57

# UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

## FACULTAD DE CIENCIAS

# ESTUDIO DE LAS SECUENCIAS EXTRA-GENICAS PALINDROMICAS REPETIDAS EN LOS GENOMAS PROCARIONTES

T E S I S

QUE PARA OBTENER EL TITULO DE

B I O L O G O

P R E S E N T A

DIEGO CLAUDIO CORTEZ QUEZADA

FACULTAD DE CIENCIAS
UNAM

DIRECTOR DE TESIS
DR. ENRIQUE MERINO PEREZ

2002

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

**M. EN C. ELENA DE OTEYZA DE OTEYZA**
Jefa de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunico a usted que hemos revisado el trabajo escrito: Estudio de las Secuencias Extragénicas Palindrómicas Repetidas en los Genomas Procariontes

realizado por Diego Claudio Cortez Quezada

con número de cuenta 9852783-2, quién cubrió los créditos de la carrera de: Biología

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

| | |
|---|---|
| Director de Tesis Propietario | Dr. Enrique Merino Pérez |
| Propietario | Dr. Lorenzo Segovia Forcella |
| Propietario | Dr. Arturo Carlos II Becerra Bracho |
| Suplente | Dr. Antonio Eusebio Lazcano-Araujo Reyes |
| Suplente | Biol. Luis José Delaye Arredondo |

**Consejo Departamental de** Biología

M. en C. Juan Manuel Rodríguez Chávez

FACULTAD DE CIENCIAS
U.N.A.M.

DEPARTAMENTO
DE BIOLOGIA

**AGRADECIMIENTOS**

Gracias a la UNAM por permitirme formar parte de ella.

Gracias a la Facultad de Ciencias por tantas y tantas horas de clases.

Gracias al IBT por tantas y tantas horas de trabajo.

Gracias a Enrique Merino, mi tutor, por todo su apoyo, confianza, dedicación, atención, enseñanza, motivación, aplicación, perseverancia, entusiasmo, conocimiento, paciencia, esperanza y entretenimiento.

Gracias también a mi primo segundo cercano Lorenzo Segovia: ¡Gracias por todo primo!

Gracias a Luis, a Arturo y a Toño por aceptar ser mis sinodales y aceptar, no sin miedo, leer esta engrandecida tesis.

Gracias a Gabriel por facilitarme la vida al permitirme usar sus bases de datos.

Gracias a Patricia por su taller.

Gracias al IE y a la Ermilo, especialmente a Amira, a Ramón y a Georgette, por permitirme ser.

Gracias a los miembros "especiales" de **Nihil Obstat**: El Carlete y el Zacky.

Gracias entrañables a todos los miembros de las ballenitas.

Gracias cariñosas a: Kach--Eugene, Marquitos, Alonso, Cuervito, Lak, Andrés, El Yupi, Francisco, Beto, Carlito, Rocío, Arlene, Argel y ??? ¡gracias amigos, gracias!

Gracias llenas de atención, paciencia, meditación, ataque sorpresa, forma Xi y Primavera-verano-otoño-invierno a mis queridísimos Shaolin-Knights: Emilio y Andrea. No se me preocupen que patadas no nos faltarán.

Gracias transatlánticas para el Gay...board.

Gracias Mariana por la felicidad que me diste... lograste siempre que saliera el sol.

Gracias Jer y Aimeé por tanto cariño... ustedes dos tendrán siempre un espacio reservado en mí.

Gracias especiales llenas de cariño para mi Paulita... sigue soñando ya que de tus sueños brotan las alas de mi felicidad. Te quiero.

Gracias hermanito por tus regaños, pese a ellos eres tú la persona que más importa.

Y por todo lo demás, gracias mamá y papá. Las siguientes 156 hojas de tesis que tienen 11 tablas, 17 figuras, 5 749 248 caracteres, 20ml de tinta, 2 kilos de papel, y representan 50 horas de escritura, 2920 horas de trabajo y 42 programas de cómputo **VAN POR USTEDES.**

... Gracias a la Vida que me ha dado tanto... me dio dos luceros... que cuando los abro... perfecto distingo perfecto distingo el negro del blanco...

## ÍNDICE

## 1. INTRODUCCIÓN

Durante la década de los ochenta y principios de los noventa se llevaron a cabo numerosas investigaciones que buscaron entender las peculiares características que presentaban las secuencias REP (**R**epetitive **E**xtragenic **P**alindromic sequences) dentro del genoma de *Escherichia coli*. Su gran abundancia, su marcada conservación, su forma doblemente palidrómica, su exclusiva presencia en regiones extragénicas y su asociación con algún gen tres-prima aledaño, son algunos de los atributos que distinguen a esta secuencia y que intrigaron por muchos años a los científicos.

Considerando los importantes avances que se han hecho en el tema, este trabajo plantea una novedosa aproximación al problema, utilizando herramientas bioinformáticas, para complementar la información disponible y tratar de resolver las interrogantes que aún existen al respecto:

a) ¿Cuál es la distribución de los elementos REP en los genomas de los organismos procariontes?

b) ¿Cuáles son las características de los elementos REP entre los distintos organismos que los presentan?

c) ¿Cuáles son las relaciones que guardan los elementos REP entre organismos?

d) ¿Es posible encontrar indicios de los procesos que conservan, producen y distribuyen a las secuencias REP, a partir de las características que presentan?

e) ¿Es factible reconstruir la historia evolutiva de las secuencias REP a partir de las características que presentan en los diferentes organismos?

f) ¿Existe una relación significativa entre las secuencias REP y las secuencias IHF? ¿Es posible entender las características de los elementos REP gracias a esta relación?

g) ¿Es posible detectar secuencias con las características de los elementos REP en otros organismos, aparte de *Escherichia coli*?

El objetivo general de la investigación es el análisis de las secuencias REP en los diferentes genomas totalmente secuenciados de procariontes. Siendo los objetivos particulares los siguientes: diseño de un programa específico de búsqueda de elementos REP; estudio de las secuencias REP en *Escherichia coli* y en otros genomas procariontes para determinar su distribución, abundancia y características; comparar los resultados obtenidos entre los diferentes organismos con secuencias REP; implementar mecanismos que permitan la detección de secuencias con las características de los elementos REP en otros organismos, además de *Escherichia coli*; analizar las relaciones entre las secuencias REP y las secuencias IHF.

Este trabajo está estructurado en el siguiente orden: en el marco teórico se muestran los avances logrados y los resultados obtenidos al respecto de las secuencias REP en las distintas investigaciones que se han realizado hasta la fecha. En la metodología se detallan todas las herramientas diseñadas, construidas y utilizadas para resolver los objetivos planteados. En la sección de resultados y discusión se exponen y analizan todos los datos encontrados. Finalmente, en la conclusión se confrontan los resultados obtenidos en este trabajo y los datos disponibles de las otras investigaciones que se han hecho.

## 2. MARCO TEÓRICO

### 2.1 La Bioinformática y el Estudio de las Secuencias de DNA

Con la determinación de la estructura del DNA por J. Watson y F. Crick[1], en 1953, y la secuenciación de la primera proteína, la insulina, en 1955 por E. Thompson[2], se sentaron las bases para el desarrollo de las ciencias de análisis de secuencias moleculares.

Al paso de los años, el desarrollo de novedosas y ágiles técnicas de secuenciación de macromoléculas, junto con el incremento y el perfeccionamiento de las herramientas de análisis, han desencadenado un avance a un ritmo extraordinario de las "ciencias genómicas", dejando a su paso una marea de asombrosos hallazgos que han permitido la construcción de modelos que nos acercan cada vez más al entendimiento de los seres vivos.

Recientemente, con el esclarecimiento de la secuencia de genomas completos, empezando por *Haemophilus influenzae* en 1995[3], se ha dejado al descubierto el más celoso de los secretos de los organismos: la base mínima a partir de la cual se estructuran. El crecimiento actual de las bases de datos de genomas completos es prácticamente exponencial, esta acumulación de información ha permitido la generación de nuevas preguntas y el diseño de novedosas estrategias de estudio.

Con el aumento en la información no procesada, los análisis de secuencias debieron sufrir una diversificación en sus campos de estudio. Algunos de estos trabajos se orientaron

[1] Watson, J.D; Crick, F.H.C. (1953) Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. *Nature*. 171: 964-967.

[2] Thompson, E.O.P. (1955) The insulin molecule. *Scientific American*. 192(5): 36-41.

[3] Fleischmann et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*. 269(5223):496-512.

hacia la búsqueda de firmas nucleotídicas[4] que permitieran descubrir patrones dentro de los genomas. De este último tipo de estudio se desprendieron los análisis de secuencias reiteradas.

## 2.2 La Repetición de Secuencias Dentro de los Genomas

Las secuencias repetidas han sido siempre un tema interesante dentro de la genómica. El conocimiento de su presencia, su distribución, su forma, su abundancia, su origen y sus posibles funciones, son algunos de los aspectos que han intrigado durante años a los investigadores. Su presencia dentro de los genomas es, al parecer, una característica común a éstos, principalmente en lo que se refiere a secuencias cortas, es decir, de 2 a 6 nucleótidos[5].

Dentro de los genomas hay múltiples eventos no-Mendelianos que provocan la diseminación y reiteración de secuencias y genes. No es novedad el saber que, en procariontes, procesos como la recombinación, la redistribución y la transferencia horizontal han modificado constantemente el arreglo interno de sus genomas e incluso se han visto involucrados en procesos de especiación[6]. Para el caso específico de la repetición de secuencias, el mecanismo

[4] Heringa, J. (1998) Detection of internal repeats: how common are they?. *Current Opinion in Structural Biology*. 8(3):338-45.

[5] Heringa, J. (1998) Detection of internal repeats: how common are they?. *Current Opinion in Structural Biology*. 8(3):338-45.

Karlin, S; Burege, C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics*. 11(7):283-90. Review.

Gelfand, M; Koonin, E. (1997) Avoidance of palindromic words in bacterial and archeal genomes: a close connection with restriction enzymes. *Nucleic Acids Research*. 25(12): 2430-2439.

[6] Koonin, E; Aravind, L; Kondrashov,A.(2000).The impact of comparative genomics on our understanding of evolution. *Cell*. 102:573-576.

Ochman, H; Lawrence, J; Grolsman, E. Lateral gene transfer and the nature of bacterial innovation. *Nature*. 405:299-304.

Lawrence, J. Gene transfer, speciation, and the evolution of bacterial genome. *Current Opinion in Microbiology*. 2:519-523.

de la Cruz, F; Davies, J. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends in Microbiology*. 8(3):128-129.

principal por el cual comúnmente aparecen es el de duplicación. Este fenómeno en procariontes puede producirse por la reintroducción de fragmentos de mRNA hacia el genoma por la acción de reversas transcriptasas o por cambios a DNA de híbridos DNA-RNA[7].

Las secuencias reiteradas se han clasificado en tres diferentes grupos de acuerdo a su forma y a las conformaciones secundarias, no-B-DNA, que pueden adoptar:

a) Las repeticiones en "tandem" pueden formar estructuras cruciformes, estructuras "slipped-loop" y Z-DNA, sólo si hay alternancia de pirimidina-purina.
b) Las repeticiones invertidas perfectas de homopurinas-homopirimidinas que pueden formar triples hélices o conformaciones H.
c) Las repeticiones largas invertidas que pueden formar estructuras cruciformes[8].

Una segunda clasificación de las secuencias repetidas las ha agrupado de acuerdo al tamaño de sus elementos más que a la forma y estructura que guardan[9]. Los grupos generados son:

---

[7] Gilson, E; Perrin, D; Hofnung, M. (1990) DNA polymerase I and a protein complex bind specifically to *Escherichia coli* palindromic unit highly repetitive DNA: implications for bacterial chromosome organization. *Nucleic Acids Research.* 18(13):3941-3952.

