UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

INSTITUTO DE BIOTECNOLOGÍA

Dinámica de los genes transferidos horizontalmente al clado de *E.coli*; localizados por no estar presentes en otras enterobacterias.

T E S I S

QUE PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS BIOQUIMICAS

P R E S E N T A

BIOL. SANTIAGO CASTILLO RAMÍREZ

TUTOR
DR. ALEJANDRO GARCAIRRUBIO GRANADOS.

Cuernavaca, Mor. 2005

INDICE

ABREVIATURAS.

TV	transferencia vertical
TH	transferencia horizontal
GH	grupos de homólogos
PAR	genes parásitos
NOPR	genes no parásitos
INC	genes inciertos
ORFans	genes huérfanos
HGT-DB	"Horizontal Gene Transfer Database"
GHV	grupos de homólogos transferidos verticalmente
dS	tasa de substitución sinónima
dN	tasa de substitución no sinónima
M	mediana
DS	desviación estándar

RESUMEN

El análisis de secuencias, y más recientemente el de genomas, ha permitido valorar a la transferencia horizontal como una de las principales fuerzas evolutivas en los procariotes. Actualmente hay un fuerte debate respecto a la importancia de la transferencia horizontal (TH). Este debate se debe, en parte, la identificación poco confiable de la TH. En este trabajo se planteó una estrategia para localizar genes transferidos horizontalmente con confianza pues se quería saber que les ha pasado a dichos genes después de la TH. La estrategia fue la siguiente: todos aquellos genes que estuvieran en alguna cepa del clado de *E. coli* y no estuvieran en especies pertenecientes a la familia Enterobacteriaceae se asume que llegaron por TH. Las transferencias horizontales fueron subsecuentes a la divergencia de *E. coli* y *Salmonella*. Se encontró que los casos de transferencia horizontal, aunque siempre presentes, variaron a lo largo del clado; siendo las cepas con genomas más grandes (*E. coli* CFT073, *E. coli* O157:H7 y *E. coli* O157:H7 EDL933) las que presentaron más genes transferidos horizontalmente. La inmensa mayoría de los genes transferidos horizontalmente han sido poco estudiados y parecen presentar muy pocos homólogos fuera del clado de *E. coli*. Este trabajo dividió a los genes en categorías, por un lado los genes relacionados con secuencias de inserción, transposones, fagos (PAR). Por otro lado, genes que representan alguna función útil para la bacteria (NOPR). Al comparar los genes aquí detectados con los anotados en la base de datos HGT-DB (esta tiene genes atípicos en el contenido de GC y/ó el uso de codones por lo que se infiere que son transferidos horizontalmente) dicha base de datos sólo detectó un 30% del total de genes posibles del conjunto de aquí detectados. Muchos de los homólogos más cercanos fuera del clado de *E. coli* de los genes transferidos horizontalmente no provinieron de los géneros más cercanos posibles. Sin embargo conforme más cercano sea el género mayor es la frecuencia de genes transferidos horizontalmente. La dinámica de los genes transferidos horizontalmente fue diferente de la de los genes transferidos verticalmente. Los genes transferidos horizontalmente presentaron menor proporción de genes duplicados que los genes transferidos verticalmente. Por otra parte, los genes transferidos horizontalmente presentaron tasas más altas de cambio, tanto de mutaciones sinónimas como no sinónimas, en comparación con los genes transferidos verticalmente. Los genes transferidos horizontalmente presentaron un régimen selectivo más laxo modificando más rápido sus proteínas. Además también modifican con mayor frecuencia el tamaño de su secuencia. Por otra parte, los genes transferidos horizontalmente presentaron un alto porcentaje de pérdida. Estas diferencias entre los genes transferidos horizontalmente y los genes transferidos verticalmente seguramente se deben a que las restricciones funcionales de los genes transferidos horizontalmente son muy bajas. Este trabajó presento un análisis global y comparativo de los genes transferidos horizontalmente. Dicho trabajo descubrió nuevos aspectos de los genes transferidos horizontalmente y, además, permitió integrar estos nuevos aspectos con los previamente conocidos. A diferencia de otros trabajos que solo toman en cuenta pocos aspectos este trabajo no solo tomó varios aspectos sino que trató de darles conexión; dando una visión más coherente de los genes transferidos horizontalmente.

INTRODUCCIÓN

La reproducción de cualquier organismo, independientemente de las formas particulares, involucra transmitir el material genético del organismo parental, o los organismos parentales, al nuevo organismo; este proceso es llamado transferencia vertical (TV). La transferencia horizontal (TH) de genes es el paso de material genético de un genoma a otro (22). La TH se puede dar tanto entre organismos de la misma especie como entre organismos de especies diferentes. Dado la existencia de la TH, la TV no es el único mecanismo por el cual se pueden transmitir los genes. A diferencia de los eucariotes, que evolucionan principalmente por la modificación de la información genética existente (mutación y selección), las bacterias han obtenido una proporción significativa de su diversidad genética por medio de la adquisición de secuencias de organismos distantemente relacionados (10). Son tres los mecanismos que permiten el intercambio de material genético. Estos son la conjugación: la cual implica transferencia de material genético por medio de contacto directo, generalmente por pillus sexuales, entre el donador y el receptor; la transducción: la transferencia se lleva a cabo por partículas virales y la transformación: donde material genético libre en el medio es asimilado por una bacteria. Casi todas las bacterias conocidas presentan al menos uno de estos tres mecanismos (18). La inserción del material genético en el genoma es llevada a cabo por la recombinación, ya sea homóloga o no homóloga. La recombinación homóloga media el intercambio entre individuos estrechamente relacionados (23). Como resultado, la recombinación homóloga posibilita la dispersión de alelos ventajosos que incrementaran su frecuencia por medio de eventos de selección periódica (23). La TH puede introducir material genético nuevo en un genoma proveniente de linajes no relacionados por medio de la recombinación homóloga y no homóloga. El intercambio genético por esta última ruta sirve para distribuir genes entre clados no relacionados y por lo tanto puede conferir nuevas habilidades (23).

El análisis de secuencias, y más recientemente el de genomas, ha permitido valorar a la transferencia horizontal como una de las principales fuerzas evolutivas en los procariotes. Actualmente hay un fuerte debate respecto a la importancia de la TH. Hay quienes proponen que la TH es la fuerza evolutiva principal en los procariotes. Que incluso no se puede representar la historia de los procariotes como un árbol (el árbol del gen 16S), que ésta se debe representar como un entramado donde los linajes están en constante intercambio de material genético con otros linajes (1). En contra parte, algunos investigadores afirman que la TH sólo afecta a algunos grupos de bacterias y que ésta involucra en su gran mayoría a genes que no son indispensables. Más aún, ellos sostienen que el grueso de la historia de los procariotes puede ser representado como un sólo árbol, el determinado por el gen 16S (2). Gran parte de este debate se debe a las herramientas utilizadas para la identificación de la TH. Ninguna metodología es infalible. De hecho, los grupos de genes identificados por cada metodología tienen poca coincidencia entre si (25). Este conflicto se debe a que cada método reconoce diferentes características en sus genes blanco y prueban diferentes tipos de hipótesis (5). Los métodos más utilizados para la detección son aquellos que no necesitan de la comparación de genes de diferentes organismos. Éstos sólo utilizan el genoma del organismo en cuestión. Entre estos métodos encontramos aquellos que utilizan la composición de bases, la frecuencia de dinucleótidos, el sesgo en el uso de codones y patrones inferidos por análisis de cadenas de Markov para di y trinucleótidos. En general estos métodos localizan aquellos genes que son atípicos

respecto al grueso de los genes del genoma, de acuerdo a la característica utilizada. Se asume que los genes atípicos no son originarios del genoma en cuestión sino que llegaron por TH (21). Uno de los problemas de estos métodos es que genes provenientes de genomas con características similares a las del genoma receptor, no son detectados (15). Por otra parte, si la TH ocurrió hace tiempo, el gen podría no ser detectado como atípico pues los procesos de mutación y selección lo habrán adaptado al nuevo genoma (5). Por último, hay genes que son atípicos debido a la composición de aminoácidos y no por ser genes foráneos (10). Un trabajo reciente que analizó 103 genomas evidenció que para que se dé una TH exitosa (entendida como la integración de un gen en un genoma) se requiere compatibilidad entre el uso de codones del genoma receptor y los genes foráneos (24). Los resultados de dicho trabajo no concuerdan con que los genes TH sean atípicos, al menos en el uso de codones. De las metodologías utilizadas para la identificación de la TH la construcción de árboles filogenéticos es la más confiable.

No todos los genes son igualmente propensos a ser transferidos. Una extensa TH ha ocurrido para los genes operacionales (aquellos involucrados en "housekeeping"), mientras que los genes informacionales (los involucrados en transcripción, traducción y procesos relacionados) rara vez son transferidos horizontalmente (6). Resultados de otro trabajo proveen evidencia de una dicotomía más pronunciada entre genes, consistiendo en las clases "adquiridos", aquellos que llegaron por TH, y "ortólogos", los árboles filogenéticos construidos con estos genes son congruentes con el del gen 16S, esto quiere decir que son transferidos verticalmente. Estas clases difieren en un aspecto fundamental: más de 70% de los genes "adquiridos" codifican para proteínas de funciones no caracterizadas, mientras cerca del 80% de los genes de la clase "ortólogos" posee anotación funcional (del cual solo de un 10 a 15% sería clasificado como genes informacionales) (3). En un estudio reciente fue determinado cuantitativamente que las funciones de los genes adquiridos por TH, con excepción de los elementos móviles, están sesgadas hacia tres aspectos: superficie celular, unión a DNA y funciones relacionados con la patogenicidad (26). Por otra parte, la duplicación de genes está significativamente sobrerrepresentada en los posibles genes transferidos horizontalmente (27).

El grupo de Peter R. Reeves observó, usando como marcadores moleculares genes "housekeeping" y enzimas multilocus, que las especies del género *Shigella* se agrupaban de manera sistemática con *E. coli* en los árboles filogenéticos construidos (19,20). Así, dicho grupo concluyo que las *Shigella* deben ser consideradas como parte de la especie *E. coli* (19, 20). Hay 4 cepas secuenciadas de *E. coli* : la frecuentemente usada *E. coli* K12; dos enterohemorrágicas *E. coli* O157:H7 (cepa de Sakai) y *E. coli* O157:H7 EDL933; y la uropatógena *E. coli* CFT073 (28, 29, 30, 31). Por su parte, *Shigella flexneri* 2a tiene dos cepas secuenciadas y son *Shigella flexneri* 2a 301 y *Shigella flexneri* 2a 2457T (32, 33). Cuando se compararon las cepas K12, O157:H7 EDL933 y CFT073 de *E. coli* se observó que hay un total de 2996 genes comunes, mientras que 585 genes fueron únicos para K12, 1346 para O157:H7 EDL933 y 1623 para CFT073 (31). En un estudio más reciente, hecho con microarreglos, y que incluyo 22 cepas, tanto de *E.coli* como de *Shigella*, se calculó que el conjunto de genes comunes era de 2800 genes (34). Como muchos de los segmentos únicos presentan contenidos de GC atípico y uso de codones distinto del conjunto de genes comunes, se ha dicho que estos genes fueron adquiridos por TH (31).

