trabajos y autores que usan indistintamente todos los métodos descritos en el presente trabajo con resultados muy parecidos. Lo más importante es conocer la muestra de datos que vamos a utilizar y todas las peculiaridades que exhibe, tal como, el cociente de transversiones/transiciones y el uso de codones. También es importante tener una reconstrucción filogenética adecuada. De esta forma aseguraremos que el modelo que escojamos para determinar las tasa de sustitución y calcular la relación de dS/dN es el que más se ajusta a nuestros datos. Otra sugerencia es la comparación de resultados utilizando varios métodos y hacer acopio de toda la información de evolución molecular y biológica a la que tengamos acceso lo que siempre garantizará la obtención de mejores resultados.

En la tabla 3.3 se resume la mayoría de los ejemplos registrados en los que se ha detectado la participación de la selección positiva mediante la determinación del cociente dN/dS y los autores respectivos (modificado de Yang y Bielawski, 2000).

Tabla 3.3. Genes de diversos organismos involucrados en actividades celulares de varias clases que se encuentran bajo selección positiva determinado por diferentes métodos moleculares donde el parámetro dN/dS > 1.

GEN	ORGANISMO	REFERENCIA	MÉTODO
Genes involucrados con sistema inmune			
Quitinasas clase I	*Arabis* y *Arabidopsis*	Bishop y col., 2000.	Li-Wu-Luo
Colicinas	*Escherichia coli*	Riley MA, 1993.	Nei-Gojobori
Defensinas	Roedores	Hughes y Yeager, 1997.	Nei-Gojobori
Inmunoglobulinas V_H	Mamíferos	Tanaka y Nei, 1989.	Nei-Gojobori
Genes del Complejo Mayor de Histocompatibilidad	Mamíferos	Hughes y Nei, 1988.	Nei-Gojobori
Inhibidor de la poligalacturonasa	Dicotiledóneas y legumbres	Stotz y col., 2000.	---
Genes de grupo sanguíneo RH (RH50)	Primates y roedores	Kitano y col., 1998.	Nei-Gojobori
Ribonucleasas	Primates	Zhang y col., 1998.	Nei-Gojobori
Gen de la transferrina	Salmones	Ford y col., 1999.	Nei-Gojobori
Interferon-ω tipo I	Mamíferos	Hughes, 1995.	Nei-Gojobori
Inhibidor de la proteinasa α-1	Roedores	Goodwin y col., 1996.	Nei-Gojobori
Genes involucrados en la evasión de la respuesta a sistema inmune			
Gen de la cápside	Virus de la fiebre aftosa	Haydon y col., 2001.	ML*
Genes CSP, TRAP, MSA-2 y PF83	*Plasmodium falciparum*	Hughes y Hughes, 1995.	Nei-Gojobori

Región codificadora del antígeno delta	Virus de la hepatitis D	Wu y col., 1999.	Comeron
Gen de la envoltura de la cápside	Virus HIV	Yamaguchi-Kabata y Gojobori, 2000.	Ina
Glicoproteína gH	Virus de la pseudorabia	Endo y col., 1996.	Nei-Gojobori
Genes del antígeno de la invasión del plásmido	*Shigella*	Endo y col., 1996.	Nei-Gojobori
Antígeno-1 de la superficie del merozoito (MSA-1)	*Plasmodium falciparum*	Hughes, 1992.	Nei-Gojobori
Proteína de la membrana externa	*Chlamydia*	Endo y col., 1996.	Nei-Gojobori
Porina 1 (porB)	*Neisseria gonhorreae* y *N. meningitidis*	Smith y col., 1995.	Nei-Gojobori
Glicoproteínas S y HE	Coronavirus del ratón	Baric y col., 1997.	Nei-Gojobori
Sigma 1	Reovirus	Endo y col., 1996.	Nei-Gojobori
Gen determinante de la virulencia	*Yersinia*	Endo y col., 1996.	Nei-Gojobori
Genes involucrados en la reproducción			
Proteína involucrada en la fertilización 18 kDa	*Haliotis* (Abulón)	Vacquier y col., 1997.	Nei-Gojobori
Gen Acp26Aa	*Drosophila*	Tsaur y Wu, 1997.	Li-Wu-Luo
Proteína que une al andrógeno	Roedores	Karn y Nachman, 1999.	Li
Hormona de puesta de huevos	*Aplysia californica*	Endo y col., 1996.	Nei-Gojobori
Gen de caja homeótica Ods	*Drosophila*	Ting y col., 1998.	---

Gen de caja homeótica Pem	Roedores	Sutton y Wilkinson, 1997.	Li
Protamina P1	Primates	Rooney y Zhang, 1999.	Nei-Gojobori
Lisina del esperma	*Haliotis* (Abalone)	Vacquier y col., 1997.	Nei-Gojobori
RNAsa S	Rosaceae	Ishimizu y col., 1998.	Nei-Gojobori
Gen de la diferenciación testicular Sry	Primates	Pamilo y O'Neill, 1997.	Pamilo-Bianchi-Li
Genes involucrados en la digestión			
Caseína *K*	Bovinos	Ward y col., 1997.	---
Lisozima	Primates	Messier y Stewart, 1997.	---
Toxinas			
Conotoxina	Gástropodo *Conos*	Duda y Palumbi, 1999.	Ina
Fosfolipasa A_2	Serpientes crotalinas	Nakashima y col., 1995.	Miyata y Yasunaga
Genes relacionados a transporte de electrones y síntesis de ATP			
Subunidad F_0 de la ATP sintetasa	*E. coli*	Endo y col., 1996.	Nei-Gojobori
Isoforma COX7A	Primates	Schmidt y col., 1999.	---
COX4	Primates	Wu y col., 1997.	---
Citoquinas			
SF de Granulocito-macrófago	Roedores	Shields y col., 1996.	---
Interleucina-3	Primates	Shields y col., 1996.	---
Interleucina-4	Primates	Shields y col., 1996.	---

Misceláneos			
CDC6	*Saccharomyces cerevisiae*	Endo y col., 1996.	Nei-Gojobori
Hormona de crecimiento	Vertebrados	Wallis, 1996.	Nei-Gojobori
Cadena β de la hemoglobina	Peces antárticos	Bargelloni y col., 1998.	ML
Gen Jingwei	*Drosophila*	Long y Langley, 1993.	---
Péptido C3 de la prostateína	Rata	Endo y col., 1996.	Nei-Gojobori

*ML : Método de Máxima verosimilitud

De esta tabla es evidente que existen pocos ejemplos a nivel molecular de genes que se encuentren sujetos a selección positiva. En la literatura la mayoría de los trabajos indican que la mayor parte de la evolución molecular está dominada por los procesos neutros y la selección negativa o purificadora, como Kimura (1983) y otros autores (Ohta, 1992; Li, 1997) han discutido. Pero ¿qué sucede cuando estudiamos los genomas completos? ¿acaso la selección actúa de la misma forma que a nivel génico e intragénico?

3.6 La selección a nivel genómico

La gran cantidad de secuencias almacenadas en las bases de datos (GenBank, EMBL) representan tanto un reto como una oportunidad para la genética de poblaciones y la genómica en cuanto al desciframiento de la historia evolutiva que contienen, así como, para entender los mecanismos evolutivos que han moldeado la diversidad molecular observada. De la misma forma han implicado la necesidad del desarrollo de nuevos métodos de análisis que puedan manejar esta gran cantidad de datos y que nos permitan llevar a cabo interpretaciones efectivas acerca de los procesos evolutivos. Esta tarea es especialmente difícil si tomamos en consideración que muchos procesos evolutivos nos pueden llevar a observaciones similares.

El advenimiento de la era de la genómica ha generado la disponibilidad de una gran cantidad de genomas completos, lo que se ha convertido en una herramienta poderosa para el estudio de la actuación de las distintas fuerzas evolutivas a gran escala. Este hecho nos

brinda la posibilidad de hacer comparaciones de alta resolución y poder estadístico debido al número de genes involucrados en el estudio.

Se han descrito hasta la fecha dos trabajos importantes que parecen describir los papeles que tiene la selección a nivel genómico, curiosamente ambos trabajos implican el estudio de dos dominios diferentes, el de las bacterias y el eucarionte. Y de esta misma forma los resultados parecen ser antagónicos. El primer estudio corresponde a la comparación entre las tasas de sustitución de los denominados genes esenciales y no esenciales de tres especies bacterianas: *E. coli*, *Helicobacter pylori* y *Neisseria meningitidis* (Jordan y col., 2002) (Tabla 4). En este trabajo los autores asignan la categoría de genes esenciales apoyados en los datos experimentales (sobre todo de mutantes) que existen de *E. coli*, extrapolándolos a los genes homólogos en las otras especies bacterianas. Las tasas de sustitución fueron obtenidas mediante el método de Pamilo-Bianchi-Li (Li, 1993 y Pamilo y Bianchi, 1993). Mostramos a continuación una modificación de la tabla descrita en el trabajo de Jordan (Jordan y col., 2002) (Tabla 4), donde podemos ver la comparación entre las distintas tasas de sustitución y el cociente de relación entre ambas Ka/Ks (equivalente a dN/dS).

Tabla 3.4. Comparación de las distintas tasas de sustitución de nucleótidos y el cociente (Ks/Ka) entre genes esenciales y no esenciales de tres especies bacterianas, obtenidas mediante el método de Pamilo-Bianchi-Li (Pamilo y Bianchi, 1993 y Li, 1993).

	Ks (±se)	**Ka** (±se)	**Ka/Ks** (±se)
Escherichia coli			
Genes esenciales	0.02699 ±0.0021	0.00111 ±0.0010	0.0450 ±0.007
No esenciales	0.051 ±0.0010	0.00360 ±0.0002	0.0840 ±0.002
Helicobacter pylori			
Genes esenciales	0.11133 ±0.0041	0.01289 ±0.0011	0.1132 ±0.009
No esenciales	0.03524 ±0.0031	0.02164 ±0.0014	0.1614 ±0.008
Neisseria meningitidis			
Genes esenciales	0.06537 ±0.0069	0.00476 ±0.0007	0.0732 ±0.010
No esenciales	0.09156 ±0.0059	0.00960 ±0.0009	0.1765 ±0.019

Como podemos observar, al parecer existe una diferencia significativa entre los genes esenciales y no esenciales dentro del genoma de *E. coli* (Tabla 4). De igual manera observamos que este patrón se repite para *H. pylori* y *N. meningitidis*. Además, al parecer las tasas de sustitución son mayores en éstos dos últimos patógenos en comparación con las observadas para *E. coli*, lo que parece indicar que estos organismos evolucionan más rápidamente (Tabla 4). Uno de los resultados más importantes de este estudio es el hecho de que los autores no detectan selección positiva en todo el análisis. Aunque no descartan la posibilidad de que en algunas regiones específicas de algunos genes exista este tipo de selección. Concluyen que, en el caso particular de las bacterias, la selección negativa o purificadora es casi la regla en la evolución a nivel proteínas. Es decir, que la tasa evolutiva es determinada por la proporción de sitios en la proteína que poseen altos coeficientes selectivos contra las mutaciones deletéreas (Jordan et al, 2002). En otras palabras, estos resultados sugieren que la selección purificadora es la principal fuerza moldeando la historia bacteriana a nivel genómico.

El segundo trabajo analiza el papel que juega la adaptación a nivel genómico dentro del género *Drosophila* (Smith y Eyre-Walker, 2002). Se basan en la determinación de la proporción de aminoácidos que se sustituyen debido a selección positiva durante la divergencia de las especies *D. simulans* y *D. yakuba* utilizando todos los genes homólogos existentes para ambas especies. Los valores de las sustituciones sinónimas y no sinónimas son calculados a partir del modelo de máxima verosimilitud desarrollado por Yang y colaboradores (Goldman y Yang, 1994, Nielsen y Yang, 1998, Yang y Nielsen, 2000 y Yang y Bielawski, 2000), que ya han sido descritos ampliamente. Se correlacionan estos parámetros con el polimorfismo de tipo sinónimo o no sinónimo a lo largo de la secuencia de los distintos genes. Las conclusiones de este trabajo muestran que aproximadamente un 24% de las sustituciones de aminoácidos entre ambas especies de mosca de la fruta son producto de la selección positiva. Finalmente, determinan que existen aproximadamente 270,000 sustituciones de aminoácidos seleccionadas positivamente durante la divergencia de las dos especies. Lo que implica (tomando el tiempo de divergencia entre ambas que es de 6 millones de años) que estas dos especies han sufrido en promedio una sustitución del tipo adaptativo cada 45 años ó cada 450 generaciones (si contamos 10 generaciones por

año). Este último resultado es consistente con lo que Haldane describe como el costo de la selección positiva de tipo Darwiniano. Lo que indicaría que para el caso de este grupo de eucariontes la selección natural de tipo darwiniano posee el papel principal dentro de la evolución molecular, a diferencia de su contraparte procarionte.

3.7. Discusión

La selección natural ha sido, desde la propuesta inicial hecha por Darwin que la centraba como fuerza evolutiva principal dentro del proceso evolutivo y el origen de nuevas especies, centro de continuo debate. Existe hoy en día un interés especial en el estudio de la selección positiva ya que es la que nos brinda evidencias directas acerca del proceso adaptativo a nivel molecular, lo que nos ayuda en la comprensión de las relaciones entre genotipo-fenotipo. El papel que juega ha sido discutido desde la perspectiva macroevolutiva, donde al parecer es más clara su acción, hasta la microevolutiva donde los postulados de la teoría neutra parecen gobernar la evolución molecular. Sin embargo, esto no necesariamente la ubica en un segundo plano dentro de los procesos de especiación y polimorfismo a nivel molecular. Sólo es hasta el presente que se han desarrollado nuevas herramientas metodológicas que nos permiten cuantificar la selección incluso hacia dentro de los mismos genes. Esto cambia las unidades en estudio y nos invita a reflexionar en las unidades mínimas evolutivas o de selección que tenemos que utilizar con fines de detectar la selección a nivel molecular. Los métodos clásicos suponen que la mínima unidad evolutiva es el gen. Por lo tanto, la cuantificación de la selección se lleva a cabo como un promedio de todos los sitios dentro del mismo, sin distinción alguna entre ellos. Los métodos más novedosos basados en sustitución de codones y máxima verosimilitud toman como unidad evolutiva el codón lo que nos permite registrar detalladamente si existen presiones de selección distintas dentro de un gen lo que nos brinda una mayor resolución. Es por esta razón, que se han convertido en herramientas indispensables para el análisis molecular, ya que no sólo nos permiten cuantificar la selección sino además el cálculo mismo de las tasas de sustitución molecular puede ser aplicable a estudios de reloj molecular y tasas de especiación, lo que nos estaría abriendo un puente entre los procesos macroevolutivos y microevolutivos, en lo que creemos será la segunda gran síntesis dentro de la biología moderna. Por otra parte, el hecho de contar con la descripción de cada uno de

los sitios y el tipo de selección bajo la cual están sujetos, serán la base de futuros experimentos de mutación dirigida, terapias génicas, diseño de proteínas, predicción de estructura secundaria y terciaria, así como, de evolución experimental.

El debate continúa sobre todo en el sentido de la metodología, sin embargo, las evidencias acerca del papel que la selección positiva tipo Darwiniana juega dentro del proceso de evolución molecular apuntan a que éste es más importante de lo que se ha pensado pero que ha sido relevante en sitios muy específicos dentro de las proteínas. Estos resultados arrojan nuevas discusiones acerca de las unidades mínimas de selección que al parecer pueden ser tan pequeñas como un sitio dentro de un codón.

* Se anexa una descripción detallada de los programas más utilizados para el cálculo de las tasas de substitución molecular y selección, además de otros paquetes útiles para el estudio de evolución molecular, en el Apéndice I.

Capítulo 4. A genomic population genetics analysis of the pathogenic enterocyte effacement island in *Escherichia coli*: the search for the unit of selection.

Resumen

El estudio de la genómica comparada ha resultado ser una herramienta poderosa para la comprensión de la evolución y la organización de los genomas. Las herramientas matemáticas y el marco teórico de la genética de poblaciones, en conjunto con el análisis genómico, nos brindan una poderosa aproximación al estudio de las heterogeneidades dentro de la evolución del genoma. En este trabajo se presenta un análisis jerárquico de la isla LEE (Locus of enterocyte and effacement) (35kb), que se encuentra presente en las cepas enteropatogénica y enterohemorrágica de *Escherichia coli* y *Citrobacter rodentium*. Esta isla en *E. coli* se considera una unidad clonal dentro de un organismo clonal y se espera que evolucione como una unidad genética. El análisis del presente estudio prueba la hipótesis clonal mediante la determinación de la diversidad genética, el contenido de GC, así como, las tasas de substitución nucleotídica en varios niveles funcionales de organización: (i) la isla genómica, (ii) los cinco operones en los que la isla se encuentra organizada y (iii) cada uno de los 41 genes que comprenden la isla. Se encontró que existe una región conservada que se compone de los genes pertenecientes a un sistema de secreción tipo III y que son producto de transferencia horizontal. Una región más diversa comprende genes de proteínas secretoras y genes que al parecer son originales del genoma de *E. coli*. Este mosaico genético parece estar moldeado de manera diferencial por la mutación y la selección. Nuestros resultados sugieren que tanto la recombinación como la selección pueden estar rompiendo la estructura clonal de la isla por lo que la mayoría de sus elementos se encontrarán débilmente ligados en su evolución. Estas observaciones sugieren que las unidades de selección no son las islas genómicas, sino unidades mucho más pequeñas dentro de los genes que la integran.

* Se anexa la parte que corresponde a Supporting Materials and Methods en el Apéndice II.

A genomic population genetics analysis of the pathogenic enterocyte effacement island in *Escherichia coli*: The search for the unit of selection

Amanda Castillo, Luis E. Eguiarte, and Valeria Souza*

Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Ap. 75-275, Coyoacán, 04510, Mexico

Communicated by M. T. Clegg, University of California, Irvine, CA, November 29, 2004 (received for review March 17, 2004)

Comparative genomic analysis is a powerful tool for understanding the history and organization of complete genomes. The mathematical tools of population genetics combined with genomic analysis provide a powerful approach to dissect heterogeneities in genome evolution. This study presents a hierarchical analysis of the enterocyte and effacement island (35 kb), which is found in the enteropathogenic and enterohemorrhagic strains in *Escherichia coli* and in *Citrobacter rodentium*. The locus of enterocyte and effacement in *E. coli* is considered to be a clonal unit inside a clonal organism and is expected to evolve as a single unit. This analysis examines the clonal assumption by determining genetic diversity, GC content, and the substitution rates at the different functional levels of (*i*) the complete pathogenic island, (*ii*) the five operons in which the island is organized, and (*iii*) for each of the individual 41 genes that comprise the locus. We find that there is a conserved region that is composed of genes that belong to the type III secretion system and that may be products of horizontal transfer. A more diverse region is composed of genes for secreted proteins and genes that we infer to be original components of the *E. coli* genome. This genetic mosaic seems to be differentially affected by selection and mutation. Our results suggest that recombination and selection may be breaking this structure so that different elements are, at best, weakly coupled in their evolution. These observations suggest that the units of selection are not the complete island, but rather, much smaller units that comprise the island.

natural selection | pathogenicity island | positive Darwinian selection | mutation | GC content

Escherichia coli is a diverse bacterial species living in multiple habitats, including the intestine of mammals and other vertebrates as a free organism, commensal organism, or pathogenic organism (1). The comparative analysis of complete genome sequences of four *E. coli* serotypes, the enterohemorrhagic *E. coli* (EHEC) pathogens EHEC O157:H7 strain EDL933 (2) and O157 Sakai (3), the uropathogenic strain CFT073 (4), and the nonpathogenic laboratory K-12 MG1655 strain (5), reveals that this bacteria exhibits substantial genome diversity, where only 39.2% of proteins are shared between the four strains (4). This result strongly suggests that the genome is a mosaic that includes a conserved backbone considered to be the core *E. coli* genome, together with genomic islands comprised of groups of genes interleaved throughout the genome (2). Numerous studies of population genetics of human-related *E. coli* have strongly suggested that this enteric bacteria is a clonal organism (6, 7), where periodic selection is the cohesive evolutionary force. Yet, when viewed at the whole-genome level, *E. coli* is a mosaic characterized by different units (islands, operons, and genes) with different evolutionary histories.

Genomic islands and operons are considered units where groups of genes are transcribed together and whose products contribute to a specific function (8). Typical examples of genomic islands are the pathogenic islands (PAIs) present in pathogenic bacteria that form the principal molecular component responsible for the development of a specific disease (9–11). Evidence supports the idea that PAIs are considered genetic units horizontally transferred through bacterial species during evolution (11). PAIs have common features including a preference for insertion at tRNA sites and atypical GC content (9, 10). On the other hand, it has been suggested that operons are also mobile elements originated by horizontal transfer events (8). These genetic units (PAIs and operons), where genes are acting in concert, are expected to evolve at homogeneous rates owing to their mutual interdependence in producing a phenotype (8). Thus, evolutionary parameters of these genetic units such as GC content, genetic diversity, codon usage, and substitution rates are expected to be homogeneous. However, if selection is sufficiently weak and if the magnitude of recombination is sufficiently large, the PAIs may become decoupled in their evolution (8). A second important feature of genomic islands is their hierarchical organization because islands are themselves composed of groups of operons, which are in turn composed of groups of genes that are interdependent in their regulation. The goal of this study is to ask whether an important PAI evolves in a homogeneous fashion, and if not, whether patterns of evolution are homogeneous within the operons and/or within the genes that comprise the PAI.

A PAI of interest is the locus of enterocyte and effacement (LEE) that is likely to encode almost all of the genes necessary to produce an intestinal attaching/effacing (A/E) lesion (12); the acquisition of this PAI probably transforms nonpathogenic *E. coli* strains into pathogenic strains (13). The average size of LEE is ≈35 kb with a GC content of 38% (14, 15), which is very different from the housekeeping genes of *E. coli* (GC content = 50%) (4). The locus comprises 41 genes that include a type III secretion system (TTSS) (16), an adhesin-denominated intimin (*eae*) (17), its receptor (*tir*) (18), several secreted proteins (*espA*, *espD*, *espB*, and *espF*), and their chaperones (19). The TTSS apparatus directs the transfer of specific proteins across the bacterial envelope, where the secreted proteins function to transfer effector proteins into host cells (16). The adhesin receptor *tir* is transferred into host cells, where it is modified by host kinases, and becomes inserted into the plasma membrane to orchestrate cytoskeletal rearrangements; this activity depends on its interaction with the adhesin (*eae*) and tyrosine phosphorylation (18). The secreted proteins are required for the translocation of other proteins into the host cell; in the specific case of *espA*, this protein forms a filamentous conduit along which secreted proteins travel before they arrive at the translocation pore in the plasma membrane of the host cell, comprised of *espB* and *espD* (19, 20). Many secreted proteins before secretion are maintained in the bacterial cytoplasm by association with a specific chaperone (19–21). The 41 genes are organized in five polycystronic operons known as LEE1, LEE2, LEE3, TIR, and LEE4, all of them positively regulated by *ler* (12), which is localized at LEE1.

Abbreviations: EPEC, enteropathogenic *Escherichia coli*; EHEC, enterohemorrhagic *E. coli*; TTSS, type III secretion system; PAI, pathogenicity island; LEE, locus of enterocyte effacement; A/E, attaching/effacing; dS, synonymous substitutions per synonymous site; dN, nonsynonymous substitutions per nonsynonymous site; PDS, positive Darwinian selection; ILD, incongruence length difference.

*To whom correspondence should be addressed. E-mail: souza@servidor.unam.mx.

www.pnas.org/cgi/doi/10.1073/pnas.0408633102

Additionally, recent evidence suggests that there are two more regulators inside LEE: *orf10* seem to encode a negative regulator and *orf11* seem to encode a positive regulator (22). However, nearly half of LEE genes (with the exception of the TTSS) seem to have no homologs in other bacteria and have no identified function. LEE islands can be found in diverse range of A/E pathogens with different host specificity and evolutionary history. These pathogens include *E. coli* strains that are natural pathogens of animals, such as rabbits, pigs, cats, dogs, and *Citrobacter rodentium*, a mouse pathogen.

We present a comparative bioinformatic analysis of the LEE island from six epidemic pathogenic strains of *E. coli* that includes descriptions of the genetic structures of the islands and a determination of the role of selection and mutation on the present structure of LEE.

Methods

Sequences. For this study, we used six complete sequences for LEE that are available at GenBank; they correspond to the A/E pathogens described in Table 1, which is published as supporting information on the PNAS web site. We carried out three different scales of analysis: (*i*) we used the complete LEE as a unit, (*ii*) we dissected LEE into its five operons and analyzed them separately, and (*iii*) we studied each of the 41 genes that comprise the island individually. A statistical analysis of ANOVA and a Tukey's test were also performed to test for significant differences of GC content and substitution rates between operons and genes.

Alignment. CLUSTALW (23) was used to produce a multiple alignment of the six LEE islands. The corresponding delimitation of the coding and noncoding regions was performed by using BIOEDIT (24) (the complete alignment is available on request). Both LEE islands of EHEC strains are reported in the reverse orientation compared with the other strains, so we hand-corrected them to be in the same direction. Shiga toxigenic *E. coli* corresponds to LEE locus II, which contains LEE as generally described, and a region of 23,586 bp that carries additional elements of pathogenesis characteristic of this strain, but that we discarded for this study. A special case is *orf3* that is not considered to be present in EPEC strain RDEC-1 and Shiga toxigenic *E. coli* strains because a mutation of T to C at the first codon interrupts the initial methionine; however, because the rest of the sequence remains unchanged and homologous to the other islands, we included it for the study as additional informative sites. *C. rodentium* has an inversion of the two first genes of the island (*rorf1* and *espG*), and they are localized at the end of LEE. For the purpose of this study, we analyzed them as if they were in the same order as the rest of the strains. For the *espF* gene, we included only the partial sequence (606 bp) because the last base pairs of the gene are hypervariable (data not shown).

Genetic Diversity. The genetic diversity of LEE was assessed by the estimation of pi (π), which was calculated from the total number of sites and determined with DNASP (25).

GC Content. The GC content determination was carried out by using MEGA 2.1 (28). We presented the average GC content for the complete LEE, for each of the five operons, and for every gene. We also included the distribution of GC content at first, second, and third position of LEE genes.

Codon Adaptation Index. The codon adaptation index for the 41 genes of LEE was calculated by using CODONW, which was written by J. Peden (Institut Pasteur, Paris), and can be accessed at www.bioweb.pasteur.fr/seqanal/interfaces/codonw.html, based on the work of Ikemura (27).

Phylogenetic Analyses. We constructed the genealogy of the six islands by using the alignment described above that corresponds to the complete core consensus sequence homologous between them, including coding and noncoding sites by using MEGA 2.1 (26). The genetic distances were generated under neighbor joining (28) with Tamura–Nei distance (31) (data not shown). The same method was used for the genealogy construction of every individual LEE gene. Character support for the genealogies was assessed by 5,000 bootstrap resamplings of the data (30), in the case of LEE genes, and 10,000 iterations for the complete six islands. In addition to the genealogy construction, congruence between genes was assessed by using the incongruence length difference (ILD) test (31), which is available in PAUP V.4.0B (32). The ILD was performed for combined data matrices of some selected genes, and the test was performed on all matrices with 1,000 data partitions by using branch and bound searches (33). One additional method was used to support the ILD results. Split-decomposition analysis detects conflicting phylogenetic signals by allowing the genealogies to be expressed in a network rather than a tree-like representation of relatedness. This analysis was performed by using SPLITSTREE V.2.4 (34) with Hamming distances and using informative sites only.

Statistical Test for Adaptive Selection. The determination of the nucleotide substitution changes and the nonsynonymous (dN) and synonymous (dS) substitutions per site ratio is frequently used as a test for positive Darwinian selection (PDS). When this ratio is equal to 1, the gene is under the neutral model of evolution, if this ratio is >1, PDS is inferred, whereas if this ratio is <1, purifying selection is inferred (35, 36). To assess the role of selection among LEE, we used two methods. First, we performed a general estimation of the dS and dN, respectively, for the whole-sequence sample; in this case, we used the analysis developed by Nei–Gojobori (37) implemented by using the DNASP package (25). Second, we used the integrative approach for detecting selection at specific amino acid sites included at HYPHY 0.90IB, which can be accessed at www.hyphy.org (38) from the site www.datamonkey.org developed by Kosakovski-Pond and Frost (39). The integrative analysis for detecting selection includes three different analyses called single likelihood-derived ancestor counting, approximate likelihood ratio at a site, and full likelihood (see refs. 39–41 for more detailed information on the methodology). All of the analyses start with a given estimate of the genealogy (described above), and, fitting a codon substitution model, the number of changes occurred along each genealogy are estimated with different methodology. We present the average dN/dS ratio and the sites for each LEE gene that are inferred to be under PDS and/or purifying selection.

Results

General Analysis of LEE Islands. A total of 32,148 bp comprises the core sequence shared by the six LEE islands, including coding and noncoding sites. From the general alignment, it is evident that the core region of LEE is conserved among A/E pathogens in sequence and structure. The variable parts include the flanking regions, the insertion sites, some intergenic regions of genes such as the intimin (*eae*), *rorf1*, and *espG*, and insertions between the operons TIR and LEE4, which is especially evident in *C. rodentium* (data not shown). From the core sequence of the six islands, 78% (25,067) corresponds to conserved sites, and 22% (7,051) are polymorphic sites. Although there is high degree of conservation in structure, the genetic diversity is relatively high ($\pi = 0.10$). The GC content is low at 38.6% ($\pm$ 0.46 SE), which is congruent with previous reports (38%) (14, 15). The complete genealogy for the six islands was in agreement with the pathotypes relations described (42) (Table 1 and Fig. 1), where the EHEC-1 and enteropathogenic *E. coli* (EPEC)-1 groups are more divergent and more related between each other, and the EHEC-2 group is more conserved and less divergent (42).

LEE Operons. The number of genes varies among LEE operons. From Table 2, which is published as supporting information on the

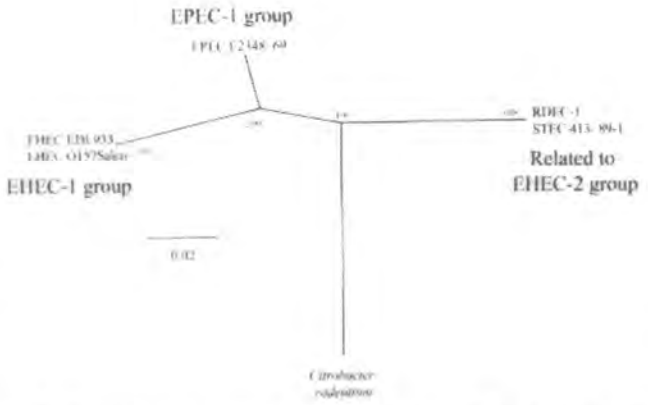


Fig. 1. Genealogy of the six LEE islands used in the study. The genealogy was constructed by using the 32,148 bp that comprise the core consensus of the six islands (including coding and noncoding sites), under the neighbor-joining method with Tamura–Nei distance, and 10,000 bootstrap resamplings. STEC, Shiga toxigenic *E. coli*.

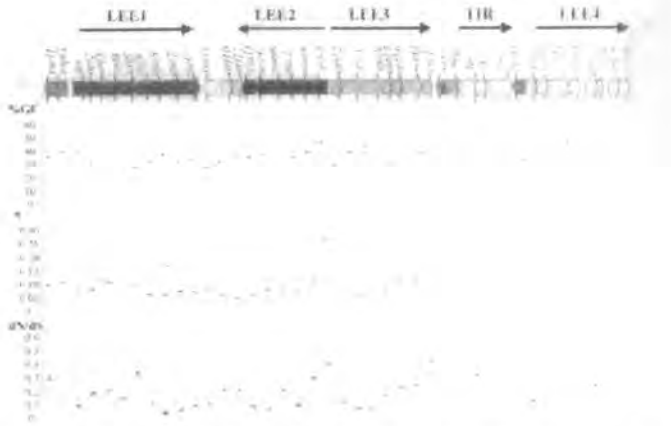


Fig. 2. GC content, genetic diversity, and dN/dS ratio distribution for the 41 genes of LEE.

PNAS web site, we observe that the first three operons (LEE1, LEE2, and LEE3) have similar number of genes, genetic diversity, and GC content. These three operons comprise almost the entire TTSS of the LEE locus. On the other hand, the last two operons (TIR and LEE4) exhibit more variation in gene number and sequence size. LEE1 comprises nine genes (from *ler* to *escU*) with an average genetic diversity of $\pi = 0.060$. This operon has the lowest GC content (33.1%) of the group, which is significantly lower ($P < 0.0001$) than the average of the complete LEE locus (38.6%). This operon contains the main regulator recognized for LEE (*ler*). LEE2 includes six genes (from *cesD* to *sepZ*) with a genetic diversity of $\pi = 0.090$ and a GC content of 37.8% close to the average of the complete island. This operon includes the gene that contains the highest genetic polymorphism (*sepZ* $\pi = 0.24$) and has not been functionally studied yet. LEE3 includes six genes (from *orf12* to *espH*) with an average genetic variation of $\pi = 0.073$ that, with the exception of *espH* ($\pi = 0.18$), is conserved among the genes that integrate the operon. The average GC content is 39.2% (including *espH*). TIR comprises only three genes (*Tir*, *cesT*, and *eae*) that have the highest genetic variation observed for the LEE locus ($\pi = 0.133$). The average GC content (43.3%) is also significantly higher ($P < 0.0001$) than the average of the LEE locus. This operon includes the adhesin denominated intimin (*eae*) and its receptor (*tir*) that together are fundamental for the development of the A/E lesion, giving the close attachment to the host membrane that characterize the A/E pathogens. LEE4 includes eight genes (from *sepL* to *espF*) with a genetic diversity similar to TIR ($\pi = 0.129$) and a GC content of 41.9%. This operon contains several secreted proteins as *espA*, *espD*, *espB*, and *espF* that are responsible for the signal transduction system and for the melting of the microtubules of the host cell. The ANOVA and Tukey's tests show that GC content between the five operons is significantly different ($P < 0.0001$).

Genetic Diversity of LEE. The 41 genes of LEE have an average genetic diversity distribution that ranges from $\pi = 0.03$ (*orf11* SE ± 0.008) to 0.24 (*sepZ* SE ± 0.027) (Table 3, which is published as supporting information on the PNAS web site). This distribution characterizes the broad range of genetic variation contained within the island. A comparison of the diversity of LEE genes with genes that are part of the conserved backbone of *E. coli* like *mdh* ($\pi = 0.01$; $n = 46$), *putp* ($\pi = 0.02$; $n = 12$), *fimA* ($\pi = 0.06$; $n = 7$), and *trpA* ($\pi = 0.03$; $n = 25$) (43) reveals the high genetic diversity characteristic of the pathogenic genes. This finding is especially evident if we consider the sample size of the present study ($n = 6$), and that all of the strains used belong to epidemic clones. It will be of future interest to explore this variation in nonepidemic strains from a broad range of *E. coli* natural hosts (A.C., unpublished work).

GC Content of LEE Genes and Codon Adaptation Index. The average GC content of LEE genes also exhibits a broad distribution from 28.3% SE ± 0.33 (*rorf3*) to 53.1% SE ± 0.22 for *espF* (Fig. 2). To describe in more detail how this GC content is distributed among the different coding positions of the genes, we divided it in first, second, and third positions. The GC content is significantly different ($P < 0.0001$) at the first position (45.09%), as compared with the second (34.4%) and third (33.5%) positions that are similar to each other. It has been proposed that sequences introduced by horizontal transfer, as suspected for the complete LEE locus, will be affected by the same mutational processes as the recipient genome and eventually, will converge to the base composition and codon usage of the resident genome (44). This process should occur most rapidly at sites with little or no functional constraint, particularly the third position, where most changes are synonymous. Accordingly, we expect the third position to have a higher GC content. However, in this case, the first position is the most similar to the resident genome in GC content. Perhaps more interesting is the observation that the third position has an even a lower average GC content than the second position. If mutational processes and selection affect all genes homogeneously, we might expect GC content to have increased toward the *E. coli* average (55.4% for the third position, 40.7% for the second position, and 58.8% for the first position) (45), instead, we observe a very low average GC content (33.5%). These results establish a high codon bias for LEE genes. This result could be interpreted as a signal of conservation from its original source, even when the island has been horizontally transferred in multiple occasions for a considerable time (11, 13), or it could be a signal of regulation, because LEE is only expressed on the logarithmic growth phase during infection to the host (17). We calculated the codon adaptation index (0.200 ± 0.0016) for the genes of the LEE locus (excluding *C. rodentium*) and found that they were considerably lower than for the average present in *E. coli* genes (0.485 ± 0.051) (4). This index indicates the use of rare codons when it has low values. In this case, is showing that the genes of LEE are biased and differ from the reminder of the *E. coli* genome.

Genealogy of LEE Genes. The phylogenetic study of LEE island genes highlighted differences in the evolutionary histories of some of its members. Almost all of the 41 genes give the same branch order between each strain in congruence with the described for the

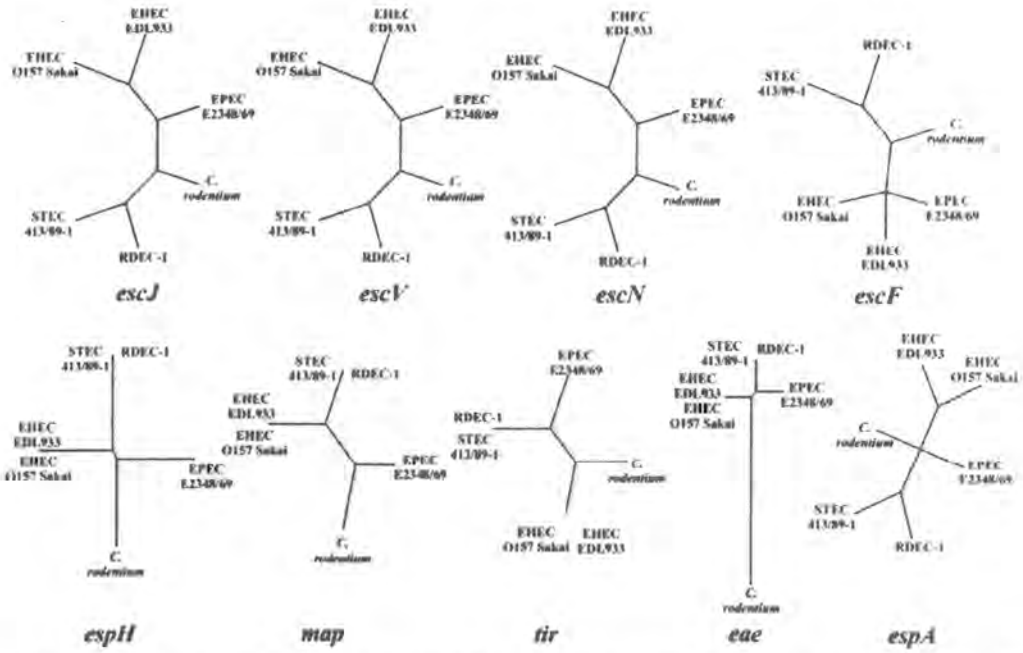


Fig. 3. Genealogy of nine LEE genes used in the study. The genealogy of each gene was constructed under the neighbor-joining method with Tamura–Nei distance and 5,000 bootstrap resamplings. The genes *escJ*, *escV*, *escN*, and *escF* belong to the TTSS. The genes *espH*, *map*, and *espA* are secreted proteins, and *tir* is the receptor for the adhesin *eae*.

genealogy based on the pooled LEE genes sequence data. Interesting exceptions are *orf3*, *cesD*, *rorf6*, *rorf8*, *sepZ*, *orf16*, *espH*, *cesF*, *map*, *tir*, *eae*, *espA*, and *espB* that give different branch orders and lengths (the genealogy of each gene is available on request). In these genes, the EPEC strain E2348/69 is very divergent, not closely related to the EHEC-1 group as described before, but is grouped with *C. rodentium*. We selected a sample of these genes (Table 4, which is published as supporting information on the PNAS web site) and constructed matrices in a pairwise manner by using some of the genes that presented a different genealogy with a sample of those that have the same consensus LEE genealogy for the ILD test. The ILD test statistically confirmed the topologic incongruence that supports the idea that some genes have a different phylogenetic history than the rest of LEE. Significant differences ($P = 0.001$) were observed when we compare some genes of the TTSS (*escJ*, *escR*, *escS*, *escU*, *escV*, *escN*, *escD*, *sepZ*, and *escF*) with *espH*, *map*, *tir*, *eae*, *espA*, and *espB* (Fig. 3). We also use these genes to test for the split-decomposition analysis, and the incongruence, together with possible signals of recombination. We observed recombination for *espH*, *map*, *tir*, *eae*, *espA*, and *espB* (Fig. 4). This result suggests that recombination may be breaking the linkage between these genes and allowing them to diverge.

Nucleotide Substitutions of LEE Genes. We determined the dS and dN for the 41 genes of the LEE locus. The dS is assumed to be primarily related to mutational processes because they do not alter the amino acid composition. The dS estimates range from 0.10 (SE ± 0.02) for *orf11* to 1.12 (SE ± 0.20) for *cesF* (Fig. 2). From this analysis, we can observe that dS estimates are not uniform among the members of LEE.

The dN describes the substitution rates at sites with amino acid changes, so they are an index of both selective and neutral events. The dN estimates for the LEE genes are more restricted in distribution than the dS. This result is especially true in the case of the TTSS, where we find values as low as dN = 0.011 SE ± 0.001(*escS*), indicating that most mutations at these positions are eliminated. These genes have homologs in *Salmonella* and *Yersinia* (46) and are clearly a product of horizontal gene transfer. In contrast, dN estimates for some genes are high (Table 3), suggesting that some LEE members may show a signal of PDS (*sepZ*, dN = 0.221, SE ± 0.024; *tir*, dN = 0.202, SE ± 0.024; *espH*, dN = 0.176, SE ± 0.019; *espB*, dN = 0.165, SE ± 0.020; *espF*, dN = 0.121, SE ± 0.016; *map*, dN = 0.108, SE ± 0.013; *cesF*, dN = 0.107, SE ± 0.011; *espD*, dN = 0.107, SE ± 0.015; *eae*, dN = 0.096, SE ± 0.011; *escF*, dN = 0.086, SE ± 0.032; and *espA*, dN = 0.082, SE ± 0.010). Some of these genes have direct interaction with the host.

dN/dS ratios were averaged over all of the sites in the gene sequence and they are given for each site of the gene sequence. Thus, we obtained the average dN/dS and also an indication of sites that may be under purifying and/or PDS. For LEE, although there is high polymorphism, most members appear to be under purifying selection (dN/dS <1), whereas some genes are close to neutrality (dN/dS = 1), but none of the genes appear to be under PDS (dN/dS >1) (Fig. 2). The gene that has the highest dN/dS ratio is *espG* (0.54), a secretion protein, this gene in *C. rodentium* is localized at the end extreme of the island, possibly as a product of a rearrangement. Other genes with a high dN/dS ratio are *espF*, *espH*, and *sepZ* with 0.43, 0.41, and 0.39, respectively. The lowest dN/dS ratio is present at some members of the TTSS with values as low as 0.05 for *escS* or 0.07 for *escT* and *escC*.