[8] Heringa, J. (1998) Detection of internal repeats: how common are they?. *Current Opinion in Structural Biology.* 8(3):338-45.

[9] Heringa, J. (1998) Detection of internal repeats: how common are they?. *Current Opinion in Structural Biology.* 8(3):338-45.

a) Microsatélites:

Estas repeticiones están formadas por dos o tres nucleótidos de expansión numérica variable, es decir, pueden estar repetidas de manera seguida ("tandem") de decenas hasta centenares de veces.

Un reconocido investigador en bioinformática, Dr. Samuel Karlin, pionero en el análisis de microsatélites, comprobó que los patrones y las frecuencias que presentaban éstos eran señas particulares de los genomas[10], es decir, cada genoma presentaba ciertas secuencias y ciertas frecuencias únicas, irrepetibles y características. A partir del trabajo realizado por el Dr. Karlin se desprendieron múltiples estudios de análisis de huellas de reconocimiento de los genomas[11].

b) Minisatélites:

Los minisatélites consisten en motivos de uno a seis nucleótidos que están repetidos en "tandem" (uno seguido del otro), en números que pueden variar de dos a varias docenas de repeticiones.

Estas secuencias simples repetidas (SSR, Simple Sequences Repeats), como también se les conoce, fueron las primeras secuencias repetidas que se estudiaron. El enorme polimorfismo que guardaban estas secuencias atrajo

[10] Karlin, S; Burege, C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics*. 11(7):283-90. Review.

[11] Versalovic. (1991). Distribution of repetitive DNA sequences in eubacteria and applications to fingerprint of bacterial genomes. *Nucleic Acids Research*.19(24)6823:6831.

numerosas investigaciones que intentaron infructuosamente, durante años, encontrar explicaciones convincentes al respecto[12].

Las investigaciones sobre estas secuencias se enfocaron, en un inicio, en los genomas eucariontes y más recientemente en genomas procariontes[13]. Las conclusiones de los trabajos les asignaron diferentes funciones aparentes: afectación de la expresión genética de algunos genes, firmas genómicas y, gracias a su marcado polimorfismo, merecieron el honor de figurar como mediadoras en la evolución de la regulación genética. De las muchas funciones que se les han concedido, pocas se han comprobado; su presencia y polimorfismo sigue sin entenderse en muchos casos (esto, claro, si es que hay algo que entender).

c) Secuencias repetidas de gran tamaño:

Las secuencias repetidas que superan las seis bases de nucleótidos y que pueden llegar a ser de hasta 50 bases, son comúnmente series largas invertidas de nucleótidos[14]. Estas secuencias forman estructuras secundarias crucifromes que pueden interaccionar con diversas proteínas[15].

[12] Karlin, S; Burege, C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics*. 11(7):283-90. Review.

[13] Gur-Arie, R; Cohen, C; Eitan, Y; Shelef, L; Hallerman, E; Kashi, Y. (2000) Simple sequence repeats in *Escherichia coli*: Abundance, distribution, composition, and polymorphism. *Genome Research*. 10:62-71.

[14] Heringa, J. (1998) Detection of internal repeats: how common are they?. *Current Opinion in Structural Biology*. 8(3):338-45.

[15] Yang, Y; Ames, G. (1988) DNA gyrase binds to a family of prokariotic repetitive palindromic sequences. *Proceedings of the National Academy of Sciences*. 85(23):8850-8854.

## 2.3 Las Secuencias Palindrómicas Extragenéticas Repetidas (REP)

### 2.3.1 Descubrimiento

A principios de la década de los ochenta un nuevo descubrimiento revolucionó las ideas sobre las secuencias reiteradas en los genomas. Dos grupos de investigación, el encabezado por el doctor Higgins en 1982 [16] y el dirigido por el doctor Gilson en 1984[17], al analizar los pocos genes secuenciados de *Escherichia coli* que se tenían en ese entonces, se toparon con la presencia de una serie de nucleótidos que se repetían abundantemente dentro del genoma, estaban altamente conservados, eran exclusivos de regiones extragénicas de genes tres-prima (3') y, además, presentaban un arreglo palindrómico (Fig.1). Los denominaron secuencias extragénicas repetidas palindrómicas, REP por sus siglas en inglés (también conocidas como unidades palindrómicas, PU en inglés también)[18].

### 2.3.2 Características de las Secuencias REP

Los elementos REP son secuencias invertidas cuyo tamaño varía de las 33 a las 53 bases de largo. Se reconocen dos tipos de elementos REP (Fig.1 y 2), cada uno de los cuales es exactamente la imagen opuesta del otro, es decir, son palindrómicos. Estos elementos, a su vez, presentan una estructura palindrómica interna compuesta por 14 bases y separada a la mitad por una región variable "RV" (Fig.1). Este arreglo doblemente palindrómico (Fig.2A y B) les

[16] Higgins, C.F; Ames, G.F; Barnes, W; Clément, J-M; Hofnung, M. (1982) A novel intercistronic regulatory elment of procaryotic operons. *Nature*. 298:760-762.

[17] Gilson, E; Clément, J-M; Brutlag, D; Hofnung, M. (1984) A family of dispersed repetitive extragenic palindromic DNA sequences in *E. coli*. *EMBO Journal*. 3(6): 1417-1421.

[18] Gilson, E; Clément, J-M; Brutlag, D; Hofnung, M. (1984) A family of dispersed repetitive extragenic palindromic DNA sequences in *E. coli*. *EMBO Journal*. 3(6): 1417-1421.

permite a los elementos formar diferentes estructuras de tallo y asa tanto en DNA como en RNA (Fig.3), dependiendo de la orientación de éstos y de su cercanía[19].

Cada elemento está formado por dos secciones, una de quince y la otra de dieciocho bases nucleotídicas, separadas por una región variable **"RV"** cuyo tamaño oscila entre las dos y las veinte bases[20] (Fig. 1 y2). Las secciones constantes son las que forman la estructura del tallo y son secuencias ricas en contenido de G-C (guanina-citocina)[21].

Esta peculiar secuencia se encuentra repetida en el genoma de *Escherichia coli* varios cientos de veces. Se ha estimado que su abundancia es de aproximadamente el 1% del genoma total[22], un número asombrosamente grande para una secuencia reiterada. Además de estar presentes en esta bacteria han sido inferidas, por estudios de unión de oligómeros, en varias otras enterobacterias[23]: *Shigella*, *Salmonella*, *Klebsiella*, *Enterobacter*, *Serratia*, *Erwinia* y *Proteus*[24].

---

[19] Gilson, E; Clément, J-M; Brutlag, D; Hofnung, M. (1984) A family of dispersed repetitive extragenic palindromic DNA sequences in *E. coli*. *EMBO Journal*. 3(6): 1417-1421.

[20] Goberdhan, D; Kenneth, R; Morgan, M; Bayat, H; Ames, G. (1992) Physical mapping of repetitive extragenic palindromic sequences in *Escherichia coli* and phylogenetic distribution among *Escherichia coli* strains and other enteric bactera. *Journal of Bacteriology*. 174(14):4583-4593.

[21] Gilson, E; Perrin, D; Saurin, W; Hofnung, M.(1987) Species specificity of bacterial palindromic unit. *Journal of Molecular Evolution*. 25(4):371-373.

[22] Newbury, S; Smith, N; Robinson, C; Hiles, I; Higgins, C. (1987) Stabilization of translationally active mRNA by procariotic REP sequences. *Cell*. 48: 297-310.

[23] Gilson, E; Saurin, W; Perrin, D; Bachellier, S; Hofnung, M. (1991) The BIME family of bacterial highly repetitive sequences. *Research Microbiology*.137B (2-3):217-222.

Bachellier, S; Perrin, D; Hofnung, M; Gilson, E. (1993). Bacterial interspersed mosaic elements (BIMEs) are present in the genome of *Klebsiella*. *Molecular Microbiology*. 7(4):537-544.

[24] Goberdhan, D; Kenneth, R; Morgan, M; Bayat, H; Ames, G. (1992) Physical mapping of repetitive extragenic palindromic sequences in *Escherichia coli* and phylogenetic distribution among *Escherichia coli* strains and other enteric bactera. *Journal of Bacteriology*. 174(14):4583-4593.

Más recientemente se ha identificado la presencia de los elementos REP en el "cluster" del gen fimbria de *Haemophilus influenzae* tipo b[25]. Al parecer, esta aparición de los elementos se debe a un evento de transferencia horizontal del gen al que están unidos, mas no por la transferencia exclusiva de la secuencia. Su localización restringida en el genoma de *Haemophilus influenzae* es una señal clara de la ausencia de los mecanismos relacionados con la aparición y diseminación de los elementos REP que existen en las enterobacterias.

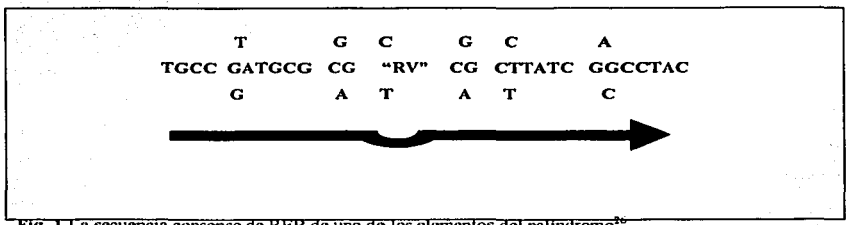



Fig. 1 La secuencia consenso de REP de uno de los elementos del palindromo[26].

Lo más interesante de estas secuencias son la serie de cualidades que las distinguen: además de ser inferida su presencia en números muy elevados dentro de los genomas de las enterobacterias, estas secuencias son exclusivamente extragénicas relacionadas forzosamente con la región tres-prima (3') de algún gen[27]. La inmensa

---

[25] Van Ham, S.M; van Alphen, L; Mooi, F.R; van Putten, O.M. (1994) The fibrial gene cluster of *Haemophilus influenzae* type b. *Molecular Microbiology*. 13(4): 673-684.

Vasconselos, A.T; Mattoso,M.A.G;de Almeida, D.F. (2000) Short interrupted palindromes on the extragenic DNA of *Escherichia coli K-12, Haemophilus influenzae* and *Neisseria meningitidis*. *Bioinformatics*. 16(11):968-977.

[26] Merino, E; Bolivar, F. (1989) The ribonucleoside diphosphate reductase gene (nrdA) of *Escherichia coli* carries a repetitive extragenetic palindromic (REP) sequence in its 3'structural terminus. *Molecular Microbiology*. 3(6): 839-841.

[27] Gilson, E; Bachellier, S; Perrin, S; Perrin, D; Grimont, P; Grimont, F; Hofnung, M. (1990) Palindromic units highly repetitive DNA sequences exhibit species specificity within Enterobacteriaceae in bacteria. *Research in Microbiology*. 141(9):1103-16.

mayoría de los elementos que se han encontrado están altamente conservados, es decir, hay pocas bases nucleotídicas distintas entre dos de ellos. La constancia de la secuencia ha permitido formar un consenso, en el cual se aprecia que sólo seis de las treinta y tres bases permiten dos opciones con igual probabilidad de aparición (Fig.1 y 2). Para tratar de entender estas características exclusivas e intrigantes de las secuencias REP se ha postulado la existencia de mecanismos moleculares específicos relacionados a su aparición, conservación, selección y movimiento dentro de los genomas.