De las cepas secuenciadas, *E. coli* K12 es en la que más se ha estudiado la TH a gran escala. En 1998 con el genoma completo de *E. coli* K12 se infirió, por medio de composición atípica de nucleótidos, que hasta el 18 % del genoma había sido obtenido por TH (21). Con la misma herramienta, pero corrigiendo para aquellos casos donde la composición en aminoácidos impone una composición atípica de nucleótidos del gen y entonces el no sería TH, en el 2000 se determinó que un 13 % del genoma de *E. coli* K12 era foráneo (10). Usando contenido de GC y uso de codones en "The Horizontal Gene Transfer DataBase" se encontró que un 9 % del genoma de *E. coli* K12 llegó por TH (15). Como se puede ver el rango entre los diferentes estudios es muy amplio. Estos estudios adolecen de los problemas de las metodologías que no se basan en la comparación de genes entre distintos organismos.

La pregunta que guió este trabajo fue ver qué le pasa a los genes adquiridos por TH después de la transferencia; cuál es su dinámica: su tasa de substitución, la facilidad de pérdida tanto de una parte del gen como del gen completo, su proceso de duplicación. Además de la dinámica se analizó el proceso de la llegada de los genes transferidos horizontalmente; si llegaron en grupo y qué tan cercano esta el clado que posiblemente donó los genes transferidos horizontalmente. La idea fue hacer un análisis global, que integrara diversos aspectos de los genes transferidos horizontalmente, contrario a la mayoría de los trabajos previos, en los que sólo se analizaba unos pocos aspectos. De tal manera que se pueda hacer un perfil de características particulares de los genes TH.

La estrategia de identificación de TH de nuestro trabajo fue la siguiente: todos aquellos genes presentes en el clado de *E. coli* y no presentes en ninguna otra especie de la familia Enterobacteriaceae, son genes TH al clado de *E. coli*. Dichos genes no tienen homólogos en el clado de la familia Enterobacteriaceae - esta familia se encuentra dentro de la subdivisión Gamaproteobacteria y es una de las familias más sobrerrepresentadas en las bases de datos y, además, es la que tiene más especies secuenciadas-. Las TH fueron subsecuentes a la divergencia de *E. coli* y *Salmonella*, hace aproximadamente 100 millones de años (45), ya que los genes solo están presentes en *E. coli*. A diferencia de los métodos utilizados para detecciones masivas de TH esta estrategia detecta genes transferidos horizontalmente independientemente de si estos resultan ser atípicos en el contenido de GC o el uso de codones. Por otra parte al basarse en relaciones filogenéticas creemos que la estrategia identifica de manera confiable genes transferidos horizontalmente. En contraste con otros estudios este hace un análisis con varias cepas de un clado y no con una sola; de tal manera que se puede ver que esta ocurriendo no solo en una cepa sino en el clado.

HIPÓTESIS

Los genes transferidos horizontalmente, debido a sus funciones, presentan un comportamiento diferente al de los genes transferidos verticalmente.

OBJETIVOS

Analizar la dinámica de los genes transferidos horizontalmente al clado de *E. coli*.

Objetivos particulares.

1) **Detectar los genes transferidos horizontalmente al clado de *E. coli* y hacer una categorización funcional de ellos.**
2) **Determinar el clado de procedencia de los genes transferidos horizontalmente.**
3) **Analizar los procesos de agrupación y pérdida en los genes transferidos horizontalmente.**
4) **Determinar un conjunto de genes transferidos verticalmente.**
5) **Analizar la presencia de parálogos en los genes transferidos horizontal y verticalmente.**
6) **Determinar las tasas de substitución y la presencia de "indels" en los genes transferidos horizontalmente y en los transferidos verticalmente.**

METODOLOGÍA.

Filogenia del clado *E. coli.*

Con el fin de conocer como se relacionan las *E. coli* entre ellas y ver si efectivamente las *Shigella* forman parte de *E. coli* se hicieron dos filogenias. Para construir la filogenia del clado *E. coli* se tomaron 325 genes que presentaban una solo copia para las 4 cepas de *E. coli* secuenciadas, las 2 *Shigella flexneri* secuenciadas y *Salmonella typhimurium* LT2 (apéndice Cepas 1). Esta última se tomó como grupo externo pues se sabe que es mucho más distante a *E. coli* y *Shigella* de los que son *E. coli* y *Shigella* entre ellas. Con los 7 ortólogos (uno por cada cepa) se hizo el alineamiento para cada gen con el programa ClustalW (37). Se tomaron los 325 genes porque cuando mucho hasta dos pares de los 7 ortólogos presentaron secuencias idénticas. Los 325 alineamientos fueran editados quitando todas aquellas columnas que tuvieran algún "gap" o que el nucleótido de la columna estuviese totalmente conservado. Se juntaron todos los alineamientos para tener uno solo, el cual tuvo un total de 68092 nucleótidos. Por medio del programa SEQBOOT (35, Phylip 3.57c) se generaron 1000 repeticiones del alineamiento. Con el programa DNAPARS (Phylip 3.57c) se sacó el árbol más parsimonioso (aquel que involucra el menor número de cambios evolutivos) de cada repetición y con el programa CONSENSE (Phylip 3.57c) se vio cuales fueron los grupos más frecuentes que tuvieron dichos árboles. Se hizo una segunda filogenia. Con SEQBOOT se generaron otras 1000 repeticiones del alineamiento. Para cada repetición se sacó su matriz de distancia, con el programa DNADIST (Phylip 3.57c). De cada matriz se obtuvo un árbol con el programa NEIGHBOR (Phylip 3.57c) y por último con CONSENSE se determinaron los grupos más frecuentes. Así, se obtuvieron dos filogenias; una por parsimonia y otra por distancia. En estas filogenias el largo de las ramas no representa el número de cambios ocurrido por rama.

Genes transferidos horizontalmente y grupos de homólogos (GH).

EnteroDB fue formada por las proteínas de los genomas completamente secuenciados de bacterias pertenecientes a Enterobacteriaceae con excepción de *E. coli* y *Shigella.* Escogimos todas aquellas proteínas de las cepas de *E. coli* y *Shigella* (apéndice Cepas 1) que al hacerles un BlastP (39), con un valor de expectancia de 0.1, contra EnteroDB no tuvieron "hits". Este valor de expectancia es poco estricto e implica un criterio de homólogia laxo; de tal manera que incluso homólogos con muy poco parecido son detectados. Si con un criterio tan laxo de homología no se localizan homólogos los más probable es que no existan. A continuación esas proteínas se organizaron en grupos de homólogos (GH), esto se hizo con un valor de expectancia 0.001con BlastP, el cual conlleva un criterio de homólogia más estricto que el anterior. EnteroDB no incluye todas la proteínas conocidas de la familia Enterobacteriaceae, solo las de los genomas completamente secuenciados; por lo que establecimos un filtro. Este consistió en eliminar todos aquellos GH que al hacer un BlastP, con un valor de expectancia de 0.1, tuvieran un "hit" con alguna de bacteria de la familia Enterobacteriaceae (que no fuera *E. coli* o *Shigella*) en la base de datos no redundante de "genbank" del "National Center for Biotechnology Information" (NCBI).

Categorización funcional de los GH.

Para clasificar los GH, nos basamos en la anotación de los genes en "genbank" y de las proteínas correspondientes en Swiss-Prot. La categorías fueron las siguientes:

1)Genes parásitos (PAR); comprendido por fagos, secuencias de inserción y transposones.

2)Genes no parásitos (NOPR); todos aquellos casos de los que se tuviera algún indicio de función pero que no caen en la categoría anterior.

3)Genes inciertos (INC); casos que solo se sabe que son marcos de lectura abierta.

Las categorías son mutuamente excluyentes; un GH solo puede estar en una. Sin embargo se espera que los GH de la tercera categoría pertenezcan a una de las otras dos categorías, el problema es que no se tiene información para asignarlos a una de ellas y por eso se estableció INC.

También se vió cuántos GH eran huérfanos (ORFans);es decir, secuencias menores a 150 AA, que se encuentran en un solo genoma, que no tienen función y no presentaron parálogos; son sólo proteínas hipotéticas y probablemente son errores de anotación. Todos ellos pertenecen a INC.

Comparación de los GH detectados con "Horizontal Gene Transfer Database" (HGT-DB).

La HGT-DB es una base de datos de acceso público de genes transferidos horizontalmente basada en criterios estadísticos del contenido de GC, uso de codones y aminoácidos. La base de datos cuenta con tres de las cepas de *E. coli* que nosotros utilizamos y ninguna de *Shigella*; de tal manera que no se compararon todos los GH sino sólo aquellos que presentaron alguna cepa en común con la HGT-DB. Así, el total de GH encontrados en HGT-DB es normalizado por el número de GH comparados, y no por el total de GH.

Casos con homólogos más allá de las cepas de *E. coli* y *Shigella*.