The result from the site-by-site analysis shows that few genes have sites under PDS. These genes are *espG* (one site), *map* (one site), *tir* (one site), *eae* (three sites), *espD* (one site), and *espF* (one site) (Table 3); none of them are part of the TTSS. These are interesting sites for the study of directed mutagenesis and gene therapy because they are genes that have an important role on the virulence of A/E pathogens, especially the intimin (*eae*) and its receptor (*tir*). Two of the regulators of LEE seem to be neutral (*orf10* and *orf11*), with no sites under adaptive or purifying selection. Most of the genes with the highest number of sites under purifying selection belong to the TTSS (*orf4*, *orf5*, *escR*, *escC*, *escJ*, *escV*, and *escN*) (Table 5, which is published as supporting information on the PNAS web site). TTSS are involved in the development of a complex structure that crosses the membrane so maybe any change at the sites is purged to preserve the structure. The receptor of the intimin (*tir*) is a special case, that, along with *escV* (TTSS) and *escC* (TTSS), present the

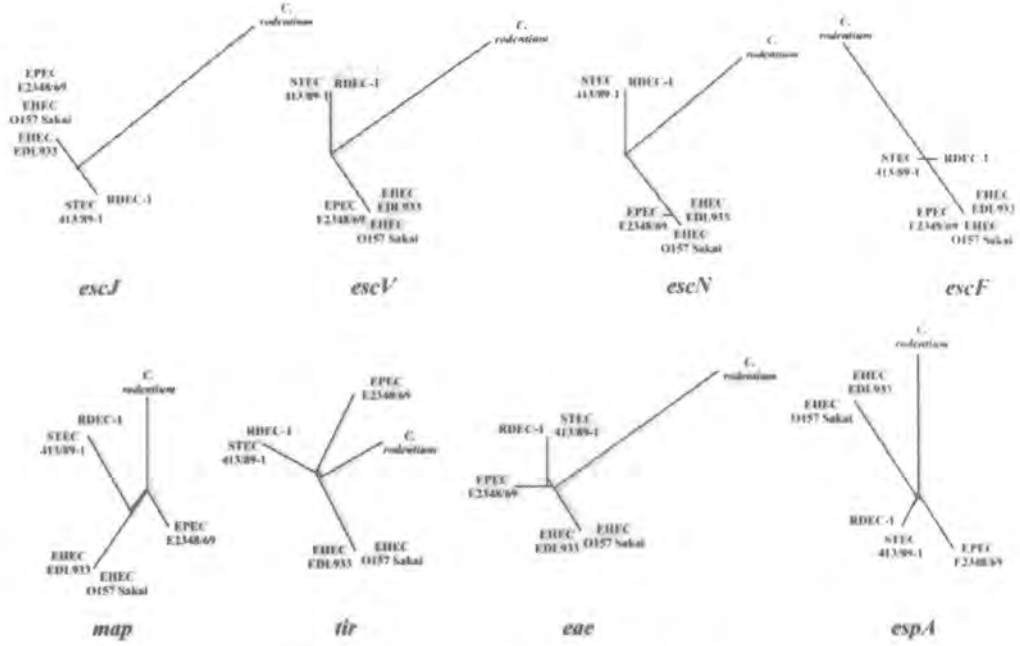


Fig. 4. Split-decomposition analysis of eight LEE genes.

highest number of sites under purifying selection with 102, 110, and 115 sites, respectively.

Discussion

The pathogenic LEE island has been the focus of numerous epidemiological and molecular studies because it represents an excellent model for the evolution of pathogenesis. The acquisition of this PAI is thought to confer pathogenic characteristics upon a normal commensal *E. coli* strain (13). From the evolutionary point of view, LEE has been considered a genetic unit that has been horizontally transferred through A/E pathogens evolution (13). However, from the analysis of wild hosts of *E. coli*, it has been suggested that this island could be more dynamic, involving an ongoing process of construction and disruption (47). This PAI confers different fitness to some pathogenic strains of *E. coli* as in the case of the serotype O157:H7, a successful epidemic clone responsible of important epidemic outbreaks that have caused the death of adults and children (15).

From the present study, we observe an increase of genetic diversity and GC content along the LEE island as we analyze it from its 3' end through the 5' end (Fig. 2). The TTSS, largely represented in the first three operons, is the group of genes with lower levels of genetic diversity and GC content. On the other hand, genes such as *map*, *tir*, *eae*, *espA*, *espD*, *espB*, and *espF* present the highest levels of genetic diversity and GC content closer to the average of *E. coli* genes. From the GC content, substitution rates, and phylogenetic analyses, we infer that the TTSS travel together as a cluster of genes linked by function. The nonsynonymous substitution rate for these genes is low and purifying selection is eliminating diversity. This group of genes that have the lowest GC contents, genetic diversities, and conserved nucleotide substitution rates may preserve the phylogenetic signal of the early formation of the island. These results, together with the finding of other regulators of the secretion system localized outside the LEE island (22), strongly support the idea that the TTSS may also be participating in other processes not necessarily related to virulence. The parallel acquisition of some other virulence factors as an invasin (*eae*) by nonpathogenic *E. coli* strains that already have a TTSS may increase its virulence, completing the molecular scenario for the appearance of a new pathogen. We suggest that the TTSS could be a good candidate for the calculation of the time of its integration to *E. coli* genome and compare it with the adhesin (*eae*) or some secreted proteins as *espB*. This result may reveal the time when the LEE island was originally assembled. Otherwise, because the TTSS is involved in pathogenesis, this finding may provide targets for future therapies.

The fact that genes like the adhesin and its receptor are more divergent, less conserved, and with different genealogies than the rest of the island, supports the hypothesis that virulence is a recent derived state that may be a result of the parallel acquisition of virulence factors. The presence of genes like the adhesin, which is a mosaic product of recombination (48, 49), suggests that the origin and evolution of the LEE island is a complex process.

From the observed polymorphism present at some LEE genes and the nucleotide substitutions results, we expected that some genes will have a dN/dS ratio >1, a clear sign of the participation of PDS. However, none of the loci seem consistent with this assumption, and again this ratio is highly variable across the genes of the island. The second part of the analysis highlighted the specific sites under PDS or purifying selection. Interestingly, it seems that PDS had, at best, a minor role in shaping the history of the LEE island, but this does not mean that was not an important role. A few sites of LEE genes are under PDS in genes that are fundamental for the correct development of the A/E lesion, such as *map* or the adhesin and its receptor. These sites are very important for directed mutagenesis analysis and antibiotic treatment because these genes are key virulence factors in A/E lesion. On the other side, the participation of purifying selection is evident and important for several genes of LEE, like some members of the TTSS, where any change is purged to preserve the protein structure and function. Thus, purifying selection seems to be delimitating important regions for the protein structure.

There are two different types of genealogy for LEE. The TTSS has a different genealogy than other genes like *map*, *tir*, or *eae*, indicating that at least two different transfer events originated the island or that recombination is weakening the assemblage. If recombination is common, we might expect to reconstruct many different genealogies within the island. Curiously, we only have two types of relatedness inside LEE, the general consensus LEE

genealogy and the alternative genealogy where the EPEC strain is more related to *C. rodentium*. The ILD test shows that there is congruence between the genes that belong to the TTSS, so it is clear that this group of genes have a shared ancestry and have been traveling together since the origin of LEE. On the other hand, internal recombination is shaping the genealogies of genes such as *map*, *tir*, *cue*, *espA*, and *espB*, and may be promoting the development of new pathotypes. Thus, the ancient formation of the LEE island may be a product of two different transfer events, the first was the acquisition of the TTSS, and the second the acquisition of genes as the adhesin. However, although they are not part of the conserved backbone of genes in *E. coli*, some unique genes in LEE (like *tir*) may have been generated *de novo*. This result is possible, especially if we considered pathogenesis as a derived state in *E. coli*.

There is a high degree of heterogeneity present among LEE island genes in genetic diversity, GC content, and nucleotide substitution rates. We dissect these heterogeneities beginning with the complete LEE island level on through its constituent operons and genes. A group of genes linked by function and coregulation might be expected to experience common mutational and selective processes. Consequently, if the island is a genetic unit evolving in concert, this fact will be reflected in the conservation of a signature present in GC content, genetic diversity, and nucleotide substitution rates, especially if they share common ancestry and are regulated in synchrony. In this study, we find that diversity, GC content, and nucleotide substitution rates are variable, suggesting that mutation and selection are acting with different intensity within the PAI, generating a mosaic. The results derived from the present study suggest that the assumption that LEE was assembled and has been evolving as a unit is not supported by the data. There probably have been recombination events and different selection pressures for different parts of LEE generating a genetic mosaic, which will create different coalescent times for different regions of the PAI.

The present study included only pathogenic strains from epidemic clones, where selection is believed to be sufficiently high to maintain LEE as a unit during horizontal transfer. However, even in this group of strains, the phylogenetic signal points to evidence for different transfer events in the origin of the LEE island. The results of this study suggest that the origin and evolution of LEE is a more complex process than previously thought. The unit of selection in this specific case, is not the whole island, but smaller modules inside the island and within its constituent genes. This study may also delimitate the minimal LEE unit needed for A/E to became pathogens at the first instance. Future work will need to explore both atypical strains of EPEC and EHEC, and also wild strains of *E. coli* not related to humans, to better understand the evolution of pathogenesis.

We thank L. Martínez-Castilla for technical assistance, C. Silva and L. Forney for carefully reviewing the manuscript, and two anonymous reviewers and Dr. Mike Clegg, who made interesting suggestions that considerably improved this study. This work was supported by Consejo Nacional de Ciencia y Tecnologia Genomic Project 0028 and Direccion General de Asuntos del Personal Académico Grant IN-208601.

1. Souza, V., Rocha, M., Valera, A. & Eguiarte, L. E. (1999) *Appl. Environ. Microbiol.* **65**, 3373–3385.
2. Perna, N. T., Plunkett, G., Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A., *et al.* (2001) *Nature* **409**, 529–533.
3. Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C. G., Ohtsubo, E., Nakayama, K., Murata, T., *et al.* (2001) *DNA Res.* **8**, 11–22.
4. Welch, R. A., Burland, V., Plunkett, G., III, Redford, P., Roesch, P., Rasko, D., Buckles, E. L., Liou, S. R., Boutin, A., Hackett, J., *et al.* (2002) *Proc. Natl. Acad. Sci. USA* **99**, 17020–17024.
5. Blattner, F. R., Plunkett, I. G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., *et al.* (1997) *Science* **277**, 1453–1462.
6. Selander, R. K. & Levin, B.R. (1980) *Science* **210**, 545–547.
7. Selander R. K., Caugant D. A. & Whittam, T. S. (1987) in *Escherichia coli and Salmonella typhimurium: Genetic Structure and Variation in Natural Populations of Escherichia coli.*, eds. Neidhardt, F. C., Ingraham, J. L., Low, K. B., Schaeter, M. M. & Umbager, H. E. (Am. Soc. Microbiol., Washington, DC), Vol. 2.
8. Lawrence, J. G. & Roth, J. R. (1996) *Genetics* **143**, 1843–1860.
9. Blum, G., Ott, M., Lischewski, A., Ritter, A., Imrich, H., Tschape, H. & Hacker, J. (1994) *Infect. Immun.* **62**, 606–614.
10. Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R. & Goebel, W. (1990) *Microb. Pathog.* **8**, 213–225.
11. Groisman, E. A. & Ochman, H. (1996) *Cell* **87**, 791–794.
12. Elliott, S. J., Sperandio, V., Girón, J. A., Shin, S., Mellies, J. L., Wainwright, L., Hutcheson, S. W., McDaniel, T. K. & Kaper, J. B. (2000) *Infect. Immun.* **68**, 6115–6126.
13. Reid, S. D., Herbelin, C. J., Bumbaugh, A. C., Selander, R. K. & Whittam, T. S. (2000) *Nature* **406**, 64–67.
14. Elliot, S. J., Wainwright, L. A., McDaniel, T. K., Jarvis, K. G., Deng, Y. K., Lai, L. C., MacNamara, B. P., Donnenberg, M. S. & Kaper, J.B. (1998) *Mol. Microbiol.* **28**, 1–4.
15. Perna, N. T., Mayhew, G. F., Posfai, G., Elliot, S., Donnenberg, M. S., Kaper, J.B. & Blattner, F. R. (1998) *Infect. Immun.* **66**, 3810–3817.
16. Jarvis, K. G., Girón, J. A., Jerse, A. E., McDaniel, T. K., Donnenberg, M. S. & Kaper, J. B. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 7996–8000.
17. Jerse, A. E., Yu, J., Tall, B. D. & Kaper, J. B. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 7839–7843.
18. Kenny, B., DeVinney, R., Stein, M., Reinscheid, D. J., Frey, E. A. & Finlay, B. B. (1997) *Cell* **91**, 511–520.
19. Creasey, E. A., Friedberg, D., Shaw, R. K., Umanski, T., Knutton, S., Rosenshine, I. & Frankel, G. (2003) *Microbiology* **149**, 3639–3647.
20. Clarke, S. C., Haigh, R. D., Freestone, P. P. E. & Williams, P. H. (2003) *Clin. Microbiol. Rev.* **16**, 365–378.
21. Franke, G., Phillips, A. D., Rosenshine, I., Dougon, G., Kaper, J. B. & Knutton, S. (1998) *Mol. Microbiol.* **30**, 911–921.
22. Deng, W., Puente, J. L., Gruenheid, S., Li, Y., Vallance, B. A., Vázquez, A., Barba, J., Ibarra, J. A., O'Donnell, P., Metalnikov, P., *et al.* (2004) *Proc. Natl. Acad. Sci. USA* **101**, 3597–3602.
23. Thompso, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22**, 4673–4680.
24. Hall, T. A. (1999) *Nucleic Acids Symp. Ser.* **41**, 95–98.
25. Rozas, J. & Rozas, R. (1999) *Bioinformatics* **15**, 174–175.
26. Kuma, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001) MEGA2: Molecular Evolutionary Genetics Analysis software (Arizona State Univ., Tempe).
27. Ikemur, T. (1985) *Mol. Biol. Evol.* **2**, 13–34.
28. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4**, 406–425.
29. Tamura, K. & Nei, M. (1993) *Mol. Biol. Evol.* **10**, 512–526.
30. Felsenstein, J. (1985) *Evolution (Lawrence, Kans.)* **39**, 783–791.
31. Farris, J. S., Kallersjo, M., Kluge, A. G. & Bult, C. (1995) *Cladistics* **10**, 315–319.
32. Swofford, D. L. (1999) PAUP*: *Phylogenetic Analysis Using Parsimony (* and Other Methods)* (Sinauer, Sunderland, MA), Version 403b.
33. Allard, M. W., Farris, J. S. & Carpenter, J. (1999) *Cladistics* **15**, 75–84.
34. Huson, D. H. (1998) *Bioinformatics* **14**, 68–73.
35. Goldman, N. & Yang, Z. (1994) *Mol. Biol. Evol.* **11**, 725–736.
36. Yang, Z., Nielsen, R., Goldman, N. & Krabbe-Pedersen, A. M. (2000) *Genetics* **155**, 431–449.
37. Nei, M. & Gojobori, T. (1986) *Mol. Biol. Evol.* **3**, 418–426.
38. Kosakovski-Pond, S. L. & Muse, S. V. (2004) *Mol. Biol. Evol.*, in press.
39. Kosakovski-Pond, S. L. & Frost, S. D. W. (2004) *Mol. Biol. Evol.*, in press.
40. Suzuki, Y. & Gojobori, T. (1999) *Mol. Biol. Evol.* **16**, 1315–1328.
41. Felsenstein, J. (1981) *J. Mol. Evol.* **17**, 368–376.
42. Donnenberg, M. S. & Whittam, T. S. (2001) *J. Clin. Invest.* **107**, 539–547.
43. Boyd, E. F. & Hartl, D. L. (1998) *J. Mol. Evol.* **47**, 258–267.
44. Lawrence, J. G. & Ochman, H. (1997) *J. Mol. Evol.* **44**, 383–397.
45. Nakamura, Y., Gojobori, T. & Ikemura, T. (1998) *Nucleic Acids Res.* **26**, 334.
46. Cornelius, G. L. (2000) *Philos. Trans. R. Soc. London B* **355**, 681–693.
47. Sandner, L., Eguiarte, L. E., Navarro, A., Cravioto, A. & Souza, V. (2001) *Microbiology* **147**, 3149–3158.
48. McGraw, E. A., Li, J., Selander, R. K. & Whittam, T. S. (1999) *Mol. Biol. Evol.* **16**, 12–22.
49. Tarr, C. L. & Whittam, T. S. (2002) *J. Bacteriol.* **184**, 479–487.

Capítulo 5. Genética de poblaciones.

5.1 Breve introducción a la genética de poblaciones

Una de las mayores y más interesantes preguntas que se hacía la biología a principios del siglo XX era la reconciliación de la teoría evolutiva propuesta por Darwin-Wallace y las leyes de Mendel sobre la herencia. Es durante los años de 1920 a 1930 cuando Ronald Fisher, J.B.S. Haldane y Sewall Wright demuestran que ambas teorías son compatibles dando origen a la disciplina denominada genética de poblaciones (Eguiarte, 1999), la cual se encarga del estudio de las bases genéticas de la evolución. Asimismo, la genética de poblaciones nos brinda una mejor comprensión del proceso adaptativo, así como del origen y mantenimiento de la diversidad genética en las poblaciones naturales. En consecuencia, determina los mecanismos mediante los cuales ocurre la evolución, interpretada como el cambio en las frecuencias alélicas. Esta conjunción de ideas se le conoce como la teoría sintética que marcó una de las etapas de mayor desarrollo y producción científica dentro de la biología. La teoría sintética en conjunto con la teoría neutra propuesta por Kimura (Kimura, 1968 y 1983), marcan un nuevo paradigma y brindan el marco teórico bajo el cual se desarrolla la biología moderna. En la actualidad, con el surgimiento del estudio de los genomas completos y la gran acumulación de información genética, es aún más evidente la necesidad de aplicar el marco teórico de la genética de poblaciones a nivel genómico. De esta manera, abarcaremos desde los procesos macroevolutivos hasta los microevolutivos en lo que parece ser la segunda gran síntesis.

5.2 Estructura genética de las poblaciones de E. coli y el paradigma clonal

En el caso de las bacterias, la reproducción no se encuentra ligada a la recombinación o sexualidad como en los eucariontes. Las bacterias se dividen por fisión binaria, lo que genera en principio individuos clonales, donde la variación únicamente es producida por la mutación. Sin embargo, existen diferentes procesos denominados parasexuales (conjugación, transformación y transducción) los cuales promueven la transferencia horizontal de información genética que, además de la mutación, van a generar variación en las poblaciones. Por lo tanto, una de las preguntas centrales para la genética de

poblaciones bacterianas, es acerca del grado de clonalidad que poseen sus poblaciones, así como la definición de la especie bacteriana. Si las poblaciones son clonales estarían constituidas por linajes evolutivos independientes. Siendo así difícil hablar de especies ya que no tendríamos una poza génica común. Para este caso es difícil aplicar la teoría clásica de la genética de poblaciones; además, la evolución estaría dada por sustituciones de linajes completos ya sea por selección periódica o deriva génica. Por el contrario, si las especies bacterianas presentan altos niveles de recombinación, se tienen estructuras cercanas a la panmixia y se pueden aplicar aproximadamente la teoría de genética de poblaciones que usamos en los organismos diploides. A pesar de que la existencia de la recombinación se demostró hace más de 50 años en *E. coli* (Lederberg y Tatum, 1946), el papel que tiene dentro de las poblaciones bacterianas no se encuentra totalmente establecido y se encuentra bajo continuo debate.

El primer intento por determinar cuál era la estructura de las poblaciones bacterianas y en específico de *E. coli* fue realizado mediante la técnica de electroforesis multilocus (MLEE, por sus siglas en inglés) (Milkman, 1973). Se utilizaron 829 aislados obtenidos principalmente a partir de humanos. En este estudio se determinó que la diversidad genética de la especie era de $\pi = 0.23$ (Milkman, 1973). Estudios posteriores utilizando la misma técnica MLEE (Selander y Levin, 1980) además de demostrar la existencia de desequilibrio de ligamiento en numerosos loci (Selander y Whittam, 1983; Whittam y col., 1983), sugirieron que la estructura de las poblaciones de *E. coli* era clonal. Trabajos posteriores basados en polimorfismo de enzimas en la colección denominada ECOR (la colección de referencia que se compone de 72 aislados a partir de humanos y no humanos y que se considera representa la variación genética natural de *E. coli*) (Ochman y Selander, 1984), han mostrado que la diversidad genética de esta bacteria se encuentra organizada dentro de un número limitado de clonas diferentes, lo que sugiere que la recombinación es baja (Whittam y col., 1983b). A partir de estos resultados se llegó a establecer lo que se denominó el paradigma clonal para las poblaciones de *E. coli*. Sin embargo, las primeras sospechas acerca de este resultado provinieron de los estudios con secuencias de DNA, como es el caso de análisis realizados a partir de las genealogías de algunos loci que han demostrado la existencia de recombinación intragénica (Dykhuizen y Green, 1986; Guttman y Dykhuizen, 1994; Nelson y Selander, 1994). De acuerdo a los análisis anteriores, existe

una contradicción evidente entre la aparente diversidad que las poblaciones bacterianas exhiben y la clonalidad, reflejada en la recuperación del mismo haplotipo en repetidas ocasiones. Estas observaciones sólo podrían ser reconciliadas asumiendo que los tamaños efectivos de las poblaciones son extremadamente pequeños, lo cual resulta muy controversial (Maynard-Smith, 1991; Maynard-Smith y col., 1993). Una alternativa para la explicación de estos resultados contradictorios puede ser que las poblaciones bacterianas poseen una estructura ecológica, es decir, que las diferentes cepas dentro de la misma especie se encuentran adaptadas a nichos ecológicos específicos (Souza y col., 1992; Gordon y Lee, 1999; Souza y col., 1999; Maynard-Smith y col., 2000), como es el caso de *E. coli*. La estructura ecológica en las especies bacterianas tendría consecuencias significativas para dos procesos fundamentales como son la selección periódica y la recombinación (Souza y col., 1992; Gordon y Lee, 1999; Souza y col., 1999; Maynard-Smith y col., 2000). La selección periódica purga la variación genética permitiendo que, en el caso de las clonas bacterianas, una mutación en una de ellas pueda hacer que un mismo haplotipo desplace o sustituya a todos los demás (Levin, 1981). Es entonces que, la estructura ecológica de las especies limitaría el proceso de la selección periódica al nicho ecológico específico (Maynard-Smith, 1981). Si una especie consiste en numerosas poblaciones con múltiples nichos ecológicos distintos es muy poco probable que una sola mutación altere a todas las poblaciones y sean desplazados todos los haplotipos, o que esta mutación en sí tenga el mismo significado adaptativo para todos los casos, es decir, no va a afectar a toda la especie en su conjunto (Maynard-Smith, 1981; Cohan, 1994). De igual manera, la estructura ecológica va a limitar las asociaciones aleatorias generadas por recombinación. Si las distintas cepas se encuentran adaptadas a nichos específicos, entonces es posible que exista una tasa de recombinación alta intrapoblacional pero no necesariamente entre poblaciones de diferentes nichos (Gordon y Lee, 1999). Bajo este escenario los estudios de simulación indican que el desequilibrio de ligamiento será mantenido incluso aunque exista recombinación a una tasa mucho mayor que la tasa de mutación (Guttman, 1997). Por lo tanto, la existencia de estructura ecológica tendrá impacto en la estructura genética y la evolución de las especies bacterianas.

Es entonces, que a luz de estas nuevas propuestas, existe la necesidad de reevaluar el paradigma clonal en *E. coli*, pues al parecer se ha subestimado el papel de la

recombinación y el intercambio genético en esta bacteria, ya que no se ha tomado en cuenta su estructura ecológica y el mosaicismo presente al interior de su genoma. Existen varios trabajos que han profundizado en la ecología de *E. coli* más allá de la colección ECOR (Whittam y col., 1983; Whittam, 1989; Boyd y col., 1994; Souza y col., 1999; Gordon y Lee, 1999). La mayor parte coincide en que esta bacteria presenta una mayor diversidad que la sugerida con anterioridad, y que al parecer el grupo de las enterobacterias si posee una estructura ecológica (Gordon y Lee, 1999). Por lo tanto, *E. coli* no es tan clonal como se había supuesto. De hecho, se ha encontrado que la mayoría de las bacterias poseen un amplio espectro de estructura poblacional que va desde altamente clonales como *Salmonella* (Maynard-Smith y col., 1993) y *Borrelia* (Maynard-Smith y col., 2000), hasta casi panmícticas como *Helicobacter* (Go y col., 1996) y *Neisseria* (O'Rourke y Spratt, 1994).

En la literatura existen cuatro tipos de evidencia que han sido usados para demostrar la participación de la recombinación: 1) la presencia de desequilibrio de ligamiento entre alelos de diferentes loci dentro de una población (Dykhuizen y Green, 1986; Souza y col., 1992; Haubold y col., 1998); 2) la distribución de los sitios polimórficos en la secuencia de un gen (Maynard-Smith, 1992; Maynard-Smith y Smith, 1998); 3) la existencia de incongruencias filogenéticas entre genealogías de genes (Dykhuizen y Green, 1991; Lecointre y col., 1998; Brown y col., 2002); 4) la construcción de relaciones filogenéticas permitiendo que se construyan a manera de red y no necesariamente con estructura de árbol usando el test denominado Split-Decomposition (Huson, 1998; Reid y col., 2000; Brown y col., 2002).

5.3 Estructura epidémica en las poblaciones patógenas

Al investigar las relaciones entre las poblaciones de las cepas comensales y patógenas de distintas especies bacterianas que conviven en el mismo nicho ecológico, como puede ser el tracto digestivo de los mamíferos en el caso de *E. coli*, se ha observado que éstas últimas presentan lo que se denomina estructura epidémica (Maynard-Smith y col., 1993). Esta estructura se refiere a que ocasionalmente un genotipo patógeno es selectivamente favorecido incrementando su frecuencia en la población y desplazando a los demás, con la consecuencia de que se recupera el mismo genotipo en numerosas ocasiones

(Maynard-Smith y col., 1993). Existen abundantes ejemplos de bacterias patógenas que presentan ésta estructura poblacional como es el caso de *Neisseria* (Maynard-Smith y col., 1993) y la cepa de *E. coli* EDL933 y O157 Sakai serotipo O157:H7, responsable de numerosos brotes epidémicos en Estados Unidos y Japón, respectivamente. El descubrimiento de este tipo de estructura en las cepas patógenas ha generado cierta confusión en relación al paradigma clonal pues se ha sugerido que, el hecho de que existan poblaciones patógenas con estructura epidémica en donde se recupera un mismo serotipo, apoya fuertemente la hipótesis clonal para las poblaciones de *E. coli*. Sin embargo, este fenómeno se encontraría igualmente explicado si consideramos que *E. coli* posee estructura ecológica, como se había mencionado con anterioridad. Además, se ha propuesto que en las especies medianamente clonales (que parece ser el caso de *E. coli*), la recombinación dominará la evolución a largo plazo pero no previene el surgimiento rápido de las clonas epidémicas (Maynard-Smith y col., 2000).

Por otra parte, la mayoría de los estudios acerca de poblaciones patógenas en *E. coli* no realizan comparaciones entre genes relacionados al desarrollo de la patogénesis y genes no relacionados a patogénesis, en especial acerca de la contribución de los factores de virulencia a la historia de la especie y no sólo acerca de la diversificación de las cepas patógenas. Las comparaciones genéticas entre estas poblaciones ayudarán a explicar el origen y aparición de nuevas cepas patógenas e identificar las diferencias involucradas en su desarrollo. Asimismo, la evaluación de la participación de la selección y la mutación, nos demostrará cuales son las principales fuerzas evolutivas que moldean la historia de las cepas patógenas en comparación con las no patógenas.

5.4 Marcadores moleculares en el estudio de genética de poblaciones de E. coli

Durante la historia del estudio de la genética de poblaciones en *E. coli* se han utilizado diversos marcadores moleculares, en su mayoría de genes metabólicos, que han resultado ser útiles para entender las relaciones filogenéticas entre las distintas cepas así como para cuantificar la diversidad existente.

En el presente estudio utilizamos para el análisis poblacional cinco marcadores moleculares pertenecientes a genes metabólicos (*mdh* y *gapA*), transportadores (*putP*), genes que participan en el desarrollo de estructuras que promueven la adhesión al

hospedero denominadas fimbrias (*fimA*) y un gen que participa en el mecanismo de reparación del DNA (*mutS*). Estos marcadores cubren un amplio rango de actividades celulares en *E. coli* por lo que en su conjunto nos brindan la posibilidad de entender con mayor detalle la historia evolutiva de esta bacteria. Además, se usaron paralelamente tres genes pertenecientes a la isla patogénica LEE (*tir*, *eae*, *espB*) como marcadores de la virulencia, con el objetivo de llevar a acabo un estudio comparativo entre genes cromosomales y genes localizados en una isla patogénica, lo que cubriría un amplio espectro de la ecología de *E. coli*.

5.4.1 Malato deshidrogenasa (mdh)

Esta proteína cataliza la oxidación reversible del malato a oxalacetato (involucrados en el ciclo de Krebs). Es un homodímero y se encuentran bien identificados dos dominios funcionales, el dominio de unión al NAD^+ que comprende los amino ácidos 1 al 150, y el dominio catalítico que consiste en los amino ácidos 151-313. Esta enzima ha sido el marcador molecular más utilizado en los estudios de genética de poblaciones de *E. coli* y es uno de los marcadores responsables del establecimiento del paradigma clonal, debido a su bajo polimorfismo y tasas de sustitución molecular, además de que se han reconocido pocos haplotipos (Boyd y col., 1994). Los dos dominios funcionales de este gen poseen tasas de sustitución diferentes como se puede ver a continuación (Boyd y col., 1994):

	Dominio unión al NAD^+		Dominio Catalítico		Ambos dominios	
	dS	dN	dS	dN	dS	dN
mdh	1.88±0.65	0.09±0.09	5.53±1.18	0.20±0.15	18.53±1.61	0.15±0.09

A partir de estos resultados podemos observar que el dominio catalítico es más susceptible de aceptar sustituciones sinónimas en comparación con el de unión al NAD^+, lo que sugiere que al interior de este gen existen diferentes presiones selectivas.

5.4.2 Gliceraldehido 3-fosfato deshidrogenasa (gapA)

Esta proteína cataliza la formación del D-glicerol-3-fosfato durante el primer paso de la segunda fase de la glucólisis y su localización celular es en el citoplasma. Posee una región codificante de 330 aminoácidos a partir de la cual se han identificado dos dominios estructurales, el dominio de unión al $NADH^+$ (aminoácidos 5-148) y el catalítico (aminoácidos 149-313). Ambos dominios exhiben diferencias en cuanto a sus tasas de sustitución molecular (Nelson y col., 1991):

	Dominio unión al NAD^+		Dominio Catalítico		Ambos dominios	
	dS	dN	dS	dN	dS	dN
gapA	0.46±0.26	0.09±0.07	1.01±0.47	0.08±0.05	0.78±0.28	0.09±0.04

Se ha propuesto que el ancestro de *E.coli* adquirió este gen a partir de un eucarionte (Doolittle y col., 1990). Sin embargo, diversos estudios indican que los genes *gapA* de *Salmonella*, *E. coli* y *Klebsiella* poseen un ancestro en común (Nelson y col., 1991). Este gen ha sido ampliamente utilizado para la determinación de tiempos de divergencia entre *E. coli* y *Salmonella*, así como para sus homólogos eucariontes (Nelson y col., 1991). Otras evidencias sugieren que la baja diversidad que este gen posee se debe a un evento de cambio selectivo reciente (Guttman y Dykhuizen, 1994b).

5.4.3 Prolina-permeasa (putP)

Esta es una proteína transportadora que cataliza la transferencia activa dependiente de sodio de la L-prolina, que es degrada a glutamato para ser utilizada como fuente de carbono y nitrógeno; también es capaz de transportar litio de manera alternativa. Es una proteína transmembranal localizada principalmente en la cara interna en donde se pueden distinguir dos regiones, una corresponde a la que se encuentra completamente inmersa en la membrana y la segunda corresponde a una estructura de tipo asa y la cola del péptido. Al igual que para los marcadores moleculares descritos con anterioridad, se ha registrado

variabilidad en las tasas de sustitución molecular hacia dentro del gen (Nelson y Selander, 1992):

	Dominio transmembranal		Dominio asa y cola		Ambos dominios	
	dS	dN	dS	dN	dS	dN
putp	9.18±1.76	0.03±0.03	9.84±2.52	0.34±0.25	9.04±1.46	0.15±0.11

Los estudios comparativos de genética de poblaciones de este gen en conjunto con *gapA*, han indicado que existe baja recombinación intragénica, sugiriendo que este mecanismo es más importante de lo que se había supuesto con anterioridad. Además, las reconstrucciones filogenéticas con ambos genes demostraron ser congruentes con las propuestas utilizando otros marcadores como *mdh* (Nelson y Selander, 1992).

5.4.4 Fimbria tipo 1 cadena A (fimA)

Esta proteína es una fimbria cuya función principal es permitir la colonización del epitelio del hospedero, así que juegan un papel muy importante para el desarrollo de la patogénesis. Se compone de filamentos polares que salen de la superficie de la membrana de la bacteria con un tamaño aproximado de 0.5-1.5 micras y en números de 100 a 300 por célula. Los estudios de genética de poblaciones han demostrado que las tasas de sustitución sinónimas son mayores para este gen que para los marcadores metabólicos (dS- 0.157±0.025), mientras que las tasas no sinónimas (dN- 0.034±0.007) sí son equiparables a las de los genes metabólicos (Boyd y Hartl, 1998b). Otros estudios utilizando un mayor número de haplotipos han demostrado que el polimorfismo de este gen es más alto que el de los genes metabólicos. Además, la recombinación es más frecuente y diferencial en distintas regiones de la proteína y la diversidad del gen parece estar mantenida por selección diversificadora (Positiva del tipo Darwiniano) (Peek y col., 2001).

5.4.5 Proteína reparadora del DNA (mutS)

Este gen se encuentra involucrado en la reparación de errores en el apareamiento de las bases del DNA en el momento de la replicación, lo más probable es que lleve a cabo la

fase de reconocimiento del error, posee baja actividad de ATPasa. Ha sido involucrada en el proceso de la recombinación mitótica y meiótica en sus homólogos eucariontes y en la inhibición del intercambio genético ente especies (Yang, 2000). Homólogos de *mutS* han sido reconocidos tanto en procariontes como en eucariontes lo que nos habla de que son parte de un mecanismo celular muy conservado en los tres dominios. Las deficiencias en esta proteína interrumpen la reparación generando un fenotipo mutante con altas tasas de mutación y recombinación. Se ha propuesto que esto es lo que sucede en las cepas patógenas y provoca su rápida expansión y divergencia (LeClerc y col., 1996). Esta aparición de alelos mutantes en las poblaciones naturales implica que tienen un papel muy importante en la emergencia de nuevos patógenos (LeClerc y col., 1996; Li y col., 2003). Se ha propuesto que *mutS* ha sido transferido horizontalmente en múltiples ocasiones en distintos patógenos de *E. coli*, y que el sitio de su localización cromosómica se encuentra bajo recombinación constante ('hot-spot'), ya que sus reconstrucciones filogenéticas son incongruentes con las de otros marcadores como *mdh* (Brown y col., 2001). A continuación se señalan los sitios de localización dentro del genoma de *E. coli* de cada uno de los marcadores usados en el presente estudio:

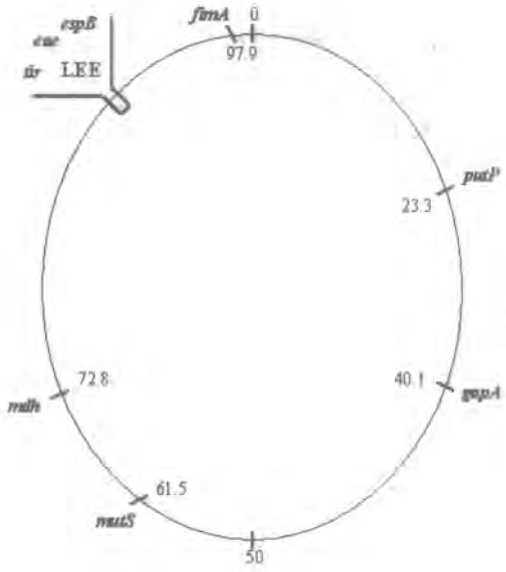


Fig. 5.1. Localización de los distintos marcadores en estudio dentro del genoma de *E. coli*.

5.5.Metodología

Para el análisis comparativo de genética de poblaciones se determinaron parámetros clásicos como es la diversidad representada tanto por π como por θ, ambos son dos parámetros poblacionales importantes que nos describen la distribución del polimorfismo en las poblaciones y funcionan bajo diferentes supuestos que en su conjunto nos permiten reconocer si los marcadores utilizados se encuentran bajo algún tipo de presión selectiva o bajo el modelo neutro (ver capítulo 3, prueba de Tajima). Además se determinó el contenido promedio de GC y para cada una de las posiciones codificantes (primera, segunda y tercera). De igual manera, se calculó el CAI y las tasas de sustitución molecular sinónimas (dS) y no sinónimas (dN), así como el cociente entre ambas (dN/dS). La determinación de estos parámetros se realizó bajo la misma metodología descrita en Castillo y col.(2005) (capítulo 4) y utilizando los mismos programas computacionales. A continuación se presentan los resultados preliminares obtenidos.

5.6 Resultados

En la tabla 5.1 se encuentran los resultados de la determinación de los principales parámetros evolutivos obtenidos a partir de las muestras completas de genes para cada marcador molecular que se encuentran descritas en el GenBank hasta marzo 2005 (ver Apéndice III). Como se puede observar esta muestra no es equivalente para cada marcador y refleja el sesgo que existe en los estudios de genética de poblaciones de *E. coli*. Los marcadores más pobremente representados son los que están relacionados con la virulencia y desarrollo de la patogénesis (*eae*, *tir* y *espB*) y el más representado es el gen metabólico *mdh*. Este sesgo limitó un poco las comparaciones entre los distintos marcadores pero no fue impedimento para la determinación de la determinación de las tasas de substitución molecular y el test de neutralidad (D de Tajima).

Genes	N	Sec. (bp)	H*	Sitios Cons.	Sitios polim.	π (±se)	θ (±se)	D Tajima (p)	%GC (±se)	%GC 1a pos (±se)	%GC 2a pos (±se)	%GC 3a pos (±se)	CAI	dS (±se)	dN (±se)	dN/dS	Sitios bajo selección/ adaptativa (+) o Purificadora (-) (p<0.1) (Bayes factor > 50)
putp	47	696	26	629	67	0.023 (0.0002)	0.021 (0.0009)	0.19 No sig	53.2 (0.005)	44.1 (0.011)	63.8 (0.186)	51.6 (0.052)	0.330 (0.0020)	0.096 (0.006)	0.0009 (0.0003)	0.0320	No + 43 -
gapA	67	663	20	508	155	0.013 (0.0006)	0.048 (0.0016)	-2.54 (p<0.001)	51.4 (0.041)	58.8 (0.021)	43 (0.016)	33.1 (0.041)	0.828 (0.0022)	0.051 (0.015)	0.0036 (0.0015)	0.1162	No + 50 -
mutS	29	453	19	341	111	0.009 (0.0003)	0.029 (0.0018)	-2.53 (p<0.001)	53.1 (0.131)	54.4 (0.158)	40.9 (0.078)	65.1 (0.388)	0.194 (0.0208)	0.014 (0.002)	0.0046 (0.0012)	0.0893	No + 47 -
mdh	98	825	50	576	249	0.020 (0.0005)	0.058 (0.0014)	-2.17 (p<0.01)	52.3 (0.079)	65.6 (0.034)	43.2 (0.011)	48.1 (0.0253)	0.573 (0.0021)	0.097 (0.024)	0.0068 (0.0006)	0.0572	1 + 123 -
fimA	52	555	34	422	133	0.068 (0.0004)	0.050 (0.0020)	1.23 No sig	51.4 (0.101)	59.07 (0.133)	51.2 (0.102)	44.05 (0.207)	0.415 (0.0026)	0.183 (0.012)	0.0373 (0.0023)	0.2378	3 + 43 -
tir	16	1707	14	910	794	0.203 (0.0054)	0.143 (0.0128)	0.65 No sig	48.9 (0.168)	59.8 (0.327)	50.8 (0.133)	36.3 (0.679)	0.222 (0.021)	0.618 (0.102)	0.1754 (0.0263)	0.3677	5 + 92 -
eae	32	2811	18	1673	1138	0.186 (0.0017)	0.109 (0.0074)	2.87 (p<0.01)	43.06 (0.122)	48.6 (0.129)	40.9 (0.124)	39.5 (0.208)	0.273 (0.0013)	0.317 (0.030)	0.146 (0.0137)	0.5166	14 + 95 -
espB	20	798	10	439	359	0.185 (0.0034)	0.126 (0.0096)	1.94 No sig	46.3 (0.060)	51.5 (0.143)	49.8 (0.639)	37.4 (0.530)	0.308 (0.0049)	0.646 (0.111)	0.1505 (0.0242)	0.3439	No + 45 -

Tabla 5.1. Diversidad genética, D de Tajima, contenido de GC, CAI, tasas de substitución molecular y sitios sobre selección para los 8 marcadores moleculares utilizados en el presente estudio.

*H- haplotipos

Es evidente a partir de la tabla anterior que existe una gran heterogeneidad en parámetros dentro de *E. coli*, cada marcador molecular posee características específicas y al parecer los procesos evolutivos que moldean a cada uno son diferentes. La tabla de resultados completos del estudio de selección con el soporte estadístico y señalando los codones para cada gen se describe ampliamente en el Apéndice III.

5.6.1 Análisis de diversidad y prueba de neutralidad

El análisis de diversidad arrojó resultados muy interesantes. Es evidente que existen diferencias entre los factorez de virulencia (*tir*, *eae* y *espB*) y los demás (Tabla 5.1 y Figura 5.1). Son mucho más diversos a pesar de que las muestras son substancialmente más pequeñas que para el resto de los marcadores. En especial el caso de la *mdh* que posee 98 secuencias descritas en el GenBank, y sin embargo, su diversidad es de 0.020 con 50 haplotipos identificados, lo que demuestra que la mayoría de las secuencias son casi idénticas, a pesar de que fueron obtenidas a partir de diferentes hospederos de *E. coli* (Apéndice III). El marcador que posee mayor diversidad (0.20) así como mayor número de haplotipos (14) por el total de la muestra es *tir*.

Podemos observar (Figura 5.2) que existe una gran diferencia entre los estimados de π y θ y sus patrones cambian de acuerdo al tipo de gen en estudio. Por ejemplo, para *mdh*, *gapA* y *mutS*, θ es mayor que π, lo que se ve reflejado en el resultado de la prueba de neutralidad (Tabla 5.1y Figura 5.2) donde la D aparece como negativa. Esto nos señala que la diversidad se está purgando y que existen mutaciones deletéreas. Este resultado es de esperarse sobre todo para estos genes que desarrollan funciones que se encuentran altamente conservadas como son procesos metabólicos y reparación del DNA. Para *putP* al parecer no se registra diferencia significativa entre ambos estimados, lo que lo hace ser un gen cercano a la neutralidad. En cambio, para los genes *fimA*, *tir*, *eae* y *espB* el patrón de diversidad cambia por completo, y π es mayor que θ, dando como resultado que la D sea positiva. Además, los tres marcadores de LEE son un orden de magnitud más diversos que el resto de los genes. Esto nos habla de que existe evidencia de algún tipo de selección diversificadora actuando sobre ellos. Más interesante aún, es que todos ellos son marcadores de virulencia, es decir, participan directamente en el proceso de desarrollo de la patogénesis.

5.6.2 Contenido de GC y CAI

El contenido de GC promedio del genoma de *E. coli* es de aproximadamente 50% (Blattner y col., 1997), podemos observar (Tabla 5.1) que en el caso de los distintos marcadores existe una distribución heterogénea, más notoria dentro de cada una de las posiciones codificantes. En general, la mayoría de los genes poseen una composición cercana al promedio total del genoma de *E. coli*, siendo *eae* (43.06%) el gen con menor contenido de GC.

Para el caso de cada una de las posiciones codificantes, tenemos diversidad de patrones. Para la primera posición el contenido de GC promedio en todos los genes es alto, mayor al promedio total. La segunda posición muestra una mayor variabilidad, en la mayoría de los casos es menor que la primera posición pero mayor que la tercera, las excepciones son *mdh* y *mutS*, donde el contenido de GC es menor en la segunda posición en comparación con las otras posiciones. La tercera posición es aún más variable, y tenemos que en la mayoría de los casos es menor que la segunda y la primera, siendo la excepción *mutS* cuyo contenido de GC en la tercera posición es muy alto (65.13%).

La distribución de contenidos de GC para las posiciones codificantes se encuentra ampliamente discutida en la literatura (Sharp y col., 1986; Shields y col., 1988; Wright, 1990 y Morton, 1993), sin embargo en la práctica los patrones registrados no parecen coincidir necesariamente con lo que se ha postulado. En el caso de la tercera posición se supone que es la que va a acumular con mayor probabilidad mutaciones y que tiende a asemejarse o parecerse al uso de codones óptimo del organismo en cuestión (Lawrence y Ochman, 1998), junto con la primera posición. En el caso de *E. coli* se esperaría que las terceras y primeras posiciones fueran las que tuviesen mayores contenidos de GC, alrededor del 50%, ya que este es el promedio registrado para esta bacteria. Es claro que ésto no es lo que se observa a partir de la muestra de marcadores que usamos. De hecho es la tercera posición la que registra los menores contenidos de GC en general. Se ha especulado que este fenómeno tiene que ver con eventos de regulación, pero no está del todo claro y aún existe discusión al respecto.

Por otra parte, en cuanto al índice de uso de codones que utilizamos aquí (CAI) (Figura 3) (Ikemura, 1985; Sharp y Li, 1987), de acuerdo a lo observado para las terceras

posiciones es de esperarse que exista un gran sesgo al uso óptimo que se encontraría cerca del 1 en la gráfica. El único gen que posee un uso de codones cercano al óptimo es *gapA* (0.828). Los demás genes en general se encuentran por debajo de este valor, en especial los genes pertenecientes a la isla LEE (*tir*,*eae* y *espB*). El caso especial es *mutS* que es el que mayor sesgo posee (0.194), estos datos apoyan la idea de que este gen es de origen extranjero como se ha sugerido en algunos trabajos (LeClerc y col., 1996).