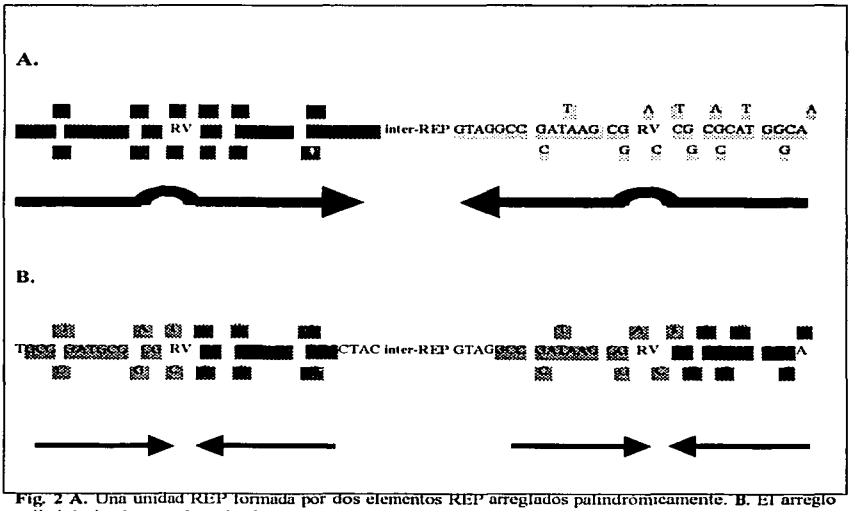



Fig. 2 **A.** Una unidad REP formada por dos elementos REP arreglados palindrómicamente. **B.** El arreglo palindrómico interno de cada elemento.

Asimismo, se ha observado que en algunos casos las unidades REP están rodeadas de otras secuencias repetidas menos abundantes que varían su disposición, arreglo de bases, tamaño y presencia. A estas asociaciones entre unidades REP de dos o más elementos y secuencias repetidas adyacentes se les ha denominado mosaico bacteriano "entremezclado" (BIME por sus siglas en inglés)[28]. Se ha supuesto que estas agrupaciones desempeñan una función trascendental al interior de los genomas de las enterobacterias.

### 2.3.3 Posibles Mecanismos de Aparición

Y. Yang y G. Ames, a finales de la década de los ochenta, comentaron que "la naturaleza palindrómica y la conservación de las secuencias REP sugieren que éstas sean reconocidas por una o varias proteínas"[29]. Esta idea ya había sido asumida por muy diversas personas que habían trabajado en el tema, ya que forzosamente algún mecanismo molecular proteico debía estar generando y favoreciendo la conservación de las características de los elementos REP.

---

[28] Bachellier, S; Perrin, D; Hofnung, M; Gilson, E. (1993). Bacterial interspersed mosaic elements (BIMEs) are present in the genome of *Klebsiella*. *Molecular Microbiology*. 7(4):537-544.

Bachellier, S; Saurin,W; Perrin, D; Hofnung, M; Gilon, E. (1994) Structural and functional diversity among bacterial interspersed mosaic elements(BIMEs). *Molecular Microbiology*. 12(1):61-70.

Gilson, E; Saurin, W; Perrin, D; Bachellier, S; Hofnung, M. (1991) Palindromic units are part of a new bacterial interpersed mosaic element (BIME). *Nucleic Acids Research*. 19(7):1375-1383.

Gilson, E; Saurin, W; Perrin, D; Bachellier, S; Hofnung, M. (1991) The BIME family of bacterial highly repetitive sequences. *Research Microbiology*.137B (2-3):217-222.

[29] Yang, Y; Ames, G. (1988) DNA gyrase binds to a family of prokariotic repetitive palindromic sequences. *Proceedings of the National Academy of Sciences*. 85(23):8850-8854.

En 1990 nuevamente el equipo de trabajo encabezado por el doctor Gilson, encontró, tras varios análisis de unión proteína-DNA, que la polimerasa I de *Escherichia coli* tenía una cierta afinidad por los fragmentos de DNA que contenían a las secuencias REP[30]. Este descubrimiento podría ser la clave para explicar la aparición de los elementos REP pues como se ha visto, la DNA polimerasa I de *Escherichia coli* puede tener actividad de reversa transcriptasa[31].

Con respecto a la aparición de elementos REP cabe mencionar que las primeras evidencias de eventos de repetición datan de 1989 cuando, investigadores del Instituto de Biotecnología de la Universidad Nacional Autónoma de México, hallaron elementos REP contiguos que, por su carácter de idénticos, eran clara muestra de eventos recientes de duplicación[32]. Aunque no es posible descartar otros mecanismos genéticos que pudieran aumentar la presencia de las secuencias REP, al parecer, la duplicación es el más común de ellos, o por lo menos, es del que se tienen evidencias más contundentes[33]

[30] Gilson, E; Perrin, D; Hofnung, M. (1990) DNA polymerase I and a protein complex bind specifically to *Escherichia coli* palindromic unit highly repetitive DNA: implications for bacterial chromosome organization. *Nucleic Acids Research.* 18(13):3941-3952.

[31] Ricchetti, M; Buc, H.(1993) *E.coli* DNA polymerase I as a reverse transcriptase. *EMBO Journal.* 12(2):387:396.

[32] Merino, E; Bolivar, F. (1989) The ribonucleoside diphosphate reductase gene (nrdA) of *Escherichia coli* carries a repetitive extragenetic palindromic (REP) sequence in its 3'structural terminus. *Molecular Microbiology.* 3(6): 839-841.

[33] Shyamala, V; Schneider, E; Ames, G.F. (1990) Tandem chromosomal duplications: role os REP sequences in the recombination event at the join-point. *EMBO Journal.* 9:939-946.

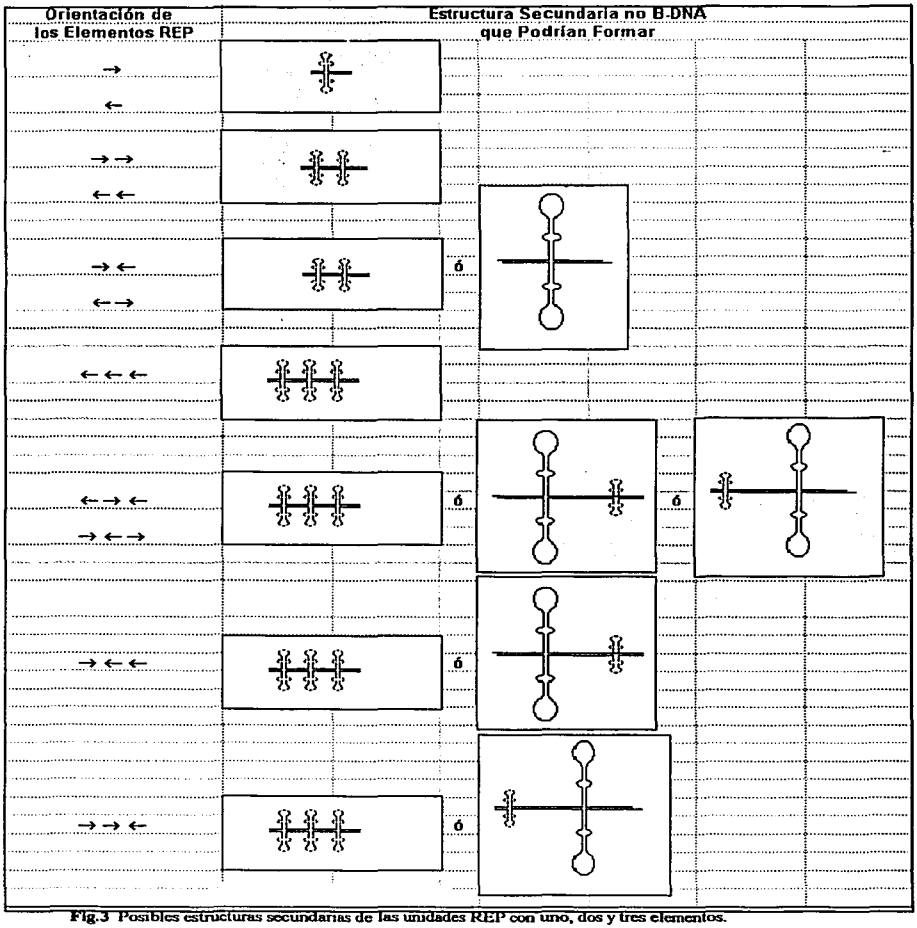



**Fig.3** Posibles estructuras secundarias de las unidades REP con uno, dos y tres elementos.

### 2.3.4 Terminación de la Transcripción y Estabilización del Mensajero

Con el transcurso de los años, las especulaciones sobre las posibles funciones de estas secuencias tan conservadas y tan abundantes han variado. En un inicio se supuso que las secuencias REP presentaban una correlación con la presencia de operones. Por lo tanto, la primera labor que se les asignó fue la de servir como terminadores de la transcripción[34]. Sin embargo, posteriores análisis detectaron que no había correlación entre la presencia de elementos REP, el final de un operón y el producto genético del mismo, lo que significaba que REP no influía en el desempeño del operón y por lo tanto no podía estar relacionada con los procesos de terminación[35]. Además, pese a las pocas secuencias REP con las que se contaba, se observó que gran parte de ellas variaba su posición con respecto a los genes entre *Escherichia coli* y *Salmonella typhimurium*, lo que contradecía la relación secuencias-REP-operones[36].

Algunos años después, los estudios de los mRNA mostraron que la secuencia REP era transcrita junto con el producto del operón[37]. Rápidamente las investigaciones se enfocaron en la importancia de las secuencias como estabilizadores del mensajero. Los experimentos que se efectuaron se centraron en la eliminación de las secuencias

[34] Gilson, E; Rousset, JO; Clement, JM; Hofnung, M.(1986) A subfamily of *E.coli* plindromic units implicated in transcription termination?. *Annales d'Institut Pasteur Microbiology*. 137B(3):259-270.

[35] Gilson, E; Bachellier, S; Perrin, S; Perrin, D; Grimont, P; Grimont, F; Hofnung, M. (1990) Palindromic units highly repetitive DNA sequences exhibit species specificity within Enterobacteriaceae in bacteria. *Research in Microbiology*. 141(9):1103-16.

Gilson,, E; Clément, J.M; Perrin, D; Hofnung, M. (1987) Palindromic units: a case os highly repetitive DNA sequences in bacteria. *Trends in Genetics*. 3(8): 225-230.

[36] Goberdhan, D; Kenneth, R; Morgan, M; Bayat, H; Ames, G. (1992) Physical mapping of repetitive extragenic palindromic sequences in *Escherichia coli* and phylogenetic distribution among *Escherichia coli* strains and other enteric bactera. *Journal of Bacteriology*. 174(14):4583-4593.

[37] Gilson, E; Bachellier, S; Perrin, S; Perrin, D; Grimont, P; Grimont, F; Hofnung, M. (1990) Palindromic units highly repetitive DNA sequences exhibit species specificity within Enterobacteriaceae in bacteria. *Research in Microbiology*. 141(9):1103-16.

Gilson,, E; Clément, J.M; Perrin, D; Hofnung, M. (1987) Palindromic units: a case os highly repetitive DNA sequences in bacteria. *Trends in Genetics*. 3(8): 225-230.

REP de los mRNA. Los resultados fueron contundentes: la desaparición de las secuencias REP producía la consiguiente caída en la vida media del mensajero[38].

Sin embargo, cabe aclarar que el efecto que tiene REP con respecto a la vida media del mensajero puede ser explicado simplemente por la estructura de tallo y asa que forma en RNA[39]. Todos los grupos de secuencias involucradas en el fenómeno de estabilización del mensajero carecen de homología, es decir, no hay conservación de bases entre estas secuencias; casi cualquier secuencia que forme una estructura de tallo y asa al final de un mensajero sirve indudablemente para su estabilización. Mientras que, las secuencias REP, presentan una clara homología debida a su alta conservación. La pregunta lógica que se desprende es: ¿por qué sólo la secuencia REP, en su función como estabilizador del mensajero, se conserva?