Para algunos de los GH se detectaron homólogos fuera de la familia Enterobacteriaceae, estos fueron detectados con BlastP con un valor de expectancia de 0.000000001. Para cada uno de estos casos se hizo su filogenia, si se podía. Primero se alinearon las proteínas con ClustalW, posteriormente con PROTDIST se saca una matriz de distancia y por último con NEIGHBOR se construyó un árbol. Este proceso se repitió una segunda vez pero en esta se editó el alineamiento quitando las columnas con "gaps" y las totalmente conservadas. La especie de la proteína homóloga más cercana representa el pariente conocido más cercano a la especie ancestral que fue el verdadero donador en el evento de TH. Lo anterior nos da una idea de la distancia filogenética de la cual provienen los genes transferidos horizontalmente. Las especies "donadoras" se agruparon en géneros "donadores". De cada género se obtuvo del "NCBI" la secuencia del gen 16S más larga. El gen 16S de todos los géneros fue alineado junto con el de *E. coli* K12 con ClustalW. De este alineamiento, excluyendo los sitios con "gaps", se sacó el número de sitios idénticos entre cada género y *E. coli* K12 y fue dividido por el total de sitios en común del alineamiento; de esta forma se obtuvo el porcentaje de identidad con K12 para cada género. El porcentaje de identidad da una idea de la cercanía con *E. coli*; entre más alto sea el porcentaje más cercano es el género a *E. coli*. Los géneros se dividieron en cuatro categorías. La primera incluye a los géneros de las Gamaproteobacteria fuera del orden Enterobacteriales (no-entero). En la segunda están los géneros de las Proteobacterias que no son Gamaproteobacteria (no-gama). Tercera, géneros de cualquier Phylum bacteriano que no sea Proteobacteria (no-proteo) y en la última

están los géneros no bacterianos (no-bacteria). De acuerdo al género "donador" cada GH se vió en que categoría cayo. Así, cada categoría se obtuvo una frecuencia de GH. Si alguna de las categoría esta sobrerrepresentada en las base de datos dicha categoría presentaría un número de GH artificialmente alto (por ejemplo, un GH podría tener el homólogo más cercano en no-gama pero al no haber sido depositada esa secuencia de no-gama en la base de datos pero si una de no-proteo el GH caería en no-proteo). Con el fin de corregir dicho sesgo, las proteínas presentes en Swiss-Prot y Trembl se clasificaron en las cuatro categorías antes descritas. Previó a la clasificación se redujo el número de "hits" de Swiss-Prot y Trembl con CD-HIT (44) a grupos de secuencias cuyo porcentaje de identidad es igual o mayor al 80%; de tal suerte que, secuencias que compartan 80% o más de identidad cuentan como un solo "hit". Los GH para cada categoría es normalizado por el número de secuencias pertenecientes a dicha categoría en Swiss-Prot y Trembl; a esto le llamamos frecuencia normalizada. Como una medida de distancia entre cada categoría y el clado de *E. coli,* se tomó el porcentaje de identidad del género más idéntico al de *E. coli* K12 en cada categoría.

Agrupación.

Un evento de TH podría en principio involucrar más de un gen. Los números "gi" de los archivos de "genbank" son consecutivos para genes contiguos en un genoma. Nosotros consideramos que dos o más genes se transfirieron en un mismo evento si los genes de 2 o más GH tenían, en algún genoma de las *E. coli* y/o *Shigella*, números "gi" consecutivos.

Llegada de los GH al clado *E. coli.*

Asumiendo que la filogenia del clado *E. coli* es la presentada en la figura 1 y analizando la distribución de homólogos en las cepas de *E. coli* , se puede tratar de inferir que rama del árbol recibió la TH. La mejor hipótesis se obtiene escogiendo la historia que minimiza el número de eventos de TH y de pérdida genética; es decir, la hipótesis más simple. Por ejemplo, si un GH tiene genes en las dos *Shigella* y en *E. coli* K12 lo más simple es asumir que este GH fue transferido al ancestro común de las cepas en cuestión, lo cual solo implicaría una sola TH (una alternativa más complicada seria una TH para cada cepa, esto daría tres TH). Siempre se prefirió la alternativa más sencilla. El caso anterior fue relativamente fácil de resolver pero hay otros que no lo son tanto. Por ejemplo, si hay un GH con genes en *E. coli* K12, *E. coli* O157:H7 EDL933, y *E. coli* CFT073 una explicación sería pensar que hubo una TH al ancestro de todas las *E. coli* y que después se perdió en el ancestro de las *Shigella* y también *E. coli* O157:H7; esos implicaría una TH y dos perdidas. La otra alternativa seria asumir 3 TH, una a cada una de las cepas. Casos como el anterior en el que hay dos alternativas las dos igual de simples (las dos implican 3 pasos) no se tomaron en cuenta pues no hay un criterio objetivo que permita decantarse por alguna de las alternativas. Por otra parte, casos que implicaron dos eventos de TH para un GH tampoco fueron tomados en cuenta, ya que cabe la posibilidad de que en realidad haya sido un solo evento pero por el proceso de pérdida de genes pueda hacer parecer que es un evento doble.

Parálogos.

Cada GH puede tener más de una proteína por genoma. Si un GH presenta más de una proteína en alguna cepa entonces este GH tiene parálogos –la lógica es pensar

que hay dos genes, donde uno de ellos se origino por duplicación del otro, que codifican para proteínas muy parecidas; por eso si un GH tiene más de una proteína por genoma se establece que tiene parálogos-. Para cada GH se determinó si presentaban más de una proteína por genoma en alguna de las *E. coli* y/o *Shigella*. De lo anterior se determinó que porcentaje de GH presentaba parálogos. Para saber si este porcentaje es particular de los genes transferidos horizontalmente, se definió un conjunto de genes transferidos verticalmente; el cual se agrupo en GH y a estos se les determinó el porcentaje que presentó parálogos. El grupo de transferencia vertical consistió en todos aquellos genes que estuvieran presentes en todas las *E. coli* y *Shigella* consideradas pero que además estuvieran en algunas de las especies de los géneros *Salmonella*, *Yersinia* y *Photorhabdus* (Apéndice Cepas 2). Esto se hizo con un valor de expectancia 0.001. Del grupo de transferencia vertical se eliminaron aquellos genes que tuvieran funciones como las descritas en PAR. El grupo de genes de transferencia vertical se agrupó en GH transferidos verticalmente (GHV), con un valor de expectancia de 0.001.

Pérdida de genes.

De los datos de la llegada de GH al clado de *E. coli* se tomaron aquellos que llegaron al último ancestro común de toda las *E. coli* (LCA) y los que llegaron al ancestro común de *E. coli* K12 y *Shigella* (AnK12_Shg). Si el GH presenta genes en todas las cepas derivadas del ancestro común es un caso completo. Por otra parte, si el GH no presenta genes en todas las cepas derivadas del ancestro común es un caso incompleto. Los casos incompletos implican pérdida de genes. Si uno divide el número de casos incompletos por el total de casos (incompletos más completos) obtendrá el porcentaje de pérdida.

Tasas de substitución y presencia de "indels".

Para 102 GH (fue el máximo número de GH que presentaron las mismas dos cepas en común) y para un grupo de GHV se analizaron las tasas de: substitución sinónima (dS) y substitución no sinónima (dN). Además se analizó el cociente dN/dS conocido como "W", que se utiliza como una medida de la selección que esta actuando en un gen. Solo los GHV que tuvieran una solo copia del gen por cepa fueron escogidos. Por medio de un alineamiento global, generado con el programa stretcher (36, EMBOSS 2.7.1), se alinearon las proteínas de las cepas *E. coli* K12 y *E. coli* O157:H7 para los GH y para los GHV. Con el alineamiento de las proteínas como referente se alinearon las secuencias de los genes, utilizando el programa tranaling (EMBOSS 2.7.1); de tal manera que al final se obtuvo un alineamiento de la secuencias nucleotídicas de los genes. Tanto los GH como los GHV presentaron un solo gen por cepa para *E. coli* K12 y *E. coli* O157:H7. Lo anterior es importante pues para hacer un buen estimado de "dS" y "dN" es necesario contar con ortólogos. Con los alineamientos de las secuencias nucleotídicas y utilizando el programa codeml (38, PAML 3.13d) se obtuvieron los estimados de "dS", "dN" y con ellos se calculó W.

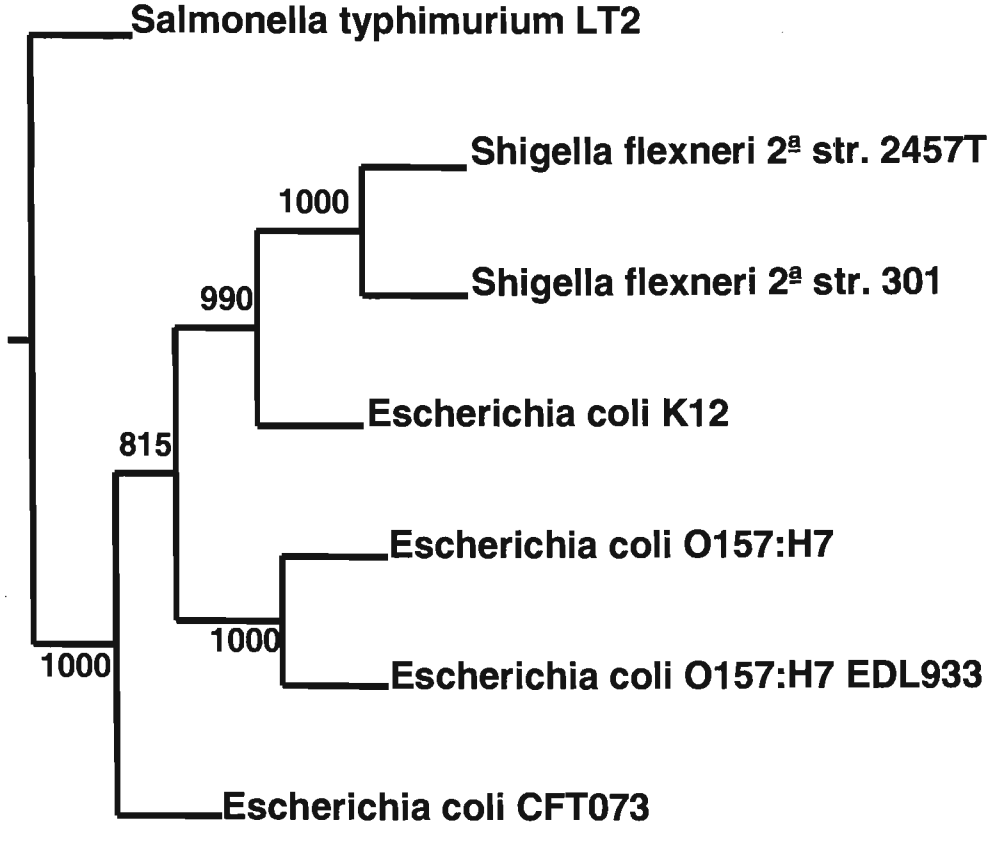
Por otra parte, se analizó cuantos de los alineamientos de las secuencias nucleotídicas, tanto para los GH como para los GHV, presentaban "indels". Si un alinemiento tiene "gaps" quiere decir que hubó "indels". La presencia de "indels" implica que la proteína ha cambiado de tamaño, ya sea por pérdida o ganancia de secuencia.

RESULTADOS.

Filogenia del clado *E. coli*.