5.6.3 Tasas de substitución molecular y prueba de selección

Las tasas de sustitución molecular se han utilizado recientemente para mostrar la participación de la selección (ver capítulo 3). Las tasas de sustitución sinónima son más variables y más altas que las sustituciones no sinónimas para el caso de los 8 marcadores. El gen que posee la tasa de sustitución no sinónima más baja es *putP* (0.0091), el que posee la más alta es *eae* (0.146) (Figura 5.3). El cociente de las tasas de sustitución (dN/dS) mostró que ninguno de los marcadores en promedio se encuentra bajo selección del tipo darwiniano, sin embargo esto no quiere decir que al interior no posean sitios bajo este tipo de selección (Figura 5.3). El cociente más bajo registrado es para *putP*, este es el marcador que más conservado se encuentra y que tiene un comportamiento cercano a la neutralidad lo que lo hace un buen candidato para resolver relaciones filogenéticas profundas. Otros marcadores útiles para registrar historia evolutiva son *mdh* y *gapA* cuyos cocientes de sustitución son también bajos. El caso de los genes relacionados a la patogénesis (*fimA* y los genes de LEE) es muy diferente estos son los que poseen el cociente de sustitución más alto, lo que nos habla de que al interior es posible encontrar sitios bajo selección darwiniana o la presencia de recombinación intragénica.

La prueba de selección mostró que existe diversidad en las tasas de sustitución dentro de los genes (ver tabla 5.1, Figura 5.3 y Apéndice III). El gen que posee más sitios bajo selección positiva del tipo darwiniano es *eae* (14 sitios), le sigue su receptor *tir* con 5 sitios, *fimA* con tres y *mdh* con uno. Estos sitios son importantes para reconocer regiones variables. En cuanto a la participación de la selección negativa *mdh* es el gen que posee un mayor número de sitios (123), le sigue *eae* con 95, *tir* con 92, y *gapA* con 50. Estos sitios son relevantes sobre todo para la conservación de la estructura.

5.6.4 Análisis de split-decomposition

Como prueba preliminar para registrar si existen eventos de relaciones no verticales dentro de los genes, utilizamos el programa SplitsTree (Huson, 1998) (Figuras 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 5.10 y 5.11). Este programa nos permite representar las relaciones genealógicas no necesariamente de manera vertical sino a manera de red. Nos parece importante este tipo de representación ya que en el presente estudio no es necesario obtener las reconstrucciones filogenéticas de los genes, es decir, no es el objetivo principal. Por el contrario nos interesa evaluar las distintas causas y consecuencias del proceso evolutivo en diferentes genes pertenecientes al genoma de un mismo organismo. Una forma de registrar esto es evaluar cual es la topología de las reconstrucciones genealógicas lo cual es posible con el programa de SplitsTree (Huson, 1998).

5.7 Discusión

El análisis comparativo de genética de poblaciones usando distintos marcadores nos permite tener un registro más amplio de los diferentes procesos evolutivos que ocurren al interior del genoma. Si el muestreo de genes es amplio y representativo de la especie, se pueden reconstruir con alta probabilidad estadística los procesos que moldean su historia. En el presente estudio se incluyeron marcadores moleculares que se encuentran localizados en diversas regiones del genoma de *E. coli*. Se incluyeron genes metabólicos, de transporte, de reparación del DNA y estructura, y se compararon detalladamente con los genes pertenecientes a la isla LEE.

Es interesante observar la diferencia que existe entre los genes relacionados a la patogénesis (*fimA*, *tir*, *eae* y *espB*) y los metabólicos (*mdh*, *gapA*, *putP* y *mutS*). Al evaluar el polimorfismo mediante los dos parámetros π y θ, vemos que existen diferencias significativas en los patrones de diversidad que están directamente relacionadas con procesos evolutivos diferenciales. La diversidad expresada por π se encuentra relacionada con el incremento de los alelos de mayor frecuencia y es independiente del tamaño de la muestra. Mientras que θ se ve más afectada por el tamaño poblacional y por alelos deletéreos. Al realizar la prueba de Tajima es aún más evidente la diferencia entre ambos estimadores (Figura 5.2). Claramente en los factores de virulencia existe un incremento de diversidad posiblemente originado a partir de recombinación, mayor frecuencia en la tasa

de mutación, y a que es posible que la selección de tipo darwiniano promueva estos cambios. Por otra parte, para los genes metabólicos la selección purificadora se encuentra purgando la diversidad. Este resultado en sí es importante ya que señala que los dos tipos de genes se encuentran sujetos a diferentes presiones selectivas dentro de las poblaciones de *E. coli*. Este fenómeno tiene repercusiones directas en la descripción de la estructura poblacional de las cepas patógenas y no patógenas de esta bacteria.

En cuanto al contenido de GC y uso de codones, sabemos que la posición que nos da información más relevante para el CAI es la tercera. En nuestro análisis podemos observar que la tercera posición es variable y no siempre cercana al contenido promedio del genoma de *E. coli*. La relación que podemos apreciar es que existe un fuerte sesgo en CAI cuando hay un fuerte sesgo en contenido de GC. Es decir, si la desviación del contendio de GC para la tercera posición en comparación con el promedio genómico es grande, tanto en un contenido menor como en contenido mucho mayor, se verá reflejado en su sesgo en el CAI. Este es el caso de genes como los pertenecientes a LEE, *gapA* cuyo sesgo es hacia un uso de codones casi óptimo y *mutS* para el cual el sesgo es aún mayor que para los factores de virulencia. Por otra parte, tenemos genes compo *putP* que es el que más se acerca al promedio tanto de CAI como contenido de GC, resultando ser el marcador más típico y parecido al promedio del genoma de *E. coli*.

El patrón de sustitución molecular indica que las tasas de sustitución sinónimas son mucho mayores a las no sinónimas en todos los casos. Sin embargo, para el caso de los factores de virulencia existe un incremento considerable en las tasas de sustitución no sinónimas. Este resultado puede explicarse si existen diferentes presiones de selección al interior de los genes. Este incremento puede deberse a que existe recombinación intragénica o al aumento de la tasa de mutación. Si nos concentramos específicamente en el caso de la adhesina (*eae*), vemos que hay un aumento en las tasas no sinónimas. Se sabe que este gen es recombinante y se han identificado cinco alelos diferentes (Tarr y Whittam, 2002). Lo más probable es que el incremento en dN refleja el aumento del polimorfismo en la región propiamente recombinante del gen. Para el caso de los demás genes relacionados a la patogénesis, no se ha registrado cuál es la participación de la recombinación en el desarrollo de nuevos alelos, los datos que aquí presentamos son preliminares pero sugieren que la

recombinación podría tener un papel importante en la diversificación de las cepas patógenas.

Las pruebas de selección confirman que dentro de los genes se encuentran regiones sujetas a presiones selectivas diferentes. El caso del gen metabólico *mdh* es interesante ya que posee un sitio bajo selección positiva lo que es sorprendente pues es un gen muy conservado, poco diverso y que parece estar sujeto en promedio a selección purificadora (posee 125 codones bajo selección negativa). Este es un claro ejemplo de que la selección de tipo darwiniano puede tener un papel importante incluso en genes cuyo promedio no parece indicarlo. Este resultado apoya la discusión acerca de las unidades de selección que pueden ser tan pequeñas como un sitio dentro del codón.

Por otra parte, como era de esperarse derivado de los análisis previos, los factores de virulencia son los que poseen un mayor número de sitios bajo selección positiva. Estos sitios son posibles candidatos a ser regiones recombinantes o con tasas de mutación elevadas. También observamos un gran número de sitios bajos selección purificadora que podrían estar relacionados con regiones estructurales importantes. Para los genes metabólicos (*mdh*, *gapA*) y transportadores (*putP*) es más evidente la participación de la selección purificadora, lo cual es congruente pues estos genes participan en funciones celulares esenciales y conservadas entre la mayoría de los organismos.

Nuevamente el caso de *mutS* merece una atención especial, de acuerdo a los resultados de las distintas pruebas, este gen lleva a cabo una función central en la reparación del DNA, lo que quiere decir que se encuentra altamente conservado. Los resultados de la prueba de selección apoyan estas ideas ya que posee 47 sitios bajo selección purificadora y ninguno bajo selección positiva. Sus tasas de sustición molecular son bajas y casi comparables al marcador *putP*. La diversidad observada no es mucho mayor que el promedio de los genes cromosomales y la prueba de Tajima parece indicar la presencia de selección purificadora. Sin embargo, el CAI parece señalar que existe un gran sesgo en el uso de codones que este gen posee, mucho mayor al observado para los genes de LEE. Esto es paradójico ya que es un gen que que realiza una función primordial y antigua dentro de la célula. Si posee un uso de codones tan diferente tendrá repercusiones durante su expresión. Para este gen se ha propuesto un origen extranjero, nuestros resultados apoyan esta hipótesis y hacen de este marcador un pésimo indicador de la historia filogenética profunda

en *E. coli* y un buen ejemplo de cómo la transferencia horizontal juega un papel primordial en la historia temprana de los procariontes. Se ha sugerido que algunas cepas patógenas poseen alteraciones en este gen lo que las convierte en hipermutantes. En el presente estudio no tenemos indicación de que este gen se encuentre alterado y de que las cepas posean un fenotipo mutante, al contrario parece encontrarse bien conservado y la única señal de su origen extranjero es el sesgo registrado en su uso de codones. Este tipo de sesgo no puede ser solamente explicado en términos de regulación, como para el caso de algunos de los genes de LEE (e.g. *eae*, *tir*).

El estudio de reconstrucción de las genealogías de los marcadores y de Split-decomposition arrojaron datos interesantes. Existe congruencia entre las relaciones genealógicas de cada marcador y sus cepas correspondientes y el análisis de Split-decomposition (Figuras 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 5.10 y 5.11). A pesar de que la señal filogenética se encuentra conservada, tenemos evidencias de relaciones no verticales al interior de los genes, lo que puede ser interpretado como una señal de recombinación intragénica, en todos los marcadores a excepción de *mdh*. Estos proceso pueden ser recientes y por eso no han afectado las relaciones filogenéticas profundas entre las cepas, tanto para los genes cromosomales como para los de LEE. Esto hace evidente que se ha subestimado el papel de la recombinación. De acuerdo a la integración de la pruebas realizadas a todos los genes, los marcadores moleculares que más información del tipo filogenético profundo en *E. coli* nos va a dar son los metabólicos *mdh* y *gapA*, y sobretodo el transportador *putP*.

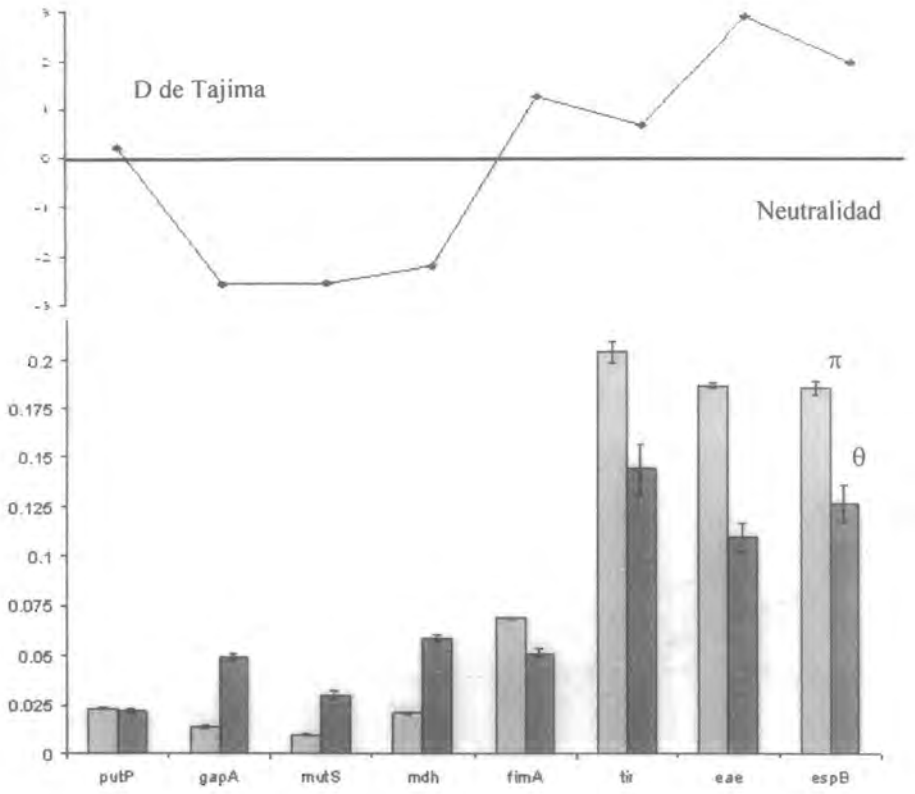


Figura 5.2. Distribución de la diversidad y D de Tajima para los ocho marcadores utilizados en este estudio. Las barras claras se refieren a la diversidad estimada por π y las barras obscuras a la representada por θ.

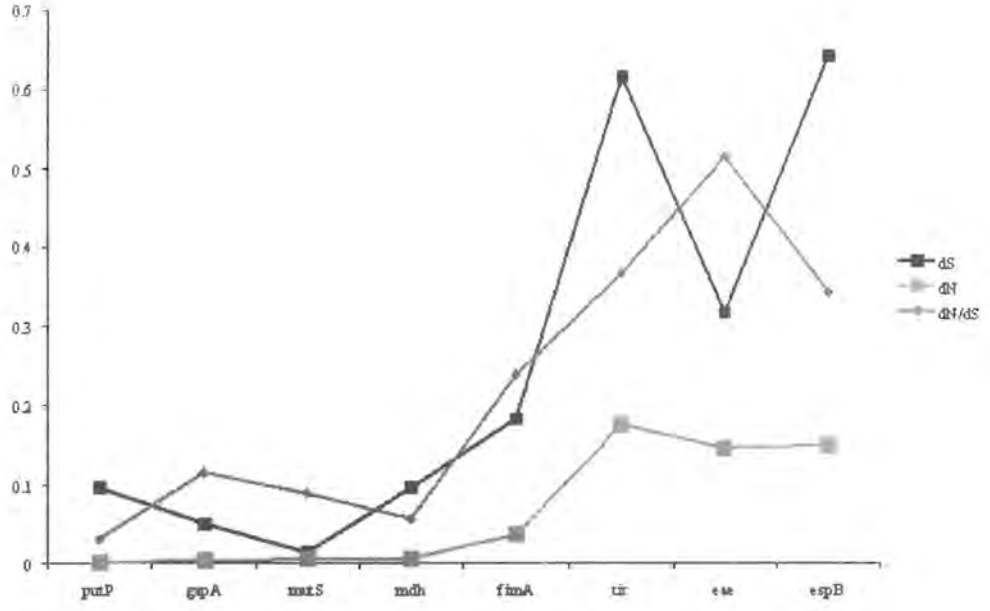


Figura 5.3. Distribución de las tasas de substitución molecular para los ocho marcadores.

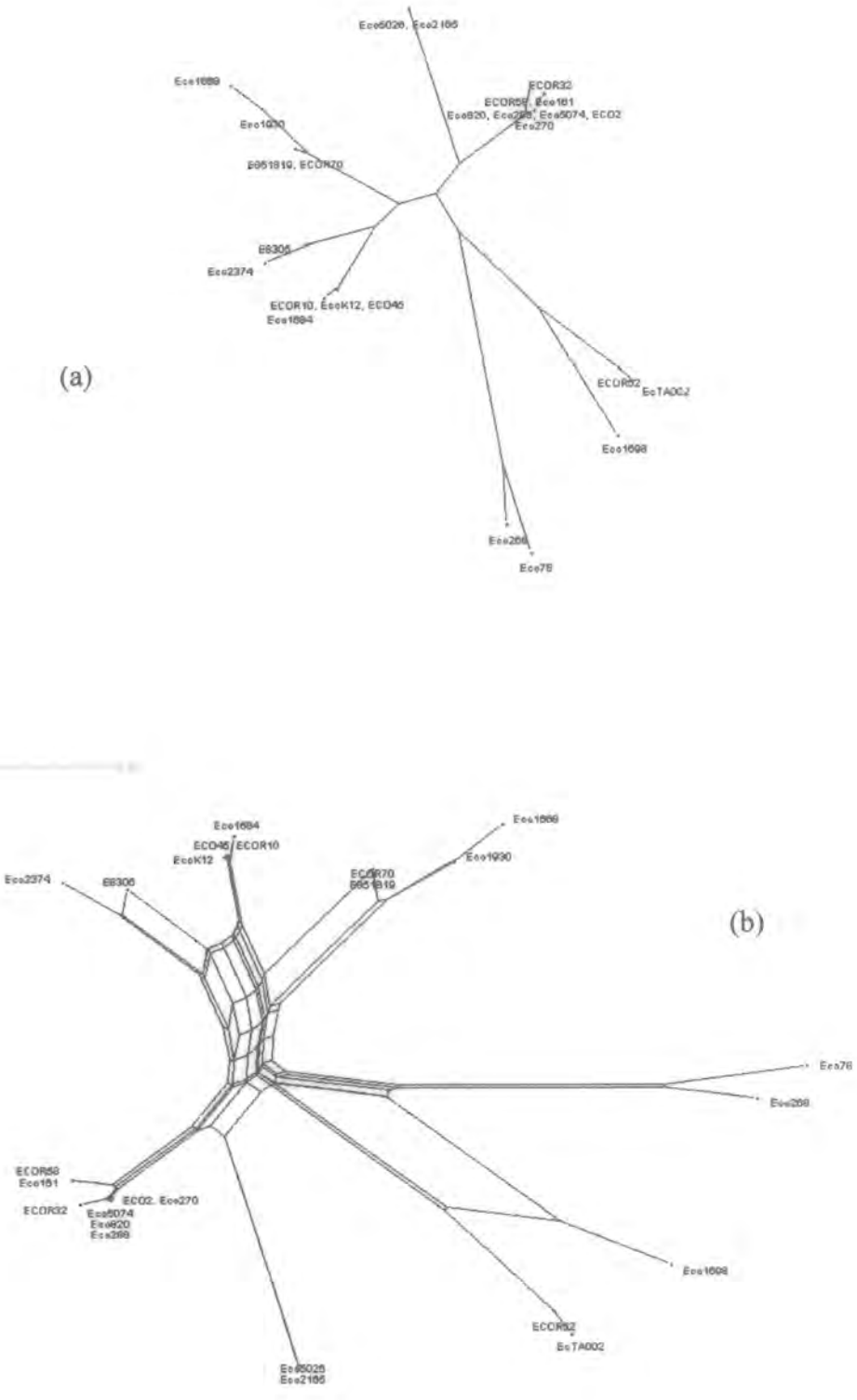
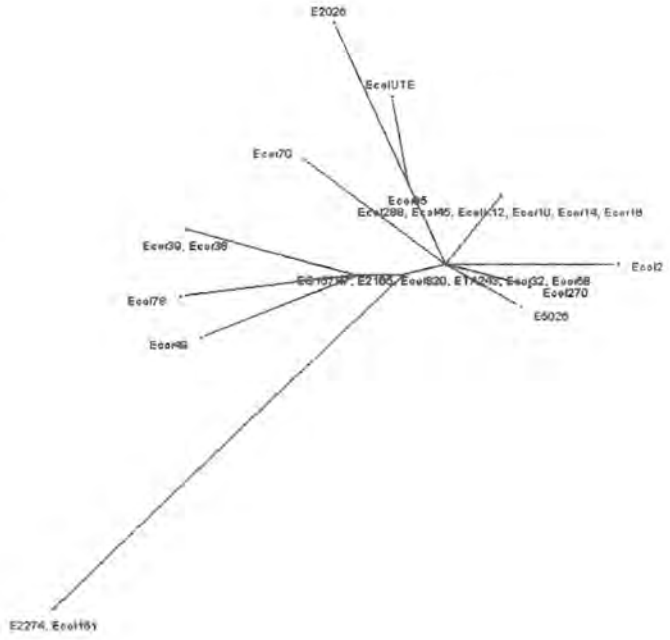


Figura 5.4. Análisis de NJ (a) y Split-decomposition (b) para el gen *putP*.

(a)



(b)

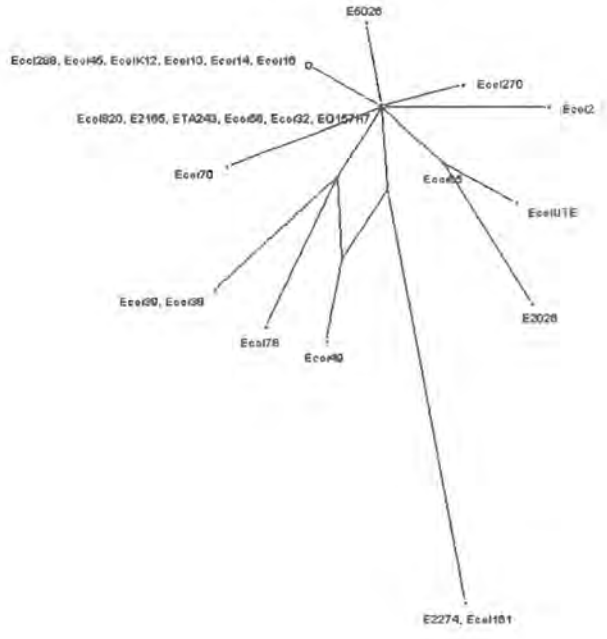


Figura 5.5. Análisis de NJ (a) y Split-decomposition (b) para el gen *gapA*.

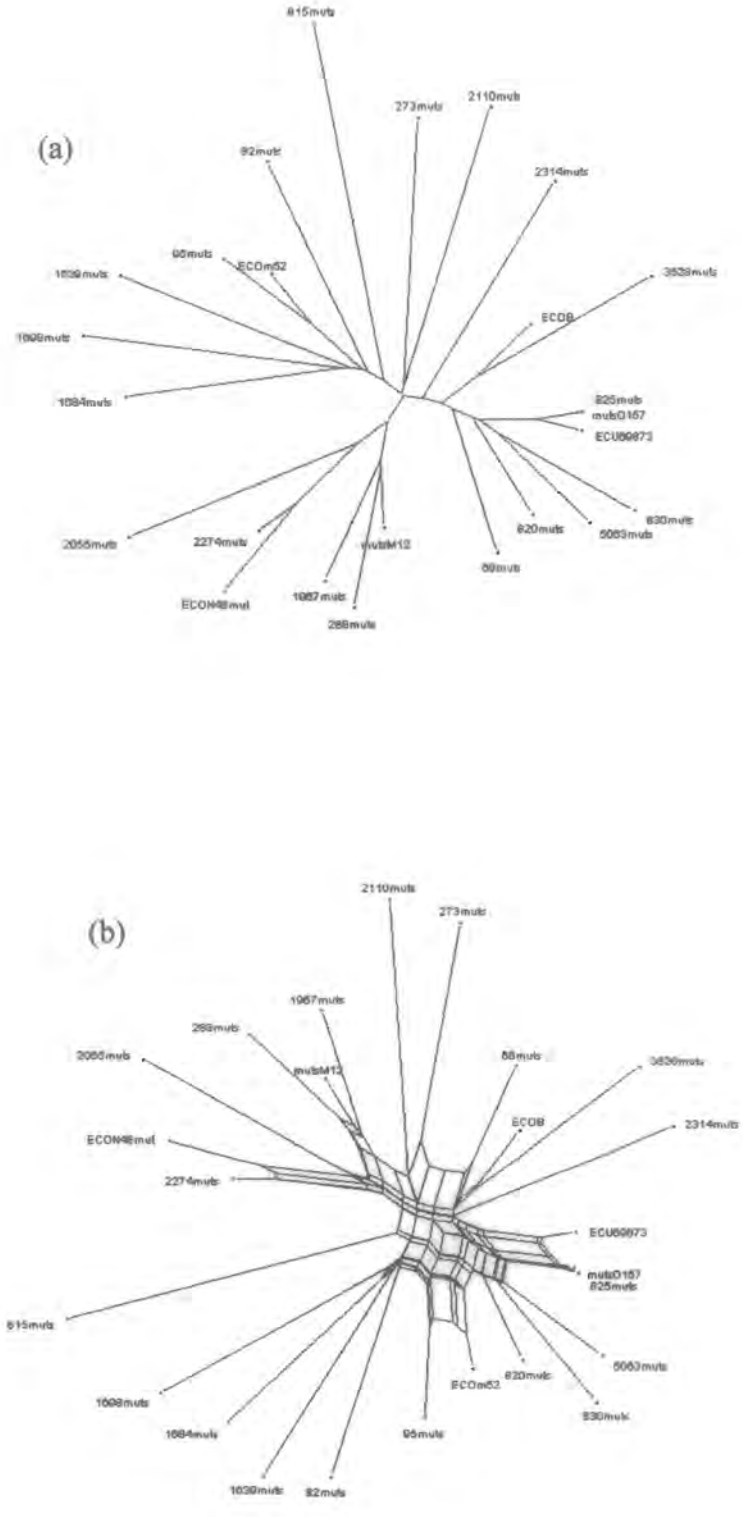


Figura 5.6. Análisis de NJ (a) y Split-decomposition (b) para el gen *mutS*.

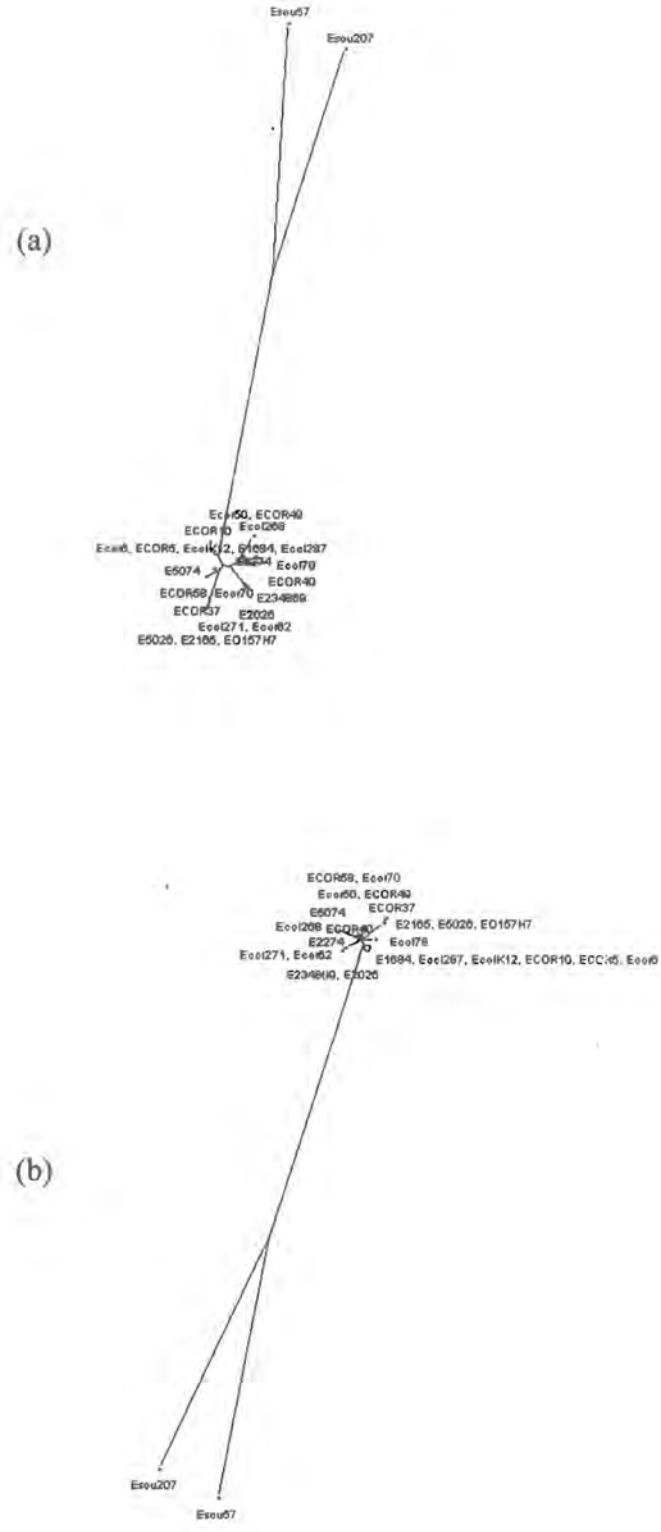


Figura 5.7. Análisis de NJ (a) y Split-decomposition (b) para el gen *mdh*.

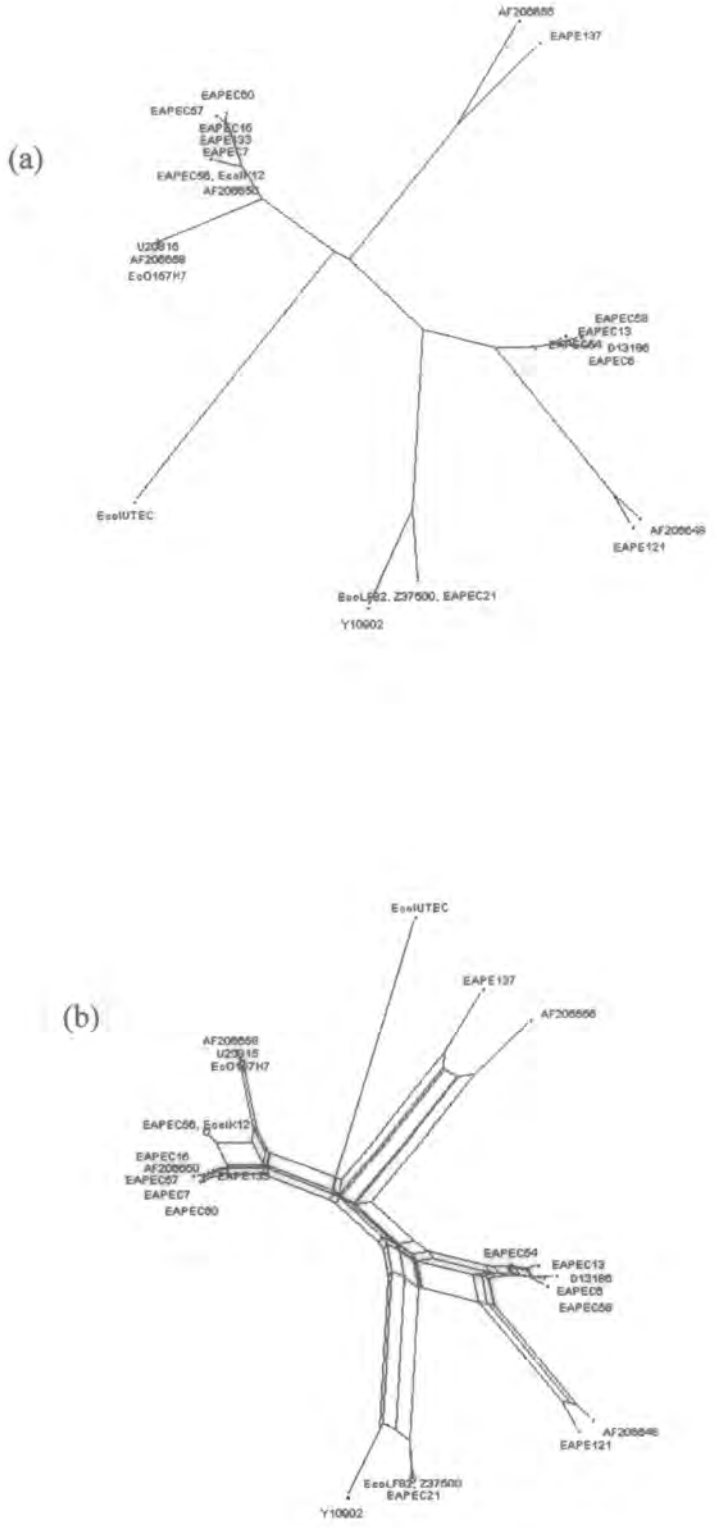


Figura 5.8. Análisis de NJ (a) y Split-decomposition (b) para el gen *fimA*.

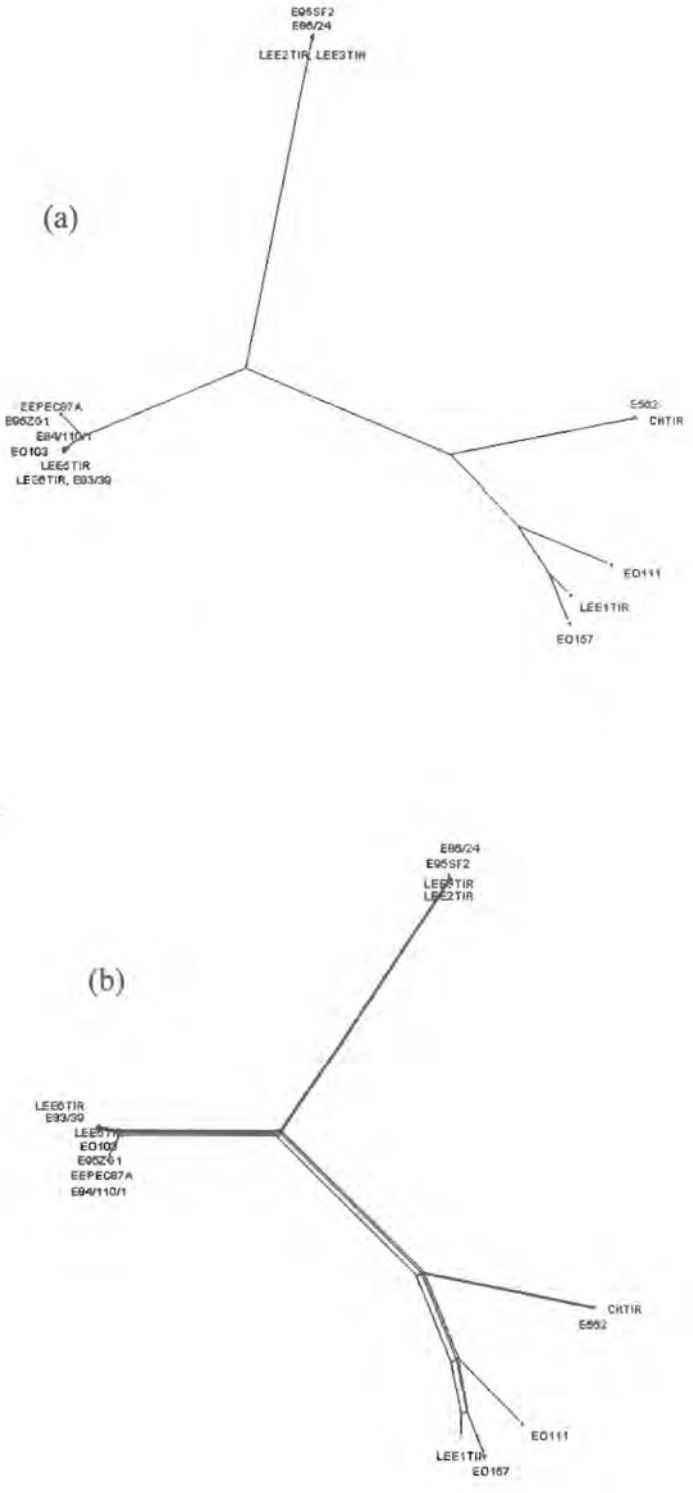


Figura 5.9. Análisis de NJ (a) y Split-decomposition (b) para el gen *tir*.

Figura 5.10. Análisis de NJ (a) y Split-decomposition (b) para el gen *eae*.

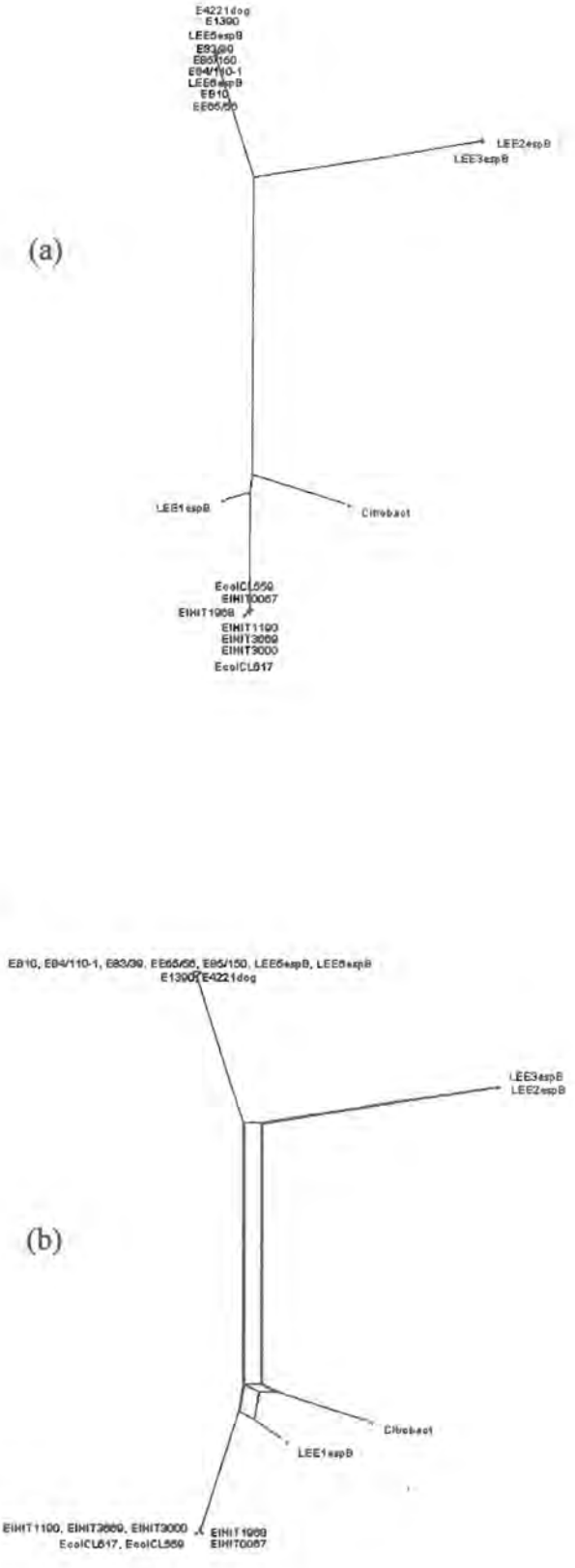


Figura 5.11. Análisis de NJ (a) y Split-decomposition (b) para el gen *espB*.

Capítulo 6. Discusión general y conclusiones

6.1 Genómica comparada de la isla patogénica LEE

La isla patogénica LEE ha sido objeto de numerosos estudios epidemiológicos y moleculares ya que es un excelente modelo para el estudio de la evolución y desarrollo de la patogénesis. Al parecer la sola adquisición de esta isla por parte de una cepa no patógena de *E. coli* la puede transformar en una patógena (Elliot y col., 1999). Desde el punto de vista evolutivo, LEE siempre ha sido considerada una unidad genética que ha sido transferida horizontalmente en múltiples ocasiones durante la evolución de los patógenos A/E. Estos resultados han sido fuertemente apoyados ya que se encuentran basados en análisis de cepas epidémicas clonales clínicas, como es el caso de la cepa O157:H7. Sin embargo, a partir de estudios en cepas silvestres de *E. coli* sabemos que el origen de LEE es más complejo y las evidencias sugieren que se encuentra en un proceso constante de construcción y disrupción (Sander y col., 2001).

A partir del estudio de genómica evolutiva comparada podemos observar que en las clonas epidémicas analizadas la estructura de la isla LEE se encuentra conservada y a pesar de ello registramos que existe una gran diversidad genética. El grupo de genes que presentan los niveles más bajos de variación genética y contenido de GC son los pertenecientes al SSTT que se encuentran conformando la mayor parte de los tres primeros operones. Por otra parte, genes como *map*, *tir*, *eae*, *espA*, *espD*, *espB* y *espF* presentan los niveles más altos de diversidad y contenido de GC más parecido a los genes típicos de *E. coli*.

La evaluación de la participación de la selección natural en la estructura de LEE arrojó datos interesantes. Debido a la diversidad genética presentada por algunos de los genes de LEE (*sepZ* , *tir* y *espB*, por ejemplo) se esperaba que algunos de ellos presentasen señales de selección positiva del tipo Darwiniano (SPD). Sin embargo, no se detectó SPD para ninguno de los miembros de la isla. Esto es cierto al avaluar el cociente de dN/dS para los genes haciendo un promedio de todos los sitios de la secuencia. Si hacemos este mismo estudio de manera detallada y hacia adentro de los genes, es posible detectar algunos sitios dentro de la secuencia que si poseen señales de SPD como *map*, *tir* y *eae*. Estos resultados reducen la participación de la SPD dentro de la estructura y dinámica evolutiva de LEE, sin

embargo, no quiere decir que este papel sea un papel menor. Los genes que presentan algunos sitios con SPD son genes clave dentro del desarrollo de la lesión A/E y son candidatos a estudios de mutación dirigida y terapias génicas. Por otra parte, es evidente el papel que la selección purificadora ha jugado en la estructura de los genes de LEE. La mayor parte de los genes de la isla poseen numerosos sitios bajo selección purificadora, en estos casos se ha purgado cualquier tipo de cambio y por lo tanto creemos que se ha preservado la estructura de regiones importantes de la proteína. Los genes con un número mayor de sitios afectados por selección purificadora son el SSTIII y *tir*. El SSTII en su conjunto forma una estructura importante para la transducción de señales y translocación de proteínas efectoras, por esta razón es que casi todos los cambios dentro de ellos alterarían un aparato que se ha mantenido conservado durante la historia de *E. coli*. El caso del receptor de la adhesina *tir* es especial ya que, además de ser el gen más representativo de la lesión A/E y único a las cepas patógenas de *E. coli* que desarrollan esta lesión, posee una estructura particular que le permite translocar a la adhesina y engañar a la célula del hospedero para que ésta a su vez lo presente y la adhesina se adhiera de la manera típica que los patógenos A/E hacen. Por lo tanto, es lógico suponer que cualquier alteración a la estructura de la proteína de *tir* se vería alterado todo el proceso de adherencia tan característico de estos patógenos.

La determinación de los sitios evolutivamente relevantes dentro de la secuencia es de suma importancia para la delimitación de regiones estructuralmente importantes, sobretodo en genes donde no se conoce la estructura de la proteína, como es el caso de la mayoría de los genes de LEE.

El estudio filogenético de LEE demuestra que existen al menos dos tipos de relación entre los genes que conforman la isla. La genealogía del SSTIII concuerda con la genealogía general propuesta para los patotipos principales de *E. coli*, en donde las cepas que pertenecen al grupo EHEC-1 son más divergentes, dentro de ellas podemos localizar a las cepas O157:H7. Mientras que las cepas pertenecientes al grupo EPEC-1 (la cepa EPEC E2348/69) se encuentran más conservadas y más relacionadas a las EHEC-1. Por otra parte, las cepas RDEC-1 y STEC 413/89-1 se agrupan dentro de la categoría de las cepas relacionadas al grupo EHEC-2. Finalmente el grupo exterior hermano de los patotipos de *E. coli* corresponde a *Citrobacter rodentium*. Este es el tipo de relación genealógica que se

había propuesto para la isla LEE en su totalidad. Sin embargo, al disectar las historias evolutivas de LEE gen por gen es evidente que esta propuesta no se sostiene, al menos no para todos los integrantes de la isla. Tal es el caso de los genes *map*, *tir*, *eae*, *espA*, *espD*, *espB* y *espF*, cuyas genealogías no corresponden a la genealogía general de los distintos patotipos de *E. coli*. En estos casos, el grupo EPEC-1 se encuentra más cercano a *C. rodentium* en lugar del grupo EHEC-1, además de que se observa una mayor divergencia entre las secuencias. Estos resultados pueden ser interpretados de dos maneras; la primera explicación, es que en los genes que poseen genealogías diferentes la recombinación se encuentra rompiendo el ligamiento y los genes están divergiendo de forma diferencial a los del núcleo de genes de LEE. Si este fuese el caso esperaríamos reconstruir diferentes historias para cada uno de los genes. La segunda explicación es que estos genes en su conjunto se originaron en un evento de transferencia diferente al de los genes del SSTIII. Si este fuese el caso esperaríamos obtener una genealogía consistente y compartida entre los genes divergentes. Esta última explicación parece ser la más cercana a lo que sucede en la realidad, ya que sólo se reconstruyen dos historias para los genes de LEE. Incluso para genes en donde ya se ha demostrado que la recombinación es la que está participando de manera principal en la generación de nuevos alelos, como es el caso de la adhesina (*eae*); la genealogía de la adhesina es congruente con la del resto de los genes divergentes. Estos resultados evidencian que LEE puede haberse originado a partir de dos eventos distintos de transferencia horizontal y por lo tanto a diferentes tiempos en la historia de *E. coli*. Para determinar con especificidad en que momento se originó LEE como isla patogénica, necesitamos usar marcadores que pertenezcan a ambos grupos de genes. Es decir, utilizar al SSTIII como el grupo de genes más antiguos y a los genes divergentes como *espB* o *tir* como marcadores de la historia reciente de la isla. Entre ambos grupos podremos determinar con mayor exactitud la integración de LEE y la aparición de los patógenos A/E. Además, el hecho de que genes clave en el desarrollo de la lesión A/E, como la adhesina y su receptor, sean más divergentes, menos conservados y con distintas genealogías, sostiene la hipótesis de que la virulencia es un estado derivado en la historia natural de *E. coli*. Es decir, *E. coli* no es un patógeno ancestral sino se hizo patógeno a lo largo de su evolución y como nicho alternativo al comensalismo o la vida libre. Por lo tanto, la patogénesis se adquirió con la integración paralela de elementos de virulencia como el SSTIII en un

principio, y se completó el escenario molecular para la aparición de un nuevo patógeno con la integración de otros elementos moleculares como la invasina y proteínas secretoras.