### 2.3.5 Estructuración del Cromosoma

Tras los resultados poco convincentes a los que se había llegado, las siguientes investigaciones se centraron en el por qué de la abundancia y la conservación de las secuencias REP.

---

[38] Newbury, S; Smith, N; Robinson, C; Hiles, I; Higgins, C. (1987) Stabilization of translationally active mRNA by procariotic REP sequences. *Cell*. 48: 297-310.

Merino, E; Becerril, B; Valle, F; Bolivar, F.(1987) Deletion of a repetitive extragenic palindromic (REP) sequence dowstream from the structural gene of *Escherichia coli* glutamate dehydrogenase affects the stability of its mRNA.*Gene*.58(2-3):305-309.

[39] Merino, E; Bolivar, F. (1989) The ribonucleoside diphosphate reductase gene (nrdA) of *Escherichia coli* carries a repetitive extragenetic palindromic (REP) sequence in its 3'structural terminus. *Molecular Microbiology*. 3(6): 839-841.

Yang, Y; Ames, G. (1988) DNA gyrase binds to a family of prokariotic repetitive palindromic sequences. *Proceedings of the National Academy of Sciences*. 85(23):8850-8854.

Entre 1988 y 1991, los investigadores Y. Yang y G. Ames, encontraron que una girasa y una proteína tipo-histona (HU) [40] reconocían y se unían significativamente a las secuencias REP. La unión de la girasa de *Escherichia coli* a las secuencias REP tenía una afinidad 1.5 x $10^{14}$ veces superior por éstas que por las secuencias con carencia de elementos REP. Como las girasas actúan durante el proceso de superenrrollamiento del cromosoma de *Escherichia coli*[41], Y. Yang y G. Ames postularon que su actividad estaría mediada por su interacción con los elementos REP. No obstante, el experimento que realizaron no demostró que hubiera una unión entre las girasas y los elementos REP durante el superenrrollamiento del cromosoma.

Las proteínas tipo histonas (HU) de *Escherichia coli* se unen a casi cualquier secuencia que pueda formar una estructura cruciforme[42] para evitar, precisamente, esta distorsión en el DNA. Aparentemente, a bajas concentraciones de proteína HU, ésta puede estimular la formación de estructuras cruciformes, pero a altas concentraciones las inhibe. La explicación a este fenómeno es el decremento en la energía libre para formar una estructura cruciforme, debido al superenrrollamiento del DNA en una estructura tipo-nucleosoma por la acción de las proteínas HU[43].

Las estructuras cruciformes que pueden formar los elementos REP (Fig.3), son favorecidas por la presencia de iones bivalentes[44] y como cualquier otra secuencia

[40] Clarkson, S; Bates, AD. (1996) Action of DNA gyrase at RIP elements in *E.coli*. *Biochemical Society Transactions*. 24(3):420.

Yang, Y; Ames, G. (1988) DNA gyrase binds to a family of prokariotic repetitive palindromic sequences. *Proceedings of the National Academy of Sciences*. 85(23):8850-8854.

[41] Pettijohn, D.(1982) Structure and properties of the bacterial nucleoid. *Cell*. 30:667-6669.

[42] Pentiggia, A; Negri, A; Beltrama, M; Blanchi, M.(1993) Protein HU binds specifically to kinked DNA. *Molecular Microbiology*. 7:343-350.

[43] Pentiggia, A; Negri, A; Beltrama, M; Blanchi, M.(1993) Protein HU binds specifically to kinked DNA. *Molecular Microbiology*. 7:343-350.

[44] Dickmann, S; Lilley, D. (1985) The anomalous gel migration of a stable cruciform: temperature and salt dependence. And some comparitions with curved DNA. *Nucleic Acids Research*.15:5765-5774.

palindrómica, pueden interaccionar con proteínas tipo histonas. La pregunta que surge es: ¿dada la abundancia de los elementos REP en el cromosoma de *Escherichia coli*, podría ser que éstos sean los que estén interactuando de forma más significativa con las proteínas HU?

A partir de los resultados anteriormente expuestos parecería factible suponer que las características de conservación, distribución y abundancia de las secuencias REP dentro del genoma de *Escherichia coli*, podrían estar relacionadas con los procesos de compactación y organización del cromosoma bacteriano.

### 2.3.6 Relación con las Proteínas IHF

Quizá, de todas las secuencias que se han asociado con los elementos REP las secuencias de reconocimiento IHF (**I**ntegration **H**ost **F**actor)[45] podrían ser parte de la clave que ayude a dilucidar los mecanismos moleculares involucrados con los elementos REP. Las proteínas IHF son heterodímeros que se unen a DNA y están involucradas con procesos de recombinación, transposición, inversión y control de la expresión genética[46]. Ambas subunidades guardan una gran homología en su secuencia de aminoácidos con la proteína HU. Las asocaciones entre secuencias REP y secuencias

---

Kohwi, Y; Kohwi-Shigematsu, T.(1988) Magnesium ion dependent triple-helix structure formed by homopurine-homopyrimidine sequences in supercoiled plasmid DNA. *Proceedings of the National Academy of Sciences*. 85:3781-3785

[45]Boccard, F; Prentki, P. (1993) Specific interaction of IHF with RIBs, a class of bacterial repetitive DNA elements located at the 3'end of transcription units. *EMBO Journal*. 12'13):5019-5027.

Clarkson, S; Bates, AD. (1996) Action of DNA gyrase at RIP elements in *E.coli*. *Biochemical Society Transactions*. 24(3):420.

Oppenheim, AB; Rudd, KE; Mendelson, I; Teff, D. (1993) Integration host factor binds to a unique class of complex repetitive extragenic DNA sequences in *Escherichia coli*. *Molecular Microbiology*. 10(1):113-122.

[46] Freundlich, M; Ramani, N; Mathew, E; Sirko, A; Tsui, P. (1992) The role of integration host factor in gene expression in *Escherichia coli*. *Molecular Microbiology*. 6(18):2557-2563.

IHF han sido denomiadas RIP por Repetitive IHF-binding Palindromic elements[47]. Han sido identificadas veintidós de estas relaciones[48] solamente en *Escherichia coli K12*. Se ha supuesto una relación importante entre estas asociaciones y la compactación del cromosoma bacteriano.

## 2.4 Análisis de Secuencias Repetidas

Los análisis de búsqueda y caracterización de secuencias reiteradas en los genomas procariontes han tenido últimamente dos enfoques distintos: uno basado en análisis por técnicas de laboratorio y el otro por técnicas computacionales o bioinformáticas[49].

El primer tipo de análisis al que se hace referencia, permite inferir indirectamente la presencia de secuencias específicas dentro de los genomas a partir de la interacción con oligómeros diseñados. Parte fundamental de este procedimiento es el conocer la secuencia que se quiere analizar en un inicio, pues el diseño de los oligómeros requiere forzosamente un mapa del cual partir. Los resultados que se obtienen por estos métodos son de presencia o ausencia, es decir, se sabe si la secuencia está o no presente en el genoma analizado y a "grosso modo" se conoce en que proporción se encuentra, pero lamentablemente, se desconocen las variaciones de la

[47] Oppenheim, AB; Rudd, KE; Mendelson, I; Teff, D. (1993) *Integration host factor binds to a unique class of complex repetitive extragenic DNA sequnces in Escherichia coli*. Molecular Microbiology. 10(1):113-122.

[48] Oppenheim, AB; Rudd, KE; Mendelson, I; Teff, D. (1993) *Integration host factor binds to a unique class of complex repetitive extragenic DNA sequnces in Escherichia coli*. Molecular Microbiology. 10(1):113-122.

[49] Saurin, W.(1987) Repetitive palindromic sequences in *Escherichia coli*.Detection and characterization with a new computer program. *Computer Applications in the Biosciences*. 3(2):121-127.

secuencia, el número exacto de éstas, su distribución y su posición relativa con respecto a los genes[50].

Todos aquellos aspectos que no pueden ser dilucidados por las técnicas tradicionales de laboratorio pueden encontrar una solución gracias a la aplicación de métodos bioinformáticos, siempre y cuando se cuente con la secuencia genómica. Hoy en día, el número de genomas totalmente secuenciados supera los cuarenta, y su número se encuentra en constante aumento, por lo que muy pronto, la falta de secuencias no será una limitante para los análisis teóricos.

Los trabajos en bioinformática parten del diseño de algoritmos computacionales para la búsqueda precisa y definida de sólo aquellos aspectos en los que el investigador está interesado. Es éste el tipo de enfoque que se utilizó en el desarrollo de esta investigación y que a continuación se detalla.

[50] Goberdhan, D; Kenneth, R; Morgan, M; Bayat, H; Ames, G. (1992) Physical mapping of repetitive extragenic palindromic sequences in *Escherichia coli* and phylogenetic distribution among *Escherichia coli* strains and other enteric bactera. *Journal of Bacteriology*. 174(14):4583-4593.

Versalovic. (1991). Distribution of repetitive DNA sequences in eubacteria and applications to fingerprint of bacterial genomes. *Nucleic Acids Research*.19(24):6823:6831.

Heringa, J. (1998) Detection of internal repeats: how common are they?. *Current Opinion in Structural Biology*. 8(3):338-45.

## 3. METODOLOGÍA

### 3.1 Generación de Bases de Datos

Para el análisis desarrollado se utilizaron todas las secuencias completas de los genomas procariontes y las anotaciones que en ellas se encuentran. Éstas se adquirieron de la base de datos de genomas completos del GenBank[51]. Los organismos analizados fueron, las **arqueobacterias:** *Aeropyrum pernix K1*[52], *Archaeoglobus fulgidus DSM4304*[53], *Halobacterium sp NRC-1*[54], *Methanothermobacter thermautotrophicus delta H*[55], *Pyrococcus abyssi*[56], *Pyrococcus horikoshii OT3*[57], *Sulfolobus solfataricus P2*[58], *Thermoplasma acidophilum*[59], *Thermoplasma volcanium GSS1*[60] y las **bacterias:** *Agrobacterium tumefaciens*[61], *Aquifex aeolicus VF5*[62], *Bacillus halodurans C-125*[63], *Bacillus subtilis 168*[64], *Borrelia burgdorferi B31*[65], *Buchnera sp. APS*[66], *Caulobacter crescentus*[67], *Campylobacter jejuni*[68], *Chlamydia muridarum*[69], *Chlamydophila*

[51] NCBI, National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/
[52] Kawarabayasi et al. (1999) Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix K1*. *DNA Research*. 6: 83-101.
[53] Klenk et al. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature*. 390:364-370.
[54] Ng et al. (2000) Genome sequence of *Halobacterium species NRC-1*. *Proceedings of the National Academy of Sciences* 97:12176-12181.
[55] Smith et al. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum deltaH*: functional analysis and comparative genomics. *Journal of Bacteriology*. 179:7135-7155.
[56] Heilig,R. *Pyrococcus abyssi* genome sequence: insights into archaeal chromosome structure and evolution. Aún no publicado
[57] Kawarabayasi et al. (1998) Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii OT3* (supplement). *DNA Research*. 5: 55-76.
[58] She et al. (2001) The complete genome of the crenarchaeon *Sulfolobus solfataricus P2*. *Proceedings of the National Academy of Sciences* 98: 7835-7840.
[59] Ruepp et al. (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature*. 407: 508-513.
[60] Kawashima et al. (2000) Determination of the complete genomic DNA sequence of *Thermoplasma volcanium GSS1*. *Proceedings of the National Academy of Sciences* 97: 14257-14262.
[61] Goodner et al. (2001) Genome Sequence of the Plant Pathogen and Biotechnology Agent *Agrobacterium tumefaciens C58*. *Science*. 294 (5550), 2323-2328.
[62] Deckert et al. (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature*. 392:353.
[63] Takami et al. (2000) Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. *Nucleic Acids Research* 28: 4317-4331.
[64] Kunst et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*. 390: 249-256.
[65] Fraser et al. (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature*. 390: 580-586.
[66] Shigenobu et al. (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera sp.* APS. *Nature*. 407: 81-86.
[67] Nierman et al. (2001) Complete Genome Sequence of *Caulobacter crescentus*. *Proceedings of the National Academy of Sciences* 98: 4136-4141.
[68] Parkhill et al. (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*. 403: 665-668.
[69] Read et al. (2000) Genome sequences of *Chlamydia trachomatis MoPn* and *Chlamydia pneumoniae AR39*. *Nucleic Acids Research*. 28 (6): 1397-1406.