Se ha sugerido que las especies del género *Shigella* deben ser consideradas como parte de la especie *E. coli* (19, 20). Para confirmar que las *Shigella* de nuestro estudio pertenecen al clado de *E. coli* se hicieron dos filogenias, una por parsimonia y otra por distancia. Si se utiliza una gran cantidad de genes concatenados se puede obtener la filogenia de las especies analizadas (16). En este estudio se utilizaron 325 genes. Los genes provenían de 4 *E. coli*, 2 *Shigella* y *Salmonella typhimurium* LT2 como grupo externo (apéndice Cepas 1). Se hizo un bootstrap de 1000 repeticiones para cada filogenia. Las dos filogenias dieron la misma topología que es la mostrada en la figura 1. Los valores de soporte fueron más altos para la filogenia hecha con parsimonia. El mismo nodo fue el que recibió menor soporte para ambos filogenias y es el que en la figura 1 presentó 815 repeticiones. En la figura 1 se puede apreciar que efectivamente las *Shigella* forman un grupo monofilético con *E. coli* y que de hecho las *Shigella* no están en la rama basal del grupo *E. coli*. Por otra parte se vé que el grupo de *E. coli* y *Shigella* se separan muy bien del grupo externo pues el soporte más alto posible (1000/1000 repeticiones en parsimonia). Dé aquí en adelante cuando se refiera al clado de *E. coli* es al de grupo de *E. coli* que incluye a las *Shigella*.

FIGURA 1



Filogenia, determinada por parsimonia, del clado *E. coli*. La topología fue la misma tanto para parsimonia como para distancia. Los números en los nodos indican el número de repeticiones que sucedió dicho grupo de un máximo posible de 1000 para parsimonia. Nótese que las *E. coli* y *Shigella flexneri* forman un grupo monofilético, y se separo de *Salmonella* con el soporte más alto posible. El largo de las ramas no representa el número de cambios ocurridos.

Genes transferidos horizontalmente y GH.

El criterio filogenético utilizado para la identificación de genes adquiridos por TH fue el siguiente: todos aquellos genes del clado *E. coli* que no tienen un homólogo en las demás bacterias de la familia Enterobacteriaceae (con un criterio de homología laxo: un valor de expectancia de 0.1) fueron adquiridos por TH. Si la explicación no fuera TH, habría que implicar un gran número de pérdidas para que un gen no se encuentre en ninguna bacteria de la familia Enterobacteriaceae y si en clado *E. coli*, lo cual seria muy poco probable. Se obtuvieron 1360 GH que no tuvieron homólogos en las bacterias de la familia Enterobacteriaceae. Un hecho importante es que de estos 595 GH son ORFans (ver metodología Categorización funcional); es decir, de este 44 % de los datos no se sabe nada salvo que quizá codifiquen para alguna proteína. El número de genes transferidos por cepa está desglosado en la Tabla 1.

TABLA 1.

Número de genes TH por cepa, con y sin ORFans.

Cepa	Con ORFans	Sin ORFans
E. coli CFT073	785	372
E. coli O157:H7	436	332
E. coli O157:H7 EDL933	391	326
E. coli K12	224	209
Shigella flexneri 301	149	136
Shigella flexneri 2457T	138	126
Total de GH	1360	765

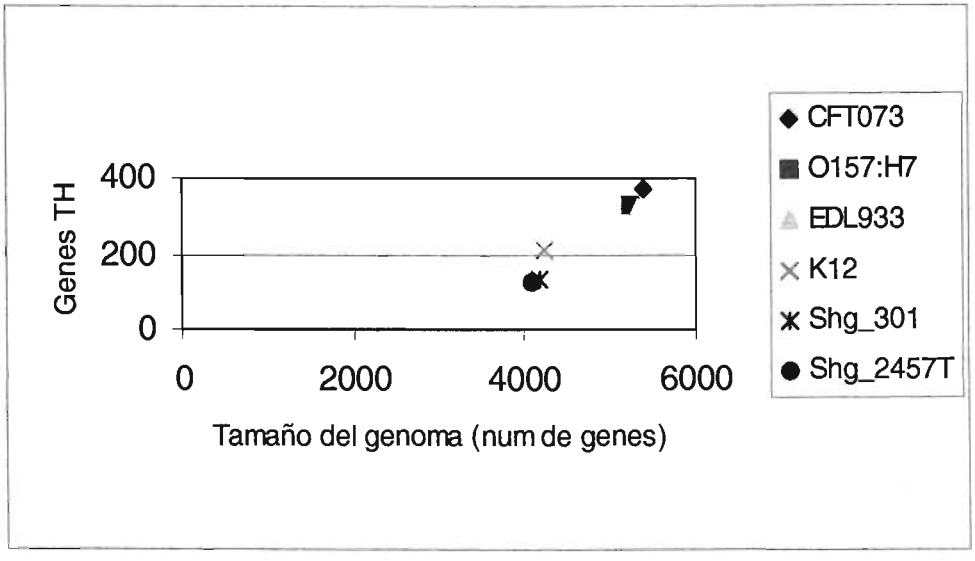
E. coli CFT073 fue la que presento más genes transferidos horizontalmente. Seguida por las cepas *E. coli* O157:H7 y *E. coli* O157:H7 EDL933 que fueron las segunda y tercera cepa con más genes TH. Estas tres cepas son las que presentan genomas con más genes. En la figura 2 se aprecia que hay una relación entre el número de genes transferidos para cada cepa y el número de genes total por cepa. El coeficiente de correlación entre el número total de genes y los genes adquiridos por TH es de $r = 0.85$, dando una $p < 0.02$. Si este mismo coeficiente se saca pero ahora sin incluir a los ORFans $r = 0.97$ y dá una $p < 0.0005$. De tal forma que las cepas con genomas más grandes presentaron mayor número de genes adquiridos por TH. La proporción de ORFans por cepa no es constante. Las cepas que presentaron una mayor proporción fueron la *E. coli* O157:H7 y la *E. coli* CFT073 (en esta última si se toman en cuenta los ORFans el número de genes transferidos se incrementaría 110%). Mientras que los genomas más pequeños, las dos *Shigella* y la *E. coli* K12, fueron los que presentaron menor proporción de genes transferidos horizontalmente.

Categorización funcional de los GH.

Como varias veces se ha hecho notar (13, 14, 21, 26), por su naturaleza móvil los fagos y los elementos transponibles participan de manera importante en los eventos de TH. La adquisición de un elemento móvil en sí misma ya es un evento de TH, pero además estos elementos pueden ser acarreadores de genes celulares. Por ello hemos decidido considerar 3 categorías de genes transferidos horizontalmente: los móviles, a los cuales llamamos parásitos (PAR), los que tienen funciones celulares, ya sea conocida ó hipotética, que llamamos no parásitos (NOPR), y otra categoría, donde

incluimos aquellos genes cuya función se desconoce, los inciertos (INC). Este análisis lo hicimos en paralelo con y sin ORFans; la única diferencia es que la categoría INC duplica su número si se consideran los ORFans y el número total de GH se incrementa en un 78%. Del total de los 765 GH sin ORFans, la mayoría, 560 casos (73 %) son tipo INC, mientras que 123 (16%) son tipo PAR y 82 (11%) casos son NOPR. Debe aclararse que la mayoría de los GH tipo NOPR no tienen una función bien determinada; son funciones probables o putativas. Lo más representativo de estos 82 casos fueron probables proteínas de membrana, proteínas periplásmicas, proteínas relacionadas con fimbrias, toxinas y proteínas adhesión, algunas otras tienen que ver con procesos de transporte y secreción. También hubo algunos activadores, represores y antiterminadores. De ser genes los 560 GH tipo INC deben pertenecer a alguna de las dos últimas categorías.

FIGURA 2



Relación entre el tamaño del genoma y los genes TH por cepa. Hay una correlación positiva significativa entre el número de proteínas totales y el número de genes transferidos horizontalmente por cepa. El coeficiente de correlación fue de r = 0.97 y obtuvo una p < 0.0005.
***E. coli* CFT073 (CFT073), *E. coli* O157:H7 *EDL933* (EDL933),**
***E.coli* O157:H7 (O157:H7), *E. coli* K12 (K12),**
***Shigella flexneri 2a* str. 2457T (Shg_2457T),**
***Shigella flexneri* str 301 (Shg_301).**

Comparación de los GH detectados con HGT-DB.

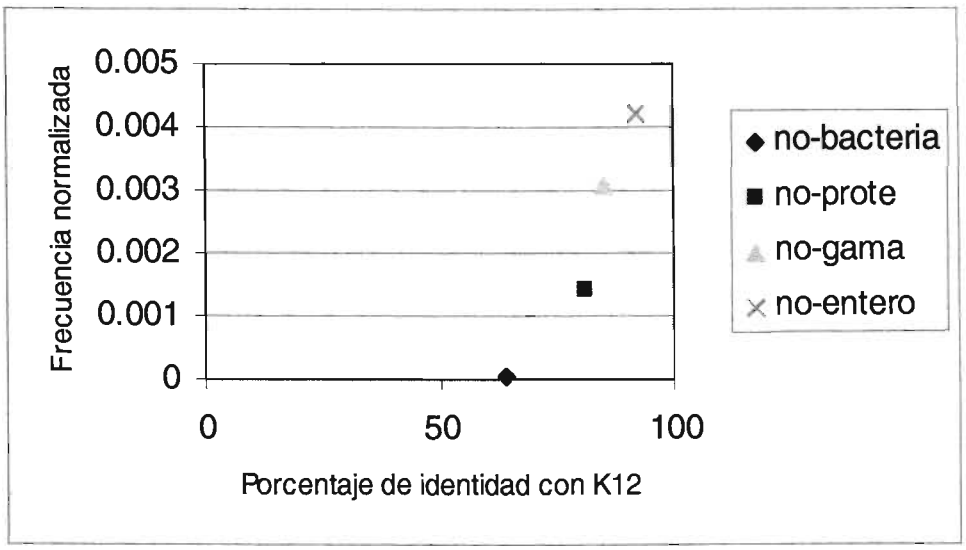
Se compararon 488 GH de los 765, pues solo estos tenían alguna cepa en común con las presentes en HGT-DB. Pero solo 146, de los 488 posibles, fueron localizados en HGT-DB; lo cual representa un 30%. Es decir, sólo un 30 % de los GH posibles son atípicos en cuanto al contenido de GC y/o uso de codones. Estos GH quizá representen eventos recientes de TH.

Casos con homólogos más allá de las cepas de *E. coli* y *Shigella*.