Los resultados del presente trabajo determinan que la estructura genética de la isla es un mosaico compuesto por al menos dos grupos de genes diferentes en origen y dinámica evolutiva. Lo que quiere decir que LEE se formó a partir de al menos dos eventos diferentes de transferencia de genes. El primer evento fue la adquisición de un sistema de secreción tipo III, que es el grupo de genes más antiguo dentro de la isla LEE. El segundo evento lo conforman la adquisición de proteínas secretoras (*espB*, por ejemplo) y algunos genes como la adhesina intimina (*eae*). Sin embargo, también existieron eventos de generación de genes *de novo*, como se sugiere para el caso del receptor de la adhesina (*tir*), que es un gen que no posee homólogos en otras especies bacterianas y es único a *E. coli* patógena.

La genómica comparada contribuye de manera significativa al mejor entendimiento de la evolución bacteriana y el desarrollo de la patogénesis. El gran incremento en la disponibilidad de información que se obtiene a partir de la secuenciación de los genomas ha mostrado que los genomas bacterianos son dinámicos y evolucionan constantemente. La adquisición de nuevas capacidades es un factor relevante para la colonización de nuevos nichos ecológicos y especialmente para el desarrollo de nuevos patógenos.

6.2 Análisis comparativo de genética de poblaciones entre una muestra de genes relacionados a la patogénesis y genes metabólicos de E. coli: ¿ ser o no ser patógeno?

El análisis comparativo entre genes relacionados a la patogénesis y marcadores metabólicos nos ha mostrado que ambos grupos de genes se encuentran bajo presiones selectivas diferentes. La diversidad es mayor para los factores de virulencia, lo cual puede deberse a que son de reciente origen y que están sometidos a distintas presiones selectivas que fomentan su variación para promover su rápida propagación o colonización a nuevos ambientes. El caso más representativo es *tir*, que es el receptor de *eae* (adhesina). Este gen parece estar cambiando en respuesta a la adaptación a nuevos hospederos. Por otra parte, se sabe que la variedad de alelos que se han encontrado para *eae* son producto de la recombinación, lo que ha generado la aparición de nuevos patotipos de patógenos A/E. Estos resultados apoyan la idea de que los genes que participan en el desarrollo de la patogénesis son de aparición reciente y poseen tasas de substitución molecular más rápidas

que los genes cromosomales que no participan en el desarrollo de la patogénesis. Estos resultados indican que los factores de virulencia han contribuido al desarrollo de nuevas capacidades metabólicas que *E. coli* ancestralmente no poseía y han sido factor fundamental en la colonización de nuevos nichos y el desarrollo de nuevos patógenos. Al ser un estado reciente en *E. coli*, la radiación hacia el desarrollo de nuevos patógenos es un proceso dinámico y parece encontrarse en continuo movimiento ya que muchos de los genes determinantes de la virulencia se encuentran presentes en elementos móviles como es el caso de las islas patogénicas. Es importante registrar la dinámica y estructura genética de estos genes ya que terapias de antibióticos puede generar presiones selectivas que pueden generar cambios rápidos y switch selectivos. Como se demuestra en el presente trabajo, los factores de virulencia han cambiado en respuesta a nuevos retos ambientales y parece que la SPD es muy importante en este proceso al igual que la recombinación.

El caso de los genes cromosomales es mejor conocido ya que se han utilizado como marcadores filogenéticos de la historia evolutiva de *E. coli*. Estos genes que pertenecen al núcleo conservado de genes de *E. coli* como especie también registran distintas presiones selectivas al interior y parece ser que la recombinación ha jugado un papel importante durante su historia. Al ser genes más antiguos, sus tasas de sustitución molecular son más lentas y pueden registrar la historia profunda de *E. coli*. Sin embargo, hay casos excepcionales como *mutS*, que a pesar de llevar a cabo una función muy conservada parece ser un gen extranjero en esta bacteria. Consideramos que es necesario hacer estudios evolutivos detallados acerca de la dinámica molecular del marcador en estudio. Esto nos llevará a identificar sus características moleculares específicas y de esta manera podremos seleccionar un marcador adecuado al tipo de estudio que pretendemos realizar. Por ejemplo, si queremos registrar la historia filogenética, los marcadores útiles serán los metabólicos *mdh* y *gapA*, así como el transportador *putP*. Mientras que si queremos registrar la dinámica reciente de cepas patógenas tendremos que utilizar genes como *fimA*. En cambio, para evaluar la dinámica de eventos de transferencia horizontal será necesario registrar la historia de genes asociados a ellos como son las islas patogénicas. Esta es la única forma de registrar la historia de la evolución y dinámica de *E. coli*.

La aparición de nuevas cepas patógenas es un proceso dinámico en donde interviene un bagaje genético predeterminado y la adquisición de nuevos factores de virulencia por

transferencia horizontal. El establecimiento de estos factores dentro del genoma se debe a que las presiones ambientales de colonización a nuevos nichos y la presencia en el ambiente genético de estos factores se conjuntan. Ser o no ser patógeno en *E. coli* parece depender del azar y la necesidad, tal como lo describió Monod. Estas ideas y los resultados del presente trabajo apoyan fuertemente el papel del azar y los supuestos de la teoría neutra, dando un rol relevante a procesos aleatorios, la deriva génica y la selección purificadora dentro de la evolución molecular de los procariontes. Mientras que la SPD en un promedio genómico parece jugar un papel menor.

6.3 El paradigma clonal

El establecimiento del paradigma clonal es resultado de análisis evolutivos realizados con genes metabólicos como *mdh*. En el caso específico de este marcador, la recombinación parece no tener un papel importante en su estructura, es altamente conservado y se tienen poco haplotipos. Consideramos que estas características y su intepretación llevaron a la concepción errónea de que la recombinación y la SPD no tienen un papel importante dentro de la estructura de las poblaciones de *E. coli*. En el presente trabajo podemos observar que no necesariamente es el caso. La determinación de la recombinación es difícil pues los métodos basados en incongruencias filogenéticas parecen estar perdiendo resolución a eventos de recombinación recientes, por lo que es necesario el desarrollo de nuevas metodologías y su combinación para detectar este proceso adecuadamente. De lo contrario estaremos perdiendo información valiosa acerca de la dinámica evolutiva de los genes y sobretodo haremos suposiciones acerca de las estructuras poblacionales basadas en un solo marcador. En la actualidad cada vez es más claro que el paradigma clonal no existe en *E. coli* pues la dinámica evolutiva de los genomas procariontes ha demostrado ser mucho más compleja de lo que se suponía con anterioridad. Tenemos que ser cuidadosos en la interpretación de los datos moleculares y su extrapolación hacia todo el genoma o la especie, pues no necesariamente registra la historia evolutiva en su totalidad. Es necesario cubrir un amplio rango de actividades celulares para tener una mejor idea de quien es *E. coli* como especie, ya que no sólo tenemos que hablar de que es un comensal importante sino un patógeno importante. A su vez, se deben ampliar

los estudios con cepas de vida libre no asociadas a un hospedero para tener el rango completo de los nichos ecológicos que esta bacteria posee y tener un mejor registro de su plasticidad genómica.

6.4 Las unidades de selección

La discusión acerca de las unidades de selección ha sido más de tipo filosófico que empírico. El presente trabajo es una primera aproximación empírica acerca de la mínima unidad de selección que después de muchos debates parecía ser el gen según Richard Dawkins y otros autores. Si un organismo es clonal su unidad mínima de selección es su genoma completo. Mientras menos clonal sea este organismo las unidades de selección a su interior serán también menores. Es decir, la especie es una unidad de selección, el genoma lo es, las islas genómicas lo son, los operones, cada uno de los genes, pero también dominios internos hasta llegar al mínimo que parece ser el codón. Estas ideas concuerdan con la propuesta para el origen de los primeros genes y la estructura de los dominios de las proteínas actuales (Gilbert y col., 1997). El proceso selectivo parece generar unidades a distintos niveles en analogía a una muñeca rusa, donde cada muñequita es una unidad en sí misma.

6.5 Conclusiones

1. La estructura genética de la isla LEE es un mosaico.
2. El origen de LEE es producto de al menos dos eventos de transferencia horizontal.
3. La selección positiva de tipo Darwiniano tiene un papel importante en algunos sitios de genes como *map*, *tir* y *eae*.
4. La patogénesis es un estado derivado en *E.coli*.
5. La recombinación parece estar formando nuevas variedades de patógenos A/E.
6. El SSTIII es el grupo de genes más antiguo y son los que guardan la señal filogenética de la formación temprana de la isla.
7. El gen *tir* es el más representativo de los patógenos A/E y sirve junto con las proteína secretoras para registrar la diversificación de los patógenos A/E, y para datar la formación inicial de la isla.
8. LEE no es una unidad evolutiva en *E. coli*.

9. LEE sólo se mantiene como unidad en clonas epidémicas.
10. La isla no es la unidad de selección, ni los operones o los genes, al parecer existen unidades de selección más pequeñas al interior de los genes.
11. La definición de la unidad de selección nos brinda la oportunidad de registrar cual es la estructura genética de la unidad que nos encontremos estudiando. En el caso de LEE, nos muestra que el mosaico observado se encuentra generado a varios niveles y desde el interior de algunos de sus genes.
12. Los factores de virulencia poseen mayor diversidad genética que los no relacionados a la patogénesis.
13. Existen presiones selectivas diferentes para cada uno de los ocho marcadores usados en el estudio y hacia el interior de sus genes.
14. Existen señales de eventos de recombinación tanto en genes metabólicos como en los factores de virulencia.
15. El paradigma clonal no se mantiene para el caso de *E. coli*.

6.4 PERSPECTIVAS

a) Ampliación de la colección de genes patógenos que se encuentra descrita en la literatura. Se necesita hacer una mayor representación de otros genes no necesariamente metabólicos que representen un mayor rango de la ecología de *E. coli* y de sus funciones celulares.
b) Realizar un par de pruebas de recombinación basándonos en la distribución de los sitios polimórficos en las secuencias de los marcadores y con algún método de coalescencia para, en su conjunto, describir claramente la participación de la recombinación en la estructura de los genes de *E. coli*. De esta manera podríamos cuantificar cual es la aportación de la recombinación a la historia evolutiva de la especie. Esto último tendrá repercusiones en el paradigma clonal de las poblaciones de *E.coli*.

c) Realizar un estudio de genómica comparada equiparable al realizado para la isla LEE pero con una isla metabólica. Esto con el objetivo de registrar cuáles son las unidades mínimas de selección en *E. coli*.

d) Extrapolar los resultados de las pruebas de selección a los sitios importantes dentro de las estructuras de las proteínas que se encuentren descritas en la literatura. Estos sitios son relevantes para estudios de mutación dirigida, terapias génicas y diseño de antibióticos.

REFERENCIAS

Baric RS, Yount B, Hensley L, Peel SA y Chen W. 1997. Episodic evolution mediates interspecific transfer of a murine coronavirus. *J Virol* 71: 1946-1955.

Bargelloni L, Marcato S, Patarnello T. 1998. Antarctic fish hemoglobins: evidence for adaptive evolution at subzero temperatures. **PNAS** 95: 8670-8675.

Bettelheim KA. 1994. Biochemical characteristics of *Escherichia coli*. En Gyles CL (Ed.). *Escherichia coli* in domestic animals and humans. CAB International. Wallingford, UK, 3-30p.

Bishop JA y Cook LM. 1975. Moth, melanism and clean air. *Sci Am* 232(1): 90-99.

Bishop JG, Dean AM y Mitchell-Olds T. 2000. Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. *PNAS* 97: 5322-5327.

Blattner FR, Plunkett 3rd G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B y Shao Y. 1997. The complete genome sequence of *Escherichia coli* K12. *Science* 277: 1453-1462.

Blum G, Ott M, Lischewski A, Ritter A, Imrich H, Tschäpe H, and Hacker J. 1994. Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an E. coli wild-type pathogen. *Infect Immun* 62: 606-614.

Boyce TG, Swerdlow DL y Griffin PK. 1995. *Escherichia coli* O157:H7 and the hemolytic-uremic syndrome. *N Engl J Med* 333(6): 364-8.

Boyd EF y Hartl D. 1998. Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution. *J Bacteriol* 180: 1159-1165.

Boyd EF y Hartl D. 1998b. Diversifying selection governs sequence polymorphism in the major adhesins proteins *fimA*, *papA*, and *sfaA* of *Escherichia coli*. *J Mol Evol* 47: 258-267.

Boyd EF, Nelson K, Wang FS, Whittam TS y Selander RK. 1994. Molecular genetic basis of allelic polymorphism in malate dehydrogenase (*mdh*) in natural populations of *Escherichia coli* and *Salmonella enterica*. *PNAS* 91: 1280-1284.

Brown EW, Kotewicz ML y Cebula TA. 2002. Detection of recombination among *Salmonella enterica* strains using the incongruence lenght difference test. *Mol Phylo Evol* 24: 102-120.

Brown EW, LeClerc JE, Baoguang L, Payne WL y Cebula TA. 2001. Phylogenetic evidence for horizontal transfer of *mutS* alleles among naturally occurring *Escherichia coli* strains. *J Bacteriol* 183(5): 1631-1644.

Burland V, Shao Y, Perna NT, Plunkett G, Sofía HJ y Blattner FR. 1998. The complete DNA sequence and analysis of the large virulence plasmid of *E. coli* O157:H7. *Nucl Acid Res* 26 (18): 4196-4204.

Clarke SC, Haigh RD, Freestone PPE y Williams PH. 2003. *Clin Microbiol Rev* 16: 365-378.

Cohan, FM. 1994. Genetic exchange and evolutionary divergence in prokaryotes. *Trends Ecol Evol* 9: 175-180.

Cohan FM. 1996. The role of genetic exchange in bacterial evolution. *ASM News* 62 (12): 631-636.

Comeron JM. 1995. A method for estimating the numbers of synonymous and non synonymous substitutions per site. *J Mol Evol* 41: 1152-1159.

Creasey EA, Friedberg D, Shaw RK, Umanski T, Knutton S, Rosenshine I y Frankel G. 2003. *Microbiol* 149: 3639-3647.

Deng W, Li Y, Vallance BA y Finlay BB. 2001. Locus of enterocyte effacement from *Citrobacter rodentium*: Sequence analysis and evidence of horizontal transfer among attaching and effacing pathogens. *Infect Immun* 69 (10): 6323-6335.

Deng W, Puente JL, Gruenheid S, Li Y, Vallance BA, Vázquez A, Barba J, Ibarra JA, O´Donell P, Metalnikov P, Ashman K, Lee S, Goode D, Pawson T y Finlay BB. 2004. *PNAS* 101: 3597-3602.

Doolittle RF, Feng DF, Anderson KL y Alberro MR. 1990. A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. *J Mol Evol* 31(5): 383-388.

Donnenberg MS y Kaper JB. 1992. Enteropathogenic *Escherichia coli*. *Infect Inmun* 60: 3953-3961.

Donnenberg MS y Whittam TS. 2001. Pathogenesis and evolution of virulence in enteropathogenic and enterohemorrhagic *Escherichia coli*. ***J Clin Invest*** 107 (5): 539-547.

Donnenberg MS, Zhang HZ y Stone KD. 1997. Biogenesis of the bundle forming pilus of enteropathogenic *Escherichia coli*: reconstitution of fimbriae in recombinant *E. coli* and role of *dsbA* in pilin stabiltiy, a review. *Gene* 192(1): 33-38.

Duda TF y Palumbi SR. 1999. Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. *PNAS* 96: 6820-6823.

Dykhuizen DE y Green L. 1986. DNA sequence variation, DNA phylogeny, and recombination in *Escherichia coli*. *Genetics* 113: S71.

Dykhuizen DE y Green L. 1991. Recombination in *Escherichia coli* and the definition of biological species. *J Bacteriol* 173: 7257-7269.

Efron B y Tibshirani RJ. 1993. **An introduction to the bootstrap**. Chapman y Hall, New york, USA.

Eguiarte LE. 1999. Una guía para principiantes a la genética de poblaciones. En : J. Núñez Farfán y L.E. Eguiarte (eds.). La evolución biológica. México, D.F., UNAM. 35-50pp.

Elliot SJ, Yu J y Kaper JB. 1999. The cloned locus of enterocyte effacement from enterohemorrhagic *Escherichia coli* O157:H7 is unable to confer the attaching and effacing phenotype upon *E. coli* K-12. *Infect Immun* 67 (8): 4260-4263.

Endo T, Ikeo K, Gojobori T. 1996. Large-scale search for genes on which positive selection may operate. *Mol Biol Evol* 13: 685-690.

Feil EJ y Spratt BG. 2001. Recombination and the population structrures of bacterial pathogens. *Annu Rev Microbiol* 55: 561-590.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17: 368-376.

Felsenstein J. 1993. Phylogeny inference package (PHYLIP).Version 3.5. University of Washington, Seattle.

Fitch WM, Bush RM, Bender CA, Cox NJ. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. *PNAS* 94: 7712-7718.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, McKenney K, Sutton G, Fitzhugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu LI, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom ME, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrman JL, Geoghanen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO y Venter JC. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496-512.

Ford MJ, Thornton PJ y Park LK. 1999. Natural selection promotes divergence of transferrin among salmonid species. *Mol Ecol* 8: 1055-1061.

Frankel G, Candy DCA, Everest P and Dougan G. 1994. Characterization of the C-terminal domains of intimin-like proteins of enteropathogenic and enterohemorrhagic *Escherichia coli*, *Citrobacter freundii*, and *Hafnia alvei*. *Infect Inmun* 62: 1835-1842.

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

Futuyma D. 1986. Evolutionary Biology. 2a. ed. Sinahuer Associates Inc. Sunderland. Massachusetts, USA.

Gilbert W, deSouza S y Long M. 1997. Origin of genes. *PNAS* 94: 7698-7703.

Go MF, Kapur V, Graham DY y Musser JM. 1996. Population genetic analysis of *Helicobacter pylori* by multilocus enzyme electrophoresis-extensive allelic diversity and recombinational population structure. *J Bacteriol* 178: 3934-3938.

Gogarten JP, Doolittle WF y Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* 19: 2226-2238.

Goldman N y Yang Z. 1994. A codon-based model of nucleotide substitution for protein coding DNA sequences. *Mol Biol Evol* 11(5): 725-736.

Gómez-Duarte OG y Kaper JB. 1995. A plasmid-encoded regulatory region activates chromosomal *eaeA* expression in enteropathogenic *Escherichia coli*. *Infect Inmun* 63: 1767-1776.

Goodwin RL, Baumann H y Berger FG. 1996. Patterns of divergence during evolution of a_1-proteinase inhibitors in mammals. *Mol Biol Evol* 13: 346-358.

Gordon DM y Lee J. 1999. The genetic structure of enteric bacteria from Australian mammals. *Microbiol* 145: 2673-2682.

Grant PR. 1986. Ecology and evolution of Darwin's finches. Princeton University Press, Princeton, New Jersey, USA.

Groisman AG y Ochman H. 1996. Pathogenicity islands: bacterial evolution in quantum leaps. *Cell* 87: 791-794.

Guttman DS. 1997. Recombination and clonality in populations of *Escherichia coli*. *Trends Ecol Evol* 12: 16-22.

Guttman DS y Dykhuizen DE. 1994. Clonal divergenece in *Escherichia coli* as a result of recombination, not mutation. *Science* 266: 1380-1383.

Guttman DS y Dykhuizen DE. 1994b. Detecting selective sweeps in naturally occurring *Escherichia coli*. *Genetics* 138: 993-1003.

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

Hacker J, Blum-Oehler G, Muhldorfer I y Tschape H. 1997. Pathogenicity islands of virulent bacteria: structure, function and impact on mirobial evolution. *Mol Microbiol* 23(6): 1089-1097.

Hartland EL, Batchelor M, Delahay RM, Hale C, Matthews S, Dougan G, Knutton S, Connerton I y Frankel G. 1999. Binding of intimin from enteropathogenic *Escherichia coli* to *tir* and to host cells. *Mol Microbiol* 32(1): 151-158.

Haubold B, Travisano M, Rainey PB y Hudson RR. 1998. Detecting linkage disequilibrium in bacterial populations. *Genetics* 150: 1341-1348.

Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, and Shinagawa H. 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K12. *DNA Res* 8: 11-22.

Haydon DT, Bastos AD, Knowles NJ y Samuel AR. 2001. Evidence for positive selection in Foot-and-Mouth disease virus capsid genes from field isolates. *Genetics* 157: 7-15.

Hentschel U, Steintert M y Hacker J. 2000. Common mechanisms of symbiosis and pathogenesis. *Trends Microbiol* 228: 226-231.

Hentschel U y Hacker J. 2001. Pathogenicity islands: the tip of the iceberg. *Microbes and Infection* 3: 545-548.

Hudson RR, Kreitman M y Aguadè M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116: 153-159.

Hughes AL. 1992. Positive selection and interallelic recombination at the merozoite surface antigen-1 (MSA-1) locus of *Plasmodium falciparum*. *Mol Biol Evol* 9: 381-393.

Hughes AL. 1995. The evolution of the type I interferon family in mammals. *J Mol Evol* 41: 539-548.

Hughes AL y Nei M. 1988. Pattern of nucleotide subtitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335: 167-170.

Hughes AL y Yeager M. 1997. Coordinated amino acid changes in the evolution of mammalian defensines. *J Mol Evol* 44: 675-682.

Hughes MK y Hughes AL. 1995. Natural selection on *Plasmodium* surface proteins.*Mol Biochem Parasitol* 71: 99- 113.

Huson DH. 1998. Splits Tree: a program for analysing and visualizing evolutionary data. *Bioinformatics* 14: 68-73.

Ina Y. 1995. New methods for estimating the numbers of synonymous and non synonymous substitutions. *J Mol Evol* 40: 190-226.

Ishimizu T, Endo T, Yamaguchi-Kabata Y, Nakamoya KT, Sakiyama F y Norioka S. 1998. Identification of regions in which positive selection may operate in S-Rnase of Rosaceae: implications for S- allele-specific recognition sites in S-Rnase. *FEBS Lett* 440: 337-342.

Jarvis KG Girón JA, Jerse AE, McDaniel TK, Donnenberg MS y Kaper JB. 1995. Enteropathogenic *Escherichia coli* contains a specialized secretion system necessary for the export of proteins involved in attaching and effacing lesion formation. *PNAS* 92: 7996-8000.

Jerse AE, Gicquelais KG y Kaper JB. 1991. Plasmid and chromosomal elements involved in the pathogenesis of attaching and effacing *Escherichia coli*. *Infect Inmun* 59(11): 3869-3875.

Jerse AE, Yu J, Tall BD y Kaper JB. 1990. A genetic locus of enteropathogenic *Escherichia coli* necessary for the production of attaching and effacing lesions on tissue cultured cells. *PNAS* 87: 7839-7843.

Jope EM. 1976. The evolution of plants and animals under domestication: the contribution of studies at the molecular level. *Philos Trans R Soc Lond B Biol Sci* 275 (936): 99-116.

Jordan IK, Rogozin IB, Wolf YI y Koonin EV. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Gen Res* 12: 962-68.

Jores J, Rumer L, Kiessling S, KaperJB y Wieler LH. 2001. A novel locus of enterocyte effacement (LEE) pathogenicity island inserted at *pheV* in bovine shiga toxin-producing *Escherichia coli* strain O103:H2. *FEMS Microbiol Lett* 204(1): 75-79.

Jukes TH y Cantor CR. 1969. *Evolution of protein molecules*. In: *Mammalian protein metabolism*. (Munro HN, ed), pp21-132. Academic Press, New York, USA.

Karmali KA, Steele BT, Petric M y Lim C. 1983. Sporadic cases of haemolytic-uremic syndrome associated with faecal cytotoxin and cytotoxin-producing *E. coli* in stools. *Lancet* 2: 619-20.

Karn RC y Nachman MW. 1999. Reduced nucleotide variability at an androgen-binding protein locus (Abpa) in house mice: evidence for positive natural selection. *Mol Biol Evol* 16: 1192-1197.

Kenny B, DeVinney R, Stein M, Reinscheid DJ, Frey EA y Finlay BB. 1997. Enteropathogenic *E. coli* (EPEC) transfer its receptor for intimate adherence into mammalian cells. *Cell* 91: 511-520.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217 (129): 624-626.

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

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16: 111-120.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, England.

Kitano T, Sumiyama K, Shiroishi T y Saitou N. 1998. Conserved evolution of the Rh50 gene compared to its homologous Rh blood group gene. *Biochem Biophys Res Commun* 249: 78-85.

Knapp S, Hacker J, Jarchau T y Goebel W. 1986. Large, unstable inserts in the chromosome affect virulence properties of the uropathogenic *Escherichia coli* O6 strain 536. *J Bacteriol* 168: 22-30.

Kumar S, Tamura K, Jakobsen I y Nei M. 2000. MEGA: Molecular evolutionary genetics analysis, ver 2. Pennsylvania State University, University Park, and Arizona State University, Tempe, USA.

Lack D. 1947. Darwin's finches. Cambridge, University Press, Massachusetts, USA.

Lawrence JG y Ochman H. 1998. Molecular archaeology of the *Escherichia coli* genome. *PNAS* 95: 9413-9417.

Lawrence JG y Roth JR. 1996. *Genetics* 143: 1843-1860.

LeClerc JE, Li B, Payne WC y Cebula TA. 1996. High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. *Science* 274: 1208-1211.

Lecointre G, Rachdi L, Darlu P y Denamur E. 1998. *Escherichia coli* molecular phylogeny using the incongruence length difference test. *Mol Biol Evol* 15(12): 1685-1695.

Lederberg J y Tatum EL. 1946. Gene recombination in *Escherichia coli*. *Nature* 158:558.

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

Li B, Tsui HCT, LeClerc JE, Dey M, Winkler ME y Cebula TA. 2003. Molecular analysis of *mutS* expression and mutation in natural isolates of pathogenic *Escherichia coli*. *Microbiol* 149: 1323-1331.

Li WH. 1997. Molecular Evolution. Sinauer Publishers, Massachusetts,USA.487 pp.

Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol* 36: 96-99.

Li WH, Wu CI y Luo CC. 1985. A new method for estimating synonymous and non synonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol* 2: 150-174.

Long M and Langley CH. 1993. Natural selection and the origin of *jigwei*, a chimeric processed functional gene in *Drosophila*. *Science* 260: 91-95.

Maynard-Smith J. 1991. The population genetics of bacteria. *Proc R Soc Lond Ser B* 245: 37-41.

Maynard-Smith J. 1992. Analysing the mosaic structure of genes. *J Mol Evol* 34: 126-129.

Maynard-Smith J, Feil EJ y Smith NH. 2000. Population structure and evolutionary dynamics of pathogenic bacteria. *BioEssays* 22: 1115-1122.

Maynard-Smith J y Smith NH. 1998. Detecting recombination from gene trees. *Mol Biol Evol* 15: 590-599.

Maynard-Smith J, Smith NH, O'Rourke M y Spratt BG. 1993. How clonal are bacteria? *PNAS* 90: 4384-4388.

McDaniel TK, Jarvis KG, Donnenberg MS y Kaper JB. 1995. A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. *PNAS* 92: 1664-1668.

McDaniel, TK and Kaper JB. 1997. A cloned pathogenicity island from enteropathogenic *Escherichia coli* confers the attaching and effacing phenotype on *E. coli* K12. *Mol Microbiol* 23: 399-407.

McDonald JH y Kreitman M. 1991. Adaptive protein evolution at the *adh* locus in *Drosophila*. *Nature* 351: 652-654.

Messier W y Stewart CB. 1997. Episodic adaptive evolution of the primate lysozymes. *Nature* 385: 151-154.

Milkman R. 1973. Electrophoretic variation in *Escherichia coli* from natural sources. *Science* 182: 1024-1026.

Miyata T y Yasunaga T. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J Mol Evol* 16: 23-36.

Moon HW, Whipp SC, Argensio RA, Levine NM y Gianella RA. 1983. Attaching and effacing activities of rabbit and human enteropathogenic *E. coli* in pig and rabbit intestines. *Infect Immun* 41 (3): 1340-51.

Morton BR. 1993. Chloroplast DNA codon use: evidence for selection at the *psbA* locus based on tRNA availability. *J Mol Evol* 37: 273-280.

Muse SV y Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* 11(5): 715-724.

Nakashima K, Nobuhisa I, Deshimaru M, Nakai M, Ogawa T, Shimohigashi Y, Fukumaki Y, Mattori M, Sakai Y, Hattori S y Ohto M. 1995. Accelerated evolution in the protein-coding regions is universal in crotaline snake venom gland phospholipase A_2 isozyme genes. *PNAS* 92: 5605-5609.

Nataro JP, Maher KO, Mackie P y Kaper JB. 1987. Characterization of plasmids encoding the adherence factor of enteropathogenic *Escherichia coli*. *Infect Inmun* 55: 2370-2377.

Nataro JP y Kaper, JB. 1998. Diarrheagenic *Escherichia coli*. *Clinical Microbiology Rev* 11(1): 142-201.

Nei M y Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and non synonymous nucleotide substitutions. *Mol Biol Evol* 3: 418-426.

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

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

Nelson KN y Selander RK. 1992. Evolutionary genetics of the proline permease gene (*putP*) and the control region of the proline utilization operon in populations of *Salmonella* and *Escherichia coli*. *J Bacteriol* 174: 6886-6895.

Nelson KN y Selander RK. 1994. Intergenic transfer and recombination of the 6-phosphogluconate dehydrogenase gene (*gnd*) in enteric bacteria. *PNAS* 91: 10227-10231.

Nelson KN, Whittam TS, Selander RK. 1991. Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (*gapA*) in natural populations of *Salmonella* and *Escherichia coli*. *PNAS* 88: 6667-6671.

Nielsen R. 2001.Statistical tests of selectively neutrality in the age of genomics. *Heredity* 86: 641-647.

Nielsen R y Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148(3): 929-936.

Ochman H y Selander RK. 1984. Standard reference strains of *Escherichia coli* from natural populations. *J Bacteriol* 157: 690-693.

Ohta T. 1992. The nearly neutral theory of molecular evolution. *Ann Rev Ecol Syst* 23: 263-286.

O'Rourke M y Spratt BG. 1994. Further evidence for the non-clonal population structure of *Neisseria gonorrhoeae*: extensive genetic diversity within isolates of the same electrophoretic type. *Microbiol* 140: 1285-1290.

Ota T y Nei M. 1994. Variances and covariances of the number of synonymous and non-synonymous substitutions per site. *Mol Biol Evol* 11: 613-619.

Otto SP. 2000. Detecting the form of selection from DNA sequence data. *TIG* 16(12): 526-529.

Pamilo P y Bianchi NO. 1993. Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. *Mol Biol Evol* 10: 271-281.

Pamilo P y O'Neill RW. 1997. Evolution of *Sry* genes. *Mol Biol Evol* 14:49-50.

Peek AS, Souza V, Eguiarte LE y Gaut B. 2001. The interaction of protein structure, selection, and recombination on the evolution of the Type-1 fimbrial major subunit (*fimA*) from *Escherichia coli*. *J Mol Evol* 52: 193-204.

Perna NT, Mayhew GF, Posfai G, Elliott S, Donnenberg MS, Kaper JB y Blattner FR. 1998. Molecular evolution of a pathogenicity island from enterohemorrhagic *Escherichia coli* O157:H7. *Infect Inmun* 66(8): 3810-7.

Perna NT, Plunkett G, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Grgor J, Kirkpatrick HA, Pósfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Wayne-Davis N, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, and Blattner FR. 2001. Genome sequence of enterohemorrhagic *Escherichia coli* O157:H7. *Nature* 409: 529-533.

Pupo GM, Karaolis RL y Reeves P. 1997. Evolutionary relantionships among pathogenic and non pathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and *mdh* sequences studies. *Infect Inmun* 65: 2685-2692.

Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK y Whittam TS. 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* 406:64-67.

Riley MA. 1993. Positive selection for colicin diversity in bacteria. *Mol Biol Evol* 10: 1048-1059.

Rocha M. 1999. Plásmidos en *Escherichia coli* y su importancia en la evolución de la patogénesis. Tesis de maestría. Facultad de Ciencias, UNAM, 69pp.

Rooney AP y Zhang J. 1999. Rapid evolution of a primate sperm protein: relaxation of functional constraint or positive darwinian selection? *Mol Biol Evol* 16: 706-710.

Rumer L, Jores J, Kirsch P, Cavignac Y, Zehmke K y Wieler LH. 2003. Dissemination of *phe V-* and *phe U-* located genomic islands among enteropathogenic (EPEC) and Enterohemorrhagic (EHEC) *E. coli* and their possible role in the horizontal transfer of the locus of enterocyte effacement (LEE). *Int J Med Microbiol* 292(7-8): 463-475.

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

Sato A, O'Huigin C, Figueroa F, Grant PR, Grant BR, Tichy H y Klein J. 1999. Phylogeny of Darwin's finches as revealed by mtDNA sequences. *PNAS* 96(9): 5101-5106.

Schmidt TR, Goodman M y Grossman LI. 1999. Molecular evolution of the COX7A gene family in primates. *Mol Biol Evol* 16: 619-626.

Selander RK, Caugant DA y Whittam TS.1987. *Genetic structure and variation in natural populations of Escherichia coli.* In: Neidhardt FC, Ingraham JL; Low KB, Magasanik B, Schaechter M y Umbarger HE, (editors). ***Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology.** Washington DC, American Society for Microbiology, 1625-48p.

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

Sharp PM, Tuohy TMF y Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucl Acid Res* 14: 5125-5143.

Shields DC, Sharp PM, Higgins DG y Wright F. 1988. Silent sites in *Drosophila* genes are not neutral: evidence for selection among synonymous codons. *Mol Biol Evol* 5: 704-716.

Shields DC, Harmon DL y Whitehead AS. 1996. Evolution of hemopoietic ligands and their receptors: influence of positive selection on correlated replacements throughout ligand and receptor proteins. *J Immunol* 156: 1062-1070.

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

Smith NH, Maynard-Smith J y Spratt BG. 1995. Sequence evolution of the porB gene of *Neisseria gonorrhoeae* y *Neisseria meningitidis*: evidence for positive darwinian selection. *Mol Biol Evol* 12: 363-370.

Smith NGC y Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415: 1022-24.

Souza V, Castillo A, y Eguiarte LE. 2002. The evolutionary ecology of *Escherichia coli*. *Am Sci* 90 (4): 332-341.

Souza V, Nguyen TT, Hudson RR, Piñero D y Lenski RE. 1992. Hierarchical analysis of linkage disequilibrium in *Rhizobium* populations: evidence for sex? *PNAS* 89: 8389-8393.

Souza V, Rocha M, Valera A y Eguiarte LE. 1999. Genetic structure of natural populations of *Escherichia coli* in wild hosts on different continents. *Appl Environ Microbiol* 65(8): 3373-3385.

Sperandio V, Kaper JB, Bortolini M, Neves BC, Keller R y Trabulsi LR. 1998. Characterization of LEE in different enteropathogenic *E. coli* (EPEC) and Shiga-toxin producing *E. coli* (STEC) serotypes. *FEMS Microbiol Lett*

Stone KD, Zhang HZ, Carlson LK y Donnenberg MS. 1996. A cluster of 14 genes from enteropathogenic E.coli is sufficient for the biogenesis of a type IV pilus. *Mol Microbiol* 20(2): 325-337.

Stotz HU y col. 2000. Identification of target amino acids that affect interactions of fungal polygalacturonases and their plant inhibitors. **Mol Physiol Plant Path** 56: 117-130.

Sutton KA y Wilkinson MF. 1997. Rapid evolution of a homeodomain: evidence for positive selection. *J Mol Evol* 45: 579-588.

Suzuki Y y Gojobori T. 1999. A method for detecting positive selection at single amino acid sites. *Mol Biol Evol* 16: 1315-1328.

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

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

Tanaka T y Nei M. 1989. Positive darwinian selection observed at the variable-region genes of immunoglobulins. *Mol Biol Evol* 6: 447-459.

Ting CT y col. 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. ***Science*** 282: 1501-1504.

Tarr CL y Whittam TS. 2002. *J Bacteriol* 184: 479-487.

Tsaur SC y Wu CI. 1997. Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of *Drosophila*. *Mol Biol Evol* 14: 544-549.

Vacquier VD, Swanson WJ y Lee YH. 1997. Positive darwinian selection on two homologous fertilization proteins: what is the selective pressure driving their divergence? *J Mol Evol* 44: 15-22.

Wallis M. 1996. The molecular evolution of vertebrate growth hormones: a pattern of near-stasis interrupted by sustained burst of rapid change. *J Mol Evol* 43: 93-100.

Ward TJ, Honeycott RL y Derr NJ. 1997. Nucleotide sequence evolution at the k.casein locus: evidence for positive selection within the family Bovidae. *Genetics* 147: 1863-1872.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Pop Biol* 7: 256-276.

Welch RA, Burland V, Plunkett G3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *PNAS* 99: 17020-17024.

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

Whittam TS, Ochman H y Selander RK. 1983. Multilocus genetic structure in natural populations of *Escherichia coli*. *PNAS* 80: 1751-1755.

Whittam TS, Ochman H y Selander RK. 1983b. Geographic components of linkage disequilibrium in natural populations of *Escherichia coli*. *Mol Biol Evol* 1: 67-83.

Whittam TS, Wolfe ML, Waschmuth IK, Orskov F, Orskov I y Wilson RA. 1993. Clonal relationships among *Escherichia coli* strains that cause hemorrhagic colitis and infantile diarrhea. *Infect Immun* 61: 1619-1629.

Wieler LH, McDaniel TK, Whittam TS, and Kaper JB. 1997. Insertion site of the LEE locus in EPEC and EHEC *Escherichia coli* differs in relation to the clonal phylogeny of the strains.

Wiesenfeld SL. 1967. Sickle-cell trait in human biological and cultural evolution. Development of agriculture causing increased malaria is bound to gene-pool changes causing malaria evolution. *Science* 157 (793): 1134-1140.

Wright F. 1990. The effective number of codons used in a gene. *Gene* 87: 23-29.

Wright SI y Gaut BS. 2004. Molecular population genetics and the search for adaptive evolution in plants. *Mol Biol Evol* 22(3): 506-519.

Wu JC, Chiang TY, Shive WK, Wang SY, Sheen J, Huang YH, y Syu WJ. 1999. Recombination of hepatitis D virus RNA sequences and its implications. *Mol Biol Evol* 16: 1622-1632.

Wu W, Goodman M, Lomak M y Grossman LF. 1997. Molecular evolution of cytochrome c oxidase subunit IV: evidence for positive selection in simian primates. *J Mol Evol* 44: 477-491.

Yamaguchi-Kabata Y y Gojobori T. 2000. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. *J Virol* 74: 4335-4350.

Yang W. 2000. Structure and function of mismatch repair proteins. *Mut Res* 460: 245-256.

Yang Z y Bielawski J. 2000. Statistical methods for detecting molecular adaptation. *TREE* 15(12): 496-503.

Yang Z, Kumar S y Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141(4): 1641-50.

Yang Z y Nielsen R. 2000. Estimating synonymous and non synonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17: 32-43.

Zhang HZ y Donnenberg MS. 1996. *DsbA* is requiered for stability of the type IV pilin of enteropathogenic *E.coli*. *Mol Microbiol* 21(4): 787-797.

Zhang J, Kumar S y Nei M. 1997. Small-sample test of episodic adaptive evolution: a case study of primate lysozymes. *Mol Biol Evol* 14: 1335-38.

Zhang J y Nei M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood and distance methods. *J Mol Evol* 44 (Suppl 1): S139-46.

Zhang J, Rosenberg HF y Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *PNAS* 95: 3708-3713.

Zhang YX, Perry K, Vinci VA, Powell K, Stemmer WP y del Cardayre SB. 2002. Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* 415 (6872): 644-646.

Zhu C, Agin TS, Elliot SJ, Johnson LA, Thate TE, Kaper JB y Boedeker EC. 2001. Complete nucleotide sequence and analysis of the locus of enterocyte effacement from rabbit diarrheagenic *Escherichia coli* RDEC-1. *Infect Immun* 69 (4): 2107-2115.

Zuckerl andl E y Pauling L. 1965. Molecules as documents of evolutionary biology. *J Theor Biol* 8(2): 357-366.

APÉNDICE I

Programas computacionales para determinación de selección positiva

Hicimos una revisión de los métodos y de los resultados directos del estudio de la participación de la selección a nivel molecular pero en la práctica ¿cómo determinamos el cociente dN/dS y cómo aplicamos las distintas pruebas estadísticas para probar que existe selección positiva en una muestra de secuencias de nucleótidos ó amino ácidos? ¿ cómo realizamos reconstrucciones filogenéticas?.

En la actualidad, existen varios paquetes de programas computacionales cuyo objetivo principal es el análisis y la determinación de distintos parámetros evolutivos, dentro de los cuáles también encontramos implementadas la mayoría de las pruebas estadísticas necesarias en cualquier estudio de genética. A continuación describo cinco de los principales paquetes que son los que se utilizan en la mayoría de los trabajos que hemos descrito en el presente capítulo y en la literatura científica.

1) PAML (Phylogenetic Analysis by Maximum Likelihood)

Su autor es Ziheng Yang y se encuentra accesible de forma gratuita por la red en la siguiente dirección : abacus.gene.ucl.ac.uk/software/paml.html. La versión actual es la 3.13 (Agosto de 2002). Puede ser instalado en computadoras con sistema operativo Windows 95/98/2000/NT, PowerMac y Unix. Este paquete de programas es en la actualidad uno de los que tiene más usuarios ya que la mayoría de los trabajos de estudio de selección positiva usan este paquete, incluye también la determinación del cociente dN/dS por el método de Nei y Gojobori. Entre los programas que incluye y el tipo de análisis que realizan se encuentran:

- Baseml – Análisis bajo el modelo de máxima verosimilitud, cálculo de topologías de árboles filogenéticos, cálculo de largo de las ramas y parámetros de sustitución bajo varios modelos evolutivos (Jukes y Cantor, Kimura-2, etc), análisis de reloj molecular, reconstrucción de secuencias ancestrales, etc.
- Basemlg – Lo mismo que el anterior sólo que incluye el modelo de distribución del parámetro gamma.
- Codonml – Análisis de máxima verosimilitud de secuencias codificantes usando modelos de sustitución de codones (Goldman y Yang, 1994), uso de codones,

cálculo de tasas de sustitución sinónimas y no sinónimas, prueba estadística de selección positiva basado en cociente dN/dS, reconstrucción de secuencias de codones ancestrales.

- aaml – Análisis de máxima verosimilitud de amino ácidos bajo varios modelos (Poisson, Dayhoff, etc), cálculo de reloj molecular, parámetro gamma, y reconstrucción de secuencias de amino ácidos ancestrales.
- Pamp – Análisis basados en parsimonia.
- Mcmctree – Estimación de filogenias de secuencias de DNA por método Bayesiano, cálculo de probabilidades posteriores por método de Monte Carlo.
- Yn00 – Método descrito por Yang y Nielsen, 2000, incluye Nei-Gojobori (1986).

2) HYPHY (Hypothesis testing using phylogenies)

Fue desarrollado por Sergei Kosakovski y Spencer Muse, sigue siendo la versión piloto, los autores no han liberado una versión formal final, pero se puede obtener de forma gratuita desde la siguiente dirección: www.hyphy.org. La versión es la .95 beta. Se puede instalar en computadoras con sistema operativo Windows 95/98/2000/NT, MacOX y Unix. Algunos de los análisis que incluye son: análisis de uso de codones, frecuencia de nucleótidos y aminoácidos, análisis de agrupamiento, comparación entre distintos modelos con parámetros W diferentes basados en los trabajos de Yang y col. (Goldman y Yang, 1994, Nielsen y Yang, 1998, Yang y Nielsen, 2000 y Yang y Bielawski, 2000), análisis de reloj molecular, descomposiciones de estrella, análisis de topología.