*pneumoniae AR39*[70], *Chlamydophila pneumoniae CWL029*[71], *Chlamydophila pneumoniae J138*[72], *Clostridium acetobutylicum ATCC 824*[73], *Deinococcus radiodurans R1*[74], *Escherichia coli K-12 Strain MG1655*[75], *Escherichia coli O157:H7 strain EDL933*[76], *Escherichia coli O157:H7 (RIMD 0509952)*[77], *Haemophilus influenzae Rd*[78], *Helicobacter pylori 26695*[79], *Helicobacter pylori J99*[80], *Lactococcus lactis IL1403*[81], *Methanococcus jannaschii DSM 2661*[82], *Mesorhizobium loti MAFF303099*[83], *Mycobacterium leprae*[84], *Mycobacterium tuberculosis CDC1551*[85], *Mycobacterium tuberculosis H37Rv*[86], *Mycoplasma genitalium G-37*[87], *Mycoplasma pneumoniae M129*[88], *Mycoplasma pulmonis UAB CTIP*[89], *Neisseria meningitidis MC58 (ATCC BAA-335)*[90], *Neisseria meningitidis serogroup A strain Z2491*[91], *Pasteurella multocida Pm70*[92], *Pseudomonas aeruginosa PAO1*[93], *Rickettsia conorii Malish 7*[94], *Rickettsia prowazekii Madrid E*[95], *Salmonella typhimurium LT2* [96], *Salmonella typhi*[97],

---

[70] Read et al. (2000) Genome sequences of *Chlamydia trachomatis MoPn* and *Chlamydia pneumoniae AR39. Nucleic Acids Research*. 28: 1397-1406.
[71] Kalman et al. (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nature* Genetics 21: 385-389.
[72] Shirai et al. (2000) Comparison of whole genome sequences of *Chlamydia pneumoniae J138* from Japan and *CWL029* from USA. *Nucleic Acids Research* 28:2311-2314.
[73] Nolling et al. (2001) Genome Sequence and Comparative Analysis of the Solvent-Producing Bacterium *Clostridium acetobutylicum*. *Journal of Bacteriology*. 183: 4823-4838.
[74] White et al. (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans R1*. *Science*. 286: 1571-1577.
[75] Blattner et al. (1997) The complete genome sequence of *Escherichia coli K-12*. *Science*. 277:1453-1474.
[76] Perna et al. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli O157:H7*. *Nature*. 409:529-533.
[77] Hayashi et al. (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli O157:H7* and genomic comparison with a laboratory strain *K-12*. *DNA Research*. 8:11-22.
[78] Fleischmann et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. *Science*. 269:496-512.
[79] Tomb et al. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*. 388:539-547.
[80] Alm et al. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*. 397:176-180.
[81] Bolotin et al. (2001) The complete genome sequence of the lactic acid bacterium *Lactococcus lactis ssp. lactis IL1403*. *Genome Research*. 11:731-753.
[82] Bult et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*. 273 (5278): 1058-1073.
[83] Kaneko et al. (2000) Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. *DNA Research*. 7:331-338.
[84] Cole et al. (2001) Massive gene decay in the *leprosy bacillus*. *Nature*. 409:1007-1011.
[85] Fleischmann et al. Whole genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. Aún sin publicar.
[86] Cole et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*. 393:537.
[87] Fraser et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*. 270:397-403.
[88] Himmelreich et al. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Research* 24:4420-4449.
[89] Chambaud et al. (2001) The complete genom- sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acids Research* 29:2145-2153.
[90] Tettelin et al. (2000) Complete genome sequence of *Neisseria meningitidis serogroup B strain MC58*. *Science*. 287: 1809-1815.
[91] Parkhill et al. (2000) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis Z2491*. *Nature*. 404: 502-506.
[92] May et al. (2001) Complete genomic sequence of *Pasteurella multocida,Pm70*. *Proceedings of the National Academy of Sciences* 98: 3460-3465.
[93] Stover et al. (2000) Complete genome sequence of *Pseudomonas aeruginosa PA01*, an opportunistic pathogen. *Nature*. 406: 959-964.
[94] Ogata et al. (2001) Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science*. 293:2093-2098.
[95] Andersson et al. (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*. 396: 133-140.

*Sinorhizobium meliloti 1021*[98], *Staphylococcus aureus Mu50*[99], *Staphylococcus aureus N315*[100], *Streptococcus pneumoniae TIGR4 (ATCC BAA-334)*[101], *Synechocystis sp. PCC 6803*[102], *Streptococcus pyogenes M1*[103], *Thermotoga maritima MSB8*[104], *Treponema pallidum Nichols*[105], *Ureaplasma urealyticum serovar 3*[106], *Vibrio cholerae serotype O1, Biotype El Tor, strain N16961*[107], *Xylella fastidiosa 9a5c*[108], *Yersinia pestis CO-92 Biovar Orientalis*[109].

El formato de la secuencia del genoma contenida en estos archivos fue modificado con el fin de mejorar el rendimiento de los programas. Se generaron dos versiones distintas a través de dos programas escritos en Perl. La primera de ellas contenía la secuencia lineal completa del genoma, a diferencia de la original, las bases nucleotídicas fueron convertidas a una representación numérica, de tal forma que las adeninas equivalieron a 1, las citocinas a 2, las guaninas a 3 y las timinas a 4 (Fig. 4A). Lo anterior se realizó para agilizar los cálculos comparativos que el programa de búsqueda de elementos REP habría de realizar. La segunda versión del genoma presentaba nuevamente al genoma completo lineal del organismo, pero con una distinción importante: las regiones extragénicas se mantuvieron en letras minúsculas, mientras que las regiones codificantes se cambiaron para que las bases nucleotídicas ahora estuvieran representadas por letras mayúsculas (Fig. 4B). Esto se hizo con el

[96] McClelland et al. (2001) Complete genome sequence of *Salmonella* enterica serovar *Typhimurium LT2. Nature*. 413 :852-856.
[97] Parkhill et al. (2001) Complete genome sequence of a multiple drug resistant *Salmonella* enterica serovar *Typhi CT18. Nature* 413 (6858), 848-852.
[98] Galibert et al. (2001) The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science*. 293: 668-572.
[99] Kuroda et al. (2001) Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet*. 357 (9264): 1225-1240.
[100] Kuroda et al. (2001) Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet*. 357(9264):1225-1240.
[101] Tettelin et al. (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science*. 293: 498-506.
[102] Kaneko et al. (1996) Sequence analysis of the genome of the unicellular *cyanobacterium Synechocystis sp. strain PCC6803. II.* Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Research*. 3: 109-136.
[103] Ferretti et al. (2001) Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proceedings of the National Academy of Sciences* 98: 4658-63.
[104] Nelson et al. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*. 399: 323-329.
[105] Fraser et al. (1998) Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science*. 281: 375-388.
[106] Glass et al. (2000) The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. *Nature*. 407: 757-762.
[107] Heidelberg et al. (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature*. 406: 477-483.
[108] Simpson et al. (2000) The genome sequence of the plant pathogen Xylella fastidiosa. The *Xylella fastidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis. *Nature*. 406: 151-157.

objetivo de identificar en que regiones se ubicaban los elementos REP y para corroborar si éstos eran preferentemente extragénicos.

Otras bases de datos tuvieron que ser construidas para eficientizar los procesos de asignación de genes y de comparación de unidades entre genomas. Para todos aquellos organismos donde se encontraron elementos REP, se formaron varios archivos que contenían el listado de genes reportados en los genomas, la posición que guardaban éstos (bases en el genoma), su posición relativa con respecto a los demás genes (número de gen, siendo el uno el primer gen secuenciado), su característica de complementariedad (en que hebra se encontraban), la función asignada a su producto proteico y las anotaciones de homología con organismos cercanos (Fig. 4C).

Finalmente, para analizar la eficiencia del programa de búsqueda de elementos REP se produjo una base de datos con los elementos REP reportados para *Escherichia coli K12*, en donde se indicaba el número de elementos y el gen al que habían sido asignados.

---

**A.** Secuencia original del genoma de algún organismo:
**actgcatcgcatctagcgatcgcatcgcgcgcatcagcgcatcgcgcatcgcgatcgctacgct**
La transformación a números quedaría:
**1243214232142413231423214232323214213232142323214232314232412324**

**B.** Secuencia original del genoma de algún organismo:
**actgcatcgcatctagcgatcgcatcgcgcgcatcagcgcatcgcgcatcgcgatcgctacgct**
Si suponemos que esta secuencia va del nucleótido 189 al 253 y que en el genoma del organismo hay dos genes indicados de la siguiente forma:
gen = "araA" (157..196) y gen = "araB" (230..270)
La transformación quedaría:
**ACTGCATCgcatctagcgatcgcatcgcgcgcatcagcgcaTCGCGCATCGCGATCGCTACGCT**
en donde las mayúsculas representan la parte de la secuencia que está dentro de alguna región codificante.

---

[109] Parkhill et al. (2001) Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*. 413: 523-527.

C. Base de datos de las anotaciones de los genes de *Escherichia coli K12*:

| Gen | Posición (nucleótidos) | Número | Función asignada | Anotaciones | Complementariedad |
|---|---|---|---|---|---|
| thrL | 190-255 | 1 | a.a biosíntesis:Treonina | 100% idéntica a LPT_ECOLI | Hebra de secuenciación |
| thrA | 337-2799 | 2 | a.a biosíntesis:Treonina | 99% idéntica a AKIH_ECOLI | Hebra de secuenciación |
| thrB | 2801-3733 | 3 | a.a biosíntesis:Treonina | 100% idéntica a KHSE_ECOLI | Hebra de secuenciación |
| thrC | 3734-5020 | 4 | a.a biosíntesis:Treonina | 100% idéntica a THRC_ECOLI | Hebra de secuenciación |

**Fig. 4** Construcciones de las principales bases de datos utilizadas.

### 3.2 Búsqueda de Elementos REP

Tomando como punto de partida el programa de predicción de promotores diseñado por Mulligan, Hawley, Entriken y McClure[110] en 1983, se llevó a cabo la construcción de un programa, escrito en Perl, de búsqueda de elementos REP.

El programa parte de la construcción de una matriz de peso a partir de la secuencia consenso de uno de los elementos REP (Fig. 5) [111]. Se decidió partir de la secuencia consenso y no de las secuencias ya conocidas de elementos REP por las razones siguientes: i) para generar completamente las bases de datos de elementos REP, evitando sesgar la búsqueda; ii) para utilizar los elementos reportados como grupo control y así corroborar el funcionamiento del programa (el programa tendría que encontrar los elementos reportados); iii) para lograr una búsqueda de elementos REP que permitiera formar grupos de similitud.

[110] Mulligan, M; Hawley, D; Entriken, R; McClure, W.(1983) *Escherichia coli* promoter sequences predict in vitro RNA polymerase selectivity. *Nucleic Acids Research.* 12(1):789-800.
[111] Merino, E; Bolivar, F. (1989) The ribonucleoside diphosphate reductase gene (nrdA) of *Escherichia coli* carries a repetitive extragenetic palindromic (REP) sequence in its 3'structural terminus. *Molecular Microbiology.* 3(6): 839-841.