Sólo para 110 GH se pudo localizar homólogos más allá del clado de *E. coli*. Los homólogos más cercanos de los 110 GH abarcaron 43 géneros. La inmensa mayoría de ellos son bacterianos pero hubo uno de arqueas y uno de eucariotes. El género de

arqueas fue *Methanosarcinia* y fue una proteína no caracterizada relacionada con el COG2120. *Schizosaccharomyces* fue el género eucariote y fue una probable deshidrogenasa de glutamato. A cada uno de los géneros se le vio el porcentaje de identidad de su 16S con respecto al de *E. coli* K12, la mayoría de los géneros se ubican entre un 74% y un 90% de identidad (ver Apéndice Identidad 16S). La media de los valores de identidad de los géneros fue 82% de identidad y la moda fue 81% de identidad . Los 110 GH de acuerdo a sus géneros fueron clasificados en las la categorías no-entero, no-gama, no-proteo y no-bacteria (ver metodología). La categoría no-entero tuvo 38 GH, no-gama 31 GH, no-proteo 39 GH y no-bacteria 2. Si se suman los GH de las tres últimas categorías se tiene que 72 de 110 GH (65%) no cayeron en la categoría más cercana al clado de *E. coli*. De tal manera que gran parte de los 110 GH provienen de géneros no muy cercanos, fuera de las Gamaproteobacteria, a *E. coli*. Hay una relación muy clara entre la frecuencia normalizada (ver metodología) de cada categoría y la distancia de la misma al clado de *E. coli* (figura 3). Entre más cercana sea la categoría más dona. El coeficiente de correlación entre la distancia y la frecuencia normalizada fue de $r = 0.95$ y da una $p < 0.05$.

FIGURA 3



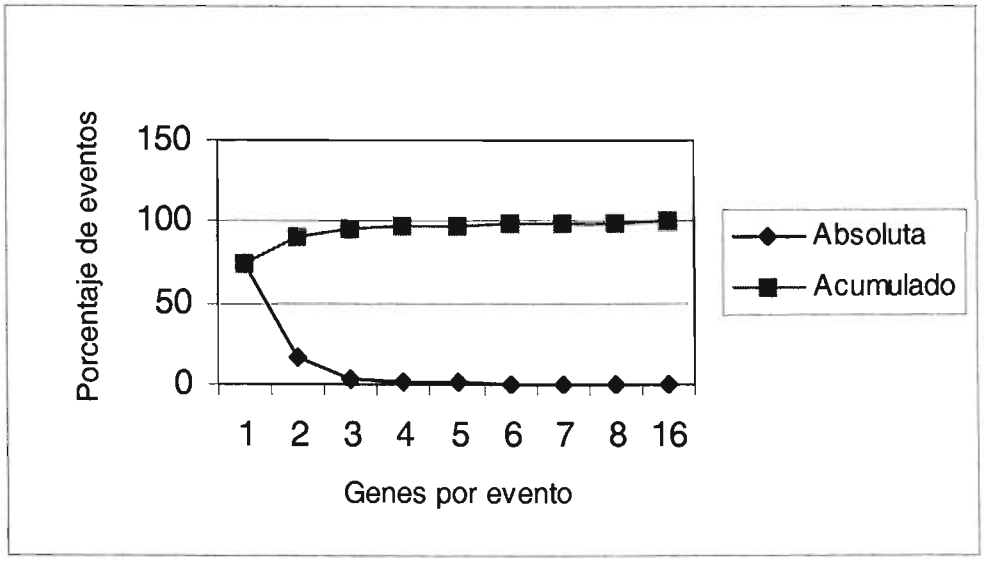
Relación entre la distancia filogenética con *E. coli* K12 y la frecuencia normalizada de la categoría. El porcentaje de identidad de la categoría con *E. coli* K12 hace referencia a la distancia filogenética que separa a la categoría del clado *E. coli*. Entre más cercana sea la categoría más dona. El coeficiente de correlación fue r = 0.95.

Agrupación.

El criterio para determinar si dos GH llegaron juntos fue ver si en alguna de las cepas estos están adyacentes. Este análisis, se hizo por pares y se compararon todos 765 los GH. Este análisis dió como resultado un total de 528 grupos; es decir, los 765 GH llegaron en 528 eventos. La mayoría de estos eventos involucran un solo gen, hubo 398 casos, lo que representa un 75 % (fig. 4). El 25 % de los eventos involucraron más de un gen. De los eventos que involucran más de un gen entre más genes involucre el evento menor es su frecuencia. Así, de los casos que involucran más de

un gen el que fue más frecuente con 83 casos fue aquel que involucra dos genes por evento y le siguió aquel de 3 genes por evento y así sucesivamente (fig. 4). El evento más grande involucró 16 GH. De dicho evento sólo dos GH de los 16 tuvieron función, uno fue una colicina y el otro una probable adhesina.

FIGURA 4



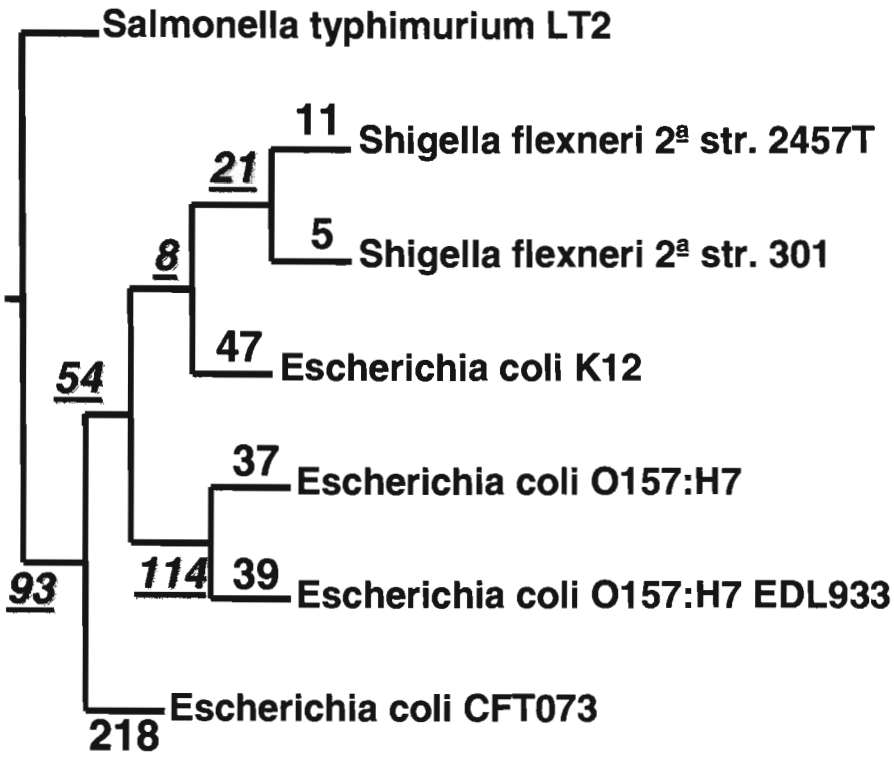
Porcentaje de eventos de TH . Los eventos involucran 1 gen, 2 genes, 3 genes, etc. Un 75% de los eventos de TH involucran un solo gen. Hay un relación inversa entre el número de genes por evento y su frecuencia. Entre más genes involucre el eventos menos frecuente es este.

Posteriormente se vió si para los eventos que involucraban más de un GH eran homogéneos en cuanto a la categoría funcional (PAR, NOPR e INC). Hubo 18 eventos que involucraron solo PAR mientras que 2 eventos involucraron solo NOPR. Otros 22 eventos involucraron a PAR y NOPR. De los 88 eventos que involucran INC con PAR o INC con NOPR no se puede concluir nada pues no se sabe que es INC. Parece entonces que cuando un evento involucra más de un gen estos en general pertenecen a más de una categoría funcional. Aunque habrá eventos que involucren eventos con genes pertenecientes a una sola categoría; esto parece acentuarse para la categoría de PAR.

Llegada de los GH al clado de *E. coli.*

Para el 85% (649 de 765) de los GH fue posible ver a que rama de la filogenia llegaron (fig. 5). Como se puede ver en la figura 5 en cada una de las ramas hubo transferencias. Mientras más interna sea la rama implica una TH más antigua. Así, nuestra estrategia fue capaz de detectar TH de diversas antigüedades. El último ancestro común (LCA) de todas las cepas del clado fue aquel organismo al cual llegaron las TH más antiguas que se pueden detectar por la estrategia aquí usada y fueron 93 GH (estas TH pudieron ocurrir desde hace poco menos de 100 millones años, que es el tiempo estimado de divergencia entre *E. coli* y *Salmonella*). Un aspecto interesante es que ninguna de las ramas tuvo ausencia de TH; es decir, la TH parece estar presente en todo el clado aunque ciertamente hay algunas ramas que tuvieron una frecuencia mucho mayor de TH.

FIGURA 5



Llegada de los GH al clado *E. coli*. Los números indican cuantos GH llegaron a cada rama. Los números subrayados implican GH que llegaron a los ancestros comunes.

Parálogos.

Son 155 de los 765 GH, 20 %, los que presentaron parálogos en alguna de las cepas del clado *E. coli*. La categoría funcional PAR tuvo un 30% de GH con parálogos mientras que NOPR un 23 %. PAR tuvo tanto un rango como una desviación estándar mayor que NOPR para el número de parálogos por GH (Apéndice Dinámica, tabla parálogos). Por su parte, de 1350 GHV 592 (44%) presentaron parálogos (ver Tabla 2). El rango y la desviación estándar de GHV para el número de parálogos fue mayor que el de PAR, presentando los valores más altos para el rango y la desviación estándar (Apéndice Dinámica, tabla parálogos). La presencia de parálogos resultó ser significativamente diferente entre PAR y GHV, y entre NOPR y GHV (ver Tabla 2). De tal manera que los GHV son mucho más susceptibles a procesos de duplicaciones.

Pérdida de genes.

Para dos ancestros comunes se determinó el porcentaje de pérdida. El primero fue LCA y presento un 65 % de casos con pérdida de genes. Por su parte, el ancestro común de las *Shigella* y *E. coli* K12 (AnK12_Shg) presentó un 44% de pérdida. Debe tomarse en cuenta que mientras más antiguo sea el ancestro común más tiempo ha pasado desde que se dio la TH y, por lo mismo, más tiempo para que se puedan perder

genes. De ahí que el ancestro común mas antiguo en esta filogenia LCA tenga casi 20 % más de pérdida que un ancestro común mucho más reciente como AnK12_Shg.

Tasas de substitución y presencia de "indels".