Este paquete de programas es muy accesible a los usuarios, con ventanas muy claras y un tutorial muy completo. Por otra parte es una forma más sencilla de aproximarse a los mismos modelos de Yang y col. (Goldman y Yang, 1994, Nielsen y Yang, 1998, Yang y Nielsen, 2000 y Yang y Bielawski, 2000) que parecen un poco más complicados en el programa de PAML y tienen menos instrucciones acerca de cómo usarse.

3) MEGA (Molecular Evolutionary Genetics analysis)

Desarrollado por Sudhir Kumar (Kumar y col., 2000). Se puede bajar de forma gratuita a través de la siguiente dirección: www.megasoftware.net. La versión actual es la 2.0. Se puede instalar en computadoras con sistema operativo Windows 95/98/2000/NT,

para Macintosh sólo si tiene un emulador de PC y para SUN Workstation con Softwindows 95. No posee un límite de tamaño y número de secuencias, éste estará dado por la capacidad de la computadora. Lo que lo hace muy adecuado incluso para el análisis de genomas completos. Una gran innovación de este programa es que posee un editor de textos que es muy fácil de usar, sin límite al número y tamaño de secuencias y que nos permite agregar la extensión necesaria para el uso de otros formatos para otros programas, cualquier formato es muy fácil de convertir en MEGA pues tiene un transformador integrado. Es bastante completo en cuanto al tipo de análisis que desarrolla, incluye: cálculo de distancias genéticas, construcción de genealogías bajo todos los modelos evolutivos existentes, incluye dos pruebas de selección, una de neutralidad (Tajima D), determinación de contenidos de GC, frecuencias de aminoácidos y uso de codones, determinación de tasas de sustitución por método de Nei-Gojobori y método modificado de Kumar.

4) DNASP

Desarrollado por Rozas y Rozas (1999), este es uno de los mejores paquetes de programas que existen en la actualidad para estudiar genética de poblaciones y algunas pruebas de evolución molecular, su única limitante es el número de nucleótidos que es de aproximadamente 16,000 pares de bases (para fines prácticos es bastante amplio pero no muy adecuado para estudios a nivel genómico). La versión actual es la 3.98.6. Incluye los siguientes análisis: varias pruebas de neutralidad como Hudson-Kreitman-Aguade (HKA), Tajima, McDonald y Kreitman, entre algunos; análisis de coalescencia, polimorfismo de DNA, varias pruebas de uso de codones, determinación de tasas de sustitución por el método de Nei-Gojobori, flujo génico, desequilibrio de ligamiento y recombinación. Se puede acceder de manera gratuita en la red a través de la siguiente dirección: www.ub.es/dnasp/ y se puede instalar para sistema operativo Windows 95/98/2000/NT y en Macintosh si tiene un simulador de PC como Softwindows, por ejemplo.

5)PHYLIP (Phylogeny inference package) (Felsenstein, 1993)

Creado por Joe Felsenstein y que se puede bajar a través del siguiente sitio: //evolution.genetics.washington.edu/phylip.html/. Este paquete de programas es uno de los pioneros en el análisis de evolución molecular y construcción filogenética, que además de

incluir métodos de distancia como el vecino más cercano (NJ) y Máxima Verosimilitud, incluye el método de parsimonia. El autor fue el primero en aplicar máxima verosimilitud a el análisis genético y casi todos los trabajos que lo utilizan derivan de su método. Es un buen sitio para consultar qué otros programas de análisis evolutivo existen dependiendo de qué modelos se quieren utilizar el autor lo mantiene muy actualizado, es bastante descriptivo y uno de los más utilizados en cualquier estudio filogenético. Se puede instalar en casi cualquier plataforma Windows todas sus versiones, Macintosh y UNIX.

APÉNDICE II.

Se incluye a continuación las figuras que se encuentran descritas en el artículo.

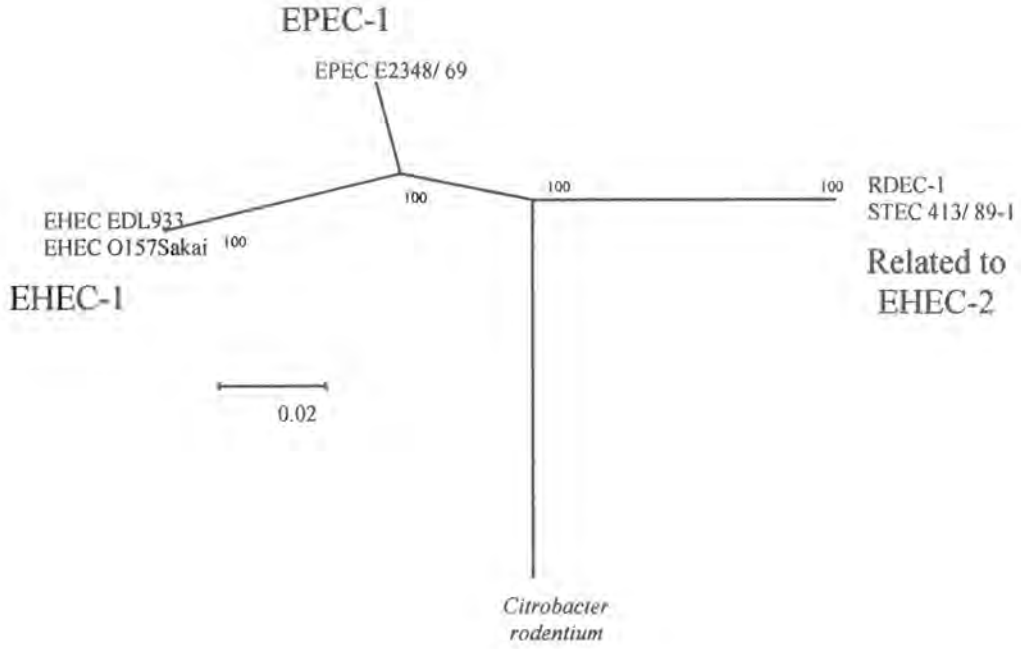


Fig.1. Genealogy of the six LEE islands used in the study. The genealogy was constructed using the 32,148bp that comprise the core consensus of the six islands (including coding and non-coding sites), under the Neighbor-Joining method with Tamura-Nei distance, and 10,000 Bootstrap.

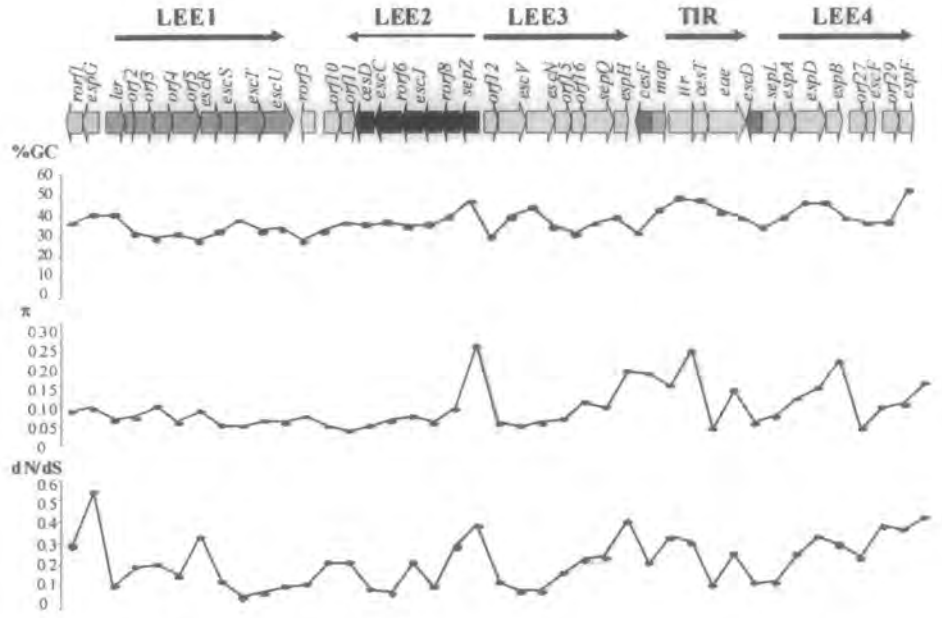


Fig.2. GC content, genetic diversity and dN/dS ratio distribution for the 41 genes of LEE.

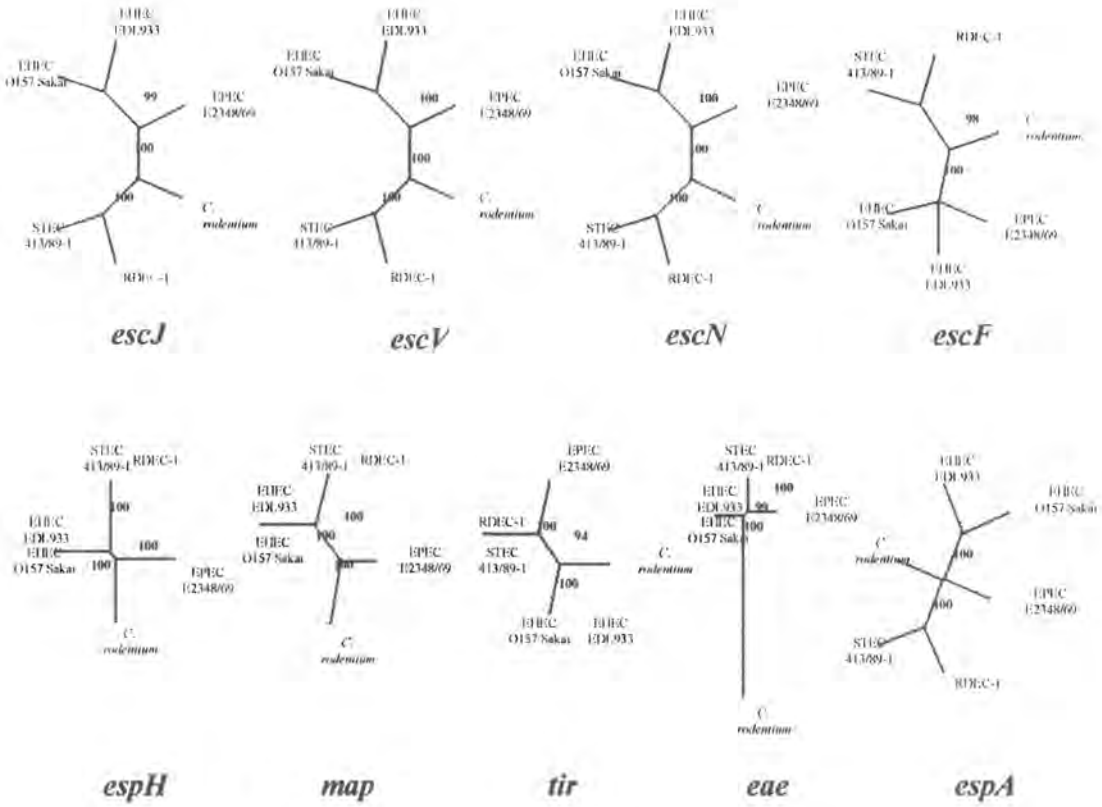


Fig. 3. Genealogy of ten LEE genes used in the study. The genealogy of each gene was constructed under the Neighbor-Joining method with Tamura-Nei distance, and 5,000 bootstrap. The genes *escJ*, *escV*, *escN* and *escF* belong to the TTSS. The genes *espH*, *map* and *espA* are secreted proteins, while *tir* is the receptor for the adhesin *eae*.

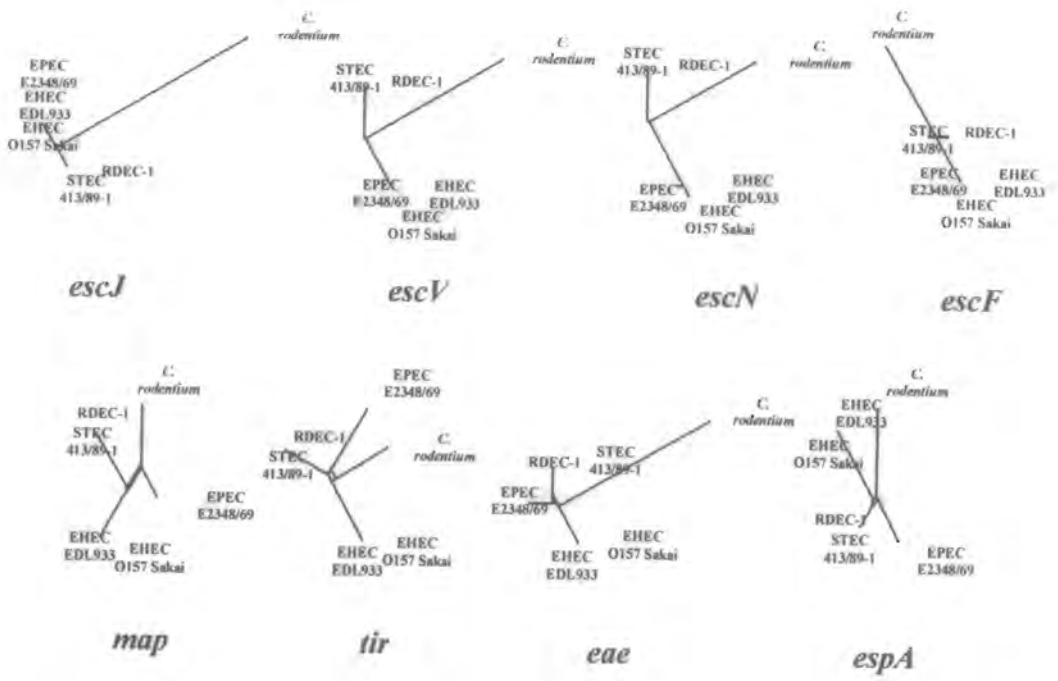


Fig. 4. Split-decomposition analysis of eight LEE genes.

A continuación se incluye la aprte correspondiente a Supporting Materials and Methods en el artículo.

Table 1.General characteristics of the six LEE islands used for the present study.

	LEE size (bp)	Sequence used for the study (bp)	Strain	Serotype	Pathotype group	Insertion site	Accesion number	Reference
EPEC	35,624	35,012	E2348/69	O127:H7	EPEC-1	selC	AF022236	(|4)
EHEC	45,325	35,214	EDL933	O157:H7	EHEC-1	selC	AF071034	(15)
EHEC	45,314	35,214	O157 Sakai	O157:H7	EHEC-1	selC		(3)
RPEC	37,889	35,043	RDEC-1	O15:H	Related to EHEC-2	unknown	AF200363	(23)
STEC	60,073	35,932	413/89-1	O26:H	Related to EHEC-2	pheu	AJ277443	(Benkel and Chakraborty, unpublished)
Citrobacter rodentium	42,001	35,191	DBS100	Not reported	Not reported	Flanked by an operon of ABC transport and IS elements	AF311901	(24)

Table 2. General analysis of the five operons of LEE.

Operon	Number of genes	π (± SE)	%GC content (± SE)
LEE1	9	0.06 (0.003)	33.1 (0.28)
LEE2	6	0.09 (0.027)	37.8 (0.20)
LEE3	6	0.07 (0.021)	39.2 (0.23)
TIR	3	0.13 (0.054)	43.3 (0.25)
LEE4	8	0.11 (0.018)	41.9 (0.18)

Table 3. General analysis of the 41 genes of LEE .

Genes	Seq. size (bp)	π (±se)	Mean %GC (±se)	%GC 1st pos (±se)	%GC 2nd pos (±se)	%GC 3rd pos (±se)	dS (±se)	dN (±se)	dN/dS	Sites under selection/ Adaptive (+) or Purifying (-) (p<0.1) (Bayes factor > 50)	Predicted function
rorf1	820	0.086 (0.009)	36.8 (0.38)	45.5 (0.84)	31.6 (0.08)	33.3 (0.82)	0.26 (0.05)	0.060 (0.009)	0.29	45 / -	Unknown
espG	1197	0.096 (0.005)	40.6 (0.29)	47.1 (0.18)	40.2 (0.87)	34.4 (0.35)	0.20 (0.04)	0.086 (0.010)	0.54	1 / + 1 / -	Secreted protein
ler	390	0.065 (0.006)	40.6 (0.21)	52.3 (0.60)	31.7 (0.92)	37.9 (1.03)	0.27 (0.06)	0.022 (0.005)	0.10	19 / -	Positive regulator
orf2	212	0.072 (0.012)	31.7 (0.74)	38.9 (0.02)	19.2 (0.72)	36.9 (1.85)	0.29 (0.05)	0.031 (0.007)	0.19	12 / -	TTSS
orf3	324	0.098 (0.006)	29.5 (1.10)	36.7 (2.05)	31.1 (0.69)	20.8 (0.74)	0.36 (0.08)	0.059 (0.01)	0.20	None	Putative chaperone of *espA* and *espB*
orf4	600	0.061 (0.006)	31.4 (0.17)	40.5 (0.10)	30.9 (0.18)	23 (0.74)	0.20 (0.04)	0.029 (0.007)	0.15	29 / -	TTSS
orf5	696	0.085 (0.005)	28.4 (0.29)	35.7 (0.29)	22.7 (0.32)	26.7 (0.35)	0.24 (0.05)	0.058 (0.013)	0.33	27 / -	TTSS
escR	654	0.051 (0.005)	32.3 (0.58)	38.2 (0.48)	27 (0.14)	31.9 (1.20)	0.21 (0.05)	0.017 (0.005)	0.12	35 / -	TTSS
escS	270	0.049 (0.008)	38.1 (0.91)	39.8 (0.94)	33.3 (0)	40.9 (2.43)	0.21 (0.06)	0.011 (0.001)	0.05	None	TTSS
escT	777	0.064 (0.003)	33.2 (0.09)	35.3 (0.30)	33.2 (0.15)	31 (0.17)	0.30 (0.06)	0.016 (0.004)	0.07	7 /-	TTSS
escU	1038	0.060 (0.005)	34.1 (0.41)	44.2 (0.50)	28.8 (0.15)	29.3 (0.76)	0.26 (0.06)	0.022 (0.005)	0.10	5 / -	TTSS
rorf3	459	0.073 (0.004)	28.3 (0.33)	29.8 (0.16)	32.5 (0.33)	22.4 (0.54)	0.42 (0.10)	0.023 (0.004)	0.11	7 / -	Putative assembly of TTSS
orf10	339	0.049 (0.004)	32.9 (0.23)	45.6 (0.58)	24.3 (0.36)	28.6 (0.53)	0.15 (0.03)	0.029 (0.005)	0.21	None	Putative negative regulator (*grlR*)
orf11	408	0.038 (0.008)	36.7 (0.23)	39.6 (0.16)	35.2 (0.51)	35.1 (0.61)	0.10 (0.02)	0.020 (0.003)	0.21	None	Putative positive regulator (*grlA*)
cesD	456	0.050 (0.003)	36 (0.12)	41.8 (0.66)	34.9 (0.24)	31.1 (0.71)	0.20 (0.03)	0.017 (0.003)	0.09	5 / -	Secretion of *espD*
escC	1539	0.067 (0.003)	37.3 (0.04)	43.3 (0.08)	37.5 (0.25)	31.3 (0.30)	0.31 (0.04)	0.014 (0.002)	0.07	115 / -	TTSS
rorf6	456	0.076 (0.003)	35.4 (0.67)	42.4 (1.40)	30.2 (0.66)	33.6 (1.37)	0.28 (0.04)	0.039 (0.006)	0.21	2 / -	Secretion of translocators
escJ	573	0.061 (0.005)	36 (0.64)	43.1 (0.77)	29.9 (0.11)	34.9 (1.24)	0.30 (0.04)	0.018 (0.002)	0.10	25 / -	TTSS
rorf8	428	0.095 (0.005)	39.9 (0.62)	50.5 (0.78)	34.6 (0.19)	34.6 (2.11)	0.28 (0.03)	0.064 (0.013)	0.29	3 / -	TTSS

sepZ	297	0.247	47.2	40.5	55.6	30.3	0.54	0.221	0.39	16 / -	Unknown
		(0.027)	(0.64)	(1.99)	(0.65)	(1.72)	(0.06)	(0.024)			
crf12	354	0.061	30.2	38.4	24.1	28.1	0.29	0.020	0.12	22 / -	TTSS
		(0.006)	(0.09)	(0.43)	(0.37)	(0.56)	(0.04)	(0.003)			
escV	2028	0.052	40.1	48.8	32.8	38.9	0.22	0.013	0.08	110 / -	TTSS
		(0.004)	(0.22)	(0.07)	(0.10)	(0.65)	(0.04)	(0.002)			
escN	1341	0.059	44.6	54.7	41.2	37.8	0.24	0.016	0.08	84 / -	TTSS
		(0.003)	(0.33)	(0.05)	(0.21)	(0.87)	(0.04)	(0.002)			
orf15	378	0.068	35.4	40.6	27.9	37.8	0.30	0.028	0.17	2 / -	TTSS
		(0.008)	(0.49)	(0.73)	(0.14)	(0.74)	(0.08)	(0.007)			
orf16	417	0.110	31.9	36.8	33.8	29.4	0.34	0.071	0.23	1 / -	Secretion of
		(0.011)	(0.31)	(0.34)	(0.91)	(0.18)	(0.05)	(0.010)			translocators
sepQ	918	0.099	37	45.2	33.2	32.4	0.32	0.059	0.24	8 / -	TTSS
		(0.008)	(0.54)	(0.62)	(0.32)	(1.19)	(0.04)	(0.008)			
espH	498	0.187	39.5	43.8	41.7	33.1	0.57	0.176	0.41	1 / -	Secreted protein
		(0.017)	(0.64)	(1.51)	(0.78)	(0.78)	(0.08)	(0.019)			
cesF	360	0.179	32.1	50.6	26.1	27.2	1.12	0.107	0.21	30 / -	Chaperone for *espF*
		(0.015)	(0.48)	(1.33)	(1.10)	(1.29)	(0.26)	(0.011)			
map	612	0.152	43	47.2	38.2	43.7	0.47	0.108	0.33	1 / +	Secreted protein
		(0.009)	(0.50)	(0.22)	(0.65)	(0.87)	(0.05)	(0.013)		5 / -	
tir	1593	0.236	49	59.7	50.9	36.4	0.59	0.202	0.31	1 / +	Intimin receptor,
		(0.011)	(0.33)	(0.74)	(0.32)	(1.52)	(0.11)	(0.024)		102 / -	secreted protein
cesT	471	0.046	47.7	42.6	31.2	29.7	0.19	0.013	0.11	2 / -	Chaperone for *tir*
		(0.003)	(0.46)	(0.20)	(0.36)	(0.79)	(0.02)	(0.002)			
eae	2784	0.138	42.3	47.6	41	38.4	0.39	0.096	0.26	3 / +	Adhesin (intimin)
		(0.015)	(0.39)	(0.46)	(0.23)	(0.69)	(0.04)	(0.011)		17 / -	
escD	1221	0.059	39.2	45.7	33.3	38.5	0.24	0.018	0.12	2 / -	TTSS
		(0.003)	(0.20)	(0.08)	(0.27)	(0.65)	(0.04)	(0.003)			
sepL	1056	0.077	34.6	42.6	28.5	32.7	0.31	0.030	0.13	68 / -	Secretion
		(0.005)	(0.35)	(0.22)	(0.18)	(1.00)	(0.04)	(0.003)			of translocators
espA	579	0.120	39.9	46.4	39.3	34	0.35	0.082	0.26	6 / -	Secreted protein
		(0.010)	(0.40)	(0.56)	(0.69)	(0.95)	(0.04)	(0.010)			
espD	1125	0.144	46.6	54.9	44.8	40	0.37	0.107	0.34	1 / +	Secreted protein
		(0.012)	(0.38)	(0.22)	(0.40)	(1.46)	(0.05)	(0.015)		67 / -	
espB	930	0.213	46.7	53.8	46.6	39.8	0.74	0.165	0.30	50 / -	Secreted protein
		(0.011)	(0.32)	(0.53)	(1.07)	(1.55)	(0.11)	(0.020)			
orf27	408	0.045	39.2	45.9	31.1	40.8	0.13	0.021	0.24	2 / -	Putative chaperone
		(0.006)	(0.31)	(0.19)	(0.30)	(0.69)	(0.02)	(0.005)			of *espD*
escF	222	0.097	36.8	44.3	30.1	35.8	0.26	0.086	0.39	None	TTSS
		(0.003)	(0.41)	(0.97)	(0.57)	(1.34)	(0.06)	(0.032)			
orf29	279	0.107	37.4	68	29	35.1	0.32	0.081	0.37	4 / -	TTSS
		(0.031)	(0.34)	(0.55)	(0.57)	(0.97)	(0.06)	(0.013)			
espF	606	0.156	53.1	59.1	61.7	38.6	0.34	0.121	0.43	1 / +	Secreted protein
		(0.008)	(0.22)	(1.40)	(0.94)	(1.02)	(0.04)	(0.016)		16 / -	

Table 4. ILD analysis results. Pairwise p values among LEE genes.

	escR	*escS*	*escU*	*escV*	*escN*	*escD*	*escF*
escJ	1	1	1	1	1	1	1
sepZ	1	1	1	1	1	1	1
espH	0.001	0.001	0.001	0.001	0.001	0.001	0.011
map	0.001	0.007	0.001	0.001	0.003	0.001	0.006
tir	0.020	0.159	0.001	0.001	0.001	0.001	0.124
eae	0.033	0.173	0.008	0.015	0.001	0.007	0.096
espA	0.001	0.194	0.001	0.001	0.001	0.002	0.114
espB	0.001	0.001	0.001	0.001	0.001	0.001	0.001

* Values represent p values for 1,000 ILD partitions using the partition homogeneity test and branch and bound search option available in PAUP v.403b (34).
p values of 1 indicates complete congruence while small p values indicates incongruence between genes.

Table 5. PDS test results.

Gene	Codon	SLAC dN/dS (p)	ARS dN/dS (p)	Full Likelihood dN/dS (Bayes factor)
rORF1				
Negative sites	10	-2.72 (0.241)	-2.49 (0.038)	-3.49 (2947.9)
	11	-2.29 (0.287)	-6.14 (0.023)	-3.47 (66099.3)
	12	-1.97 (0.333)	-1.29 (0.059)	-3.53 ($2.9x10^6$)
	33	-1.97(0.333)	-4.17 (0.033)	-3.50 (161003)
	34	-5.44 (0.101)	-2.56 (0.064)	-3.40 (2467.6)
	41	-4.73 (0.139)	-3.44 (0.031)	-3.48 (181.8)
	50	-2.60 (0.293)	-1.48 (0.080)	-3.39 (4484.9)
	52	-1.97 (0.337)	-2.20 (0.053)	-3.50 (1432.6)
	54	-3.95 (0.111)	-3.45 (0.037)	-3.50 (6903.9)
	57	-2.72 (0.253)	-1.77 (0.079)	-3.45 (813.5)
	64	-1.97 (0.333)	-1.05 (0.081)	-3.53 (21613.4)
	67	-2.72 (0.241)	-1.78 (0.074)	-3.46 (887)
	70	-1.97 (0.333)	-0.96 (0.090)	-3.53 ($2.6x10^6$)
	78	-2.72 (0.241)	-1.78 (0.054)	-3.54 (1107)
	84	-5.77 (0.060)	-3.98 (0.011)	-3.50 (63715.8)
	95	-3.95 (0.111)	-3.34 (0.039)	-3.50 (89663.8)
	97	-5.44 (0.064)	-2.56 (0.038)	-3.49 (3033.6)
	105	-1.97 (0.333)	-2.20 (0.046)	-3.53 (31715.7)
	117	-2.72 (0.241)	-1.78 (0.066)	-3.49 (922.3)
	120	-1.97 (0.333)	-0.93 (0.098)	-3.50 (1186)
	129	-2.72 (0.241)	-1.78 (0.074)	-3.46 (887)
	138	-2.72 (0.253)	-1.77 (0.079)	-3.45 (813.5)
	139	-2.72 (0.253)	-2.54 (0.046)	-3.45 (2596.5)
	142	-1.97 (0.333)	-1.08 (0.078)	-3.53 (794.2)
	145	-3.95 (0.117)	-0.98 (0.088)	-3.53 ($2.2x10^6$)
	158	-3.95 (0.111)	-4.22 (0.028)	-3.53 (180119)
	160	-2.72 (0.253)	-1.77 (0.067)	-3.48 (934.8)
	163	-1.97 (0.333)	-4.21 (0.032)	-3.50 (164233)
	169	-2.72 (0.266)	-1.78 (0.087)	-3.43 (412510)
	172	-1.97 (0.337)	-2.20 (0.053)	-3.50 (1432.6)
	174	-2.72 (0.241)	-2.49 (0.037)	-3.50 (3075)
	179	-5.44 (0.064)	-2.20 (0.073)	-3.45 (1940.6)
	182	-3.95 (0.111)	-3.45 (0.037)	-3.50 (6903.9)
	204	-2.72 (0.241)	-1.78 (0.054)	-3.54 (1107)
	211	-2.72 (0.253)	-1.77 (0.067)	-3.48 (934.8)
	214	-2.72 (0.253)	-2.17 (0.075)	-3.45 (1867)
	222	-2.72 (0.253)	-2.54 (0.039)	-3.49 (2909)
	232	-1.97 (0.333)	-0.93 (0.098)	-3.50 (1186)
	247	-3.95 (0.112)	-3.42 (0.009)	-3.54 (459595)
	255	-5.44 (0.058)	-9.80 (0.005)	-3.54 ($3.1x10^6$)
	258	-2.15 (0.386)	-15.17 (0.075)	-2.17 (934.4)
	259	-1.97 (0.407)	-15.3 (0.051)	-2.27 (6992.5)
	266	-1.97 (0.333)	-98.07 (0.010)	-3.53 ($1.1x10^6$)
	270	-3.95 (0.111)	-609.5 (0.001)	-3.50 (887201)
	272	-3.34 (0.283)	-21.2 (0.028)	-1.62 (152.3)
espG				
Positive sites	166	3.39 (0.387)	2.49 (0.097)	5.52 (139.7)
Negative sites	72	-6 (0.075)	-2.09 (0.026)	-2.18 (4)
ler				

Negative sites	13	-3.69 (0.135)	-7.41 (0.078)	-4.12 (30119.9)
	17	-5.76 (0.045)	-34 (0.025)	-4.12 (4935.1)
	22	-1.51 (0.406)	-5.70 (0.097)	-4.11 (2436.7)
	23	-6.76 (0.033)	-55 (0.012)	-4.11 (769.1)
	28	-3.69 (0.111)	-34.2 (0.023)	-4.11 (1126.9)
	30	-2.88 (0.271)	-18.8 (0.039)	-4.11 (1562.5)
	40	-2.88 (0.227)	-5 (0.089)	-4.11 (1263.7)
	49	-5.76 (0.073)	-18.8 (0.042)	-4.11 (1573.6)
	61	-2.88 (0.227)	-4.39 (0.096)	-4.11 (1308.9)
	72	-3.69 (0.111)	-5.70 (0.080)	-4.12 (5085.9)
	80	-3.69 (0.111)	-34.2 (0.005)	-4.11 (1510.6)
	81	-1.84 (0.333)	-7.03 (0.062)	-4.11 (2328.3)
	83	-2.88 (0.242)	-25.6 (0.031)	-4.12 (14075.8)
	92	-1.84 (0.347)	-5.79 (0.074)	-4.12 (15757.5)
	93	-3.69 (0.116)	-28.4 (0.018)	-4.12 (3216.1)
	94	-3.69 (0.111)	-31.4 (0.008)	-4.11 (486.1)
	119	-3.69 (0.111)	-14.6 (0.009)	-4.11 (1843.4)
	128	-2.88 (0.213)	-10.2 (0.048)	-4.11 (1111)
orf2				
Negative sites	6	-7.03 (0.125)	-18.6 (0.020)	-4.68 (368.2)
	9	-5.70 (0.075)	-477.1 (0.012)	-4.74 (407946)
	13	-7.03 (0.125)	-8.25 (0.055)	-4.59 (108.5)
	21	-5.70 (0.075)	-3.67 (0.021)	-4.92 (19542.6)
	29	-7.03 (0.073)	-5.28 (0.073)	-4.59 (103.9)
	39	-4.68 (0.111)	-1.58 (0.034)	-4.94 ($1.6x10^9$)
	42	-7.03 (0.125)	-8.25 (0.055)	-4.59 (108.5)
	54	-4.68 (0.111)	-10.3 (0.002)	-4.90 ($2.20x10^6$)
	55	-2.85 (0.282)	-3.60 (0.088)	-4.86 (4450.9)
	59	-5.14 (0.092)	-1.89 (0.092)	-4.74 (30484.1)
	62	-3.12 (0.249)	-4.66 (0.049)	-4.73 (266)
	68	-3.51 (0.244)	-13.3 (0.029)	-4.71 (379.2)
orf4				
Negative sites	2	-2.09 (0.333)	-4.06 (0.051)	-5.90 (160065)
	11	-6.01 (0.066)	-696.3 (0.002)	-5.87 ($2.1x10^{16}$)
	16	-3 (0.232)	-5.21 (0.062)	-5.89 (1004.8)
	18	-2.09 (0.333)	-4.93 (0.045)	-5.91 (90714.4)
	47	-2.09 (0.333)	-5.60 (0.043)	-5.90 (58754.8)
	52	-3 (0.232)	-15 (0.019)	-5.89 (6471.9)
	54	-4.30 (0.162)	-7.04 (0.041)	-5.88 (301.2)
	58	-2.05 (0.422)	-7193.5 (0.009)	-5.02 (21496.1)
	59	-4.30 (0.162)	-7.04 (0.041)	-5.88 (301.2)
	63	-3.15 (0.221)	-7.41 (0.046)	-5.87 (292.8)
	65	-2.09 (0.334)	-19.8 (0.012)	-5.92 (46312.4)
	69	-2.09 (0.345)	-5.61 (0.046)	-5.89 (8827.2)
	82	-2.09 (0.333)	-2.63 (0.069)	-5.92 (2150.9)
	98	-3 (0.244)	-1.96 (0.099)	-5.91 (2705.33)
	100	-3 (0.232)	-15 (0.015)	-5.91 (6678)
	103	-3 (0.232)	-8.08 (0.024)	-5.89 (8576.8)
	109	-3 (0.244)	-3.87 (0.070)	-5.90 (1014.8)
	112	-2.46 (0.282)	-5.27 (0.050)	-5.87 (17937.2)
	141	-3 (0.232)	-5.21 (0.066)	-5.89 (971.4)
	144	-3 (0.244)	-8.19 (0.050)	-5.90 (3951.7)
	145	-2.51 (0.314)	-9.78 (0.030)	-5.83 ($1.5x10^6$)
	153	-2.60 (0.290)	-19.7 (0.016)	-5.84 (44444.1)

	154	-3 (0.244)	-8.38 (0.022)	-5.90 (8391.9)
	157	-3 (0.232)	-5.21 (0.066)	-5.89 (971.4)
	161	-4.19 (0.111)	-8.73 (0.006)	-5.91 ($1.6x10^{10}$)
	164	-4.19 (0.111)	-7.16 (0.010)	-5.90 ($6.7x10^{9}$)
	170	-1.50 (0.516)	-43.5 (0.059)	-5.83 (20145.6)
	173	-3 (0.232)	-5.21 (0.048)	-5.93 (1063.9)
	177	-3 (0.244)	-3.87 (0.082)	-3.87 (956)
orf5				
Negative sites	12	-2.88 (0.262)	-1.61 (0.091)	-3.16 (96648.1)
	31	-4.10 (0.111)	-1.09 (0.098)	-3.31 (9252.6)
	37	-5.77 (0.061)	-2.62 (0.027)	-3.32 (1782.4)
	40	-2.69 (0.289)	-1.21 (0.091)	-3.11 (972.3)
	49	-2.49 (0.274)	-8.52 (0.012)	-3.21 (346.3)
	51	-2.88 (0.248)	-3.60 (0.033)	-3.19 (2566.6)
	67	-2.88 (0.248)	-3.60 (0.033)	-3.19 (2566.6)
	75	-2.05 (0.344)	-0.83 (0.085)	-3.29 (678.8)
	79	-2.40 (0.285)	-1.79 (0.048)	-3.21 (6539.4)
	80	-2.05 (0.333)	-0.83 (0.060)	-3.43 (8942.8)
	91	-1.87 (0.431)	-0.91 (0.082)	-3.28 (861.1)
	101	-2.88 (0.035)	-2.61 (0.035)	-3.32 (2100.5)
	122	-2.88 (0.238)	-0.73 (0.085)	-3.36 (255.6)
	139	-2.88 (0.236)	-5.15 (0.011)	-3.36 (2723.6)
	156	-2.88 (0.248)	-1.07 (0.077)	-3.32 (330.8)
	159	-2.05 (0.333)	-1.41 (0.040)	-3.32 ($2.4x10^{7}$)
	164	-4.10 (0.111)	-2.51 (0.009)	-3.31 ($2.3x10^{7}$)
	165	-2.05 (0.344)	-0.61 (0.082)	-3.45 (42543.5)
	178	-2.88 (0.236)	-1.61 (0.051)	-3.36 (295.9)
	180	-4.10 (0.111)	-4.32 (0.004)	-3.46 (211396)
	194	-4.10 (0.111)	-1.25 (0.052)	-3.35 ($1.1x10^{8}$)
	198	-6.53 (0.043)	-8.45 (0.014)	-3.25 (221.4)
	210	-5.77 (0.056)	-3.13 (0.018)	-3.36 (1602.8)
	211	-4.10 (0.111)	-1.09 (0.092)	-3.35 ($7.7x10^{7}$)
	213	-2.88 (0.248)	-2.61 (0.035)	-3.32 (2100.5)
	215	-4.22 (0.161)	-2.58 (0.039)	-3.22 (67.1)
	228	-2.88 (0.236)	-0.73 (0.075)	-3.41 (278.3)
escR				
Negative sites	17	-2.48 (0.333)	-5.01 (0.067)	-5.35 (170.3)
	19	-3.91 (0.211)	-6.16 (0.064)	-5.37 (373.8)
	25	-2.48 (0.333)	-25.5 (0.009)	-5.41 (934)
	31	-2.48 (0.341)	-5.40 (0.050)	-5.39 (17833.3)
	34	-3.91 (0.211)	-3.66 (0.087)	-5.37 (522.3)
	38	-2.48 (0.353)	-27.7 (0.020)	-5.37 (294.4)
	49	-2.48 (0.333)	-8.22 (0.039)	-5.40 (15955.8)
	56	-3.91 (0.211)	-19.17 (0.024)	-5.39 (767.7)
	62	-2.48 (0.333)	-9.15 (0.049)	-5.35 (99)
	67	-2.48 (0.333)	-5.01 (0.052)	-5.39 (624.3)
	68	0.869 (0.048)	0.770 (0.848)	-2.86 (61.8)
	78	-4.54 (0.181)	-8.36 (0.046)	-5.35 (107)
	82	-3.89 (0.212)	-7.28 (0.055)	-5.34 (86.9)
	85	-3.91 (0.224)	-4.14 (0.083)	-5.37 (2408.8)
	86	-2.48 (0.333)	-2.33 (0.095)	-5.41 (5058.1)
	87	-3.89 (0.214)	-4.44 (0.080)	-5.34 (107.5)
	88	-3.91 (0.211)	-6.16 (0.062)	-5.39 (1415.5)
	89	-3.91 (0.240)	-29.9 (0.013)	-5.33 (437.9)

	106	-3.91 (0.240)	-29.9 (0.031)	-5.33 (1107.8)
	113	-3.91 (0.211)	-6.16 (0.062)	-5.39 (1415.5)
	114	-2.48 (0.333)	-5.01 (0.052)	-5.39 (624.3)
	126	-2.48 (0.333)	-2.52 (0.087)	-5.41 (3104.5)
	140	-2.48 (0.350)	-5.40 (0.052)	-5.38 (913.8)
	146	-2.48 (0.333)	-9.14 (0.040)	-5.41 (2869.6)
	149	-2.48 (0.333)	-3.95 (0.063)	-5.39 (937.2)
	155	-2.48 (0.333)	-2.52 (0.087)	-5.41 (3104.5)
	157	-3.91 (0.211)	-3.66 (0.087)	-5.37 (522.3)
	160	-2.48 (0.341)	-5.40 (0.050)	-5.39 (17833.3)
	164	-3.91 (0.240)	-6.16 (0.079)	-5.33 (1098)
	166	-2.48 (0.333)	-7.42 (0.071)	-5.41 (7577.5)
	169	-2.48 (0.333)	-3.95 (0.055)	-5.41 (2132.9)
	193	-3.89 (0.212)	-7.28 (0.055)	-5.34 (86.9)
	198	-1.95 (0.450)	-61.9 (0.059)	-5.31 (209.9)
	204	-2.48 (0.346)	-5.01 (0.346)	-5.39 (12493.7)
	210	-3.91 (0.211)	-3.66 (0.087)	-5.37 (522.3)
escS	No sites			
escT				
Negative sites	3	-4.93 (0.072)	-14.4 (0.032)	-4.34 (0)
	24	-4.32 (0.088)	-8.16 (0.064)	-4.34 (0)
	38	-6.12 (0.044)	-73.9 (0.014)	-4.34 (0)
	65	-4.32 (0.088)	-8.16 (0.064)	-4.34 (0)
	78	-6.12 (0.044)	-73.9 (0.014)	-4.34 (0)
	193	-7.77 (0.082)	-63.9 (0.006)	-4.34 (0)
	249	-4.93 (0.067)	-22.6 (0.050)	-4.34 (0)
escU				
Negative sites	135	-5.22 (0.060)	-13.1 (0.015)	-4.35 (11.5)
	154	-5.22 (0.072)	-17.3 (0.056)	-4.13 (4)
	259	-4.43 (0.084)	-7.42 (0.096)	-4.09 (4.9)
	318	-5.22 (0.060)	-17.3 (0.043)	-4.15 (5.8)
	342	-5.22 (0.060)	-13.1 (0.020)	-4.33 (6.9)
rorf3				
Negative sites	7	-4.46 (0.080)	-1.74 (0.005)	-3.13 (0)
	8	-4.48 (0.076)	-4.63 (0.010)	-3.13 (0)
	32	-4.48 (0.076)	-0.72 (0.085)	-3.13 (0)
	42	-4.48 (0.076)	-0.72 (0.099)	-3.12 (0)
	107	-5.59 (0.054)	-5.45 (0.054)	-3.12 (0)
	142	-4.48 (0.086)	-0.72 (0.089)	-3.12 (0)
	147	-4.48 (0.076)	-0.72 (0.099)	-3.12 (0)
orf10	No sites			
orf11	No sites			
cesD				
Negative sites	24	-3.88 (0.214)	-7.74 (0.045)	-12.3 (1261.3)
	107	-5.19 (0.111)	-14.6 (0.003)	-12.5 (1975.8)
	112	-3.77 (0.269)	-16.6 (0.017)	-12.9 (1340.4)
	127	-3.39 (0.289)	-6.11 (0.068)	-12.3 (1144.1)
	146	-3.90 (0.209)	-4.44 (0.072)	-12.2 (951.9)
escC				
Negative sites	5	-2.62 (0.267)	-7.93 (0.047)	-6.44 (6723.7)
	17	-1.05 (0.555)	-24.8 (0.079)	-4.14 (47212.9)
	19	-2.10 (0.333)	-20.5 (0.020)	-6.44 (288269)
	32	-2.62 (0.277)	-8.88 (0.064)	-6.43 (65553.4)
	37	-2.10 (0.333)	-24.7 (0.020)	-6.42 (326249)