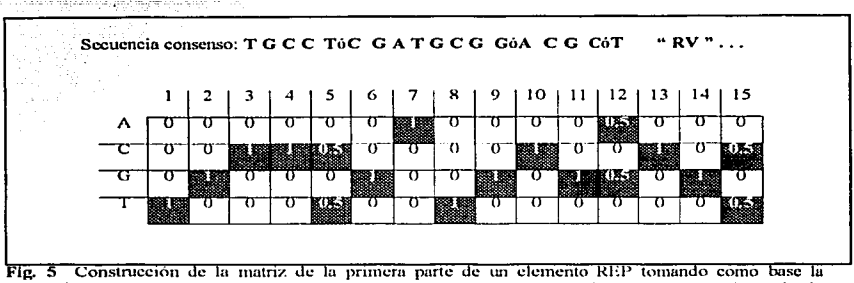



**Fig. 5** Construcción de la matriz de la primera parte de un elemento REP tomando como base la secuencia consenso de los elementos conocidos. Cuando sólo es posible encontrar un determinado nucleótido en una cierta posición el valor que adquiere éste es de 1. Por el contrario cuando determinado nucleótido no es posible encontrarlo en cierta posición el valor que adquiere éste es de cero. Para aquellas posiciones en donde es igualmente probable encontrar dos diferentes bases, el valor de éstas es de 0.5.

Se construyeron cuatro matrices iniciales, dos para cada una de las ventanas de ambos tipos de elementos REP (Fig. 2). La formación de las matrices fue realizada de acuerdo a la posición de los nucleótidos de la secuencia consenso. Dicho de otra forma, la secuencia fue ingresada a una matriz de tal manera que el valor de un nucleótido que coincidiera en tipo y posición al de la secuencia consenso presentaría el valor más alto de las cuatro posibles bases nucleotídicas (Fig.5). El valor máximo alcanzable por cualquier secuencia es, por supuesto, el total que se obtiene de la suma de los valores individuales de los nucleótidos de la secuencia consenso.

Ya que los elementos REP están formados por una región variable interna **"RV"** cuya extensión puede ir de cero a veinte nucleótidos y por dos segmentos de tamaño distinto, uno de 15 y el otro de 18 bases (Fig.1), el diseño de la búsqueda debió contemplar todas y cada una de estas características con el fin de obtener resultados óptimos al final del estudio. Para conseguirlo, se procedió a construir una primera versión del programa de búsqueda cuyo funcionamiento se detalla a continuación y que se representa en la fig. 6.

- El programa comienza la lectura del genoma del organismo a analizar en el nucleótido uno secuenciado de la base de datos que tiene la secuencia transformada a números. Forma una primera ventana con las primeras quince bases del genoma y utilizando su posición y clase (A, T, C o G) se compara la secuencia con la matriz inicial; dependiendo de cuantos nucleótidos coincidan en tipo y posición con los de la secuencia consenso el valor de la ventana variará (Fig.6A). Si el valor calculado es mayor al valor designado como umbral, el programa inicia la búsqueda de las segundas ventanas. Si por el contrario, el valor es inferior al permitido, el programa se recorre una base en la secuencia del genoma para el análisis de una nueva primera ventana (ésta iría de la segunda base del genoma a la base dieciséis) con la que efectuará nuevamente los cálculos de comparación (Fig.6A y 6B).

- Cuando una primera ventana es válida para el programa (su valor es superior o igual al valor del umbral prefijado) inmediatamente se generan treinta segundas ventanas de dieciocho bases que buscan localizar el segundo fragmento de la secuencia REP (Fig.6C). Pese a que se ha reportado que el tamaño de la región variable **"RV"** es comúnmente no mayor a veinte nucleótidos, en el estudio se decidió ampliar este valor a treinta con el fin de considerar nuevos elementos REP que pudieran presentar un mayor tamaño en esta región.

  En el ejemplo, la primera de estas ventanas iría de la base diecisiete a la treinta y cinco y la última de éstas iría de la cuarenta y siete a la sesenta y cinco. Para cada una de las treinta segundas ventanas se calcula su parecido con la secuencia consenso de la misma forma en que se describió anteriormente. La constitución nucleotídica de la región variable **"RV"** no es tomada en cuenta en el cálculo de los valores.

A continuación, al valor obtenido para la primera ventana el programa le suma, de manera independiente, cada uno de los treinta distintos valores obtenidos de las segundas ventanas (Fig.6D). Cuando la suma total de la primera ventana con alguna de las segundas supera o iguala al valor del umbral permitido, la secuencia es extraída y enviada a un archivo aparte. La extracción de la secuencia se realiza de la base de datos que permite distinguir regiones extragénicas de regiones codificantes.

- El programa recorre base por base todo el genoma del organismo, generando sucesivamente las primeras ventanas y en su caso las treinta segundas ventanas.

- Para cada secuencia válida encontrada, el programa le asigna el valor de similitud con respecto a la secuencia consenso, el tamaño de la región variable y la posición que guarda dentro del genoma (de qué nucleótido a qué nucleótido va, Fig.6E).

- Al término del primer examen por ventanas de todo el genoma se realiza una retroalimentación de la matriz inicial: a partir de la composición nucleotídica de cada una de las secuencias encontradas, el programa calcula la frecuencia de aparición de cada una de las cuatro bases para cada una de las 33 posiciones que conforman un elemento REP. Seguido de esto, cada valor obtenido es dividido entre una desviación estándar y es ingresado a una nueva matriz (Fig.6F).

- La ocurrencia igual para las cuatro bases se obtiene de la división del número de secuencias encontradas entre cuatro (por que son cuatro nucleótidos). De esta forma, una desviación estándar se calcula como la raíz cuadrada de dicha división[112] (Fig.6F).

[112] Mulligan, M; Hawley, D; Entriken, R; McClure, W.(1983). *Escherichia coli* promoter sequences predict in vitro RNA polymerase selectivity. *Nucleic Acids Research*. 12(1):789-800.

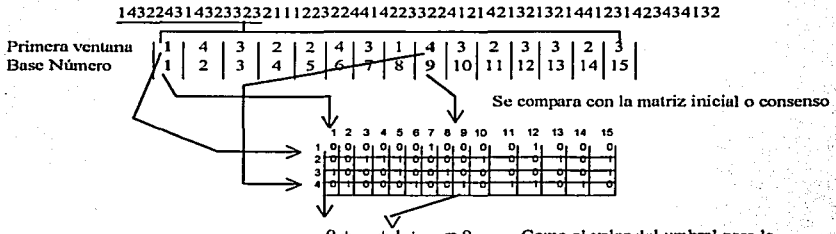
- Una vez calculados los nuevos valores para la matriz, el programa se reinicializa y comienza nuevamente la búsqueda. La retroalimentación vuelve a la matriz cada vez más robusta, permitiendo la progresiva eliminación de falsos positivos.
- Para cada ciclo los valores permitidos o de umbral de las dos ventanas y del elemento completo varían. En el primer ciclo el programa toma como válidas aquellas secuencias que tengan un 90% o más de parecido con la secuencia consenso. Esto se logra calculando el valor máximo de las matrices iniciales y calibrando la puntuación permitida al 90% de este valor máximo.

  Para cada nuevo ciclo el porcentaje de permiso baja una tercera parte entre el valor mínimo permitido en el ciclo anterior y el mínimo al que puede llegar el programa, este mínimo se situó en un valor de 68%. De esta forma, para un segundo ciclo el valor permitido se calcula como (90-68)/3, donde 90 es el valor mínimo permitido de la primera vuelta y 68 es el valor mínimo al que puede llegar el programa. Para la octava vuelta (última que se produce) el programa permite el paso de elementos cuya similitud oscila entre el 70.89% y el 69%, con respecto a la matriz previa.
- Todo el proceso se reanuda nuevamente para el segundo tipo de elemento REP (el palíndromo del ya analizado, para mayores detalles véase anexo I).

Cuando el número de turnos en el programa sobrepasaba los ocho, permitiendo la constante retroalimentación de la matriz de peso, y/o cuando se redujo el umbral a menos del 69%, el número de elementos REP en regiones codificantes se disparó mostrando la presencia de falsos elementos REP. Los datos que se generaron con estas pruebas permitieron asegurar que los parámetros óptimos a los que debía funcionar el programa para reducir a un máximo el número de falsos positivos, pero sin perder los falsos negativos, eran: un mínimo de similitud (tomando en cuenta la retroalimentación

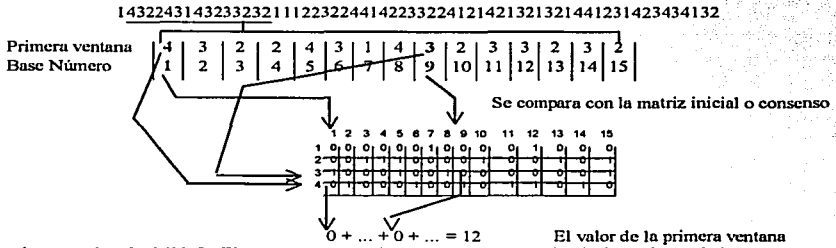
de la matriz) de un 69% con respecto a la última matriz formada, alcanzado después de un proceso de ocho turnos de búsqueda.

**A.** Secuencia del genoma en la versión de números:




Como el valor del umbral para la primera ventana es de 11.5, el valor obtenido no es suficiente y el programa la rechaza y se recorre un nucleótido en la secuencia del genoma del organismo

**B.** Secuencia del genoma en la versión de números:

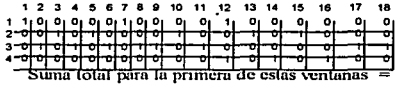



El valor de la primera ventana sobrepasa al umbral (11.5). El programa genera 30 segundas ventanas de dieciocho bases de largo para localizar (si es posible) la segunda parte del elemento REP

**C.** Secuencia del genoma en la versión de números:

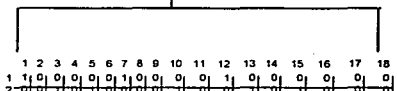
1432243143233232111223224414223322412142132132144123142343413 2

Cada segunda ventana es comparada con la matriz consenso



Suma total para la primera de estas ventanas = 13.5

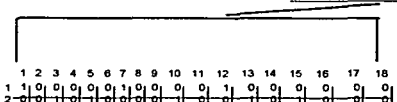
1432243143233232111223224414223322412142132132144123142343413 2



Suma total para la segunda de estas ventanas = 15

27 ventanas más

1432243143233232111223224414223322412142132132144123142343413 2



Suma total para la treintava de estas ventanas = 9

---

**D.** Se suma a la primera ventana las treinta siguientes de forma separada, dando treinta valores totales:
12 (valor de la primera ventana exitosa ) + 13.5 ( valor de la segunda ventana número uno ) = 25. 5
12 (valor de la primera ventana exitosa ) + 15 (valor de la segunda ventana número dos) = 27

27 sumas más

12 (valor de la primera ventana exitosa ) + 9 (valor de la segunda ventana número treinta) = 21

Como el valor del umbral de todo el elemento REP es de 27 y sólo la combinación de la primera ventana junto con la segunda ventana número dos iguala al valor permitido entonces, éstas ventanas son consideradas como un elemento REP exitoso, el cual quedaría como:

432243143233232 **RV** 112232244142233224

---

**E.** Es utilizada la secuencia del genoma que se encuentra modificada para distinguir las regiones extragénicas de las codificantes y el resultado final es:

Elemento REP **TGCCtgatgcggcgc tt aacgccttatccggcct**
que va del nucleótido **2** al nucleótido **37**, con una región variable **(RV)** igual a **2** y un valor de similitud con respecto a la secuencia consenso igual a **27 (90%)**.