Los análisis de las tasa de substitución y la presencia de "indels" se hicieron para tres grupos. Uno fue el de los 821 GHV, el otro son los 102 GH y el tercero es un subgrupo de 22 GH de los 102 que son NOPR. Se sacó dS, dN, W y la presencia de "indels". Para los valores de dS y dN de cada grupo se saco la mediana (M) y la desviación estándar (DS). Para dS, la M de los 102 GH, así como de los 22 GH NOPR fue mayor que la M de los 821 GHV (ver Tabla 2). Aunque la DS de los 102 GH fue mayor que la de los 821 GHV no fue así para los 22 GH NOPR; es decir, en este último grupo hubo menor variación de datos. En dN, tanto la M como DS fueron mayores en los 102 GH y en los 22 GH NOPR en comparación con los 821 GHV (ver Tabla 2). De hecho la M de los 102 GH fue casi 9 veces mayor que la de los 821 GHV mientras que la M de los 22 GH NOPR fue 6 veces mayor que la de los 821 GHV. Posteriormente, se vió si hay una diferencia significativa entre los grupos para dS y dN. Esto se hizo en comparaciones por par de grupos usando una chi cuadrada (los datos no presentaron una distribución normal). La comparación entre los 102 GH y los 821 GHV para dS fue significativa con una $p < 0.0005$, pero no fue así entre los 22 GH NOPR y los 821 GHV (ver Tabla 2). En cambio para dN hubo diferencias significativas con una $p < 0.0005$ tanto en la comparación de los 102 GH y los 821 GHV como en la de los 22 GH NOPR y los 821 GHV (ver Tabla 2), también utilizando una chi cuadrada.

TABLA 2.

Comparación de los GH, NOPR y GHV para el porcentaje de parálogos, pérdidad de genes, presencia de "indels", dS y dN.

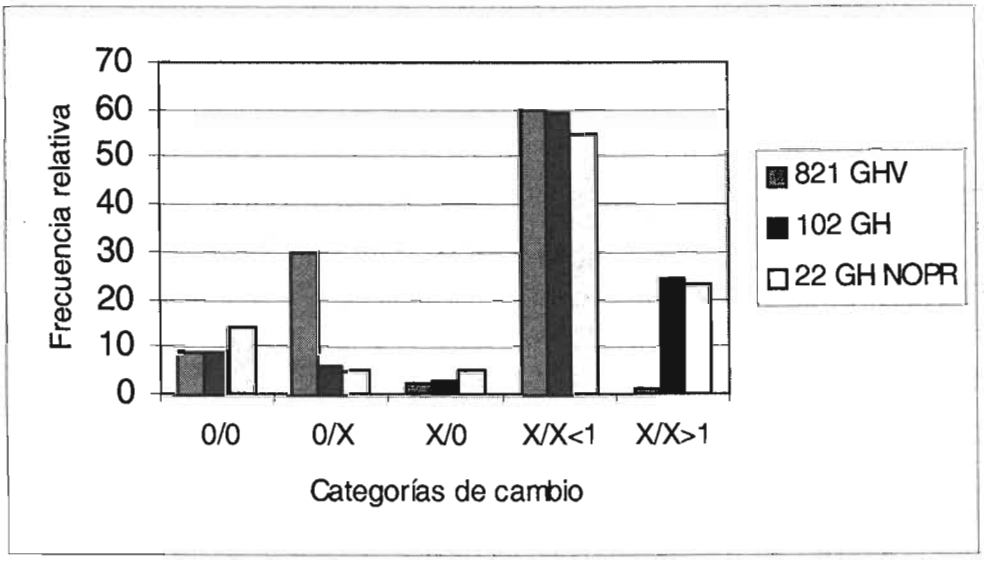
	GH con parálogos	Pérdida de genes	INDELS[a]	Mediana y DS dS[a]	Mediana y DS dN[a]
GH TH	20%	65 y 44 % *	32 %	0.0388 (0.1698)	0.0096 (0.115)
GHV	44%	------	4 %	0.03105 (0.0989)	0.0011 (0.0125)
p1	< 0.0005	------	< 0.0005	< 0.0005	< 0.0005
NOPR	23%	------	32 %	0.03515 (0.05525)	0.0069 (0.0178)
p2	< 0.0005	------	< 0.0005	No sig	< 0.0005

[a]Estos análisis fueron hechos con los alineamientos de las secuencias nucleotídicas para las cepas *E. coli* K12 y *E. coli* O157:H7 (ver metodología). Entre paréntesis las desviación estandar
*La pérdida de genes fue determinada para LCA y AnK12_Shg.
p1: probabilidad de las comparaciones entre GH TH y GHV.
p2: probabilidad de las comparaciones entre NOPR y GHV.

El cociente dN/dS es usado como una medida de selección (40). Si el cambio de un aminoácido es neutral, este será fijado a la misma tasa que la de la mutaciones sinónimas, con W = 1. Si el cambio del aminoácido es deletéreo la selección purificadora va a reducir su tasa de fijación, así W < 1. Solo cuando el cambio de aminoácido ofrece una ventaja selectiva la tasa de fijación es más alta que la tasa de mutaciones sinónimas (40). Debido a su definición, W=dN/dS, hay tres situaciones en las que W se comporta de manera extrema. Si hay dS y no dN W es igual a cero. Por otra parte si hay dN y no dS W es igual a infinito. Cuando no hay ni dS ni dN, matemáticamente W = 0. Nótese que tanto la primera situación como la tercera producen el mismo W. Con el fin de diferenciar este tipos de casos se plantearon 5

categorías de cambio para ver dN/dS. La primera contiene a los casos donde no hay ni dS ni dN (0/0); la segunda tiene los casos donde no hay dN pero si dS (0/X); tercera, casos donde no hay dS pero si hay dN (X/0); en la cuarta hay dS y dN y dN/dS es menor que 1(X/X<1); en la última hay dS y dN y dN/dS es mayor que 1 (X/X>1). Los tres grupos fueron divididos en las 5 categorías. Como se puede apreciar en la figura 6 los 102 GH y los 22 GH NOPR presentaron patrones muy similares. Los dos grupos presentaron más del 20 % de los datos en la categoría X/X > 1, esta categoría implica un régimen de selección diversificadora; es decir, están siendo seleccionados a aquellos mutaciones que cambian los aminoácidos. Por su parte los 821 GHV presentaron muy pocos datos, apenas un 1%, en esa categoría. Otra diferencia muy evidente es la frecuencia presentada en la categoría 0/X; donde los 102 GH y los 22 GH NOPR presentaron pocos datos mientras que los 821 GHV presentaron el 30%. Esta categoría es aquella que acepta solo mutaciones que no cambian la secuencia proteica. Un aspecto interesante es que los tres grupos presentaron porcentajes muy similares en la categoría X/X<1; fue en cada uno de los grupos la categoría que más datos presento. Esta categoría asume un régimen de selección purificadora, en el cual se trata de conservar la función. Al hacer una chi cuadrada entre 821 GHV y los 102 GH, usando las cinco categorías, esta dio significativa con un $p < 0.0005$. Cuando se comparó los 821 GHV con los 22 GH NOPR también se obtuvo una $p < 0.0005$.

FIGURA 6



Frecuencia de las diferentes categorías de cambio para GH, GHV y GH NOPR. Categorías de cambio. 0/0: no hay cambio alguno; 0/X: solo hay dS; X/0: solo hay dN; X/X: hay dN y dS y su cociente es menor que uno; hay dN y dS y su cociente es mayor que uno.

La presencia de "indels" fue mucho mayor en los 102 GH y los 22 GH NOPR en comparación con los 821 GHV. Los dos primeros grupos presentaron un 32 % de los alineamientos nucleotídicos con "indels" mientras que los 821 GHV presentaron un 4%. De hecho hay una diferencia significativa, con una $p < 0.0005$, en la presencia de "indels" entre los 821 GHV y los 102 GH (ver Tabla 2). También en la comparación 821 GHV con los 22 GH NOPR se obtuvo una diferencia significativa con una $p < 0.0005$ (ver Tabla 2).

DISCUSIÓN.

El conjunto de genes transferidos horizontalmente identificados en este trabajo es parcial. Esto es porque son genes que no tienen homólogos en la familia Enterobacteriaceae; es decir, se transfirieron genes que no existían en el clado. Son genes que no tienen un competidor en el genoma al que llegaron, lo cual podría ser una ventaja. Obviamente este tipo de transferencia es diferente de la que involucra a genes con homólogos en el genoma. Es muy probable que la mayoría de estas TH se dieran después de la divergencia de *E. coli* y *Salmonella*, ya que estos genes solo estuvieron presentes en el clado *E. coli*.

Todas y cada una de las cepas del clado presentó genes transferidos horizontales. Pero fueron las cepas de genomas más grandes (*E. coli* CFT073, *E. coli* O157:H7 *y E. coli* O157:H7 EDL933) las que presentaron más genes transferidos horizontalmente; de hecho hay una correlación significativa entre el número de genes transferidos horizontalmente y el tamaño del genoma. Por estudios previos se sabe que *E. coli* uropatógenas (como la CFT073) y *E. coli* con serotipo O157:H7 (en este trabajo *E. coli* O157:H7 *y E. coli* O157:H7 EDL933) tienen altas frecuencias de cepas mutadoras (41, 42). Las cepas mutadoras deficientes en el "mismatch rapair" exhiben tasas altas de mutación y recombinación homóloga (43). El tener una tasa alta de recombinación (fenotipo hiperrecombinante) incrementaría la TH. Entonces las cepas grandes quizás lo sean porque presentaron fenotipos hiperrecombinantes que facilitaron la TH y es esta la causa el incremento en el genoma de la cepa.

Dados los resultados de la categorización de los GH, tal parece ser que muchos de los genes aquí localizados han sido poco estudiados. Aun si se quita los ORFans la mayoría de los GH no están caracterizados; pues fue la categoría INC la que por mucho tuvo más GH con un 73%. Incluso para la categoría NOPR la mayoría de ellos no tienen una función bien determinada, son funciones hipotéticas o probables. Con el alto nivel de incertidumbre no se puede saber que proporción real de los genes transferidos horizontalmente son PAR y cuales NOPR. Lo anterior concuerda con un estudio previo donde se encontró que la mayoría de los genes transferidos horizontalmente codificaban para proteínas no caracterizadas (3). Además, para solo un séptimo de los GH sin ORFans fue posible encontrar homólogos más allá del clado *E. coli*. Esto indica que la distribución de estos genes es casi nula, al menos en las base de datos actuales. De tal manera, que estos GH parecen ser particulares de unos cuantos clados. Esto podría explicar porque están tan poco caracterizados estos genes.