	39	-2.62 (0.267)	-8.79 (0.040)	-6.44 (8551.4)
	41	-2.10 (0.333)	-4.41 (0.074)	-6.42 ($1.65x10^{8}$)
	46	-2.10 (0.333)	-4.24 (0.089)	-6.40 ($2.17x10^{7}$)
	48	-2.62 (0.267)	-4.46 (0.090)	-6.43 (5629.6)
	50	-2.10 (0.333)	-3.25 (0.093)	-6.44 ($5.48x10^{8}$)
	55	-2.33 (0.300)	-4.94 (0.072)	-6.41 (22081.2)
	62	-2.10 (0.333)	-4.33 (0.075)	-6.41 ($1.6x10^{7}$)
	76	-2.10 (0.333)	-24.7 (0.021)	-6.42 (577217)
	81	-2.62 (0.277)	-9.34 (0.045)	-6.43 (5702.9)
	86	-4.19 (0.112)	-9788.5 (0.002)	-6.40 ($9.6x10^{17}$)
	95	-2.62 (0.277)	-4.09 (0.089)	-6.43 (4233.5)
	109	-2.62 (0.267)	-7.63 (0.066)	-6.44 (7607.1)
	112	-2.10 (0.333)	-4.41 (0.069)	-6.42 (10922.1)
	128	-2.10 (0.333)	-4.07 (0.077)	-6.42 (7621.7)
	130	-2.10 (0.351)	-18.3 (0.033)	-6.40 (7452.6)
	136	-2.62 (0.277)	-7.43 (0.046)	-6.42 (31271)
	137	-2.62 (0.267)	-8.79 (0.054)	-6.41 (7373.4)
	143	-2.10 (0.333)	-4.07 (0.077)	-6.42 (7621.7)
	146	-2.10 (0.342)	-13.3 (0.083)	-6.41 (5658.5)
	151	-2.10 (0.333)	-3.25 (0.090)	-6.45 ($5.8x10^{6}$)
	153	-2.10 (0.333)	-20.5 (0.025)	-6.42 (276586)
	155	-2.62 (0.277)	-8.79 (0.047)	-6.42 (24424.1)
	158	-2.62 (0.286)	-8.79 (0.069)	-6.39 ($1.4x10^{7}$)
	163	-4.21 (0.122)	-21.6 (0.010)	-6.41 ($3.5x10^{6}$)
	165	-2.62 (0.267)	-7.93 (0.047)	-6.44 (6723.7)
	167	-2.62 (0.267)	-7.93 (0.052)	-6.43 (6378.6)
	169	-2.62 (0.277)	-8.68 (0.037)	-6.43 (9155.5)
	174	-4.21 (0.111)	-16.6 (0.007)	-6.44 ($7.07x10^{8}$)
	182	-2.10 (0.333)	-16.6 (0.051)	-6.43 (110549)
	185	-2.56 (0.335)	-7.71 (0.046)	-6.37 (63140.2)
	188	-4.21 (0.111)	-14.3 (0.007)	-6.44 ($6.2x10^{8}$)
	194	-2.10 (0.333)	-4.41 (0.069)	-6.42 (10922.1)
	204	-2.10 (0.333)	-17.1 (0.052)	-6.43 (6888.2)
	205	-2.10 (0.333)	-25.8 (0.018)	-6.42 (584407)
	206	-2.10 (0.333)	-3.25 (0.090)	-6.45 ($5.8x10^{6}$)
	213	-2.10 (0.333)	-4.41 (0.069)	-6.42 (10922.1)
	218	-2.10 (0.333)	-4.30 (0.059)	-6.45 ($5.03x10^{6}$)
	219	-2.10 (0.356)	-89.9 (0.028)	-6.40 ($1.8x10^{10}$)
	222	-2.62 (0.267)	-8.79 (0.040)	-6.44 (8551.4)
	223	-2.10 (0.333)	-16.6 (0.053)	-6.42 (156674)
	224	-2.10 (0.333)	-4.30 (0.059)	-6.45 ($5.03x10^{6}$)
	227	-4.21 (0.111)	-15.08 (0.007)	-6.44 ($8.6x10^{8}$)
	228	-4.21 (0.111)	-16.6 (0.009)	-6.43 ($7.1x10^{7}$)
	229	-2.62 (0.277)	-8.68 (0.037)	-6.43 (9155.5)
	230	-2.25 (0.336)	-25.2 (0.024)	-6.42 (303.8)
	232	-2.10 (0.333)	-4.30 (0.059)	-6.45 ($5.03x10^{6}$)
	235	-2.62 (0.267)	-7.93 (0.062)	-6.41 (5704.6)
	243	-2.10 (0.333)	-23.5 (0.017)	-6.44 (457221)
	249	-2.10 (0.333)	-18.3 (0.023)	-6.44 (319138)
	252	-2.10 (0.333)	-18.3 (0.026)	-6.43 (107011)
	254	-2.62 (0.267)	-4.46 (0.090)	-6.43 (5629.6)
	255	-2.10 (0.333)	-4.31 (0.069)	-6.42 ($2.1x10^{7}$)
	258	-4.21 (0.111)	-18.8 (0.009)	-6.42 ($3.1x10^{6}$)
	260	-2.62 (0.267)	-8.79 (0.045)	-6.43 (8114.4)

	261	-6.66 (0.105)	-17.16 (0.020)	-6.43 (788)
	270	-2.62 (0.267)	-7.93 (0.052)	-6.43 (6378.6)
	272	-2.62 (0.277)	-10.34 (0.038)	-6.43 (7204.2)
	277	-2.62 (0.277)	-10.34 (0.038)	-6.43 (7204.2)
	279	-2.10 (0.333)	-3.95 (0.089)	-6.41 ($5.7x10^6$)
	280	-2.10 (0.333)	-4.33 (0.065)	-6.43 ($3.2x10^7$)
	293	-2.52 (0.308)	-20.6 (0.043)	-6.38 ($7.7x10^7$)
	294	-4.21 (0.112)	-16.6 (0.007)	-6.45 ($1.6x10^6$)
	300	-2.10 (0.333)	-4.48 (0.064)	-6.42 (15580.6)
	304	-2.10 (0.335)	-3.28 (0.087)	-6.44 ($5.08x10^6$)
	311	-2.62 (0.267)	-8.79 (0.054)	-6.41 (7373.4)
	313	-2.10 (0.336)	-23.5 (0.023)	-6.41 (10612.3)
	314	-3.15 (0.259)	-16.04 (0.090)	-4.14 (81576.3)
	315	-2.10 (0.333)	-4.41 (0.059)	-6.44 ($5.03x10^6$)
	317	-1.05 (0.555)	-85.2 (0.081)	-4.14 ($3.9x10^8$)
	320	-2.62 (0.267)	-7.43 (0.039)	-6.44 (10919.8)
	321	-2.10 (0.333)	-3.95 (0.089)	-6.41 ($5.7x10^6$)
	325	-2.62 (0.267)	-8.79 (0.040)	-6.44 (8551.4)
	326	-2.62 (0.267)	-7.63 (0.085)	-6.41 (6561.8)
	330	-2.62 (0.267)	-7.63 (0.085)	-6.41 (6496)
	332	-2.33 (0.300)	-4.94 (0.072)	-6.41 (22081.2)
	333	-2.33 (0.309)	-6.50 (0.042)	-6.44 (8844)
	338	-2.62 (0.366)	-9.34 (0.084)	-6.37 (4548.4)
	343	-4.21 (0.111)	-23.5 (0.006)	-6.44 ($7.2x10^8$)
	349	-4.21 (0.111)	-3141 (0.0004)	-6.42($1.09x10^{11}$)
	365	-2.10 (0.333)	-18.3 (0.026)	-6.43 (107011)
	371	-2.10 (0.333)	-4.33 (0.065)	-6.43 ($3.3x10^7$)
	373	-2.10 (0.342)	-18.3 (0.023)	-6.44 (3992.7)
	377	-2.10 (0.333)	-4.48 (0.054)	-6.44 ($9.6x10^6$)
	378	-2.10 (0.335)	-20.5 (0.027)	-6.41 (7719.2)
	382	-2.74 (0.256)	-26.4 (0.020)	-6.41 (556.4)
	383	-2.10 (0.333)	-25.8 (0.033)	-6.42 (230931)
	384	-2.62 (0.267)	-3.55 (0.092)	-6.44 (4836.2)
	385	-2.10 (0.342)	-4.07 (0.064)	-6.44 ($2.4x10^6$)
	401	-2.62 (0.267)	-5032.1 (0.001)	-6.44 (288800)
	404	-2.10 (0.342)	-4.07 (0.064)	-6.44 ($2.4x10^6$)
	420	-2.62 (0.267)	-7.93 (0.062)	-6.41 (5793.6)
	421	-2.62 (0.366)	-9.34 (0.084)	-6.37 (4548.4)
	426	-2.62 (0.267)	-4.46 (0.083)	-6.44 (5933.2)
	436	-2.10 (0.333)	-3.95 (0.086)	-6.42 ($2.1x10^8$)
	438	-2.10 (0.333)	-18.3 (0.026)	-6.43 (107011)
	450	-2.62 (0.267)	-7.63 (0.066)	-6.44 (7607.1)
	451	-2.10 (0.333)	-21.6 (0.020)	-6.44 (455959)
	455	-2.37 (0.355)	-7.65 (0.064)	-6.37 (37601)
	456	-2.37 (0.356)	-8.28 (0.050)	-6.37 ($1.5x10^6$)
	459	-2.10 (0.342)	-4.07 (0.064)	-6.44 ($2.4x10^6$)
	460	-2.62 (0.267)	-7.93 (0.062)	-6.41 (5704.6)
	465	-2.62 (0.277)	-10.3 (0.038)	-6.43 (7204.2)
	476	-2.62 (0.267)	-4.46 (0.083)	-6.44 (5933.2)
	477	-2.62 (0.267)	-7.43 (0.043)	-6.43 (10363.6)
	484	-2.56 (0.333)	-7.17 (0.056)	-6.37 (36173.4)
	492	-2.62 (0.277)	-9.34 (0.055)	-6.41 (5068.3)
	493	-2.62 (0.267)	-8.79 (0.058)	-6.41 (7087.3)
	494	-2.62 (0.267)	-8.79 (0.040)	-6.44 (8551.4)

	503	-2.10 (0.333)	-3.28 (0.089)	-6.44 ($6.7x10^6$)
	508	-2.10 (0.336)	-23.5 (0.023)	-6.41 (10612.3)
rorf6				
Negative sites	73	-5.60 (0.079)	-3.07 (0.069)	-4.16 (0)
	122	-6.82 (0.053)	-24.09 (0.005)	-4.16 (0)
escJ				
Negative sites	18	-2.73 (0.333)	-8.10 (0.089)	-6.65 (4786.8)
	29	-2.73 (0.333)	-4.21 (0.063)	-6.74 (2376.8)
	47	-2.73 (0.333)	-18.4 (0.015)	-6.65 (91012.4)
	48	-2.73 (0.333)	-2.30 (0.091)	-6.77 ($8.9x10^6$)
	58	-2.73 (0.333)	-7.46 (0.036)	-6.74 (3448.2)
	69	-2.73 (0.333)	-4.21 (0.059)	-6.76 (900641)
	71	-5.47 (0.111)	-10.2 (0.011)	-6.74 ($4.04x10^6$)
	78	-3.36 (0.271)	-3.43 (0.099)	-6.72 (47139.3)
	79	-2.73 (0.333)	-4.12 (0.047)	-6.79 ($3.9x10^6$)
	87	-5.66 (0.192)	-14.8 (0.057)	-6.61 (20145.9)
	90	-3.36 (0.271)	-3.95 (0.083)	-6.71 (44633.4)
	91	-2.73 (0.333)	-4.21 (0.046)	-6.79 ($2.5x10^6$)
	92	-5.47 (0.111)	-37.1 (0.001)	-6.79 ($1.2x10^{13}$)
	98	-2.73 (0.354)	-4.12 (0.061)	-6.72 ($4.1x10^6$)
	104	-2.73 (0.333)	-9.49 (0.046)	-6.76 (25604.7)
	114	-4.33 (0.210)	-5.93 (0.081)	-6.65 (16626.5)
	121	-3.13 (0.291)	-25.7 (0.087)	-6.63 (10877.2)
	122	-5.47 (0.111)	-37.1 (0.001)	-6.79 ($1.2x10^{13}$)
	137	-5.66 (0.192)	-11.3 (0.067)	-6.61 (14266.3)
	140	-2.73 (0.333)	-8.39 (0.028)	-6.79 ($1.8x10^6$)
	143	-2.73 (0.333)	-8.38 (0.029)	-6.76 ($3.4x10^6$)
	158	-5.47 (0.122)	-17.01 (0.002)	-6.73 ($1.2x10^6$)
	165	-3.85 (0.247)	-349.8 (0.006)	-6.61 ($7.6x10^7$)
	169	-3.36 (0.271)	-5.51 (0.089)	-6.69 (68162)
	170	-5.47 (0.111)	-35.09 (0.007)	-6.65 ($1.2x10^{11}$)
rorf8				
Negative sites	73	-5.69 (0.046)	-13.3 (0.020)	-5.69 (13)
	78	-5.69 (0.046)	-116.2 (0.006)	-8.90 (21.4)
	120	-5.69 (0.040)	-186.3 (0.006)	-8.95 (23.9)
sepZ				
Negative sites	7	-1.25 (0.193)	-5.22 (0.020)	-0.90 (1070.3)
	14	-1.45 (0.111)	-1.63 (0.028)	-0.86 (1654.3)
	16	-1.45 (0.111)	-1.63 (0.032)	-0.85 (170.2)
	27	-1.25 (0.207)	-0.92 (0.092)	-0.88 (565.2)
	37	-1.45 (0.114)	-1.60 (0.036)	-0.84 (373.1)
	56	-2.18 (0.037)	-6.76 (0.003)	-0.85 (152.7)
	57	-0.72 (0.333)	-1.17 (0.090)	-0.85 (164.7)
	61	-0.72 (0.336)	-1.17 (0.095)	-0.85 (205.2)
	62	-1.84 (0.069)	-1.86 (0.015)	-0.89 (1391.2)
	66	-2.51 (0.043)	-2.75 (0.031)	-0.86 (2693.3)
	78	-1.21 (0.208)	-3.73 (0.009)	-0.85 (488.1)
	79	-1.45 (0.115)	-3.71 (0.044)	-0.84 (187.9)
	80	-1.84 (0.069)	-1.21 (0.056)	-0.89 (1190)
	84	-0.72 (0.333)	-1.45 (0.076)	-0.86 (246.8)
	85	-0.60 (0.458)	-1.48 (0.080)	-0.85 (232.9)
	93	-1.45 (0.111)	-1.54 (0.035)	-0.85 (449.3)
orf12				
Negative sites	8	-5.72 (0.078)	-17.01 (0.025)	-5.30 (1979.7)

	9	-2.39 (0.333)	-11.6 (0.044)	-5.29 (265.7)
	10	-2.86 (0.287)	-4.60 (0.066)	-5.31 (6298.3)
	11	-2.86 (0.287)	-3.87 (0.077)	-5.31 (7555.3)
	25	-2.86 (0.279)	-3.90 (0.070)	-5.30 (1803.6)
	27	-2.39 (0.333)	-16.08 (0.017)	-5.29 (105.9)
	33	-2.39 (0.333)	-14.2 (0.014)	-5.35 (7455.3)
	39	-2.86 (0.287)	-4.60 (0.066)	-5.31 (6298.3)
	48	-2.39 (0.333)	-2.24 (0.078)	-5.36 (120499)
	50	-4.79 (0.111)	-3.002 (0.044)	-5.38 (84654.1)
	56	-2.86 (0.279)	-3.90 (0.070)	-5.30 (1803.6)
	57	-2.39 (0.333)	-3.17 (0.048)	-5.38 (102728)
	59	-2.86 (0.279)	-3.90 (0.099)	-5.26 (599.5)
	68	-2.86 (0.287)	-6.01 (0.049)	-5.31 (4848.3)
	69	-2.25 (0.355)	-5.90 (0.059)	-5.34 (99677.9)
	72	-2.39 (0.340)	-2.47 (0.079)	-5.34 (77421.6)
	77	-2.61 (0.306)	-3.39 (0.092)	-5.28 (653)
	82	-5.72 (0.078)	-4.14 (0.060)	-5.30 (1619.3)
	95	-2.39 (0.333)	-10.9 (0.054)	-5.29 (318)
	103	-2.86 (0.388)	-6.21 (0.085)	-5.25 (10488)
	110	-2.86 (0.287)	-6.01 (0.049)	-5.31 (4848.3)
	111	-2.25 (0.459)	-4.44 (0.088)	-5.28 (44749.5)
escV				
Negative sites	34	-4.79 (0.159)	-13 (0.037)	-7.13 (6248.9)
	47	-9.71 (0.030)	-42.8 (0.006)	-7.18 (9800.4)
	50	-5.05 (0.111)	-29.4 (0.013)	-7.21 (672508)
	65	-3.43 (0.256)	-16.5 (0.029)	-7.19 (8511.8)
	69	-3.43 (0.244)	-12.2 (0.036)	-7.18 (11291.6)
	70	-5.05 (0.111)	-13.6 (0.008)	-7.19 ($1.9x10^6$)
	87	-2.52 (0.333)	-7.16 (0.052)	-7.20 (95822.8)
	112	-3.43 (0.244)	-6.48 (0.076)	-7.19 (2714.4)
	121	-5.05 (0.111)	-14.8 (0.007)	-7.20 (735292)
	126	-2.52 (0.333)	-28.8 (0.030)	-7.17 (174123)
	135	-2.52 (0.333)	-6.12 (0.046)	-7.21 ($2.5x10^6$)
	145	-3.43 (0.244)	-6.48 (0.083)	-7.18 (2590.5)
	147	-2.06 (0.407)	-8.05 (0.090)	-7.15 (46111.3)
	153	-3.43 (0.256)	-16.5 (0.030)	-7.18 (8374.1)
	167	-3.43 (0.244)	-12.2 (0.031)	-7.19 (11818.5)
	180	-3.43 (0.256)	-5.76 (0.079)	-7.18 (2122.4)
	184	-3.43 (0.244)	-6.48 (0.083)	-7.18 (2590.5)
	201	-5.05 (0.111)	-6.44 (0.044)	-7.21 ($2.5x10^6$)
	204	-2.91 (0.289)	-7.50 (0.085)	-7.18 (15344.6)
	210	-3.22 (0.280)	-31.02 (0.022)	-7.13 (67944.1)
	214	-2.52 (0.333)	-7.10 (0.070)	-7.17 (6590.7)
	217	-5.05 (0.116)	-29.4 (0.014)	-7.19 (57160.4)
	236	-2.52 (0.337)	-11.9 (0.040)	-7.19 (11047.8)
	240	-2.52 (0.333)	-3.57 (0.089)	-7.21 ($2.2x10^6$)
	241	-3.25 (0.284)	-126.7 (0.026)	-7.13 (218312)
	251	-5.05 (0.111)	-6.44 (0.049)	-7.19 ($5.2x10^6$)
	252	-2.52 (0.333)	-4.85 (0.095)	-7.17 (36913.4)
	253	-2.52 (0.333)	-6.12 (0.047)	-7.20 ($2.2x10^6$)
	254	-2.52 (0.333)	-4.65 (0.086)	-7.21 ($3.9x10^6$)
	255	-2.52 (0.333)	-21.5 (0.032)	-7.19 (47964.3)
	256	-5.05 (0.111)	-26 (0.032)	-7.17 (46211.3)
	261	-2.52 (0.333)	-26.5 (0.019)	-7.17 (4403.1)

	273	-2.52 (0.333)	-11.9 (0.040)	-7.19 (6016.1)
	285	-2.52 (0.333)	-11.9 (0.039)	-7.19 (184639)
	292	-2.65 (0.372)	-23.5 (0.021)	-7.12 (31691.4)
	300	-2.52 (0.333)	-26.5 (0.015)	-7.19 (1.2 x10^6)
	311	-2.52 (0.333)	-7.10 (0.051)	-7.21 (155706)
	316	-2.52 (0.333)	-26.5 (0.016)	-7.19 (54092.7)
	317	-6.34 (0.070)	-38.5 (0.006)	-7.18 (9945.4)
	322	-6.87 (0.105)	-16.5 (0.044)	-7.15 (8003.8)
	323	-3.43 (0.256)	-5.76 (0.079)	-7.18 (2122.4)
	333	-3.43 (0.256)	-5.76 (0.079)	-7.18 (2122.4)
	334	-6.87 (0.059)	-14.4 (0.028)	-7.19 (12168.1)
	341	-2.52 (0.337)	-11.9 (0.040)	-7.19 (11047.8)
	343	-6.87 (0.059)	-11.8 (0.049)	-7.21 (9379.7)
	345	-3.43 (0.244)	-12.2 (0.032)	-7.19 (11672.9)
	347	-2.52 (0.355)	-7.16 (0.061)	-7.18 (10465.4)
	349	-5.05 (0.111)	-7.19 (0.095)	-7.19 (3.8 x10^6)
	350	-2.52 (0.333)	-4.84 (0.083)	-7.21 (2.7 x10^6)
	352	-5.05 (0.111)	-5.6 (0.063)	-7.19 (3.9 x10^6)
	363	-6.87 (0.059)	-14.3 (0.023)	-7.21 (12884.3)
	365	-3.43 (0.244)	-12.2 (0.031)	-7.19 (11818.5)
	369	-6.87 (0.105)	-16.5 (0.044)	-7.15 (8024.3)
	372	-2.52 (0.333)	-6.12 (0.065)	-7.17 (505414)
	379	-3.43 (0.256)	-5.76 (0.076)	-7.19 (2194.7)
	392	-5.05 (0.111)	-5.67 (0.064)	-7.19 (4.2 x10^6)
	398	-3.43 (0.244)	-3.39 (0.099)	-7.21 (2529.7)
	413	-2.52 (0.333)	-11.9 (0.036)	-7.20 (99349.3)
	414	-2.06 (0.407)	-8.05 (0.090)	-7.15 (46111.3)
	419	-2.52 (0.333)	-6.12 (0.065)	-7.17 (505414)
	422	-3.43 (0.256)	-5.76 (0.076)	-7.19 (2194.7)
	423	-2.52 (0.333)	-3.57 (0.089)	-7.21 (2.2 x10^6)
	425	-3.43 (0.244)	-6.48 (0.084)	-7.18 (2630.9)
	430	-3.43 (0.244)	-12.2 (0.036)	-7.18 (11291.6)
	431	-3.43 (0.256)	-5.81 (0.076)	-7.19 (2197.3)
	436	-3.43 (0.256)	-5.76 (0.079)	-7.18 (2122.4)
	443	-5.05 (0.111)	-5.69 (0.078)	-7.17 (54577.4)
	445	-3.43 (0.256)	-16.5 (0.030)	-7.18 (8374.1)
	450	-3.43 (0.244)	-14.1 (0.054)	-7.19 (8673.5)
	452	-6.87 (0.105)	-160.7 (0.003)	-7.15 (1.4 x10^6)
	461	-2.52 (0.333)	-11.9 (0.049)	-7.17 (6877.7)
	467	-6.87 (0.059)	-14.4 (0.028)	-7.19 (12168.1)
	470	-2.52 (0.333)	-3.20 (0.098)	-7.20 (9704.4)
	480	-3.43 (0.244)	-6.48 (0.076)	-7.19 (2714.4)
	490	-2.52 (0.333)	-7.10 (0.052)	-7.20 (10436.9)
	499	-5.82 (0.086)	-9.35 (0.039)	-7.20 (8379.3)
	503	-6.87 (0.065)	-16.5 (0.029)	-7.19 (8686.5)
	507	-3.43 (0.256)	-5.76 (0.076)	-7.19 (2194.7)
	508	-4.64 (0.131)	-12.9 (0.030)	-7.15 (272992)
	516	-6.87 (0.065)	-16.5 (0.030)	-7.18 (8547.7)
	518	-5.05 (0.111)	-29.4 (0.010)	-7.17 (98017.2)
	519	-6.87 (0.059)	-11.8 (0.056)	-7.19 (8884.7)
	520	-2.91 (0.289)	-114.9 (0.011)	-7.18 (102464)
	521	-2.32 (0.362)	-12.8 (0.052)	-7.15 (299477)
	522	-3.43 (0.324)	-16.5 (0.045)	-7.15 (7842.4)
	523	-3.43 (0.346)	-8.5 (0.058)	-7.18 (11367.8)

	533	-3.43 (0.244)	-164.2 (0.004)	-7.19 (67312.2)
	534	-3.34 (0.253)	-9.75 (0.050)	-7.14 (174475)
	537	-5.82 (0.083)	-6.87 (0.087)	-7.18 (15532.1)
	539	-6.87 (0.065)	-16.5 (0.029)	-7.19 (8686.5)
	540	-2.52 (0.333)	-3.20 (0.098)	-7.20 (9704.4)
	545	-5.05 (0.119)	-24.1 (0.030)	-7.18 (93871.8)
	549	-3.43 (0.256)	-16.5 (0.030)	-7.18 (8374.1)
	560	-5.82 (0.083)	-84.7 (0.083)	-7.18 (488981)
	571	-3.43 (0.244)	-3.39 (0.099)	-7.21 (2529.7)
	575	-2.52 (0.333)	-7.16 (0.052)	-7.20 (95822.8)
	578	-3.43 (0.244)	-12.2 (0.032)	-7.19 (11672.9)
	579	-5.05 (0.111)	-6.44 (0.062)	-7.17 (505190)
	582	-3.69 (0.227)	-19.4 (0.033)	-7.18 (472.2)
	590	-4.79 (0.169)	-7.97 (0.072)	-7.12 (57247.9)
	606	-2.52 (0.355)	-7.16 (0.061)	-7.18 (10465.4)
	608	-3.43 (0.256)	-5.76 (0.076)	-7.19 (2194.7)
	613	-8.25 (0.041)	-41.1 (0.015)	-7.20 (1559.8)
	620	-5.82 (0.083)	-6.87 (0.087)	-7.18 (15532.1)
	626	-11.8 (0.059)	-11.8 (0.058)	-7.19 (8782.7)
	650	-2.52 (0.333)	-4.85 (0.095)	-7.17 (36913.4)
	657	-3.43 (0.256)	-16.5 (0.029)	-7.19 (8533.7)
	659	-2.52 (0.333)	-7.10 (0.058)	-7.19 (5697.5)
	664	-2.06 (0.059)	-8.05 (0.063)	-7.15 (46111.3)
	673	-6.87 (0.059)	-11.85 (0.063)	-7.18 (8500)
escN				
Negative sites	22	-2.33 (0.333)	-3.88 (0.085)	-6.52 (4.17×10^{7})
	31	-2.33 (0.333)	-3.88 (0.096)	-6.51 (2.05×10^{7})
	32	-2.33 (0.333)	-5.19 (0.080)	-6.52 (397875)
	48	-2.33 (0.333)	-8.77 (0.060)	-6.51 (829414)
	53	-2.33 (0.333)	-8.77 (0.060)	-6.51 (829414)
	56	-2.33 (0.333)	-4.28 (0.096)	-6.51 (1.17×10^{8})
	59	-6.14 (0.075)	-93.31 (0.003)	-6.51(1.58×10^{16})
	68	-2.33 (0.333)	-4.79 (0.072)	-6.51 (2.84×10^{9})
	81	-2.72 (0.332)	-6.58 (0.072)	-6.50 (7.69×10^{6})
	90	-3.07 (0.275)	-9.02 (0.055)	-6.51 (1.26×10^{8})
	92	-4.67 (0.111)	-15.59 (0.010)	-6.51 (2.97×10^{7})
	99	-4.67 (0.113)	-71.05 (0.010)	-6.51(1.89×10^{15})
	107	-4.67 (0.112)	-14.06 (0.010)	-6.51 (3.01×10^{7})
	110	-2.76 (0.316)	-19.79 (0.035)	-6.50 (4.43×10^{7})
	114	-2.17 (0.357)	-17.20 (0.040)	-6.50 (2.69×10^{6})
	117	-2.33 (0.333)	-4.28 (0.096)	-6.51 (1.17×10^{8})
	120	-2.65 (0.302)	-4.66 (0.074)	-6.52 (7573.2)
	124	-3.07 (0.253)	-8.29 (0.071)	-6.52 (14786.3)
	128	-3.50 (0.259)	-14.41 (0.092)	-5.99 (6.43×10^{7})
	136	-3.07 (0.253)	-8.91 (0.050)	-6.52 (25542.9)
	139	-2.33 (0.333)	-15.08 (0.035)	-6.51 (907006)
	141	-2.33 (0.333)	-22.50 (0.027)	-6.51 (6.15×10^{6})
	142	-2.33 (0.333)	-22.50 (0.027)	-6.51 (6.15×10^{6})
	157	-2.33 (0.333)	-5.19 (0.090)	-6.51 (15181)
	161	-2.85 (0.273)	-38.73 (0.008)	-6.51 (2921)
	163	-2.33 (0.333)	-19.85 (0.027)	-6.51 (47979.5)
	164	-2.98 (0.305)	-6.64 (0.068)	-6.50 (3.91×10^{6})
	165	-2.04 (0.379)	-10.50 (0.064)	-6.50 (1.61×10^{6})
	171	-3.07 (0.339)	-12.13 (0.052)	-6.50 (23655.9)

	176	-3.07 (0.253)	-8.29 (0.087)	-6.51 (13934.5)
	177	-2.33 (0.333)	-26.83 (0.014)	-6.51 ($7.14x10^{6}$)
	178	-2.33 (0.333)	-26.83 (0.015)	-6.51 (105912)
	179	-2.33 (0.338)	-19.84 (0.015)	-6.51 (105912)
	188	-2.33 (0.334)	-5.21 (0.069)	-6.51 ($1.17x10^{8}$)
	195	-2.33 (0.339)	-5.19 (0.092)	-6.51 (32743.7)
	196	-2.33 (0.333)	-4.19 (0.088)	-6.51 ($9.14x10^{7}$)
	206	-3.07 (0.264)	-11.15 (0.052)	-6.51 (14011.9)
	217	-2.33 (0.333)	-15.40 (0.036)	-6.50($5.34x10^{11}$)
	226	-3.07 (0.264)	-7.60 (0.063)	-6.51 (20291.4)
	227	-2.33 (0.333)	-72.51 (0.043)	-6.51 ($1.19x10^{7}$)
	232	-2.33 (0.333)	-8.77 (0.055)	-6.51 (36423.5)
	235	-2.33 (0.333)	-4.76 (0.084)	-6.51 ($1.04x10^{8}$)
	237	-2.33 (0.333)	-8.77 (0.060)	-6.51 (829414)
	240	-2.65 (0.303)	-6.58 (0.055)	-6.52 (13955.9)
	242	-3.07 (0.264)	-12.61 (0.029)	-6.51 (26781.1)
	248	-2.33 (0.333)	-4.43 (0.076)	-6.52 (101061)
	251	-2.33 (0.333)	-10.23 (0.053)	-6.51 (895735)
	255	-2.12 (0.365)	-8.30 (0.093)	-6.51 (69540.8)
	257	-3.07 (0.339)	-11.15 (0.074)	-6.50 (13251.6)
	258	-4.67 (0.111)	-15.59 (0.010)	-6.51 ($2.9x10^{7}$)
	266	-3.07 (0.253)	-7.60 (0.065)	-6.51 (9622.8)
	274	-2.33 (0.333)	-5.19 (0.086)	-6.51 (793265)
	275	-4.67 (0.111)	-71.05 (0.009)	-6.51($1.62x10^{16}$)
	284	-6.14 (0.069)	-124.86 (0.001)	-6.51 ($1.51x10^{8}$)
	285	-2.33 (0.333)	-15.08 (0.039)	-6.51 (915450)
	288	-2.65 (0.294)	-7.55 (0.041)	-6.52 (24344)
	297	-4.67 (0.111)	-5.76 (0.054)	-6.51 ($4.07x10^{8}$)
	305	-2.28 (0.357)	-13.28 (0.042)	-6.51 (140103)
	307	-4.67 (0.116)	-13.80 (0.011)	-6.51 ($3.01x10^{7}$)
	312	-2.33 (0.333)	-15.08 (0.036)	-6.51 (18289.6)
	315	-2.33 (0.333)	-4.79 (0.066)	-6.52 ($2.10x10^{8}$)
	321	-3.05 (0.264)	-16.98 (0.031)	-6.50 ($1.89x10^{6}$)
	325	-3.07 (0.253)	-7.60 (0.065)	-6.51 (9622.8)
	327	-2.33 (0.333)	-26.83 (0.017)	-6.51 (142604)
	330	-4.67 (0.111)	-5.76 (0.057)	-6.51($1.29x10^{10}$)
	331	-5.30 (0.86)	-7.28 (0.041)	-6.51 (70702.1)
	332	-2.33 (0.333)	-4.79 (0.075)	-6.51 ($4.66x10^{7}$)
	335	-2.33 (0.333)	-4.79 (0.084)	-6.51 ($1.25x10^{8}$)
	341	-2.33 (0.334)	-5.21 (0.069)	-6.51 ($1.17x10^{8}$)
	342	-3.07 (0.253)	-9.02 (0.052)	-6.51 (17908)
	343	-6.45 (0.058)	-17.40 (0.008)	-6.51 (42167.7)
	346	-2.33 (0.333)	-5.19 (0.080)	-6.52 (397875)
	363	-6.14 (0.064)	-11.72 (0.024)	-6.52 (37010.1)
	375	-3.07 (0.264)	-11.10 (0.047)	-6.51 (14516.5)
	379	-2.33 (0.333)	-5.21 (0.066)	-6.51 ($3.87x10^{8}$)
	381	-2.33 (0.333)	-23.51 (0.023)	-6.51 ($6.39x10^{6}$)
	391	-3.07 (0.264)	-12.61 (0.029)	-6.51 (26781.1)
	394	-3.07 (0.264)	-12.13 (0.035)	-6.51 (25060.3)
	410	-2.33 (0.333)	-4.19 (0.088)	-6.51 ($9.14x10^{7}$)
	413	-2.33 (0.333)	-10.23 (0.053)	-6.51 (895735)
	416	-6.14 (0.064)	-11.72 (0.028)	-6.52 (36033.3)
	417	-2.65 (0.296)	-4.28 (0.078)	-6.52 (7423.1)
	420	-2.65 (0.293)	-4.83 (0.079)	-6.51 (19506.9)

	427	-3.07 (0.253)	-3.41 (0.095)	-6.52 (8278.7)
orf15				
Negative sites	66	-5.27 (0.075)	-15.14 (0.023)	-3.87 (0)
	113	-5.27 (0.068)	-10.69 (0.024)	-3.87 (0)
orf16				
Negative sites	131	-5.05 (0.043)	-6.36 (0.018)	-2.68 (7.8)
sepQ				
Negative sites	13	-4.47 (0.060)	-3.37 (0.042)	-4.25 (3.5)
	62	-4.47 (0.060)	-3.37 (0.049)	-4.24 (3.3)
	200	-4.47 (0.66)	-3.37 (0.038)	-4.24 (3.7)
	237	-4.97 (0.037)	-26.81 (0.007)	-5.61 (8.5)
	248	-5.66 (0.055)	-40.04 (0.026)	-5.87 (4.9)
	262	-4.47 (0.060)	-1386.65 (0.001)	-5.66 (9.6)
	264	-4.45 (0.061)	-11.88 (0.009)	-5.33 (6.6)
	284	-5.94 (0.092)	-44.81 (0.015)	-4.11 (3.4)
espH				
Negative sites	58	-0.001 (0.065)	-8.94 (0.083)	-0.0005 (6.5)
cesF				
Negative sites	3	-1.92 (0.164)	-0.74 (0.020)	-1.42 (150.4)
	8	-1.08 (0.275)	-0.22 (0.048)	-1.42 (171.2)
	13	-1.08 (0.274)	-0.22 (0.048)	-1.43 (176)
	19	-1.39 (0.226)	-0.12 (0.092)	-1.42 (113.6)
	21	-0.89 (0.333)	-0.98 (0.015)	-1.45 (1101)
	24	-1.65 (0.129)	-0.14 (0.034)	-1.43 (294.9)
	26	-1.18 (0.251)	-0.074 (0.097)	-1.43 (212.6)
	27	-2.20 (0.073)	-0.39 (0.009)	-1.44 (858)
	28	-1.07 (0.276)	-0.36 (0.047)	-1.42 (172.8)
	30	-1.07 (0.276)	-0.36 (0.047)	-1.42 (172.8)
	33	-1.63 (0.193)	-0.20 (0.054)	-1.42 (129.1)
	37	-1.37 (0.229)	-0.14 (0.087)	-1.42 (116.4)
	42	-0.89 (0.350)	-0.22 (0.022)	-1.44 (516.2)
	43	-1.90 (0.166)	-0.60 (0.025)	-1.42 (148.8)
	55	-1.93 (0.162)	-0.98 (0.016)	-1.42 (152.7)
	66	-1.08 (0.275)	-0.36 (0.051)	-1.42 (192)
	77	-1.08 (0.302)	-0.29 (0.036)	-1.41 (62.7)
	87	-1.08 (0.275)	-0.36 (0.051)	-1.42 (192)
	90	-2.53 (0.055)	-0.33 (0.012)	-1.43 (237.8)
	91	-0.89 (0.333)	-0.04 (0.085)	-1.46 (5198.9)
	94	-1.90 (0.162)	-0.60 (0.023)	-1.42 (165.9)
	97	-1.18 (0.250)	-0.31 (0.025)	-1.42 (113.5)
	98	-0.89 (0.333)	-0.09 (0.052)	-1.45 (958.5)
	99	-1.58 (0.199)	-0.20 (0.059)	-1.42 (127.5)
	100	-1.08 (0.275)	-0.36 (0.039)	-1.43 (117)
	106	-1.08 (0.275)	-0.36 (0.051)	-1.42 (192)
	107	-0.89 (0.342)	-0.22 (0.045)	-1.42 (207.5)
	111	-1.58 (0.195)	-0.20 (0.055)	-1.42 (141.7)
	113	-1.92 (0.164)	-0.74 (0.020)	-1.42 (150.4)
	115	-1.18 (0.250)	-0.31 (0.025)	-1.42 (113.5)
map				
Positive sites	181	3.61 (0.080)	3.77 (0.100)	1.48 (155.7)
Negative sites	24	-3.78 (0.037)	-10.65 (0.010)	-2.13 (32.8)
	125	-3.67 (0.052)	-45.08 (0.008)	-2.12 (25.5)
	134	-3.67 (0.052)	-4.87 (0.042)	-2.13 (39.7)
	136	-3.67 (0.052)	-4.83 (0.064)	-2.09 (16.2)

	149	-3.67 (0.058)	-230.9 (0.005)	-2.15 (49.1)
tir				
Positive sites	441	1.73 (0.094)	0.81 (0.060)	-0.32 (0.8)
Negative sites	5	-1.13 (0.181)	-5.51 (0.013)	-0.95 (3098)
	7	-0.71 (0.333)	-0.71 (0.061)	-0.91 (1674.7)
	8	-1.01 (0.329)	-2.68 (0.093)	-0.50 (54)
	17	-2.13 (0.037)	-5.67 (0.002)	-0.90 (555.5)
	23	-0.71 (0.333)	-5.68 (0.053)	-0.90 (1400.9)
	37	-1.36 (0.120)	-0.90 (0.027)	-0.91 (3742.5)
	40	-0.71 (0.336)	-2.02 (0.042)	-0.091 (4747.1)
	42	-1.42 (0.125)	-1.40 (0.017)	-0.91 (493.8)
	45	-0.71 (0.3589	-0.54 (0.091)	-0.91 (1735.8)
	49	-1.79 (0.089)	-1.22 (0.024)	-0.90 (1878.6)
	54	-0.91 (0.265)	-0.61 (0.063)	-0.95 (10594.6)
	62	-1.30 (0.181)	-2.12 (0.032)	-0.93 (1334.4)
	64	-2.16 (0.047)	-4.38 (0.007)	-0.089 (242.4)
	69	-1.42 (0.111)	-1.23 (0.021)	-0.90 (1065.2)
	81	-1.42 (0.111)	-5.92 (0.005)	-0.089 (105.7)
	83	-1.42 (0.113)	-5.09 (0.004)	-0.91 (3820.6)
	92	-1.30 (0.096)	-5.70 (0.024)	-0.93 (1570.6)
	94	-0.71 (0.333)	-0.65 (0.073)	-0.92 (9414.1)
	98	-1.42 (0.112)	-2.13 (0.010)	-0.91 (619.4)
	100	-0.71 (0.333)	-0.48 (0.095)	-0.92 (9237.4)
	103	-2.04 (0.041)	-1.43 (0.014)	-0.91 (3397.6)
	104	-1.30 (0.181)	-0.84 (0.051)	-0.95 (4723.1)
	112	-1.30 (0.181)	-2.85 (0.012)	-0.96 (62530.2)
	113	-0.68 (0.354)	-0.60 (0.087)	-0.91 (5327.3)
	120	-0.71 (0.333)	-0.71 (0.061)	-0.91 (1674.1)
	126	-1.42 (0.112)	-2.14 (0.009)	-0.91 (614.7)
	130	-0.71 (0.333)	-0.71 (0.061)	-0.91 (1674.1)
	141	-1.30 (0.196)	-1.17 (0.045)	-0.93 (1899.4)
	146	-1.30 (0.221)	-1.17 (0.048)	-0.92 (10773.1)
	149	-1.30 (0.181)	-5.51 (0.015)	-0.93 (1142.9)
	151	-1.30 (0.196)	-5.71 (0.024)	-0.93 (1553)
	157	-0.71 (0.333)	-0.64 (0.083)	-0.92 (772.5)
	158	-1.30 (0.181)	-5.50 (0.013)	-0.96 (65108.7)
	162	-0.71 (0.333)	-0.68 (0.062)	-0.92 (10107.8)
	163	-0.71 (0.333)	-0.68 (0.073)	-0.91 (1814.8)
	164	-0.71 (0.333)	-0.71 (0.061)	-0.91 (1674.1)
	171	-1.42 (0.111)	-2.08 (0.012)	-0.92 (374.6)
	172	-0.71 (0.333)	-0.68 (0.072)	-0.92 (1521.3)
	173	-0.71 (0.333)	-0.68 (0.087)	-0.90 (217.7)
	180	-0.71 (0.333)	-0.48 (0.096)	-0.92 (11277.1)
	232	-1.42 (0.113)	-5.69 (0.006)	-0.91 (4289.4)
	233	-1.30 (0.181)	-0.56 (0.068)	-0.96 (14208.9)
	235	-0.71 (0.333)	-0.66 (0.074)	-0.92 (11127.1)
	251	-1.79 (0.083)	-0.90 (0.032)	-0.90 (2302.4)
	255	-1.42 (0.111)	-3.09 (0.006)	-0.89 (136.1)
	256	-1.42 (0.111)	-5.23 (0.019)	-0.90 (192)
	257	-1.42 (0.111)	-5.24 (0.015)	-0.92 (507.4)
	258	-0.71 (0.333)	-2.01 (0.333)	-0.91 (591.2)
	264	-1.42 (0.111)	-1.54 (0.017)	-0.90 (155.1)
	266	-1.42 (0.111)	-1.41 (0.015)	-0.92 (393.9)
	267	-1.42 (0.111)	-1.07 (0.024)	-0.90 (1953.8)