---

F. Nueva Matriz Consenso

Suponiendo que en una primera vuelta de la búsqueda se encontraron 4 secuencias que tienen una similitud del 90% o más con la secuencia consenso entonces,

**Primera parte del elemento REP (→) de las cuatro secuencias encontradas**

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. | T | G | C | C | t | g | a | t | g | c | g | g | c | g | c |
| 2. | a | g | c | c | t | g | a | t | g | c | g | g | c | c | a |
| 3. | a | g | c | c | t | g | a | t | g | c | g | g | c | g | a |
| 4. | t | g | c | c | t | g | a | t | g | c | g | g | c | g | g |
| | t=2<br>a=2 | g=4 | c=4 | c=4 | t=4 | g=4 | a=4 | t=4 | g=4 | c=4 | g=4 | g=4 | c=4 | g=3<br>c=1 | c=1<br>a=2<br>g=1 |

**RV**

| | | |
|---|---|---|
| 1. | t | t |
| 2. | t | a |
| 3. | c | t |
| 4. | c | c |

**Segunda parte del elemento REP (→) de las cuatro secuencias encontradas**

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. | a | a | c | c | g | c | c | t | a | t | g | c | g | g | c | c | t | c |
| 2. | a | a | c | c | g | c | c | t | a | t | g | c | g | g | c | c | a | c |
| 3. | a | a | c | c | g | c | c | t | a | t | c | c | g | g | c | c | a | c |
| 4. | a | a | c | c | g | c | c | t | a | t | c | c | g | g | c | c | t | c |
| | a=4 | a=4 | c=4 | c=4 | g=4 | c=4 | c=4 | t=4 | a=4 | t=4 | c=4 | c=4 | g=4 | g=4 | c=4 | c=4 | t=2<br>a=2 | c=4 |

El valor de aparición de la base por posición se divide entre una desviación estándar. La cual se calcula de la siguiente forma: $\sqrt{(4/4)} = 1$.

Los datos se acomodan en dos nuevas matrices de peso.

**Matriz de la primera parte del elemento REP (→)**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 2 | 2 | 0 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 4 | 1 | 1 |
| 3 | 1 | 4 | 0 | 0 | 0 | 4 | 0 | 0 | 4 | 0 | 4 | 4 | 0 | 3 | 1 |
| 4 | 2 | 0 | 0 | 0 | 4 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

**Matriz de la segunda parte del elemento REP (→)**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 2 | 0 | 0 | 4 | 4 | 0 | 4 | 4 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 4 | 4 | 2 | 4 |
| 3 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Fig. 6** Esquema del funcionamiento del programa de búsqueda de elementos REP

Esta primera versión del programa presentaba dos problemas importantes: la reiteración de secuencias ya encontradas y la aparición de falsos positivos producto de la combinación de varias segundas ventanas válidas. Esto último ocurría en aquellas secuencias muy conservadas cuando el valor del umbral permitido descendía.

Dos modificaciones se hicieron para resolver estos problemas. La primera de ellas fue la desaparición, al final de cada ciclo, de los elementos encontrados. Esta desaparición se decidió hacer a través de la sustitución de la secuencia por un símbolo con valor nulo para las matrices. Así se logró evitar la reiteración de secuencias muy conservadas sin afectar el tamaño de la secuencia del genoma y la posición relativa sus nucleótidos.

La segunda modificación consistió en intercalar los ciclos de búsqueda para ambas clases de elementos REP. De esta forma, cada ciclo del programa realizaba la búsqueda para los dos elementos con un mismo umbral de salida. Esto permitió eliminar los falsos positivos que, a valores de permiso muy bajos, se formaban cuando se mezclaban ambos tipos de elementos REP.

El programa final escrito en su totalidad en Perl realizaba una búsqueda muy precisa y completa de los elementos REP, pero el tiempo de cómputo era muy prolongado y por lo tanto poco práctico para ser aplicado masivamente a los más de cuarenta genomas del estudio. El retraso en el desenvolvimiento del programa ocurría en la formación de las distintas ventanas, fue por ello que se decidió transcribir esa subrutina del lenguaje Perl al lenguaje C. La compilación de la subrutina en C permitió agilizar varias decenas de veces el proceso de análisis (para mayores detalles véase anexo I).

### 3.3 Generación de Unidades y Asignación a Genes

Para el análisis más detallado de los elementos fue necesario agruparlos en unidades y asociarlos a los genes más cercanos (Fig.7). Este procedimiento sólo se realizó para aquellos organismos que presentaron elementos REP. Para llevar a cabo este trabajo se diseñó un programa, escrito en Perl, que se explica a continuación:

- El programa lee dos archivos simultáneamente, uno donde están los elementos REP encontrados previamente (Fig.7A) y un segundo archivo en el que se encuentran los genes del organismo, sus posiciones y la cadena del DNA que los transcribe (sentido, si es la hebra de DNA cuya secuencia se reporta en el archivo GenBank[113] o antisentido, complementario, en caso contrario).

- En un inicio, el programa obtiene las posiciones iniciales de cada uno de los elementos REP encontrados y hace un simple cálculo de proximidad, tomando la posición de un primer elemento y calculando la cercanía de un segundo elemento. Si ambos elementos se encuentran a menos de doscientas bases los agrupa en una misma unidad, de lo contrario el segundo elemento pasa a formar parte de una unidad distinta (Fig. 7B). Durante este proceso y para poder identificar más fácilmente las distribuciones de los elementos dentro de cada unidad, el programa le asigna a los elementos REP TGCCTGATGCG GCGC[RV]CCGCCTTATCCGGCCTAC y similares el símbolo de (→) y a su imagen opuesta, GTAGGCCGGATAAGGCG[RV] GCGCCGCAT y similares el símbolo de (←).

- Una vez que se tienen las distintas unidades, cada una con un número de elementos variable, el programa calcula la posición dentro del genoma de toda la unidad: de la primera base del primer elemento hasta la última base del último elemento (Fig.7C).

- Al mismo tiempo el programa busca, en la base de datos del genoma, la posición inicial de los genes, su nombre y si son o no complementarios (véase arriba).

- Posteriormente, para la primera unidad generada compara la posición relativa de su nucleótido de inicio con la de los nucleótidos de inicio de los genes. Cuando por primera vez la posición de inicio de la unidad es menor a la posición de inicio de algún gen, el programa asigna la unidad a la pareja de genes que se forma entre el gen que se encuentra inmediatamente después de la unidad y el que está inmediatamente antes (Fig.7D).

- La definición del gen al que será asignada la unidad se da utilizando las características de complementariedad de los genes. El programa asigna la unidad siempre que sea posible al gen 3'. Cuando la unidad se encuentra entre dos genes 3' o entre dos regiones 5', el programa no decide a qué gen asignar la unidad y deja ambos indicados (Fig.7E).

- El archivo resultante contiene el nombre del gen y la unidad que le fue asociada. El procedimiento descrito anteriormente se realiza tantas veces como elementos REP haya (para mayores detalles véase anexo II).

[11] NCBI, National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/

**A. Base de Datos:**

| Posición dentro del genoma | Orientación del elemento |
|---|---|
| (10..47) | → |
| (55..88) | ← |
| (2033..2078) | → |
| (403540..403580) | ← |

**B. Formación de unidades:**

Unidad I: (10..47) → ... menos de doscientas bases... (55..88) ← ... más de doscientas bases ↵

Unidad II: (2033..2078) ← ... más de doscientas bases ↵

Unidad III: (403540..403580) → ...más de doscientas bases ... etc. ... ↵

**C. Cálculo de posición de unidades:**

Unidad I: (10..88) → ←
Unidad II: (2033..2078) ←
Unidad III: (403540..403580) → ... etc. ...

**D. Asignación a los genes**

Gen 1 (1..9) 3'
⇦ Unidad I: (10..88) → ←
Gen 2 (120..134) 3'

Gen 300 (1980..2000) 3'
⇦ Unidad II: (10..88) ←
Gen 301 (2120..2400) 5'

Gen 10405 (403480..403530) 5'
⇦ Unidad III: (10..88) →
Gen 10406 (403626..403837) 5'

**E. Decisión a favor de un gen**

- Gen 1 3'... Unidad I ... Gen 2 3' . No se decide a favor de ninguno.
- Gen 300 3' ... Unidad II ... Gen 301 5'. Se decide a favor del Gen 300 por su carácter 3'.
- Gen 10405 5'... Unidad III ... Gen 10406 5'. No se decide a favor de ninguno.

Fig. 7 Formación de unidades y asignación de genes

### 3.4 Formación de Grupos

Una vez que se tuvieron las unidades asociadas a los diferentes genes para todas aquellas bacterias que presentaron elementos REP, fue necesario comparar los genes entre estos organismos con el fin de obtener sus relaciones de homología.

El primer paso fue buscar, en las bases de datos de los genomas, las anotaciones de homología para los genes requeridos. Posteriormente, con dicha información recabada se hicieron grupos, es decir, se separaron los genes según los homólogos que presentaban en los otros organismos y según su asociación con las unidades REP. Por ejemplo: i) grupo de similitud formado por todos aquellos genes que tienen homólogos asociados a unidades REP en todos los organismos que presentaron elementos REP; ii) grupo de similitud formado por todos aquellos genes que tienen homólogos en todos los organismos que presentaron elementos REP, pero que sólo en ellos hay unidades REP asociadas; iii) grupo de similitud formado por todos aquellos genes exclusivos del organismo con unidades REP (todos los grupos formados están contenidos en Tabla 5).

Los grupos de homología que se formaron con la información de las bases de datos de los genomas tuvieron que ser corroborados, pues las anotaciones en los genomas no siempre son exactas. Para esto fue consultada la base de datos del Dr. Gabriel Moreno, del grupo de Biología Computacional del Dr. Julio Collado del Centro de Fijación de Nitrógeno de la UNAM. Dicha base de datos fue construida utilizando los resultados del programa de búsqueda de secuencias BLAST[114] y en ella se compilan las secuencias homólogas de cada una de las proteínas codificadas en los más de

[114] Altschul, S.F; Thomas, L; Madden, A.A; Schäffer,J.Z; Zheng, Z; Webb, M; Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 25:3389-3402.

cuarenta genomas totalmente secuenciados y disponibles públicamente en el NCBI[115]. De los datos contenidos en estas bases de datos, se tomó el gen homólogo con la mejor calificación de similitud, siempre y cuando esta calificación sobrepasara un valor de $e^{-20}$.

Todos los genes a los que no les fue encontrado un gen homólogo por ninguno de los dos métodos fueron declarados genes únicos o exclusivos de ese organismo.

## 3.5 Otros Análisis

### 3.5.1 Presencia Significativa de Secuencias REP

Para averiguar si los datos obtenidos eran significativos o podían ser producto del azar, se corrió quinientas veces el programa de búsqueda de elementos REP con distintas matrices de entrada. Se partió de las cuatro matrices iniciales que se habían producido a partir de la secuencia consenso de los elementos REP y se les varió el arreglo de los nucleótidos, manteniendo constante su proporción dentro de la secuencia. Este sistema se ideó para poder distinguir la importancia *per se* de la secuencia de los elementos REP, dicho de otra forma, el mecanismo elaborado permitió averiguar si los datos encontrados eran producto de la presencia inconfundible de las secuencias REP o si eran producto del azar dada la composición de bases de los genomas y de la secuencias REP.

Los resultados significativos fueron aquellos en donde el valor real tenía una distancia de ocho o más desviaciones estándar con respecto a la media de los valores generados al azar y en donde además, el valor real no fuera menor a la varianza muestral.