Al igual que un estudio anterior (26) se encontró que las funciones que son NOPR se encuentran sesgadas: funciones relacionadas con la membrana (transportadores, proteínas periplásmicas, fimbrias, etc.) y funciones relacionadas con la patogenicidad.

Solo un 30% del total de posibles GH fue localizado en la HGT-DB. Si esto se extrapola para la mayoría de genes transferidos horizontalmente implicaría que solo una parte pequeña de los genes transferidos horizontalmente resultan ser atípicos en el contenido de GC y el uso de codones. Lo anterior contrastaría con el supuesto de que la mayoría de genes transferidos horizontalmente son atípicos en alguna de las

características utilizadas por los métodos para detectar TH masivas. Esto coincide con los resultados encontrados en otros estudios (24, 4). Uno de ellos muestra que para que haya una TH exitosa el gen en cuestión debe presentar compatibilidad en el uso de codones con el genoma receptor (24). El otro plantea que un factor interno importante en la regulación de la TH es la afinidad en la composición de GC entre los organismos (4).

Los homólogos más cercanos estuvieron en 43 géneros. Dos de estos géneros no fueron bacterianos. Hubo una parte importante de los 110 GH con homólogos fuera del clado *E. coli* que no cayeron en la categoría más cercana posible. Pero conforme se incrementa la distancia filogenética del género al clado de *E. coli* se reduce el número de genes transferidos horizontalmente. Quizá esto se deba a que entre más lejano sea el género hay menor posibilidad de recombinación.

La mayoría de eventos de TH parece involucrar un solo GH. Esto no quiere decir necesariamente que así fue la TH. Por ejemplo, si una TH involucro 3 genes provenientes de un arquea pero uno de ellos se perdió y otro tiene un homólogo en alguna enterobacteria, solo se detectaría un gen TH de acuerdo a la estrategia aquí planteada. Es probable que conforme más genes involucran un evento de TH menos frecuente es dicho evento. Cuando el evento involucra más de un GH los GH pueden pertenecer a PAR y NOPR, aunque hay casos en los que los genes del evento eran todos de la categoría PAR. En ese sentido creemos que PAR sirve muchas veces como vehículo para la otra categoría.

Todas las ramas de la filogenia presentaron eventos de TH. Esto implica que la TH ha sido continua a lo largo de la existencia del clado. Aunque la TH ha sido continua a lo largo del clado esta no se ha efectuado a las mismas tasas en las diferentes ramas.

Los genes TH resultaron tener menos proporción de parálogos por GH que los GHV. Esto contrasta con lo encontrado en el estudio de Hooper *et al.* (27) en el 2003, donde encuentran que los genes transferidos horizontalmente se duplican más que los genes nativos del genoma. Más esta diferencia no solo fue en la proporción de parálogos, los GHV también presentaron mayor dispersión en el numero de parálogos por GH. Estas diferencias se pueden deber a que los GHV son utilizados frecuentemente mientras que los genes TH solo se expresarán eventualmente. De los genes TH los PAR presentaron tanto mayor proporción de parálogos por GH como mayor dispersión en el número de parálogos. El que los PAR tengan una mayor proporción de GH con parálogos y una mayor dispersión en el número de parálogos por GH que los NOPR se puede deber a que los PAR muchas veces cuentan con sus propios medios (replicasas, transposas) para duplicarse.

La pérdida de genes en los GH fue muy frecuente. Los dos ancestros comunes analizados tuvieron un porcentaje alto de pérdida. El más antiguo de los ancestros presento el porcentaje de pérdida más alto con 65%, mientras que el más reciente tuvo un 44 %. Entre más tiempo haya transcurrido desde la TH, mayor es el porcentaje de pérdida. Estos altos porcentajes de pérdida seguramente afectan lo detectado en la agrupación de los eventos. Si hubo un evento que involucró 2 genes y uno se pierde, este evento se detectaría como de un solo gen. Estos porcentajes de pérdidas tan

altos, de cumplirse para todos los genes transferidos horizontalmente, implicarían que los genes transferidos horizontalmente son solo elementos transitorios en los genomas y que tarde o temprano se perderán. La pérdida seguramente se suscita cuando el gen en cuestión no presenta relevancia para el genoma; de tal manera que si se da la pérdida la viabilidad de la bacteria no se ve afectada. Aunque seguramente la pérdida no es el destino de todos los genes transferidos horizontalmente, puede que si sea el destino de la mayoría.

Las tasas de substitución resultaron ser más altas en los genes transferidos horizontalmente respecto a los GHV. Esto pasó tanto para las substituciones sinónimas (dS), como para substituciones no sinónimas (dN). La única excepción fueron los NOPR para dS el cual no fue diferente del dS de los GHV. Así, los genes TH presentaron más cambios que los GHV. Más no solo eso, sino que la relación entre los cambios no sinónimos y sinónimos (dN/dS) es mayor en los genes TH que en los GHV. Así, el régimen selectivo de los genes transferidos horizontalmente es más laxo que el de los GHV. Lo anterior implica que la secuencia proteica que es codificada por los genes TH cambia más que la de los GHV. Pero no solo cambian más los genes TH respecto a la secuencia sino que además parecen cambiar el tamaño de la misma; los genes TH tuvieron mayor presencia de "indels" lo cual implica o una pérdida o una ganancia de un segmento de secuencia. Entonces los genes TH no solo cambian más la identidad de su secuencia sino que además cambian el tamaño de la misma. Si los genes transferidos horizontalmente tienen tasas más altas de substitución esto implicaría que estos genes rápidamente adecuarían su contenido de GC al del genoma –en el caso de que el contenido de GC del gen no fuera similar al del genoma al que llegó- y muy pronto perderían su contenido de GC original.

La poca presencia de parálogos en los GH, los altos porcentajes de pérdida de genes y las tasas altas de cambio tanto de identidad como de tamaño de la secuencia se pueden explicar bajo un solo principio. Este es el de suponer que los genes TH tiene pocas restricciones funcionales y esto se debe a que sus funciones son accesorias. Al tener pocas restricciones funcionales el gen puede cambiar más rápido pues con tal de que no se afecte del todo la función los cambios son permitidos incluso los del tamaño de la secuencia. Si llega haber alguna duplicación de uno de estos genes cuya función es accesoria y no es relevante en ese momento para que fijar la nueva copia. La pérdida de genes quizá sea el caso extremo donde ya no hay restricción funcional y al no ser necesario el gen se puede perder. Hay diferencias entre la dinámica de los genes TH y los GHV. Éstas se pueden deber a que los genes TH tiene pocas restricciones funcionales lo cual implicaría tasa más altas de cambio, menor posibilidad de duplicación y mayor porcentaje de pérdida de genes.

Este trabajó presento un análisis global y comparativo de los genes transferidos horizontalmente. El análisis hecho descubrió nuevos aspectos de los genes transferidos horizontalmente y, además, permitió integrar estos nuevos aspectos con los previamente conocidos. A diferencia de otros trabajos que solo toman en cuenta pocos aspectos este trabajo no solo tomó varios aspectos sino que trató de darles conexión; dando una visión más coherente de los genes transferidos horizontalmente.

CONCLUSIONES.

Todas las cepas del clado presentaron genes transferidos horizontalmente. Hubo una correlación positiva entre el tamaño del genoma por cepa y el número de genes transferidos horizontalmente. La mayoría de los GH no están caracterizados y su distribución en las bases de datos (la "nr" del NCBI y Swiss-Prot) es escasa. Pocos de estos GH resultaron ser atípicos en el contenido de GC y el uso de codones. Para un séptimo de los GH fue posible encontrar un homólogo fuera del clado *E. coli* y aunque muchos de ellos no se encuentran en los géneros más cercanos posibles, conforme este más alejado el género menor es su frecuencia de donación. Un gran porcentaje de los eventos de TH involucraron un solo gen y mientras más genes involucre el evento menor es la frecuencia de tal evento. En cada una de las ramas de la filogenia del clado *E. coli* se infirieron llegadas de GH. El porcentaje de pérdida de genes resulto ser alto y se incrementa en relación al tiempo. Los GH tienen menos procesos de duplicación que los GHV. La tasa de mutaciones, sinónimas y no sinónimas, fueron más altas en los GH que en los GHV. Lo mismo sucede con la presencia de los "indels". Más aun, el régimen es mucho más laxo en los GH. Al interior de los GH los PAR presentaron más procesos de duplicación y sus tasas de mutación también fueron más altas en relación con los NOPR.

PERSPECTIVAS.

El presente análisis fue hecho para un sólo clado; el de *E. coli*. Estudios similares pueden llevarse acabo en otros clados con el fin de ver que tan generales pueden ser los comportamientos aquí descritos. Por otra parte, sería interesante analizar si la dinámica de genes transferidos horizontalmente, pero con homólogos en el clado receptor, es parecida a la aquí descrita. Haciendo uso de herramientas como el reloj molecular se podría ver en qué períodos llegaron los genes TH a este clado y entonces se podría tener una cronología de la dinámica descrita.

BIBLIOGRAFÍA

1.- Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. 2002. Prokaryotic Evolution in light of Gene Transfer. Mol. Biol. Evol. 19: 2226-2238.

2.-Kurland, C. G., Canback B., and Berg, O. G. 2003. Horizontal gene transfer: A critical view. PNAS. 100: 9658-9662.

3.-Daubin, V., Moran, N. A., Ochman, H. 2003. Phylogenetics and the Cohesion of Bacterial Genomes. 301: 829-832.

4.-Ravi, J., Rivera, M. C., Moore, J. E., and Lake, J. A. 2003. Horizontal Gene Transfer Accelerates Genome Innovation and Evolution. Mol. Biol. Evol. 20 :1589-1602.

5.-Lawrence, J. G. and Ochman, H. 2002. Reconciling the many faces of lateral gene transfer. Trends in Microbiology. 10: 1-3.

6.-Ravi, J., Rivera, M. C., and Lake, J. A. 1999. Horizontal gene transfer among genomes: The complexity hypothesis. Proc. Natl. Acad. Sci. USA . 96 : 3801-3806.

7.-Rivera, M. C., Ravi, J., Moore, J. E., and Lake, J. A. 1998. Genomic evidence for two functionally distinct gene classes. Proc. Natl. Acad. Sci. USA . 95 : 6239-6244.

8.-Lawrence, J. 1999. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. Current Opinion in Genetics & Development. 9: 642-648.

9.-Lawrence, J. G. and Hendrickson, H. 2003. Lateral gene transfer: when will adolescence end?. Molecular Microbiology. 50: 739-749.

10.-Ochman, H., Lawrence, J. G. & Groisman, E. A. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature. 405: 299-304.