	272	-0.71 (0.333)	-0.64 (0.093)	-0.90 (1376.7)
	274	-1.42 (0.111)	-0.93 (0.022)	-0.92 (503.5)
	276	-1.30 (0.181)	-0.84 (0.057)	-0.93 (1661.9)
	280	-1.42 (0.111)	-1.02 (0.027)	-0.90 (161.5)
	282	-1.01 (0.329)	-2.68 (0.078)	-0.50 (92.4)
	295	-1.30 (0.196)	-5.70 (0.024)	-0.93 (1570.6)
	296	-1.42 (0.111)	-1.07 (0.025)	-0.90 (310.3)
	299	-1.30 (0.181)	-0.56 (0.073)	-0.95 (5482.3)
	300	-0.71 (0.333)	-0.64 (0.093)	-0.90 (1376.7)
	302	-2.61 (0.032)	-4.76 (0.004)	-0.95 (2389.7)
	303	-1.30 (0.221)	-0.87 (0.062)	-0.92 (12407.2)
	306	-2.61 (0.032)	-0.90 (0.044)	-0.95 (4520)
	313	-1.42 (0.111)	-5.37 (0.010)	-0.90 (172.7)
	316	-2.13 (0.040)	-5.38 (0.002)	-0.90 (1959.8)
	319	-1.26 (0.183)	-1.01 (0.024)	-0.90 (2596.7)
	330	-2.61 (0.048)	-5.86 (0.003)	-0.92 (5055.6)
	334	-1.42 (0.111)	-5.26 (0.011)	-0.90 (174.8)
	335	-0.71 (0.333)	-0.49 (0.096)	-0.91 (1882)
	336	-1.30 (0.196)	-5.71 (0.024)	-0.93 (1553)
	345	-2.32 (0.086)	-0.77 (0.214)	-0.51 (227.9)
	350	-1.30 (0.221)	-0.87 (0.062)	-0.92 (12407.2)
	356	-1.30 (0.221)	-0.87 (0.062)	-0.92 (12407.2)
	358	-0.71 (0.333)	-0.68 (0.081)	-0.90 (460.2)
	363	-1.30 (0.196)	-1.18 (0.047)	-0.93 (1811.2)
	369	-0.71 (0.333)	-0.56 (0.091)	-0.91 (939)
	375	-1.30 (0.181)	-5.51 (0.012)	-0.96 (8588.1)
	377	-0.71 (0.333)	-0.54 (0.096)	-0.90 (374)
	388	-1.42 (0.111)	-5.69 (0.006)	-0.92 (544.2)
	389	-1.42 (0.111)	-1.14 (0.018)	-0.92 (518.9)
	394	-1.37 (0.154)	-1.68 (0.014)	-0.91 (839.9)
	396	-1.30 (0.181)	-2.85 (0.013)	-0.95 (3031.6)
	401	-1.30 (0.221)	-1.17 (0.047)	-0.92 (10832.3)
	412	-1.42 (0.111)	-5.93 (0.003)	-0.92 (299.9)
	424	-1.30 (0.181)	-2.85 (0.013)	-0.95 (3031.6)
	427	-0.71 (0.333)	-0.54 (0.096)	-0.90 (374)
	428	-1.30 (0.221)	-1.17 (0.048)	-0.92 (10773.1)
	432	-2.61 (0.032)	-4.77 (0.005)	-0.93 (877.7)
	449	-1.30 (0.181)	-0.84 (0.047)	-0.96 (12112.7)
	457	-1.30 (0.181)	-0.84 (0.057)	-0.93 (1661.9)
	464	-0.71 (0.333)	-0.47 (0.098)	-0.92 (1506.4)
	503	-0.94 (0.249)	-7.42 (0.018)	-0.95 (2426.3)
	504	-1.30 (0.221)	-1.17 (0.047)	-0.92 (10832.3)
	514	-1.42 (0.125)	-2.17 (0.009)	-0.91 (596.6)
	516	-1.61 (0.085)	-1.13 (0.025)	-0.92 (232.3)
	524	-0.71 (0.333)	-0.73 (0.075)	-0.91 (855.8)
	527	-2.61 (0.032)	-5.58 (0.002)	-0.96 (6125.1)
	535	-1.30 (0.181)	-0.49 (0.072)	-0.96 (14641.4)
	536	-1.42 (0.111)	-0.95 (0.025)	-0.91 (1548.1)
	539	-1.39 (0.118)	-1.56 (0.011)	-0.91 (3106.2)
	548	-1.42 (0.117)	-5.24 (0.016)	-0.91 (571.9)
	554	-1.01 (0.329)	-5.16 (0.068)	-0.51 (172.6)
cesT				
Negative sites	26	-7.31 (0.087)	-17.53 (0.053)	-6.98 (5.3)
	83	-7.31 (0.094)	-4.89 (0.096)	-6.96 (4.8)

eae				
Positive sites	858	2.21 (0.225)	1.27 (0.098)	2.11 (128.6)
	869	1.51 (0.428)	1.11 (0.083)	1.17 (106.2)
	884	1.72 (0.315)	0.71 (0.083)	1.21 (458.4)
Negative sites	141	-3.92 (0.037)	-4.05 (0.005)	-5.28 (15.4)
	509	-4.20 (0.32)	-3.46 (0.007)	-4.80 (11.8)
	553	-3.92 (0.55)	-5.04 (0.003)	-6.18 (16)
	607	-1.02 (0.437)	-11.72 (0.0003)	-9.89 (61.5)
	625	-3.92 (0.042)	-5.04 (0.003)	-6.19 (17)
	667	-4.20 (0.036)	-5.01 (0.003)	-6.16 (17)
	680	-3.92 (0.050)	-4.05 (0.008)	-5.25 (9.5)
	682	-3.42 (0.037)	-3.48 (0.005)	-5.71 (12.1)
	690	-3.42 (0.037)	-3.48 (0.003)	-5.73 (16.4)
	694	-3.92 (0.055)	-5.04 (0.003)	-6.18 (16)
	698	-3.92 (0.036)	-4.05 (0.005)	-5.27 (12.9)
	731	-3.42 (0.037)	-3.49 (0.004)	-5.73 (13.7)
	748	-3.42 (0.037)	-2.52 (0.0028)	-6.73 (21.1)
	791	-3.42 (0.038)	-2.78 (0.005)	-5.28 (11.2)
	793	-3.44 (0.098)	-4.006 (0.050)	-4.81 (1)
	837	-4.20 (0.045)	-3.46 (0.011)	-4.78 (8.4)
	917	-3.74 (0.086)	-4.44 (0.045)	-4.44 (1.1)
escD				
Negative sites	214	-6.29 (0.062)	-27.21 (0.006)	-5.16 (0)
	382	-8.43 (0.093)	-13.21 (0.029)	-5.17 (0)
sepL				
Negative sites	24	-2.40 (0.265)	-230.50 (0.010)	-5.14 (248542)
	27	-4.80 (0.101)	-8.54 (0.057)	-5.15 (13052.7)
	39	-3.48 (0.123)	-16.73 (0.020)	-5.17 (804219)
	49	-1.74 (0.333)	-3.41 (0.079)	-5.20 ($2.9x10^6$)
	53	-1.63 (0.425)	-21.85 (0.058)	-4.29 (2159.7)
	70	-1.74 (0.333)	-4.44 (0.052)	-5.21 ($6.4x10^6$)
	71	-2.40 (0.241)	-2.40 (0.092)	-5.23 (2920.2)
	73	-2.02 (0.290)	-2.56 (0.090)	-5.22 (1457.3)
	78	-3.48 (0.111)	-3.67 (0.047)	-5.21 ($8.4x10^6$)
	94	-1.61 (0.360)	-5.11 (0.075)	-5.18 (203813)
	99	-2.40 (0.253)	-4.08 (0.094)	-5.17 (1563.1)
	106	-2.40 (0.253)	-9.71 (0.055)	-5.17 (3860.2)
	113	-2.40 (0.318)	-9.71 (0.079)	-5.15 (3709.9)
	119	-1.74 (0.333)	-3.30 (0.093)	-5.20 (426628)
	123	-4.80 (0.058)	-6.32 (0.036)	-5.19 (14284.1)
	126	-4.80 (0.064)	-8.54 (0.037)	-5.18 (13685.7)
	128	-2.40 (0.253)	-9.71 (0.052)	-5.18 (3947.2)
	131	-3.48 (0.111)	-8.85 (0.010)	-5.20 ($2.5x10^6$)
	137	-2.40 (0.275)	-5.73 (0.067)	-5.14 (190517)
	140	-2.40 (0.253)	-15.81 (0.063)	-5.17 (5004.1)
	141	-2.40 (0.241)	-22.89 (0.016)	-5.19 (7339.4)
	150	-4.80 (0.070)	-9.42 (0.034)	-5.15 (467582)
	160	-2.40 (0.253)	-4.08 (0.090)	-5.18 (1616.5)
	161	-2.40 (0.241)	-22.89 (0.015)	-5.19 (7516.3)
	166	-1.74 (0.333)	-3.17 (0.079)	-5.20 ($3x10^6$)
	174	-1.74 (0.333)	-6.73 (0.042)	-5.21 (6992.8)
	177	-2.40 (0.241)	-22.89 (0.015)	-5.19 (7424.5)
	180	-1.74 (0.333)	-3.17 (0.075)	-5.21 ($1.6x10^6$)
	185	-2.40 (0.253)	-5.53 (0.089)	-5.17 (3286.6)

	191	-1.74 (0.333)	-4.44 (0.055)	-5.20 ($3.3x10^6$)
	199	-1.74 (0.343)	-7.06 (0.056)	-5.20 (122191)
	209	-2.02 (0.287)	-4.08 (0.079)	-5.18 (3553.5)
	213	-3.48 (0.111)	-18.74 (0.003)	-5.20 ($6.7x10^6$)
	216	-2.40 (0.253)	-15.81 (0.061)	-5.18 (5097.9)
	220	-4.80 (0.064)	-14.76 (0.016)	-5.18 (436679)
	221	-1.74 (0.333)	-5.05 (0.066)	-5.18 (862646)
	223	-2.36 (0.246)	-13.45 (0.036)	-5.18 (389.3)
	234	-2.61 (0.259)	-209.50 (0.043)	-4.28 (28284.6)
	238	-4.80 (0.058)	-6.32 (0.034)	-5.19 (14480.3)
	239	-4.80 (0.064)	-14.76 (0.016)	-5.18 (436679)
	240	-1.74 (0.333)	-3.41 (0.098)	-5.18 (32943.9)
	245	-2.02 (0.287)	-4.08 (0.079)	-5.18 (3553.5)
	247	-2.40 (0.241)	-11.28 (0.028)	-5.23 (11772.6)
	251	-3.30 (0.123)	-56.13 (0.017)	-5.20 (9627.1)
	261	-2.40 (0.266)	-8.15 (0.075)	-5.15 (416233)
	264	-2.02 (0.287)	-4.08 (0.079)	-5.18 (3553.5)
	268	-1.74 (0.333)	-5.82 (0.051)	-5.19 ($2.5x10^6$)
	269	-2.36 (0.246)	-13.45 (0.036)	-5.18 (389.3)
	281	-3.42 (0.159)	-12.32 (0.021)	-5.14 (107896)
	282	-1.74 (0.334)	-33.17 (0.039)	-5.19 (83021.3)
	287	-1.74 (0.333)	-29.08 (0.067)	-5.18 (178682)
	289	-2.27 (0.298)	-5.49 (0.083)	-5.14 (97100.7)
	290	-1.74 (0.333)	-4.27 (0.069)	-5.20 (84751.1)
	294	-3.22 (0.160)	-32.71 (0.006)	-5.15 (194426)
	296	-2.40 (0.253)	-5.53 (0.085)	-5.18 (3559.7)
	297	-4.80 (0.058)	-9.42 (0.010)	-5.23 ($1.9x10^6$)
	303	-2.36 (0.246)	-13.45 (0.036)	-5.18 (389.3)
	313	-2.40 (0.241)	-12.71 (0.027)	-5.19 (4654.2)
	323	-3.48 (0.111)	-36.36 (0.001)	-5.20 ($5.4x10^6$)
	324	-2.40 (0.241)	-12.71 (0.026)	-5.19 (4667.8)
	326	-2.40 (0.253)	-22.27 (0.017)	-5.20 (12612.4)
	331	-4.80 (0.064)	-8.54 (0.035)	-5.18 (13964)
	333	-2.02 (0.300)	-8.93 (0.038)	-5.21 (6104.5)
	340	-2.02 (0.287)	-4.08 (0.079)	-5.18 (3553.5)
	341	-1.74 (0.336)	-6.73 (0.042)	-5.21 (7342)
	343	-3.48 (0.111)	-29.13 (0.007)	-5.18 (527227)
	344	-4.22 (0.075)	-52.35 (0.004)	-5.19 (301627)
	345	-2.40 (0.241)	-22.27 (0.020)	-5.19 (5222.8)
espA				
Negative sites	23	-4.17 (0.041)	-4.51 (0.026)	-3.18 (8.9)
	25	-3.81 (0.038)	-303.98 (0.001)	-3.92 (65.9)
	94	-4.17 (0.041)	-191.02 (0.004)	-3.64 (24.2)
	140	-3.81 (0.037)	-1303.1 (0.001)	-3.91 (63.5)
	148	-4.17 (0.041)	-4.51 (0.029)	-3.17 (8.2)
	182	-4.17 (0.054)	-42.68 (0.015)	-3.48 (12.7)
espD				
Positive sites	349	3.24 (0.165)	8.81 (0.005)	3.84 (191.6)
Negative sites	2	-1.26 (0.358)	-2.44 (0.083)	-2.45 (143.7)
	3	-2.08 (0.216)	-3.37 (0.052)	-2.48 (159.6)
	14	-1.35 (0.334)	-1.77 (0.095)	-2.46 (173.8)
	16	-5.67 (0.034)	-11.85 (0.054)	-1.84 (25)
	34	-2.08 (0.216)	-12.80 (0.030)	-2.47 (144.3)
	59	-2.71 (0.111)	-704.49 (0.003)	-2.46 (178.9)

	72	-1.85 (0.241)	-3.26 (0.034)	-2.42 (91.1)
	73	-2.08 (0.216)	-12.80 (0.023)	-2.52 (325.1)
	74	-2.08 (0.230)	-15.14 (0.024)	-2.48 (188.2)
	75	-1.81 (0.255)	-3.37 (0.040)	-2.42 (83.3)
	76	-2.08 (0.216)	-12.80 (0.027)	-2.49 (238.8)
	80	-2.08 (0.216)	-12.80 (0.023)	-2.52 (325.1)
	86	-2.43 (0.188)	-5.75 (0.017)	-2.42 (106.2)
	87	-2.71 (0.111)	-15.91 (0.004)	-2.47 (160.4)
	91	-2.08 (0.230)	-2.19 (0.079)	-2.45 (104.7)
	96	-2.71 (0.111)	-17.14 (0.003)	-2.47 (160.4)
	122	-2.71 (0.111)	-3.12 (0.022)	-2.47 (164.6)
	137	-2.08 (0.230)	-8.29 (0.052)	-2.48 (167)
	146	-1.35 (0.334)	-1.77 (0.095)	-2.46 (173.8)
	155	-1.35 (0.333)	-1.65 (0.096)	-2.47 (155.1)
	160	-4.17 (0.053)	-9.96 (0.007)	-2.48 (188)
	166	-2.08 (0.276)	-14.99 (0.036)	-2.42 (107)
	183	-1.35 (0.333)	-10.91 (0.064)	-2.46 (158.5)
	186	-2.71 (0.111)	-21.62 (0.010)	-2.46 (172.7)
	187	-2.71 (0.111)	-1.59 (0.098)	-2.46 (137.8)
	188	-2.71 (0.111)	-3.49 (0.022)	-2.45 (115.5)
	190	-1.35 (0.333)	-2.44 (0.066)	-2.47 (156)
	193	-1.35 (0.333)	-1.76 (0.093)	-2.44 (112.8)
	198	-2.71 (0.111)	-2.92 (0.030)	-2.44 (129.3)
	215	-1.35 (0.333)	-2.44 (0.066)	-2.47 (136.8)
	219	-2.71 (0.111)	-17.23 (0.049)	-2.46 (137)
	225	-2.71 (0.111)	-6.18 (0.009)	-2.44 (129.2)
	229	-4.17 (0.053)	-11.97 (0.031)	-2.48 (185.1)
	230	-1.35 (0.333)	-1.76 (0.084)	-2.47 (155.7)
	232	-2.71 (0.111)	-2.92 (0.028)	-2.45 (115.4)
	233	-2.08 (0.276)	-2.19 (0.097)	-2.42 (84.6)
	237	-2.71 (0.122)	-17.23 (0.049)	-2.46 (149.2)
	255	-1.18 (0.382)	-2.005 (0.090)	-2.44 (132)
	262	-2.71 (0.111)	-3.62 (0.022)	-2.46 (133.6)
	271	-4.07 (0.037)	-2.41 (0.064)	-2.46 (132.1)
	274	-1.35 (0.333)	-1.73 (0.087)	-2.46 (137.7)
	284	-4.17 (0.047)	-12.05 (0.036)	-2.49 (213.9)
	293	-1.35 (0.333)	-1.76 (0.093)	-2.44 (112.8)
	295	-4.07 (0.041)	-18.78 (0.006)	-2.46 (162.5)
	297	-2.71 (0.111)	-17.08 (0.049)	-2.47 (159.1)
	301	-1.35 (0.333)	-1.71 (0.090)	-2.46 (200)
	304	-2.71 (0.111)	-17.08 (0.050)	-2.46 (162.1)
	308	-1.81 (0.255)	-3.37 (0.040)	-2.42 (83.3)
	326	-4.07 (0.037)	-2.32 (0.068)	-2.46 (133.8)
	328	-4.17 (0.047)	-12.05 (0.036)	-2.49 (213.9)
	332	-4.17 (0.076)	-11.97 (0.043)	-2.42 (105.2)
	334	-2.08 (0.230)	-15.14 (0.027)	-2.46 (136.2)
	336	-5.38 (0.020)	-6.01 (0.010)	-2.42 (86.6)
	343	-2.08 (0.230)	-6.40 (0.063)	-2.45 (124.1)
	347	-2.71 (0.111)	-5.50 (0.011)	-2.45 (114.5)
	350	-2.08 (0.230)	-6.40 (0.063)	-2.45 (124.1)
	351	-2.99 (0.112)	-2.67 (0.043)	-2.42 (86)
	353	-2.08 (0.216)	-1.47 (0.093)	-2.48 (146.3)
	355	-4.07 (0.037)	-12.56 (0.005)	-2.45 (115.5)
	359	-4.17 (0.076)	-11.72 (0.043)	-2.42 (104.6)

	364	-4.71 (0.035)	-9.21 (0.005)	-2.41 (79.3)
	367	-4.17 (0.047)	-13.59 (0.007)	-2.47 (158.8)
	368	-2.71 (0.136)	-17.23 (0.057)	-2.43 (105.6)
	369	-2.71 (0.111)	-1.78 (0.092)	-2.45 (115.8)
	371	-2.71 (0.111)	-1.78 (0.088)	-2.46 (131)
	376	-1.35 (0.333)	-1.73 (0.087)	-2.46 (137.7)
	377	-2.08 (0.216)	-15.92 (0.011)	-2.52 (382.4)
espB				
Negative sites	4	-1.03 (0.272)	-0.22 (0.073)	-1.26 (228.7)
	20	-0.85 (0.333)	-0.59 (0.057)	-1.23 (64)
	38	-1.28 (0.227)	-0.38 (0.030)	-1.23 (73)
	40	-0.85 (0.333)	-0.30 (0.056)	-1.23 (73.8)
	44	-1.70 (0.111)	-0.53 (0.014)	-1.24 (82.5)
	48	-1.33 (0.225)	-0.43 (0.053)	-1.25 (156.1)
	52	-1.33 (0.225)	-0.36 (0.063)	-1.24 (103.2)
	61	-0.85 (0.333)	-0.21 (0.091)	-1.24 (88)
	65	-1.33 (0.268)	-0.37 (0.076)	-1.22 (138)
	69	-1.33 (0.268)	-0.47 (0.065)	-1.22 (139.9)
	71	-1.33 (0.268)	-0.65 (0.039)	-1.22 (130.6)
	74	-1.33 (0.268)	-0.65 (0.039)	-1.22 (130.6)
	79	-1.33 (0.268)	-0.43 (0.070)	-1.22 (135.1)
	86	-1.33 (0.268)	-0.37 (0.076)	-1.22 (138)
	92	-1.33 (0.225)	-0.36 (0.063)	-1.24 (103.2)
	94	-2.67 (0.050)	-2.59 (0.002)	-1.25 (134.1)
	98	-0.85 (0.333)	-0.24 (0.091)	-1.23 (67.5)
	103	-0.85 (0.333)	-0.23 (0.082)	-1.24 (221.1)
	104	-0.85 (0.333)	-0.28 (0.075)	-1.25 (143.1)
	106	-1.03 (0.272)	-0.22 (0.073)	-1.26 (228.7)
	112	-1.03 (0.272)	-0.18 (0.089)	-1.26 (232.1)
	120	-1.33 (0.211)	-0.25 (0.079)	-1.25 (166.5)
	125	-1.70 (0.111)	-3.01 (0.002)	-1.25 (135)
	185	-1.70 (0.112)	-3.01 (0.011)	-1.22 (106.8)
	190	-1.33 (0.211)	-0.75 (0.023)	-1.24 (98.4)
	207	-2.67 (0.044)	-2.23 (0.004)	-1.24 (94.3)
	233	-0.85 (0.333)	-0.28 (0.073)	-1.23 (69.6)
	238	-1.70 (0.111)	-29.6 (0.0003)	-1.24 (84.7)
	240	-1.33 (0.211)	-0.31 (0.051)	-1.27 (634.9)
	244	-0.85 (0.333)	-0.30 (0.052)	-1.25 (219.2)
	246	-1.33 (0.225)	-0.65 (0.029)	-1.25 (148.2)
	254	-1.33 (0.225)	-0.43 (0.057)	-1.24 (104.2)
	255	-1.70 (0.113)	-0.89 (0.015)	-1.22 (79.4)
	257	-1.33 (0.225)	-0.36 (0.058)	-1.25 (154.4)
	258	-1.70 (0.113)	-0.46 (0.025)	-1.22 (88.2)
	264	-1.33 (0.225)	-0.36 (0.063)	-1.24 (103.2)
	265	-1.33 (0.268)	-0.47 (0.065)	-1.22 (139.9)
	268	-1.33 (0.225)	-0.37 (0.057)	-1.25 (155.3)
	269	-1.70 (0.124)	-3.36 (0.002)	-1.22 (75.9)
	271	-1.33 (0.225)	-0.65 (0.031)	-1.24 (99.6)
	272	-1.33 (0.211)	-0.25 (0.079)	-1.25 (166.5)
	278	-1.28 (0.227)	-0.38 (0.030)	-1.23 (73)
	281	-1.33 (0.241)	-0.31 (0.082)	-1.22 (137)
	282	-1.33 (0.225)	-0.65 (0.029)	-1.25 (148.2)
	283	-1.33 (0.265)	-0.36 (0.077)	-1.22 (135)
	295	-0.85 (0.333)	-0.31 (0.073)	-1.24 (468.8)

	299	-0.28 (0.018)	-0.04 (0.870)	-1.12 (76.9)
	305	-1.33 (0.211)	-2.42 (0.018)	-1.24 (109.4)
	312	-0.85 (0.333)	-0.41 (0.041)	-1.24 (464.5)
	314	-0.85 (0.333)	-0.59 (0.054)	-1.25 (145.8)
orf27				
Negative sites	42	-11.16 (0.056)	-10.94 (0.033)	-14.50 (0)
	81	-11.16 (0.069)	-10.94 (0.040)	-14.45 (0)
escF				
No sites				
orf29				
Negative sites	20	-4.96 (0.051)	-3.15 (0.040)	-2.15 (23)
	51	-4.02 (0.078)	-1.90 (0.063)	-2.15 (21.8)
	74	-2.48 (0.227)	-1.60 (0.068)	-2.22 (99.1)
	79	-2.48 (0.227)	-3.14 (0.037)	-2.22 (92.6)
espF				
Positive sites	136	2.13 (0.247)	1.67 (0.049)	4.34 (57.4)
Negative sites	21	-4.17 (0.038)	-2.32 (0.018)	-1.42 (413.3)
	23	-2.45 (0.111)	-0.73 (0.083)	-1.39 (197.3)
	33	-2.45 (0.111)	-1.51 (0.015)	-1.39 (163.8)
	37	-3.68 (0.037)	-2.35 (0.011)	-1.34 (38.7)
	38	-2.45 (0.111)	-3.73 (0.005)	-1.39 (173.6)
	59	-1.54 (0.283)	-0.82 (0.088)	-1.38 (168.7)
	65	-2.08 (0.196)	-1.51 (0.035)	-1.42 (428.8)
	69	-1.22 (0.333)	-0.90 (0.088)	-1.36 (56)
	103	-4.17 (0.038)	-4.07 (0.014)	-1.39 (215.2)
	104	-2.08 (0.198)	-0.96 (0.042)	-1.37 (97)
	122	-2.45 (0.111)	-0.85 (0.084)	-1.36 (94.7)
	150	-2.23 (0.196)	-416.68 (0.001)	-1.38 (158.7)
	167	-2.45 (0.111)	-2.52 (0.012)	-1.36 (106.7)
	197	-4.17 (0.038)	-4.07 (0.014)	-1.39 (215.2)
	198	-2.08 (0.198)	-0.96 (0.042)	-1.37 (97)
	202	-2.08 (0.210)	-3.19 (0.041)	-1.38 (152.7)

APÉNDICE III.

Tabla 4.4. Resultados del test de SPD para cada marcador utilizado en el estudio.

Gen	Codon	SLAC (p)	FEL (p)	REL (Bayes Factor)
mdh				
Positive sites	207	2.8781 (0.7514)	83.7293 (0.0001)	14.8531 (1.8317×10^{11})
Negative sites				
	9	-1.4591 (0.1289)	-9.0446 (0.0080)	-0.7359 (3.1180×10^{6})
	10	-1.3633 (0.1909)	-5.9458 (0.0514)	-0.7291 (3.4074×10^{10})
	12	-1.3633 (0.2297)	-6.8092 (0.0596)	-0.7490 (3.4047×10^{9})
	20	-1.4390 (0.1127)	-6.3639 (0.0105)	-0.7153 (9.1551×10^{9})
	24	-0.8449 (0.2838)	-2.6899 (0.0964)	-0.7038 (5.7687×10^{9})
	26	-0.7195 (0.3333)	-3.0561 (0.0843)	-0.6977 (1.431×10^{9})
	28	-0.7195 (0.3333)	-2.7433 (0.0791)	-0.7001 (2.9120×10^{9})
	29	-2.1586 (0.0370)	-9.1433 (0.0039)	-0.7297 (7.1879×10^{7})
	32	-2.8781 (0.0123)	-30.6251 (0.0002)	-6.3902 (3.5367×10^{7})

	34	-1.0232 (0.2343)	-4.4494 (0.0697)	-0.7398 (1.8644×10^{8})
	35	-0.9228 (0.2724)	-5.3316 (0.0311)	-0.6980 (1.0298×10^{7})
	37	-1.0232 (0.2343)	-4.4492 (0.0603)	-0.7405 (3.0697×10^{9})
	39	-0.7195 (0.3333)	-4.6348 (0.0695)	-0.7052 (3.2136×10^{8})
	40	-2.1586 (0.0370)	-9.3714 (0.0037)	-0.7499 (2.5667×10^{8})
	42	-1.4390 (0.1111)	-8.0109 (0.0150)	-0.7095 (4.7815×10^{6})
	47	-1.0232 (0.2343)	-2.9560 (0.0787)	0.7291 (1.9827×10^{11})
	48	-1.4390 (0.1114)	-5.3141 (0.0205)	-0.7462 (2.0508×10^{10})
	52	-2.1586 (0.0370)	-10.816 (0.0039)	-0.7479 (6.2171×10^{6})
	54	-0.7195 (0.3333)	-5.2480 (0.0635)	-0.7064 (2.9797×10^{8})
	55	-0.7195 (0.3333)	-3.2160 (0.0907)	-0.7002 (4.1162×10^{7})
	56	-0.9345 (0.2634)	-6.7722 (0.0209)	-0.7010 (8.2167×10^{6})
	59	-1.4390 (0.1111)	-8.1866 (0.0120)	-0.7107 (1.4420×10^{7})
	61	-1.4390 (0.1111)	-7.9626 (0.0127)	-0.7122 (2.0477×10^{6})
	66	-1.4390 (0.1111)	-9.3459 (0.0118)	-0.7234 (8.2619×10^{6})
	67	-1.4390 (0.1111)	-10.5728 (0.0286)	-0.9030 (5.8108×10^{8})
	68	-1.4390 (0.1111)	-8.1865 (0.0138)	-0.7099 (4.3021×10^{6})
	69	-0.7195 (0.3333)	-2.9787 (0.0960)	-0.6972 (9.9726×10^{7})
	71	-2.7266 (0.0364)	-15.8105 (0.0042)	-0.9352 (6.721×10^{9})
	72	-0.7195 (0.3333)	-2.7953 (0.0911)	-0.6971 (1.5051×10^{9})
	73	-1.4390 (0.1111)	-3.9063 (0.0617)	-0.7013 (1.8928×10^{8})
	79	-0.8809 (0.3038)	-4.1538 (0.0459)	-0.6947 (2.7563×10^{7})
	80	-1.0232 (0.2343)	-4.4494 (0.0383)	-0.7424 (1.4571×10^{11})
	82	-1.4390 (0.1111)	-6.2370 (0.0240)	-0.7153 (3.6780×10^{6})
	87	-0.7195 (0.3333)	-3.1857 (0.0973)	-0.6988 (1.4934×10^{7})
	89	-2.0465 (0.0549)	-6.7188 (0.0263)	-0.8505 (1.1258×10^{10})
	91	-1.8286 (0.1074)	-11.5407 (0.0494)	-0.5639 (4605.7)
	93	-2.7266 (0.0528)	-15.8105 (0.0064)	-0.9332 (9.4794×10^{8})
	95	-1.0791 (0.2593)	-9.6093 (0.0664)	-0.5218 (5487.7)
	99	-0.7195 (0.3333)	-2.7834 (0.0910)	-0.7005 (6.2877×10^{9})
	101	-1.4390 (0.1111)	-9.8951 (0.0105)	-0.7159 (1.0287×10^{7})
	103	-1.6898 (0.0805)	-4.8176 (0.0236)	-0.7213 (2.6030×10^{9})
	105	-0.8449 (0.2838)	-3.4745 (0.0609)	-0.7100 (3.8569×10^{9})
	106	-1.6898 (0.0805)	-4.6214 (0.0275)	-0.7421 (7.9275×10^{9})
	110	-1.4390 (0.1111)	-8.3632 (0.0161)	-0.7264 (2.1782×10^{6})
	112	-1.4390 (0.1111)	-6.7300 (0.0190)	-0.7226 (1.8652×10^{9})
	113	-2.1586 (0.0370)	-16.7126 (0.0009)	-1.0059 (5.4708×10^{8})
	114	-1.4390 (0.1111)	-6.3638 (0.0201)	-0.7128 (4.6384×10^{6})
	115	-1.4390 (0.1111)	-17.5873 (0.0077)	-0.8220 (2.2537×10^{7})
	116	-0.8449 (0.2888)	-3.4745 (0.0609)	-0.7100 (3.8569×10^{9})
	117	-1.4390 (0.1111)	-7.8924 (0.0128)	-0.7216 (1.0291×10^{7})
	118	-1.4390 (0.1111)	-7.4643 (0.0148)	-0.7187 (1.13×10^{7})
	125	-1.4390 (0.1111)	-3.8514 (0.0625)	-0.7010 (1.9084×10^{8})
	127	-2.0465 (0.0676)	-6.3040 (0.0219)	-0.7832 (8.1170×10^{9})
	128	-2.0465 (0.0549)	-5.6554 (0.0342)	-0.7443 (1.5726×10^{8})
	129	-1.3633 (0.1909)	-9.9814 (0.0335)	-0.7757 (1.9436×10^{10})
	131	-1.3633 (0.1909)	-5.9458 (0.0514)	-0.7291 (3.4074×10^{10})
	133	-2.0465 (0.0549)	-5.4305 (0.0148)	-0.7429 (1.3579×10^{11})
	135	-2.1586 (0.0370)	-9.3716 (0.0066)	-0.7463 (1.6507×10^{6})
	136	-1.4390 (0.1111)	-4.5806 (0.0308)	-0.7185 (3.3983×10^{9})
	137	-0.7195 (0.3333)	-2.7834 (0.0879)	-0.7016 (1.5099×10^{10})
	142	-1.4390 (0.1111)	-3.8401 (0.050)	-0.6984 (3.1347×10^{9})
	145	-1.4390 (0.1111)	-3.7902 (0.0382)	-0.7450 (9.5167×10^{9})
	146	-1.0232 (0.2343)	-4.7252 (0.0567)	-0.7477 (1.5250×10^{11})

	147	-2.1586 (0.0370)	-9.3716 (0.0066)	-0.7463 (1.6507×10^{6})
	148	-1.4390 (0.1111)	-10.5178 (0.0098)	-0.7278 (7.8312×10^{9})
	149	-1.3633 (0.1909)	-5.9459 (0.0616)	-0.7273 (6.7449×10^{8})
	152	-1.4390 (0.1111)	-3.6796 (0.0662)	-0.6977 (1.9335×10^{8})
	160	-0.7195 (0.3333)	-3.1857 (0.0973)	-0.6988 (1.4934×10^{7})
	163	-0.8449 (0.2838)	-3.4745 (0.0609)	-0.7100 (3.8569×10^{9})
	164	-2.1586 (0.0370)	-22.048 (0.0066)	-1.5916 (5.8423×10^{8})
	165	-1.4390 (0.1111)	-3.8514 (0.0625)	-0.7010 (1.9084×10^{8})
	168	-1.4390 (0.1111)	-3.8514 (0.0625)	-0.7010 (1.9084×10^{8})
	169	-1.4390 (0.1111)	-6.3961 (0.0228)	-0.7159 (3.6822×10^{6})
	173	-1.4390 (0.1111)	-8.8127 (0.0097)	-0.7208 (1.9121×10^{8})
	174	-0.8809 (0.3038)	-4.1538 (0.0459)	-0.6947 (2.7563×10^{7})
	176	-1.4390 (0.1111)	-7.7751 (0.0101)	-0.7121 (8.1077×10^{8})
	181	-1.4390 (0.1111)	-9.2683 (0.0120)	-0.7332 (1.7830×10^{6})
	182	-3.0697 (0.0128)	-31.5276 (0.0006)	-4.7464 (1.8431×10^{11})
	183	-1.0232 (0.0787)	-2.9560 (0.0787)	-0.7291 (1.9827×10^{11})
	185	-1.3633 (0.1909)	-7.7091 (0.0494)	-0.7637 (9.4469×10^{8})
	186	-1.3633 (0.2297)	-7.7088 (0.0513)	-0.7634 (6.1276×10^{9})
	187	-1.3633 (0.1909)	-5.9459 (0.0616)	-0.7273 (6.7449×10^{8})
	188	-0.7195 (0.3333)	-3.1857 (0.0973)	-0.6981 (1.6109×10^{7})
	193	-1.3633 (0.1909)	-6.8104 (0.481)	-0.7511 (2.5379×10^{10})
	194	-1.4390 (0.1111)	-5.6612 (0.0259)	-0.7187 (4.4408×10^{9})
	195	-1.6898 (0.0805)	-4.3905 (0.0268)	-0.7101 (3.6950×10^{9})
	196	-2.7266 (0.0528)	-17.3104 (0.0056)	-1.0325 (1.5333×10^{9})
	198	-2.1586 (0.0370)	-21.2539 (0.0020)	-1.1528 (6.5646×10^{7})
	201	-6.8165 (0.0002)	-308.59(3.6169×10^{8})	-10.5962 (1.0066×10^{9})
	202	-0.7195 (0.3333)	-3.3075 (0.0928)	-0.6963 (2.2834×10^{7})
	203	-0.7195 (0.3333)	-4.6349 (0.0826)	-0.7037 (6.0240×10^{6})
	204	-1.3633 (0.1909)	-5.9459 (0.6162)	-0.7273 (6.7449×10^{8})
	205	-0.7195 (0.3333)	-3.3074 (0.0663)	-0.6971 (6.9636×10^{7})
	206	-2.7266 (0.0364)	-29.7064 (0.0011)	-1.9673 (8.8863×10^{9})
	210	-2.1586 (0.0383)	-10.8157 (0.0077)	-0.7475 (2.6208×10^{7})
	211	-1.4390 (0.1129)	-9.1257 (0.1129)	-0.7688 (9.5816×10^{9})
	214	-0.9345 (0.2624)	-6.5254 (0.0245)	-0.6994 (9.7393×10^{6})
	219	-2.1586 (0.0370)	-17.1178 (0.0013)	-1.0614 (1.0506×10^{7})
	220	-1.4390 (0.1111)	-8.0136 (0.0125)	-0.7227 (9.8716×10^{6})
	223	-2.0465 (0.0549)	-6.1422 (0.0125)	-0.7818 (9.6295×10^{10})
	224	-2.1586 (0.0370)	-10.6978 (0.0134)	-1.0509 (3.4288×10^{8})
	226	-1.4390 (0.1142)	-9.7123 (0.0083)	-0.7435 (3.2944×10^{9})
	228	-3.5976 (0.0041)	-40.4244 (6.31×10^{5})	-9.0649 (7.9002×10^{6})
	229	-1.4390 (0.1111)	-5.1715 (0.0151)	-0.7086 (5.7391×10^{9})
	230	-1.4390 (0.1111)	-7.1907 (0.0154)	-0.7089 (1.5219×10^{7})
	232	-1.3633 (0.2297)	-9.3215 (0.0478)	-0.9147 (1.1398×10^{10})
	234	-1.3633 (0.1909)	-9.9815 (0.0403)	-0.7739 (4.0446×10^{8})
	237	-1.4390 (0.1111)	-8.3437 (0.0261)	-0.7246 (2.1862×10^{6})
	238	-1.4390 (0.1111)	-8.2129 (0.0110)	-0.7144 (1.9185×10^{6})
	240	-2.0465 (0.0608)	-6.3038 (0.0169)	-0.7859 (5.5498×10^{11})
	241	-0.7195 (0.3333)	-3.6013 (0.0995)	-0.6983 (2.3166×10^{7})
	242	-3.0697 (0.0176)	-11.7132 (0.0042)	-1.3617 (8.6127×10^{9})
	250	-4.3171 (0.0013)	-46.7899 (5.36×10^{6})	-10.0772 (3.1409×10^{7})
	251	-1.4390 (0.1111)	-3.8401 (0.0508)	-0.6984 (3.1347×10^{9})
	254	-0.7195 (0.3383)	-2.4095 (0.0946)	-0.7008 (1.9821×10^{10})
	255	-2.7266 (0.0528)	-29.9327 (0.0075)	-1.7283 (1.1387×10^{9})
	257	-1.2999 (0.1712)	-9.4587 (0.0073)	-0.7522 (1.4766×10^{7})

	266	-1.3633 (0.1909)	-13.4422 (0.0215)	-0.8373 (5.4823×10^{8})
	269	-1.0626 (0.2788)	-7.5974 (0.0697)	-0.5201 (3.9677×10^{6})
	270	-1.6898 (0.0805)	-4.5395 (0.0256)	-0.7113 (3.5706×10^{9})
	271	-1.4390 (0.1111)	-5.1579 (0.0497)	-0.7368 (3.6083×10^{8})
	272	-3.3539 (0.0158)	-31.3969 (0.0007)	-8.6447 (776006)
	275	-0.7195 (0.3333)	-3.0561 (0.0937)	-0.6970 (8.9578×10^{7})
gapA				
Negative sites	3	-2.2081 (0.1906)	-6.3967 (0.0214)	-2.9057 (3327.09)
	6	-4.1348 (0.0962)	-8.2450 (0.2317)	-2.1778 (487.011)
	9	-2.5982 (0.1111)	-14.9395 (0.0075)	-2.8792 (6.2448×10^{9})
	13	-2.2263 (0.2270)	-29.0646 (0.0206)	-2.9173 (2389.04)
	14	-1.7774 (0.2674)	-4.5192 (0.0673)	-2.9089 (1.8906×10^{7})
	27	-1.2979 (0.3333)	-5.8744 (0.0565)	-2.8972 (2560.45)
	31	-1.7774 (0.2433)	-4.5202 (0.0804)	-2.9004 (2568.28)
	44	-2.8437 (0.1625)	-12.8605 (0.0328)	-2.9204 (934.66)
	47	-1.2979 (0.3333)	-4.5195 (0.0837)	-2.8731 (3.3634×10^{6})
	51	-2.921 (0.1730)	-19.8711 (0.0259)	-2.9203 (2172.73)
	60	-2.5958 (0.1111)	-8.8028 (0.0173)	-2.8973 (8.4277×10^{8})
	69	-2.5958 (0.1111)	-5.8506 (0.0400)	-2.8991 (4.6679×10^{13})
	73	-2.5958 (0.1111)	-7.6953 (0.0316)	-2.8792 (1.1503×10^{7})
	75	-2.8262 (0.1788)	-12.1984 (0.0310)	-2.9286 (877.93)
	78	-3.2447 (0.1111)	-9.8478 (0.0911)	-2.1785 (2.0963×10^{23})
	80	-2.5958 (0.1111)	-9.8365 (0.0434)	-2.8854 (1.7735×10^{16})
	81	-1.2979 (0.3333)	-5.8744 (0.0700)	-2.8792 (29861.6)
	82	-8.1562 (0.0064)	-175.28(9.4415×10^{6})	-2.9263 (5.7744×10^{10})
	84	-2.5958 (0.1111)	-23.1505 (0.0075)	-2.8973 (7.4086×10^{8})
	89	-2.2979 (0.3333)	-13.2044 (0.0316)	-2.8731 (268806)
	94	-2.2979 (0.3333)	-3.2159 (0.0959)	-2.8991 (9.6225×10^{7})
	95	-2.5958 (0.1111)	-12.8433 (0.0172)	-2.8731 (1.2818×10^{8})
	97	-1.6625 (0.2601)	-5.5863 (0.0675)	-2.9180 (1694.88)
	106	-3.8705 (0.0630)	-7.1098 (0.0268)	-2.9089 (1.2598×10^{11})
	109	-2.2035 (0.2665)	-13.8267 (0.0401)	-2.8966 (640.07)
	115	-2.5958 (0.1111)	-5.6140 (0.0409)	-2.8791 (5.8956×10^{12})
	119	-2.5958 (0.1111)	-5.1198 (0.0407)	-2.8991 (3.2164×10^{12})
	123	-3.8937 (0.0370)	-14.4506 (0.0084)	-2.8792 (3.1225×10^{15})
	124	-1.2979 (0.3333)	-4.4611 (0.0848)	-2.8731 (2.3263×10^{6})
	134	-2.5958 (0.1111)	-7.6733 (0.0500)	-2.8854 (2.0988×10^{14})
	141	-1.2979 (0.3333)	-3.4761 (0.0916)	-2.8991 (9.6540×10^{7})
	144	-3.3791 (0.0655)	-17.8689 (0.0218)	-2.8977 (9.8751×10^{6})
	146	-2.5958 (0.1111)	-8.5427 (0.0197)	-2.8991 (1.6085×10^{17})
	148	-2.5958 (0.1111)	-7.0480 (0.0306)	-2.8792 (1.1951×10^{10})
	152	-2.5958 (0.1111)	-6.9888 (0.0309)	-2.8991 (2.1258×10^{18})
	154	-3.9186 (0.0487)	-15.3284 (0.0200)	-2.9073 (2.5575×10^{6})
	165	-2.5958 (0.1111)	-7.8331 (0.0401)	-2.8796 (3.3742×10^{15})
	167	-3.8937 (0.0370)	-14.7757 (0.0065)	-2.8792 (9.7122×10^{21})
	169	-2.1093 (0.2784)	-15.9798 (0.0415)	-2.8935 (41496.4)
	173	-2.5958 (0.1111)	-10.8917 (0.0132)	-2.8732 (4.2527×10^{7})
	174	-2.5958 (0.1111)	-14.2269 (0.0069)	-2.8944 (2.7343×10^{15})
	178	-2.5958 (0.1111)	-11.4992 (0.0362)	-2.8792 (3.3646×10^{11})
	201	-2.5958 (0.1111)	-19.6018 (0.0050)	-2.8732 (4.4733×10^{8})
	202	-2.5958 (0.1111)	-5.1290 (0.0407)	-2.8991 (3.2449×10^{12})
	203	-1.2979 (0.3333)	-13.2044 (0.0316)	-2.8731 (268806)
	206	-2.5958 (0.1111)	-6.5348 (0.0249)	-2.8960 (3.1663×10^{17})
	215	-2.8262 (0.1635)	-12.1984 (0.0341)	-2.9201 (821.21)

	216	-2.8262 (0.1788)	-12.1984 (0.0310)	-2.9286 (877.93)
	218	-3.8937 (0.0370)	-23.3828 (0.0053)	-2.8792 (1.2765×10^{16})
	220	-3.4100 (0.0760)	-7.4021 (0.0257)	-2.9089 (1.2745×10^{11})
fimA				
Positive sites	19	1.3900 (0.2962)	3.3100 (0.0410)	2.5925 (466.20)
	89	2.1756 (0.1803)	3.8480 (0.0680)	2.2362 (65.50)
	134	2.6215 (0.1248)	5.6681 (0.0287)	2.2617 (70.57)
Negative sites				
	8	-2.1522 (0.0824)	-3.1341 (0.0224)	-3.7820 (939.64)
	18	-1.8534 (0.1111)	-2.8042 (0.0346)	-3.4403 (174.86)
	23	-1.8534 (0.1111)	-2.5092 (0.0597)	-3.2319 (119.29)
	30	-3.7068 (0.0123)	-5.8484 (0.0036)	-4.4142 (4699)
	43	-3.4805 (0.0315)	-3.4688 (0.0181)	-3.9348 (2168.58)
	59	-3.1197 (0.4271)	-6.4875 (0.0025)	-4.3621 (2627.28)
	60	-2.7801 (0.0459)	-3.9647 (0.0316)	-4.1247 (527.38)
	65	-2.7801 (0.0370)	-5.3484 (0.0108)	-4.3166 (1099.34)
	69	-1.7272 (0.2174)	-6.3927 (0.0297)	-3.1629 (131.39)
	72	-1.8534 (0.1111)	-2.5840 (0.0567)	-2.8201 (88.98)
	74	-1.7402 (0.1775)	-2.0856 (0.0916)	-2.688 (97.81)
	76	-1.8534 (0.1111)	-3.8475 (0.0338)	-3.8723 (252.03)
	82	-3.7576 (0.0399)	-12.6292 (0.0021)	-4.2003 (2569.85)
	84	-1.7402 (0.1775)	-2.9930 (0.0561)	-3.0415 (127.30)
	87	-1.7402 (0.1775)	-2.9928 (0.0675)	-3.0381 (101.55)
	94	-1.8534 (0.1111)	-2.5066 (0.0597)	-3.1795 (113.93)
	98	-1.7402 (0.1775)	-1.7999 (0.0827)	-2.4816 (75.23)
	99	-1.4455 (0.2876)	-6.7185 (0.0055)	-4.3538 (97.66)
	100	-1.8534 (0.1111)	-4.0707 (0.0413)	-3.9807 (388.93)
	101	-3.7068 (0.0123)	-10.6845 (0.0004)	-4.4449 (22527.1)
	104	-3.4805 (0.0315)	-5.3275 (0.0126)	-5.3275 (1233.75)
	110	-2.7801 (0.0370)	-4.2103 (0.0123)	-4.1836 (702.73)
	111	-1.2749 (0.2437)	-3.4516 (0.0259)	-3.4947 (165.32)
	113	-1.2553 (0.2488)	-3.7933 (0.0227)	-3.6776 (247.31)
	114	-1.7862 (0.2102)	-5.8575 (0.0318)	-3.1019 (123.34)
	116	-2.7801 (0.0554)	-4.1194 (0.0149)	-4.1559 (279.06)
	118	-1.8534 (0.1111)	-4.5011 (0.0244)	-3.6942 (192.6)
	120	-3.4805 (0.0315)	-3.7909 (0.0210)	-3.9833 (2021.98)
	125	-1.8534 (0.1111)	-3.6467 (0.0478)	-3.8503 (309.11)
	127	-1.7266 (0.2175)	-6.0196 (0.0316)	-3.1199 (126.59)
	128	-3.2679 (0.0228)	-4.4019 (0.0061)	-4.3309 (4764.77)
	132	-3.7068 (0.0123)	-7.5697 (0.0014)	-4.4346 (13559.6)
	133	-3.7068 (0.0123)	-13.2932 (0.0012)	-4.4179 (27438.5)
	139	-5.2208 (0.0055)	-6.1436 (0.0044)	-4.3911 (7258.7)
	143	-1.7402 (0.1775)	-3.2698 (0.0479)	-3.0898 (107.08)
	152	-3.7068 (0.0145)	-9.3342 (0.0029)	-4.4440 (17614.7)
	153	-2.7801 (0.0370)	-3.8800 (0.0125)	-4.2195 (1423.78)
	154	1.7402 (0.1775)	-1.6002 (0.0997)	-2.3107 (82.14)
	166	-3.2435 (0.0452)	-9.0119 (0.0310)	-3.8849 (342.70)
	170	-5.0965 (0.0068)	-19.334 (0.0001)	-4.0561 (28344.8)
	171	-2.7801 (0.0370)	-5.7671 (0.0147)	-4.3603 (2032.18)
	172	-3.7068 (0.0123)	-9.5563 (0.0020)	-4.4382 (10856)
	182	-3.4559 (0.0472)	-13.5843 (0.0021)	-4.2032 (2601.83)
putP				
Negative sites	1	-5.8740 (0.0633)	-36.1012 (0.0109)	-16.0588 (2.3649×10^{8})
	2	-2.9370 (0.2516)	-12.4832 (0.0922)	-16.0683 (427082)

	10	-5.8740 (0.0633)	-29.1273 (0.0154)	-16.0695 (2.5747×10^{9})
	22	-2.2177 (0.3333)	-15.2999 (0.0699)	-16.0657 (4.9691×10^{8})
	23	-2.2177 (0.3469)	-15.8761 (0.0768)	-16.0577 (3.5054×10^{6})
	34	-4.4354 (0.1111)	-33.9802 (0.0098)	-16.0628 (8.9571×10^{8})
	36	-6.6531 (0.0370)	-47.6252 (0.0021)	-16.0657 (1.6663×10^{9})
	45	-5.8740 (0.0633)	-28.0715 (0.0150)	-16.0692 (4.4569×10^{8})
	57	-7.5815 (0.0250)	-48.6477 (0.0026)	-16.0663 (1.8593×10^{9})
	60	-2.9370 (0.2624)	-42.826 (0.0313)	-16.0707 (7.0369×10^{6})
	79	-2.8762 (0.2851)	-23.8318 (0.0455)	-16.0507 (3.3387×10^{7})
	98	-4.4354 (0.1111)	-21.3472 (0.0254)	-16.0609 (5.2838×10^{8})
	101	-4.4354 (0.1111)	-26.5737 (0.0230)	-16.0612 (7.2674×10^{7})
	112	-2.2177 (0.3480)	-18.3008 (0.0650)	-16.0575 (1.8731×10^{8})
	135	-2.2177 (0.3333)	-20.5298 (0.0475)	-16.0657 (1.6059×10^{9})
	136	-6.6531 (0.0381)	-52.5213 (0.0025)	-16.063 (4.3458×10^{8})
	156	-2.9370 (0.2516)	-11.9839 (0.0958)	-16.0684 (808208)
	161	-2.9370 (0.2516)	-12.747 (0.0891)	-16.0689 (536610)
	170	-2.2177 (0.3333)	-15.1658 (0.0651)	-16.0687 (6.0212×10^{6})
	171	-2.2177 (0.3333)	-18.301 (0.0587)	-16.0628 (9.1481×10^{8})
	174	-4.4354 (0.1111)	-24.2758 (0.0254)	-16.0612 (2.4148×10^{8})
	175	-5.0543 (0.0855)	-21.9106 (0.0217)	-16.0673 (8.4536×10^{8})
	176	-5.0543 (0.0855)	-24.2749 (0.0186)	-16.0662 (6.7268×10^{8})
	177	-5.8740 (0.0633)	-23.617 (0.0193)	-16.0695 (2.9236×10^{8})
	178	-4.4354 (0.1111)	-18.6261 (0.0374)	-16.0612 (2.9236×10^{8})
	182	-4.4354 (0.1111)	-17.7774 (0.0410)	-16.0639 (7.7859×10^{8})
	187	-2.2177 (0.3333)	-18.301 (0.0726)	-16.0611 (558598)
	188	-2.5271 (0.2925)	-11.5541 (0.0974)	-16.0658 (628086)
	193	-5.2243 (0.0800)	-27.7509 (0.0159)	-16.0663 (501293)
	196	-6.6531 (0.0370)	-112.622 (0.0009)	-16.0612 (2.7130×10^{8})
	201	-2.2177 (0.3333)	-14.8929 (0.0934)	-16.0639 (1.7151×10^{6})
	203	-2.9370 (0.2516)	-12.747 (0.0831)	-16.0709 (661611)
	205	-4.4354 (0.1111)	-19.7416 (0.0342)	-16.0612 (2.9387×10^{8})
	209	-6.6531 (0.0444)	-36.9363 (0.0036)	-16.0573 (1.8785×10^{8})
	210	-2.5617 (0.3357)	-15.6624 (0.0908)	-16.048 (2.0899×10^{7})
	212	-8.8708 (0.0123)	-45.5942 (0.0027)	-16.0639 (7.0964×10^{8})
	214	-2.2177 (0.3335)	-20.5288 (0.0627)	-16.0638 (6.5582×10^{8})
	219	-2.2177 (0.3333)	-15.4143 (0.0903)	-16.0638 (889952)
	220	-2.9370 (0.3363)	-36.332 (0.0519)	-16.0615 (371178)
	221	-2.2177 (0.3333)	-15.4143 (0.0719)	-16.0628 (2.2199×10^{7})
	222	-2.9370 (0.2516)	-12.0192 (0.0988)	-16.0661 (646770)
	223	-2.5271 (0.2925)	-12.3915 (0.0944)	-16.0663 (1.9450×10^{6})
	230	-4.4354 (0.1111)	-29.5263 (0.0089)	-16.0657 (1.6803×10^{9})
muts				
Negative sites	2	-2.2948 (0.0748)	-10.5423 (0.0129)	-4.3120 (7914.98)
	3	-6.5992 (0.0002)	-50.68 (7.7048×10^{6})	-4.7339 (9.5208×10^{11})
	7	-2.2763 (0.0599)	-6.1332 (0.0708)	-3.0836 (70.59)
	17	-1.1824 (0.2933)	-10.4886 (0.0624)	-3.4829 (86.24)
	19	-5.9645 (0.0054)	-11.6354 (0.1633)	-4.2856 (1.5247×10^{6})
	26	-7.6552 (0.0003)	-25.2459 (0.0021)	-4.6911 (4.2376×10^{6})
	30	-6.3117 (0.0013)	-14.6308 (0.0136)	-4.4837 (14005.4)
	33	-3.7915 (0.0089)	-4.9978 (0.0873)	-2.1125 (39.96)
	39	-4.6914 (0.0019)	-9.9061 (0.0169)	-4.0881 (5497.77)
	40	-5.1699 (0.0016)	-19.9239 (0.0024)	-4.6681 (1.9124×10^{6})
	42	-5.2217 (0.0055)	-14.5614 (0.0105)	-4.4872 (14422.9)
	43	-4.7670 (0.0015)	-13.9797 (0.0053)	-4.6391 (1.2271×10^{6})

	49	-6.7716 (0.0002)	-40.2759 (0.0001)	-4.7312 (1.0014x10^{9})
	50	-4.2566 (0.0231)	-32.7361 (0.0061)	-4.3038 (1.6666x10^{7})
	51	-5.5258 (0.0004)	-28.85 (8.0339x10^{5})	-4.7319 (3.0513 x10^{11})
	53	-6.0293 (0.0036)	-28.6978 (0.0061)	-4.2894 (1.0617x10^{6})
	55	-3.2884 (0.0137)	-12.5802 (0.0053)	-4.6473 (1.3795x10^{6})
	58	-3.4443 (0.0486)	-18.8298 (0.0381)	-4.2991 (4.5555x10^{7})
	64	-1.2144 (0.2443)	-17.0474 (0.0225)	-3.9669 (176.27)
	66	-6.1638 (0.0004)	-27.2205 (0.0001)	-4.7328 (1.5694x10^{11})
	72	-1.3629 (0.2436)	-4.8746 (0.0852)	-2.5372 (0.0852)
	77	-6.6049 (0.0011)	-17.6073 (0.0027)	-4.6794 (2.1707x10^{6})
	78	-4.0894 (0.0162)	-17.1729 (0.0108)	-4.5121 (16199.9)
	82	-3.9528 (0.0172)	-24.5725 (0.0009)	-4.7106 (1.1802x10^{11})
	85	-1.1572 (0.2390)	-9.7150 (0.0493)	-3.4306 (89.05)
	86	-3.3907 (0.0243)	-15.4436 (0.0222)	-4.2578 (326439)
	88	-3.6453 (0.0122)	-7.5765 (0.0306)	-3.8037 (3512.27)
	90	-2.4213 (0.0384)	-5.8041 (0.0721)	-2.1837 (42)
	92	-2.3845 (0.0377)	-4.9496 (0.0829)	-2.2819 (44.63)
	101	-3.0389 (0.0214)	-4.5755 (0.0951)	-1.7691 (33.78)
	102	-3.4522 (0.0152)	-15.5572 (0.0063)	-4.5462 (22783)
	103	-2.0287 (0.0913)	-13.7772 (0.0102)	-4.2033 (6641.29)
	110	-1.5773 (0.1706)	-4.7008 (0.0865)	-2.6705 (52.76)
	111	-5.5345 (0.0028)	-47.8257 (0.0002)	-4.7183 (8.5588x10^{6})
	112	-4.0115 (0.0201)	-10.7788 (0.1085)	-4.2150 (182755)
	115	-1.9829 (0.0963)	-14.2771 (0.0067)	-4.6843 (8.1902x10^{7})
	117	-3.3084 (0.0432)	-13.1131 (0.0811)	-4.2444 (280646)
	119	-1.1572 (0.2261)	-5.7520 (0.0880)	-2.9734 (60.12)
	124	-4.5973 (0.0030)	-18.0553 (0.0009)	-4.7217 (3.2157x10^{8})
	125	-1.1824 (0.2933)	-10.3486 (0.0628)	-3.4773 (83.22)
	126	-1.7229 (0.1515)	-7.5361 (0.0356)	-4.3040 (135023)
	133	-1.6483 (0.1127)	-8.2072 (0.0255)	-3.8907 (3616.86)
	134	-2.0287 (0.0913)	-9.0859 (0.0192)	-4.0226 (4785.62)
	136	-1.6851 (0.1128)	-11.5489 (0.0142)	-4.1128 (5171.84)
	140	-2.5031 (0.0395)	-9.5782 (0.0092)	-4.4874 (417255)
	146	-1.7792 (0.1108)	-8.8280 (0.0181)	-4.0349 (4979.32)
	147	-2.9581 (0.0241)	-13.779 (0.0034)	-4.6820 (2.1170x10^{6})
tir				
Positive sites				
	73	1.6328 (0.1393)	5.2922 (0.0151)	1.7156 (105.97)
	116	1.6183 (0.2253)	10.6655 (0.0541)	1.9067 (151.26)
	143	2.6744 (0.0659)	19.2092 (0.2180)	2.4510 (2187.19)
	560	0.9818 (0.3540)	27.4443 (0.0767)	1.4540 (65.85)
	566	1.5419 (0.2484)	5.4677 (0.0664)	1.5444 (79.51)
Negative sites				
	5	-3.5275 (0.0064)	-845.963 (4.81x10^{06})	-1.0005 (5338.64)
	7	-0.6575 (0.3333)	-1.7783 (0.0686)	-0.9563 (991.234)
	17	-1.9726 (0.0370)	-7.0011 (0.0037)	-0.9552 (842.101)
	23	-0.6575 (0.3333)	-1.9109 (0.0913)	-0.9549 (1023.67)
	32	-1.3201 (0.1181)	-4.2737 (0.0184)	-0.8984 (81.74)
	37	-1.1867 (0.1364)	-2.0810 (0.0347)	-0.9597 (1671.39)
	40	-0.6575 (0.3378)	-2.5989 (0.0638)	-0.9562 (1260.97)
	42	-1.3151 (0.1248)	-2.7148 (0.0231)	-0.9504 (481.96)
	45	-0.6575 (0.3571)	-1.3918 (0.0978)	-0.9476 (731.11)
	49	-0.9728 (0.2273)	-2.6610 (0.0287)	-0.9421 (489.43)
	54	-0.8434 (0.2691)	-1.5889 (0.0694)	-0.9916 (4704.6)

	62	-1.1758 (0.1864)	-3.5133 (0.0390)	-0.9779 (1451.6)
	64	-0.4027 (0.5250)	-5.6776 (0.0157)	-0.9368 (232.41)
	69	-1.3151 (0.1111)	-2.6241 (0.0277)	-0.9550 (888.91)
	81	-1.3151 (0.1111)	-5.9179 (0.0100)	-0.9467 (334.59)
	83	-1.9726 (0.0443)	-7.6419 (0.0029)	-0.9464 (787.14)
	92	-1.1758 (0.2011)	-3.9076 (0.0421)	-0.9737 (1408.74)
	94	-0.6575 (0.3333)	-1.6017 (0.0820)	-0.9662 (2448.8)
	98	-1.9726 (0.0404)	-5.3714 (0.0037)	-0.9532 (564.73)
	100	-0.6575 (0.3333)	-1.3277 (0.0996)	-0.9662 (2424.46)
	103	-1.9071 (0.0409)	-5.7059 (0.0035)	-0.9640 (2056.51)
	104	-1.1758 (0.1864)	-2.0691 (0.0569)	-1.0002 (5419.12)
	112	-1.1758 (0.1864)	-4.4214 (0.0179)	-1.0015 (12630.6)
	113	-0.6384 (0.3522)	-1.5532 (0.3522)	-0.9568 (1608.69)
	120	-0.6575 (0.3333)	-1.7783 (0.0686)	-0.9563 (991.23)
	126	-1.3151 (0.1143)	-3.6955 (0.0128)	-0.9538 (580.85)
	130	-0.6575 (0.3333)	-1.7783 (0.0686)	-0.9563 (991.23)
	141	-1.1758 (0.2011)	-2.4897 (0.0542)	-0.9741 (1411.84)
	146	-1.1758 (0.2291)	-2.4897 (0.0598)	-0.9684 (2216.69)
	149	-1.1758 (0.1864)	-6.6008 (0.1864)	-0.9780 (1478.58)
	151	-1.1758 (0.2011)	-4.3508 (0.0339)	-0.9740 (1408.4)
	157	-0.6575 (0.3333)	-1.4996 (0.0922)	-0.9586 (796.17)
	158	-1.1758 (0.1864)	-6.6015 (0.0189)	-1.0013 (13527.7)
	162	-0.6575 (0.3333)	-1.7577 (0.0701)	-0.9639 (2505.74)
	163	-0.6575 (0.3333)	-1.7855 (0.0838)	-0.9563 (1036.88)
	164	-0.6575 (0.3333)	-1.7783 (0.0686)	-0.9563 (991.23)
	171	-1.3151 (0.1111)	-3.2960 (0.0167)	-0.9719 (906.22)
	188	-2.0687 (0.0982)	-4.6061 (0.0982)	-0.5524 (45.58)
	233	-1.1785 (0.1864)	-1.3366 (0.0775)	-1.0104 (13529.1)
	235	-0.6575 (0.3333)	-1.7622 (0.0853)	-0.9641 (2596.59)
	251	-0.9728 (0.2329)	-2.1727 (0.0387)	-0.9384 (415.04)
	255	-1.3151 (0.1111)	-4.4131 (0.0117)	-0.9466 (346.98)
	256	-1.3151 (0.1111)	-6.5530 (0.0105)	-0.9467 (351.02)
	257	-1.3151 (0.1111)	-3.1044 (0.0194)	-0.9718 (1037.69)
	258	-0.6575 (0.3333)	-5.7874 (0.0668)	-0.9564 (926.01)
	264	-1.3151 (0.1111)	-3.1303 (0.0215)	-0.9466 (331.75)
	266	-1.3151 (0.1111)	-2.8238 (0.0201)	-0.9719 (901.55)
	267	1.3151 (0.1111)	-2.5676 (0.0284)	-0.9549 (937.76)
	274	-1.3151 (0.1111)	-2.2376 (0.0270)	-0.9719 (954.01)
	276	-2.3516 (0.0347)	-38.6131 (0.0026)	-0.9802 (1603.76)
	280	-1.9726 (0.0370)	-5.0853 (0.0059)	-0.9468 (345.84)
	295	-2.3516 (0.0404)	-42.0417 (0.0011)	-0.9738 (1323.76)
	296	-1.3151 (0.1111)	-2.5463 (0.0286)	-0.9465 (381.91)
	299	-3.2581 (0.0222)	-22.5848 (0.0028)	-0.5993 (123.48)
	302	-2.0820 (0.914)	-4.2162 (0.0611)	-0.5990 (119.42)
	303	-2.3516 (0.0524)	-9.3769 (0.0069)	-0.9689 (2224.69)
	307	-1.6868 (0.0675)	-2.7085 (0.0169)	-0.9947 (4831.22)
	316	-1.9726 (0.0502)	-15.5879 (0.0009)	-0.9397 (703.33)
	319	-1.1720 (0.1872)	-2.7169 (0.0265)	-0.9403 (775.87)
	330	-3.5275 (0.0120)	-199.578 (0.0001)	-0.9678 (1735.95)
	334	-1.3151 (0.1111)	-2.8250 (0.0283)	-0.9465 (387.71)
	335	1.3151 (0.1111)	-3.8534 (0.0148)	-0.9565 (950.22)
	336	-1.1758 (0.2011)	-4.3508 (0.0334)	-0.9740 (1408.4)
	350	-1.1758 (0.2291)	-2.0841 (0.0701)	-0.9690 (2361.28)
	356	-1.1758 (0.2291)	-2.0841 (0.0701)	-0.9690 (2361.28)

	358	-0.6575 (0.3333)	-1.6169 (0.0907)	-0.9465 (408.91)
	363	-1.1758 (0.2011)	-2.4855 (0.0545)	-0.9733 (1331.01)
	369	-0.6575 (0.3333)	-1.3960 (0.0988)	-0.9564 (664.22)
	275	-1.1785 (0.1864)	-4.1367 (0.0217)	-1.0067 (9530.06)
	376	-1.3613 (0.1725)	-41.8312 (0.0245)	-0.5617 (120.77)
	388	-1.3151 (0.1111)	-4.2017 (0.0106)	-0.9588 (843.73)
	389	-1.3151 (0.1111)	-2.5619 (0.0226)	-0.9719 (958.21)
	392	-1.3151 (0.1111)	-2.5490 (0.0232)	-0.9663 (2433.47)
	394	-1.2982 (0.1459)	-3.2891 (0.0175)	-0.9506 (586.01)
	396	-1.1758 (0.1864)	-4.4213 (0.0190)	-0.9982 (4314.73)
	401	-2.3516 (0.0524)	-10.6157 (0.0046)	-0.9681 (1948.22)
	412	-1.3151 (0.1111)	-5.2932 (0.0086)	-0.9719 (972.87)
	424	-1.1758 (0.1864)	-4.4213 (0.0190)	-0.9982 (4314.73)
	428	-1.1758 (0.2291)	-2.4897 (0.0598)	-0.9684 (2216.69)
	432	-2.3516 (0.0347)	-6.3975 (0.0058)	-0.9803 (1408.01)
	443	-0.8434 (0.2802)	-1.7203 (0.0692)	-0.9886 (4946)
	444	-0.6575 (0.3333)	-1.5895 (0.0921)	-0.9641 (2599.01)
	449	-1.1758 (0.1864)	-2.0691 (0.0526)	-1.0079 (10536.6)
	457	-2.3516 (0.0347)	-38.6131 (0.0026)	-0.9802 (1603.76)
	488	-1.7953 (0.1412)	-22.6794 (0.0161)	-0.5433 (84.97)
	503	-0.8438 (0.2597)	-2.2556 (0.0568)	-0.9960 (5429.08)
	504	-2.3519 (0.0524)	-10.6157 (0.0046)	-0.9681 (1948.22)
	514	-1.3151 (0.1254)	-3.5752 (0.0151)	-0.9490 (527.92)
	516	-1.5175 (0.0834)	-2.6597 (0.0231)	-0.9714 (442.26)
	524	-1.3151 (0.1111)	-5.7211 (0.0084)	-0.9568 (675.27)
	527	-2.3516 (0.0347)	-7.6563 (0.0029)	-1.0069 (8059.29)
	540	-1.2975 (0.1168)	-3.5680 (0.0144)	-0.9595 (1777.6)
eae				
Positive sites				
	8	1.3667 (0.0877)	2.6851 (0.0105)	0.5131 (4214.5)
	145	0.7956 (0.1878)	1.2862 (0.0957)	0.4631 (133.7)
	152	1.1309 (0.1364)	1.9267 (0.0415)	0.5005 (1087.4)
	467	0.5792 (0.4866)	1.3844 (0.0901)	0.4622 (121.1)
	491	1.2512 (0.1120)	2.7197 (0.0155)	0.5288 (4032.5)
	574	1.8327 (0.1378)	3.8645 (0.0444)	0.6789 (51.05)
	597	0.8531 (0.2333)	1.4501 (0.0681)	0.4935 (359.7)
	606	0.9095 (0.1989)	1.5620 (0.0896)	0.4743 (180.4)
	616	1.0928 (0.1619)	2.0171 (0.0675)	0.4690 (146)
	849	1.9930 (0.1499)	13.196 (0.0716)	2.5429 (4589.9)
	851	2.3748 (0.0928)	13.5336 (0.01479)	2.6799 (59335.9)
	854	2.7976 (0.0407)	7.2511 (0.0548)	2.6186 (635.7)
	862	1.0189 (0.3242)	4.0909 (0.0788)	0.4626 (52.9)
	863	2.5015 (0.0568)	14.3566 (0.0228)	2.6621 (2146.1)
Negative sites				
	34	-3.7949 (0.0006)	-12.6735 ($3.27x10^{05}$)	-1.3978 (79508.8)
	36	-2.8491 (0.0040)	-10.9146 (0.0004)	-1.3326 (1024.4)
	49	-1.5523 (0.0672)	-3.2596 (0.0316)	-1.1455 (28.1)
	79	-2.2779 (0.0074)	-4.5916 (0.0009)	-1.3569 (2643.3)
	80	-1.9191 (0.0369)	-3.8226 (0.0067)	-1.0487 (102.4)
	82	-2.4168 (0.0039)	-6.2374 (0.0002)	-1.3754 (19713.2)
	89	-1.3667 (0.0525)	-2.4347 (0.0125)	-0.8208 (112.5)
	90	-1.6459 (0.0248)	-3.7205 (0.0018)	-1.2673 (1208.3)
	92	-2.6272 (0.0068)	-7.8735 (0.0004)	-1.3297 (2946.1)
	96	-1.9287 (0.0247)	-6.3061 (0.0024)	-1.2236 (268)

	97	-1.3667 (0.0370)	-3.1343 (0.0045)	-1.1389 (502.6)
	99	-1.7346 (0.0306)	-1.9306 (0.0151)	-0.7361 (46.7)
	104	-2.7335 (0.0013)	-5.2285 (0.0002)	-1.3953 (39753)
	117	-1.2582 (0.0474)	-3.0634 (0.0052)	-1.1427 (491.6)
	121	-3.5214 (0.0017)	-4.9788 (0.0007)	-1.3596 (6763.2)
	123	-2.0501 (0.0178)	-4.4389 (0.0070)	-1.3047 (298.7)
	124	-2.0452 (0.0109)	-5.3952 (0.0004)	-1.3800 (6308.7)
	125	-1.8223 (0.0123)	-3.5497 (0.0020)	-1.2037 (1618.4)
	138	-2.7335 (0.0013)	-7.8513 ($2.89x10^{05}$)	-1.4052 (189223)
	141	-5.3353 (2.4810^{05})	-146.928 ($6.69x10^{05}$)	-1.4075 (644004)
	142	-1.8223 (0.0123)	-3.5740 (0.0018)	-1.2404 (2397.7)
	151	-1.7331 (0.0307)	-1.9333 (0.0187)	-0.7383 (44)
	162	-1.8223 (0.0123)	-3.5719 (0.0019)	-1.2018 (1636.9)
	170	-1.8223 (0.0123)	-4.6965 (0.0008)	-1.3170 (4025.7)
	172	-3.7949 (0.0006)	-12.6739 ($2.23x10^{05}$)	-1.3984 (117061)
	173	-1.3667 (0.0307)	-2.0612 (0.0110)	-0.6866 (138.5)
	180	-2.1831 (0.0057)	-3.8963 (0.0009)	-1.3159 (5766.9)
	181	-1.3667 (0.0400)	-3.2938 (0.0048)	-1.1391 (444.7)
	195	-1.3667 (0.0370)	-3.2470 (0.0049)	-1.1608 (563.5)
	205	-1.3667 (0.0370)	-1.9794 (0.0130)	-0.6281 (123.8)
	261	-1.3667 (0.0370)	-2.5992 (0.0073)	-0.9628 (273.9)
	270	-4.4158 (0.0001)	-7.9833 ($6.84 x10^{05}$)	-1.3995 (133058)
	303	-3.4743 (0.0009)	-6.0199 (0.0002)	-1.3654 (25437.3)
	326	-1.3667 (0.0370)	-2.1743 (0.0109)	-0.6284 (120.1)
	329	-1.3667 (0.0370)	-2.0339 (0.0105)	-0.6706 (151.8)
	330	-1.7424 (0.0303)	-1.9979 (0.0144)	-0.7675 (47)
	333	-1.8223 (0.0123)	-4.1637 (0.0009)	-1.3255 (6415.1)
	337	-1.9347 (0.0363)	-9.3495 (0.0021)	-1.2716 (332)
	338	-1.9049 (0.0141)	-4.0271 (0.0010)	-1.3628 (3682.5)
	341	-3.5169 (0.0017)	-5.2195 (0.0007)	-1.3623 (7187.4)
	345	-1.3667 (0.0370)	-2.2930 (0.0081)	-0.8322 (219.1)
	350	-1.9191 (0.0368)	-6.0841 (0.0046)	-1.2135 (206.3)
	351	-1.7424 (0.0375)	-4.1812 (0.0053)	-1.0861 (119.8)
	363	-1.5945 (0.0452)	-3.1585 (0.0210)	-1.2083 (88.7)
	379	-2.8623 (0.0071)	-9.8857 (0.0005)	-1.3526 (3169.8)
	419	-1.2222 (0.0617)	-2.7987 (0.0098)	-0.9742 (78.7)
	493	-1.8223 (0.0123)	-4.0611 (0.0011)	-1.3063 (4962.9)
	494	-4.0481 (0.0009)	-7.9082 (0.0005)	-1.3651 (6763.1)
	497	-2.2779 (0.0041)	-5.7213 (0.0002)	-1.3867 (29655)
	509	-3.6420 (0.0007)	-12.6561($4.01 x10^{05}$)	-1.3963 (62152.7)
	520	-1.3667 (0.0370)	-2.9609 (0.0049)	-1.1007 (496.4)
	521	-1.8223 (0.0123)	-5.2570 (0.0005)	-1.3571 (10987)
	528	-1.3667 (0.0370)	-2.8698 (0.0054)	-0.9828 (310.6)
	536	-1.3667 (0.0375)	-2.092 (0.0100)	-0.6885 (157)
	538	-1.7424 (0.0303)	-1.9979 (0.0142)	-0.7686 (50.2)
	540	-1.7424 (0.0303)	-1.9978 (0.0170)	-0.7679 (48)
	544	-1.3667 (0.0046)	-3.1011 (0.0046)	-1.1325 (500.4)
	549	-1.2477 (0.0607)	-2.4087 (0.0086)	-0.8756 (248.2)
	553	-4.5502 (0.0003)	-25.8282($2.18 x10^{06}$)	-1.4058 (228408)
	554	-2.2779 (0.0041)	-9.0308 ($7.94 x10^{05}$)	-1.3973 (45208.9)
	565	-1.8223 (0.0123)	-4.8453 (0.0009)	-1.3406 (5485.7)
	568	-1.8223 (0.0123)	-6.1954 (0.0006)	-1.3486 (7402.8)
	579	-1.3667 (0.0370)	-4.0911 (0.0027)	-1.2368 (945.6)
	584	-1.8223 (0.0123)	-5.1873 (0.0007)	-1.3406 (5468.2)

	596	-1.3667 (0.0370)	-3.6760 (0.0035)	-1.1849 (636.3)
	601	-1.3667 (0.0370)	-3.6760 (0.0035)	-1.1849 (636.3)
	602	-1.3667 (0.0567)	-2.2973 (0.0144)	-0.6945 (83.6)
	603	-1.3667 (0.0484)	-2.2260 (0.0107)	-0.6673 (133.3)
	607	-0.9010 (0.2344)	-8.1549 (0.0003)	-1.3964 (1547.8)
	614	-1.8223 (0.0123)	-4.0491 (0.0013)	-1.2969 (3767)
	621	-1.3667 (0.0370)	-2.6274 (0.0067)	-0.9850 (311.6)
	625	-1.8075 (0.0340)	-4.8818 (0.0038)	-1.1386 (155.5)
	630	-2.2779 (0.0053)	-5.4027 (0.0003)	-1.3729 (17650)
	639	-1.3667 (0.0370)	-2.5912 (0.0076)	-0.9276 (239.5)
	659	-4.7251 (0.0001)	-43.6819(7.42×10^{07})	-1.4080 (856126)
	667	-2.8313 (0.0055)	-7.4009 (0.0004)	-1.3326 (3296.5)
	669	-2.0990 (0.0162)	-6.3546 (0.0029)	-1.3347 (729.9)
	670	-1.3667 (0.0370)	-2.4543 (0.0083)	-0.8224 (187.7)
	677	-1.1389 (0.1111)	-3.6044 (0.0338)	-1.1664 (68.6)
	680	-3.7108 (0.0015)	-11.3022 (0.0001)	-1.3909 (15710.7)
	682	-1.3667 (0.0370)	-3.8336 (0.0032)	-1.2069 (728.1)
	684	-1.3667 (0.0370)	-2.8489 (0.0052)	-1.0255 (344.3)
	686	-2.7798 (0.0058)	-11.509 (0.0001)	-1.3533 (4753.3)
	690	-1.8223 (0.0123)	-4.4282 (0.0011)	-1.3181 (3999.8)
	691	-1.8390 (0.0664)	-4.4531 (0.0429)	-0.8628 (20)
	698	-2.6788 (0.0049)	-4.3806 (0.0016)	-1.2603 (611.4)
	705	-1.1977 (0.1104)	-3.8195 (0.0213)	-1.2234 (65.9)
	730	-1.7886 (0.0851)	-8.8787 (0.0235)	-0.8346 (225.4)
	733	-2.4423 (0.0625)	-8.9931 (0.0243)	-0.8204 (30.2)
	734	-1.5945 (0.0452)	-4.1337 (0.0129)	-1.2758 (174.8)
	742	-2.7326 (0.0189)	-13.886 (0.0025)	-1.2226 (106.5)
	749	-2.6897 (0.0244)	-12.3167 (0.0037)	-0.9075 (95.7)
	754	-1.8551 (0.0615)	-7.5285 (0.0138)	-0.8788 (62.4)
	780	-2.0345 (0.0885)	-9.5869 (0.0451)	-0.8280 (97.9)
	786	-1.8647 (0.0999)	-7.1381 (0.0543)	-0.8250 (57.5)
espb				
Negative sites	4	-1.8262 (0.0739)	-2.6135 (0.0192)	-1.1357 (1687.99)
	20	-0.7448 (0.3333)	-1.9671 (0.0889)	-1.0937 (242.39)
	38	-1.1364 (0.2259)	-2.7252 (0.0268)	-1.0775 (339)
	40	-0.7448 (0.3333)	-2.0099 (0.0609)	-1.0937 (288.91)
	44	-1.4897 (0.1111)	-2.9406 (0.0174)	-1.1093 (460.45)
	52	-1.1796 (0.2242)	-2.3249 (0.0645)	-1.1142 (771.90)
	65	-1.1796 (0.2665)	-2.3186 (0.0793)	-1.0929 (1733.33)
	69	-1.1796 (0.2665)	-2.4290 (0.0768)	-1.0958 (1973.96)
	71	-1.1796 (0.2665)	-3.8804 (0.0410)	-1.0902 (1464.69)
	74	-1.1796 (0.2665)	-3.8804 (0.0410)	-1.0902 (1464.69)
	79	-1.1796 (0.2665)	-2.7162 (0.0698)	-1.0899 (1503.36)
	86	-1.1796 (0.2665)	2.3186 (0.0793)	-1.0929 (1733.33)
	89	-0.7448 (0.3333)	-2.0101 (0.0657)	-1.0853 (195.49)
	92	-1.1762 (0.2242)	-2.3249 (0.0654)	-1.1114 (771.90)
	94	-2.3592 (0.0502)	-10.3893 (0.0027)	-1.1198 (900.82)
	98	-0.7448 (0.3333)	-1.5036 (0.0985)	-1.0937 (248.69)
	99	-1.4897 (0.1111)	-2.1693 (0.0404)	-1.0855 (198.61)
	103	-0.7448 (0.3333)	-1.5576 (0.0853)	-1.1291 (2555.58)
	104	-0.7448 (0.3333)	-1.6474 (0.0850)	-1.1296 (1518.03)
	106	-0.9131 (0.2719)	-1.376 (0.0809)	-1.1335 (1556.67)
	112	-0.9131 (0.2719)	-1.2826 (0.0890)	-1.1339 (1559.4)
	117	-0.7448 (0.3333)	-1.7551 (0.0936)	-1.0853 (165.71)

	120	-1.1762 (0.2104)	-1.7437 (0.0789)	-1.1245 (1192.98)
	122	-1.4897 (0.1111)	-3.3822 (0.0213)	-1.0854 (156.68)
	125	-2.2345 (0.3753)	-16.183 (0.0004)	-1.1290 (1114.74)
	155	-0.7448 (0.3333)	-3.9572 (0.0470)	-1.0853 (170.96)
	160	-1.4897 (0.1111)	-5.0918 (0.0122)	-1.0855 (164.86)
	171	-1.1796 (0.2665)	-2.3896 (0.0847)	-1.0971 (2321.89)
	180	-1.4897 (0.1111)	-3.1816 (0.0238)	-1.0853 (186.39)
	194	-0.7462 (0.3335)	-20.6478 (0.0122)	-1.0129 (59.36)
	198	-0.7448 (0.0976)	-1.5576 (0.0976)	-1.0852 (195.89)
	206	-0.7448 (0.3333)	-1.7551 (0.0936)	-1.0853 (165.71)
	209	-0.7448 (0.3333)	-1.5576 (0.0976)	-1.0852 (195.89)
	233	-1.4897 (0.1111)	-2.2308 (0.0499)	-1.0938 (264.67)
	238	-1.4897 (0.1111)	-45.5799 (0.0035)	-1.1093 (457)
	240	-1.1796 (0.2104)	-1.8300 (0.0607)	-1.1647 (9740.82)
	242	-0.7448 (0.3333)	-2.0101 (0.0657)	-1.0853 (195.49)
	244	-1.4897 (0.1111)	-28.8899 (0.0043)	-1.1293 (2473)
	246	-1.1796 (0.2242)	-3.8800 (0.0307)	-1.1187 (985.38)
	254	-1.1796 (0.2242)	-2.7162 (0.0572)	-1.1090 (689.66)
	255	-1.4897 (0.1236)	-4.2649 (0.0206)	-1.0548 (220.97)
	257	-1.1796 (0.2442)	-2.3249 (0.0605)	-1.1207 (1160.02)
	258	-1.4897 (0.1129)	-2.5523 (0.0349)	-1.0583 (256.68)
	264	-1.1796 (0.2242)	-2.3249 (0.0645)	-1.1114 (771.90)
	265	-1.1796 (0.2665)	-2.4290 (0.0768)	-1.0958 (1973.96)

Lista parcial de las secuencias obtenidas y su número de acceso en el GenBank de cada marcador usado en los estudios de selección y de reconstrucción filogenética.

putp

ECO2 AF230602 Escherichia coli 2 putP gene, partial cds
ECO45 AF230630 Escherichia coli 45 putP gene, partial cds
Eco78 AF230620 Escherichia coli 78 putP gene, partial cds
Eco161 AF230614 Escherichia coli 161 putP gene, partial cds
Eco268 AF230603 Escherichia coli 268 putP gene, partial cds
Eco270 AF230604 Escherichia coli 270 putP gene, partial cds
Eco288 AF230623 Escherichia coli 288 putP gene, partial cds
Eco820 AF230616 Escherichia coli 820 putP gene, partial cds
Eco1668 AF230625 Escherichia coli 1668 putP gene, partial cds
Eco1684 AF230605 Escherichia coli 1684 putP gene, partial cds
Eco1698 AF230606 Escherichia coli 1698 putP gene, partial cds
Eco1930 AF230626 Escherichia coli 1930 putP gene, partial cds
Eco2165 AF230607 Escherichia coli 2165 putP gene, partial cds
Eco2374 AF230624 Escherichia coli 2374 putP gene, partial cds
Eco5026 AF230611 Escherichia coli 5026 putP gene, partial cds
Eco 5074 AF230633 Escherichia coli 5074 putP gene, partial cds
EcoTA002 AF230634 Escherichia coli TA002 putP gene, partial cds
EcoK12 AE000203 E.coli K12 from human
ECOR10 L01150 ECOR10 from human O6:H10
ECOR32 L01152 ECOR32 from giraffe O7:H21

ECOR52 L01154 ECOR52 from orangutan O25:H1
ECOR58 L01155 ECOR58 from lion O112:H8
ECOR70 L01158 ECOR70 from gorilla O78:NM
E8305-87 L01157 E8305-87 from horse O157:NM
E851819 L01159 E851819 from pig O157:NM

gapa

Ecol2 AF230542
Ecol45 AF230562
Ecol46 AF230563
Ecol47 AF230564
Ecol55 AF230540
Ecol78 AF230547
Ecol161 AF230555
Ecol268 AF230551
Ecol270 AF230552
Ecol271 AF230548
Ecol287 AF230549
Ecol288 AF230550
Ecol298 AF230541
Ecol820 AF230543
El1668 AF230557
E1684 AF230553
E1698 AF230554
E1930 AF230558
E2026 AF230559
E2165 AF230534
E2274 AF230535
E2314 AF230544
E2339 AF230545
E2355 AF230560
E2357 AF230561
E2374 AF230556
E2382 AF230536
E4981 AF230546
E4999 AF230537
E5026 AF230538
E5028 AF230535
E5074 AF230565
ETA002 AF230566
ETA243 AF230567
ESou57 AF230652
ESou207 AF230653
ESou273 AF230654
EcolK12 AE000273
Ecor10 M66870

Ecor14 M66871
Ecor32 M66872
Ecor40 M66873
Ecor52 M66874
Ecor58 M66875
Ecor64 M66876
Ecor70 M66877
E3406 M66878
E266674 M66879
E830587 M666880
E851819 M666881
A8190M8 M666882
Ecor4 UO7773
Ecor8 UO7754
Ecor16 UO7772
Ecor39 UO7771
Ecor65 UO7768
Ecor68 UO7765
Ecor38 UO7752
Ecor49 UO7770
Ecor50 UO7769
EO157H7 AE005401
EcolUTEC AE016761
ETA79 AY301113
ETA57 AY301112
ETA157 AY301109
ETA479 AY301111
ETA234 AY301110

mdh

Ecol45 AF230596
Ecol78 AF230582
Ecol268 AF230586
Ecol271 AF230583
Ecol287 AF230584
E820 AF230578
E1684 AF230588
E2026 AF230593
E2165 AF230568
E2274 AF230569
E5026 AF230572
E5074 AF230599
ETA002 AF230600
Esou57 AF230656
Esou207 AF230657
Esou273 AF230658

EcolK12 AE000403
ECOR10 UO4742
ECOR40 UO4770
ECOR58 UO4753
E830587 UO4757
ECOR49 UO4750
ECOR5 AF004204
ECOR37 AF00420
E234869 AF071028
EO157H7 AF071027
EM501 AF004171
EA8190 UO4759
ERT030A AF091767
ERT008A AF91762
E851819 UO4758
Ecor70 UO4755
ERT272 AF091775
ERT082 AF091770
Ecor6 AF004209
ERT213 AF091774
ERT104 AF091772
Ecor50 UO4751
EC5458 AF071030

fima

D13186 pathogenic O78
M27603 J96 from human
U20815 O157:H7
Y10902 562 uropathogenic
Z37500 MT78 avian strain
AF206642 Souza 2 dog
AF206639 55 golden eagle
AF206643 78 painted spiny pocket mouse
AF206644 161 Souza
AF206645 268 Souza
AF206646 270 Souza
AF206647 271 Souza
AF206648 287 Souza
AF206640 288 Souza
AF206649 298 Souza
AF206650 820 Souza
AF206651 1684 Souza
AF206641 1698 Souza
AF206652 2165 Souza
AF206653 2274 Souza
AF206654 2314 Souza

AF206655 2339 Souza
AF206656 4981 Souza
AF206657 4999 Souza
AF206658 5026 Souza
AF206659 5028 Souza
EcolK12
EcolO157H7
EcolUTEC CFT073
EAPEC56 AF490880
Estruis AF490890
EAPEC7 AF490886
EAPEC5 AF490883
EAPEC16 AF490876
EAPEC15 AF490875
EAPEC133 AF490872
EAPEC60 AF490884
EAPEC57 AF490881
EAPEC53 AF490878
EAPEC137 AF490873
EAPEC54 AF490879
EAPEC13 AF490874
EAPEC6 AF490885
EAPEC128 AF490871
EAPEC58 AF490882
EAPEC121 AF490869
EcolMT203 AF490888
EcolLF82 AF286465
EAPEC125 AF490870
EcolMT512 AF490889
EAPEC21 AF490877
EAPEC115 AF490868

tir

Ecol O157:H45 AB036053
Ecol O111:H AF025311
Ecol 562 murine AB026719
Ecol O103:H2 AF113597
Ecol RPEC 84 U59502
Ecol STEC AF070068
Ecol E65/56 AF132728
Ecol E83/39 RPEC U59504
Ecol EPEC 87[a] AF070069
Ecol 95SF2 AF070067
Ecol 86/24 AF125993

eae

AF530556 EAE562beta intiminbeta Bovine
AY223510 EAE562zeta intimin zeta
AF081185 ECOR37 gamma intimin isolated from marmoset
AF081184 EPECDEC5d gamma intimin
AF530554 EAESTEC ovine O2 related H19 intimin epsilon
Af530557 EAElambda non STEC bovine
AF449418 EAEDEC8a theta intimin O111:NNM 2198/77
AY186750 EAEepsilon O121:H19
AF081187 EAEDEC12a beta intimin EPEC
AF530555 EAEalpha2 from healthy H.sapiens baby

espb

AF254455 EcolCL617
AF254454 EcolCL559
AJ287422 EcolIHIT3669
AJ28741 EcolIHIT3000
AJ287419 EcolIHIT1190
AJ287420 EcolIHIT1968
AJ287418 EcolIHIT0067
AF059713 Ecol85/150
U65681 Ecol4221dog
AF144010 EcolE65/56
AF144009 Ecol83/39
AF144008 Ecol84/110-1
AF054421 EcolB10
AF064683 Ecol1390