11.-Boucher, Y., Douady, C., Papke, R., Walsh, D., Boudreau, M., Nesbo, C., Case, R., and Doolitle, W. 2003. Lateral gene transfer and the origins of prokaryotic groups. Annu. Rev. Genet. 37: 283-328.

12.-Brown, J. R. 2003. Ancient horizontal gene transfer. Nature Reviews. 4: 121-132.

13.-Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillan, M., and Brüssow, H. 2003. Phage as agents of lateral gene transfer. Current Opinión in Microbiology. 6: 417-424.

14.-Ohnishi, M., Kurokawa, K., and Hayashi T. 2001. Diversification of *Escherichia coli* genomes : are bacteriophages that major contributors?. Trends in Microbiology. 9: 481-485.

15.-Garcia-Vallve, S., Guzmán, E., Montero M. A., and Romeu, 2003. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Research. 31: 187-189.

16.-Rokas, A., Williams, B. L., King, N., & Carroll, S. B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425. 798-804.

17.- Ochman, H. and Jones, I. B. 2000. Evolutionary dynamics of full genome content in *Escherichia coli*. The EMBO Journal. 19: 6637-6643.

18.-McGee J. D., Coker C., Harro J. M. and Mobley H. LT. 2001. Bacterial genetic exchange. Encyclopedia of life sciences. Nature Publishing Group.

19.- Pupo, G. M., Reeves, P. R. 2000. Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics. Proc. Natl. Acad. Sci. USA. 97: 10567-10572.

20.-Lan, R. Reeves, P. R. 2002. Escherichia coli in disguise: molecular origins of Shigella. Microbes and Infection. 4: 1125-1132.

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

22.-Li, W-H. 1997. Molecular Evolution. Sinauer Associates, Sunderland, Mass.

23.- Lawrence, J. G. 2002. Gene Transfer in Bacteria: Speciation without Species?. Theoretical Population Biology. 61: 449-460.

24.-Medrano-Soto, A., Moreno-Hagelsieb, G., Vinuesa, P., Christen, J. A., and Collado-Vides, Julio. 2004. Successful Lateral Transfer Requires Codon Usage Compatibility Between Foreign Genes and Recipient Genomes. Molecular Biology and Evolution. 21: 1884-1894.

25.-Ragan, M. A. 2001. On surrogate methods for detecting lateral gene transfer. FEMS Microbiol Lett. 11: 187-191.

26.-Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nature Genetics. 36:760-766.

27.-Hooper, S., D., and Berg, O., G. 2003. Duplication is more common among laterally transferred genes than among indigenous genes. Genome Biology. 4:R48.

28.-Blattner, F. R., *et al.* 1997. The complete genome sequence of *Escherichia coli* K12. Science. 277: 1453-1474.

29.- Hayashi, T., *et al.* 2001. Complete genome sequence of enterohemorraghic *Escherichia coli* O157:H7 and genomic comparison with laboratory strain K12. DNA Res. 8:11-22.

30.-Perna, E. S., *et al.* 2001. Genomic sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature. 409: 529-533.

31.-Welch, R. A., *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli* . Proc. Natl. Acad. Sci. USA. 99: 17020-17024.

32.-Jin, Q., *et al.* 2002. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparision with genomes of *Escherichia coli* K12 and O157. Nucleic acids Res. 30: 4432-4441.

33.-Wei, J., *et al.* 2003. Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2ª strain 2457T. Infect. Immun. 71: 2775-2786.

34.-Fukiya, S., Mizoguchi, H., Tobe, T., and Mori, H. 2004. Extensive Genomic Diversity in Pathogenic *Escherichia coli* and *Shigella* Strains revealed by Comparative Genomic Hybridization Microarray. Journal of Bacteriology. 186: 3911-3921.

35.-Felsenstein, J. 1989. PHYLIP: phylogeny inference package. Version 3.2. Cladistics. 5:164-166.

36.-Olson, S. A. 2002. EMBOSS opens up sequence analysis. European molecular biology open software suite. Brief Bioinform. 3:87-91.

37.-Thompson, J. D., Gibson, T. J., Plewniak, F. J., and Higgins, D. G. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25:4876-4882.

38.-Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood . CABIOS. 13:555-556.

39.-Schaffer, A. A., *et al.* 2001. Improving the accuracy of PSI-Blast protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 29:2994-3005.

40.-Yang, Z., and Bielawski, J. P. 2000. Statistical methods for detecting molecular adaptation. Trends in Ecology and Evolution. 15:496-503.

41.-LeClerc, J. E., Li, B., Payne, W. L., Cebula, T. A. 1996. High Mutation Frequencies Among *Escherichia coli* and *Salmonella* Pathogens. Science. 274: 1208-1211.

42.-Denamur, E., *et al* . 2002. High Frequency of Mutator Strains among Human Uropathogenic *Escherichia coli* Isolates. Journal of Bacteriology. 184:605-609.

43.- Denamur, E., *et al* . 2000. Evolutionary Implications of the Frequent Horizontal Transfer of Mismatch Repair Genes. Cell. 103:711-721.

44.-Weizhong Li. CD-HIT: Sequence clustering software. http://bioinformatics.org/cd-hit/.

45.-Battistuzzi, F. U., Feijao, A., and Hedges, S. B. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. BMC Evolutionary Biology. 4:44.

Apéndice Cepas 1.
Genomas de las cepas de *E. coli* y *Shigella flexneri* utilizados.
Escherichia coli CFT073
Escherichia coli K12
Escherichia coli O157:H7
Escherichia coli O157:H7 EDL933
Shigella flexneri 2a str. 2457T
Shigella flexneri 2a str. 301

Apéndice Cepas 2.
Genomas de las cepas utilizadas en el grupo de transferencia vertical.
Salmonella typhimurium LT2
Salmonella enterica subsp. enterica serovar Typhi str. CT18
Salmonella enterica subsp. enterica serovar Typhi Ty2
Yersinia pestis CO92
Yersinia pestis KIM
Photorhabdus luminescens subsp. laumondii TT01

Apéndices Cepas 3.
Genomas que conformaron la base de datos EnteroDB.
Genomas del apéndice anterior más:
Yersinia pestis biovar Mediaevalis str. 901
Yersinia pseudotuberculosis
Buchnera aphidicola str. APS
Buchnera aphidicola str. Sg (Schizapis graminum)
Buchnera aphidicola (Baizongia pistaciae)
Erwinia carotovora subsp. atroseptica SCRI1043
Wigglesworthia glossinidia (endosimbionte de *Glossina brevipalpis*)

42.-Denamur, E., *et al* . 2002. High Frequency of Mutator Strains among Human Uropathogenic *Escherichia coli* Isolates. Journal of Bacteriology. 184:605-609.

43.- Denamur, E., *et al* . 2000. Evolutionary Implications of the Frequent Horizontal Transfer of Mismatch Repair Genes. Cell. 103:711-721.

44.-Weizhong Li. CD-HIT: Sequence clustering software. http://bioinformatics.org/cd-hit/.

45.-Battistuzzi, F. U., Feijao, A., and Hedges, S. B. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. BMC Evolutionary Biology. 4:44.

Apéndice Cepas 1.
Genomas de las cepas de *E. coli* y *Shigella flexneri* utilizados.
Escherichia coli CFT073
Escherichia coli K12
Escherichia coli O157:H7
Escherichia coli O157:H7 EDL933
Shigella flexneri 2a str. 2457T
Shigella flexneri 2a str. 301

Apéndice Cepas 2.
Genomas de las cepas utilizadas en el grupo de transferencia vertical.
Salmonella typhimurium LT2
Salmonella enterica subsp. enterica serovar Typhi str. CT18
Salmonella enterica subsp. enterica serovar Typhi Ty2
Yersinia pestis CO92
Yersinia pestis KIM
Photorhabdus luminescens subsp. laumondii TT01

Apéndices Cepas 3.
Genomas que conformaron la base de datos EnteroDB.
Genomas del apéndice anterior más:
Yersinia pestis biovar Mediaevalis str. 901
Yersinia pseudotuberculosis
Buchnera aphidicola str. APS
Buchnera aphidicola str. Sg (Schizapis graminum)
Buchnera aphidicola (Baizongia pistaciae)
Erwinia carotovora subsp. atroseptica SCRI1043
Wigglesworthia glossinidia (endosimbionte de *Glossina brevipalpis*)

Apéndice Identidad 16S.

Porcentaje de identidad entre *E. coli* K12 y los géneros para el 16S. La tercera columna indica el número de genes donado por cada género.

Género	% de identidad con K12	Frecuencia
Vibrio (G)	92	8
Aeromonas (G)	92	1
Pasteurella (G)	91	1
Shewanella (G)	90	4
Haemophilus (G)	90	3
Actinobacillus (G)	90	1
Pseudomonas (G)	87	16
Azotobacter (G)	87	2
Xanthomonas (G)	86	1
Xylella (G)	86	1
Bordetella (P)	85	3
Burkholderia (P)	85	2
Geobacter (P)	85	2
Azoarcus (P)	85	1
Nitrosomonas (P)	85	1
Ralstonia (P)	84	5
Magnetospirillum (P)	84	3
Magnetococcus (P)	84	1
Rhodospirillum (P)	84	1
Mesorhizobium (P)	83	3
Brucella (P)	83	2
Sinorhizobium (P)	82	3
Listeria (B)	81	5
Desulfitobacterium (B)	81	4
Bacillus (B)	81	1
Enterococcus (B)	81	1
Streptomyces (B)	81	1
Streptococcus (B)	81	1
Bacteriovorax (P)	81	1
Wolinella (B)	81	1
Leptospira(B)	80	6
Corynebacterium(B)	80	4
Clostridium (B)	80	3
Campylobacter (P)	80	1
Lactococcus (B)	79	1
Nostoc (B)	78	4
Synechocystis (B)	78	1
Bacteroides (B)	76	3
Pirellula (B)	76	2
Fusobacterium (B)	76	1
Agrobacterium (P)	74	2
Methanosarcina (NB)	64	1
Schizosaccharomyces (NB)	53	1

G: género de las Gamaproteobacteria que no pertenece a los Enterobacteriales.
P: género de las Proteobacterias pero que no es Gamaproteobacteria.
B: género de cualquier Phylum bacteriano que no sea Proteobacteria.
NB: género no perteneciente a las bacterias.

Apéndice Dinámica.

Tabla parálogos

Categoría	Casos	% de catg.	*Rango	*D. Estándar	*Moda
PAR	37	30%	13	2.7	2
NOPR	19	23%	8	1.97	2
GHV	592	44%	45	6.25	2

*Los tres estadísticos son para el número de parálogos por GH.