[115] NCBI, National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/

### 3.5.2 Secuencias IHF

Por los diversos reportes que relacionaban a las secuencias REP con secuencias de reconocimiento de la proteína IHF (Integration Host Factor), se decidió estudiar la periferia de las unidades REP (100 bases a cada lado) para localizar estas secuencias. Se partió de la secuencia consenso, CAATATATTGAATTT[116], para hacer los análisis. Se tomaron como buenas todas aquellas secuencias cercanas a un elemento REP que tuvieran un parecido del 80% o más con la secuencia consenso antes descrita.

Para corroborar que los resultados obtenidos eran significativos, se buscaron todas las secuencias IHF dentro de los genomas. Posteriormente, se variaron azarosamente todas ellas mil veces. Al final de cada uno de los mil turnos, se examinaron las asociaciones entre secuencias IHF y elementos REP.

### 3.5.3 Secuencia tipo-REP

Para este análisis se utilizaron bases de datos de todos los genomas procariontes totalmente secuenciados que contenían sólo a las secciones extragénicas. A cada una de ellas se le buscó, utilizando el algoritmo público MEME[117], la firma más abundante. Esto es por que las secuencias REP son la firma más abundante en las secciones extragénicas de *Escherichia coli*.

---

[116] Oppenheim, AB; Rudd, KE; Mendelson, I; Teff, D. (1993) Integration host factor binds to a unique class of complex repetitive extragenic DNA sequnces in *Escherichia coli*. *Molecular Microbiology*. 10(1):113-122.

[117] Bailey, L T; Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*. 14: 48-54.

Se obtuvo una secuencia específica para cada uno de los organismos. Cada una de ellas fue analizada para averiguar cuántas características de una secuencia REP presentaba. Se estudiaron trece características: si la secuencia buscada tenía un palíndromo de ella; si la secuencia buscada poseía un palíndromo en su interior; si los elementos encontrados se agrupaban en unidades mayores a uno; si los elementos encontrados eran más del 50% extragénicos y conservados, además de tener un número superior a 50; si conforme se reducen los parámetros de semejanza con la secuencia de inicio, el número de elementos encontrados se incrementaba exponencialmente; si la secuencia poseía un 50% o más de contenido de GC; si la secuencia era mayor a las 10 bases.

Este análisis se hizo con una modificación del programa de búsqueda de elementos REP antes descrito. El cambio hecho permitió que el programa pudiera, para cualquier secuencia inicial, hacer una búsqueda cuidadosa dentro de un organismo específico. Al final de cada examen se pudo calcular el número de características REP que presentaba cada una de las secuencias estudiadas. Con ellas fueron agrupados los organismos.

Finalmente, para averiguar si los genes que presentaron una secuencia tipo-REP eran los mismos en todos los organismos, se hizo una comparación de homología entre ellos. Para esto se utilizaron nuevamente las bases de datos desarrolladas por Dr. Gabriel Moreno, del grupo de Biología Computacional del Dr. Julio Collado del Centro de Fijación de Nitrógeno de la UNAM.

# 4. RESULTADOS Y DISCUSIÓN

## 4.1 Búsqueda de los Elementos REP en Todos los Genomas Totalmente Secuenciados de Procariontes

La primera parte del trabajo consistió en cuantificar el número de elementos REP que pudieran estar en los genomas de los organismos procariontes totalmente secuenciados (véase metodología). Antes de utilizar el programa de búsqueda de elementos REP en todos los genomas, se decidió hacer una prueba para estimar su buen funcionamiento. Este proceso se llevó a cabo analizando el genoma de *Escherichia coli K12* debido a que era el único organismo secuenciado completamente que presentaba anotaciones de elementos REP en la base de datos GenBank. La comparación se hizo entre los resultados obtenidos por el programa de búsqueda y los datos reportados en el archivo GenBank para ese genoma.

Para la comparación con la base de datos de elementos ya reportados, fue necesario utilizar también el segundo programa importante que se construyó: el programa de asignación de genes (anexo II y metodología). Esta prueba inicial implicó, por todo ello, un estudio detallado del funcionamiento completo del procedimiento empleado en la detección y análisis de los elementos REP.

El proceso se realizó de acuerdo con los grupos de similitud que el programa generó, es decir, primero se compararon los elementos obtenidos en el rango del 90% de similitud con la secuencia consenso. Posteriormente los elementos que se hallaban entre el 90% y el 83% de similitud con la matriz formada por los elementos encontrados

previamente y así sucesivamente hasta llegar al intervalo del 70% al 69% de similitud con la matriz previa, tal y como se describió en la sección de metodología.

El número de elementos, así como las unidades en las que se agruparon y la asignación del gen vecino a la unidad coincidieron perfectamente entre la base de datos generada por el programa de búsqueda y la base de datos de elementos reportados, hasta el grupo formado entre el 74.5% y el 72% de similitud. Debajo de esta marca aparecieron varias incongruencias entre las bases de datos: los datos reportados y los datos encontrados no presentaban ninguna relación; las unidades, los genes y los elementos no coincidían.

Se realizaron dos acciones para identificar la causa de tales problemas: primeramente se analizó si el proceso de retroalimentación matricial que realizaba el programa de búsqueda funcionaba correctamente, al dar positivo el análisis se procedió a examinar detalladamente los datos reportados que el programa no encontraba. Se hallaron varias inconsistencias entre éstos: reporte en la base GenBank de elementos que carecían de las características básicas de las secuencias REP, es decir, presencia de elementos con tamaños menores a los estipulados, composición y arreglo de nucleótidos que divergían en más de un 40% con respecto a la secuencia consenso e imposibilidad de formar una estructura de tallo y asa. Esta última característica se estudió a través de un programa de predicción del plegamiento, FOLDRNA del paquete de análisis Wisconsin Package Version 10.1, Genetics Computer Group (GCG), Madison, Wisc.

Finalmente, dado el buen funcionamiento lógico y práctico de los programas construidos (véase metodología, Fig. 6 y 7) y a la presencia de irregularidades en los datos reportados del GenBank, se concluyó que el procedimiento de búsqueda y caracterización de elementos REP trabajaba en el óptimo deseado y que podía ser aplicado a todos los genomas totalmente secuenciados de procariontes.

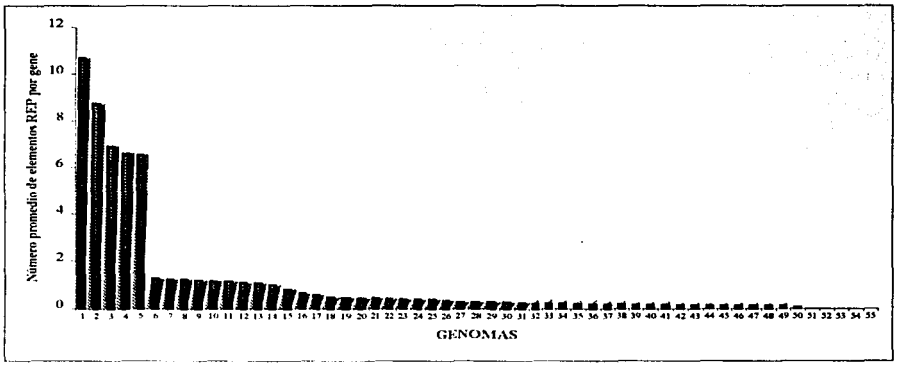



**Fig. 8** Número de elementos REP por genoma. La gráfica se ajustó al número de genes contenidos en el genoma de cada uno de los organismos. El dato de cada organismo representa el número promedio de elementos REP por gen de su genoma. (1) E.coli K12, (2) S.typhi, (3) E.coli O157H7, (4) S.typhimurium LT2, (5) E.coli O157H7 EDL933, (6) M.tuberculosis H37Rv, (7) N.meningitidis MC58, (8) M.tuberculosis CDC1551, (9) C.crescentus, (10) S.meliloti, (11) N.meningitidis Z2491, (12) M.loti, (13) D.radiodurans 1, (14) P.aeruginosa, (15) D.radiodurans 2, (16) V.cholerae 2, (17) Halobacterium sp, (18) T.pallidum, (19) P.multocida, (20) C.muridarum, (21) M.pneumoniae, (22) M.leprae, (23) X.fastidiosa, (24) A.pernix, (25) T.maritima, (26) T.volcanium, (27) P.horikoshii, (28) B.halodurans, (29) T.acidophilum, (30) M.thermoautotrophicum, (31) S.pyogenes, (32) H.pylori 26695, (33) C.pneumoniae J138, (34) C.pneumoniae CWL029, (35) C.pneumoniae AR39, (36) M.jannaschii, (37) A.fulgidus, (38) Synechocystis PCC6803, (39) V.cholerae 1, (40) H.pylori J99, (41) S.solfataricus, (42) A.aeolicus, (43) C.jejuni, (44) H.influenzae, (45) P.abyssi, (46) R.prowazekii, (47) S.aureus N315, (48) S.aureus Mu50, (49) L.lactis, (50) B.subtilis, (51) B.burgdorferi, (52) Buchnera sp, (53) M.genitalium, (54) M.pulmonis, (55) U.urealyticum.

Los primeros resultados obtenidos para todos los genomas mostraron que la presencia de elementos REP estaba aparentemente restringida a un particular grupo de bacterias, las enterobacterias, representadas por tres cepas de *Escherichia coli*, *Escherichia coli K12 Escherichia coli O157H7* y *Escherichia coli O157H7 EDL933* y por dos de sus parientes filogenéticamente más cercanos, *Salmonella typhimurium LT2 y Salmonella typhi* (Tabla 2, Fig.8).

De acuerdo a los datos obtenidos, otros organismos, aparte de los señalados anteriormente, parecieran presentar un número significativo de elementos REP en sus genomas. Para ahondar en este asunto y averiguar si era posible encontrar secuencias en proporciones equivalentes a las reportadas en los resultados (Tabla 2) sólo por cuestiones probabilísticas, dada la composición de bases de las secuencias REP y de los genomas, se hicieron quinientas matrices iniciales en las que se varió la posición de los nucleótidos de la secuencia consenso REP y para cada nueva matriz se corrió el programa de búsqueda en todos los genomas (véase metodología). Con ello se construyó un universo de datos generados al azar que pudo ser comparado con los resultados reales obtenidos para cada organismo. Las comparaciones se hicieron a través de la distancia, en desviaciones estándar, que guardaba el valor real de los elementos REP, para cada organismo, de la media de la distribución de los valores generados al azar (véase metodología).

**Tabla 1. Análisis estadístico de los elementos REP encontrados en todos los genomas de procariontes totalmente secuenciados**

| Organismo | A_aeolicus | A_fulgidus | A_pernix | B_burgdorferi | B_halodurans | B_subtilis | Buchnera_sp | C_crescentus |
|---|---|---|---|---|---|---|---|---|
| Dist.en ds | 0.38345887 | 0.5199364 | 1.40264146 | 0 | 0.25928041 | 1.47001979 | 0 | 6.32716372 |
| Valor real | 2 | 4 | 7 | 0 | 2 | 11 | 0 | 45 |
| Media | 7.27722772 | 7.61138614 | 6.14851485 | 2.2029703 | 8.02722772 | 7.56188119 | 1.48514851 | 11.6212871 |
| Des.estándar | 5.21568319 | 7.69324857 | 4.99058396 | 1.53483653 | 7.71365656 | 7.48289248 | 1.19439436 | 7.11219149 |
| Varianza | 27.2033511 | 59.1860735 | 24.9059283 | 2.35572317 | 59.5004975 | 55.9936798 | 1.42657789 | 50.5832678 |
| Mínimo | 2 | 2 | 2 | 0 | 2 | 2 | 0 | 2 |
| Máximo | 28 | 56 | 39 | 7 | 45 | 40 | 6 | 56 |

| Organismo | C_jejuni | C_muridarum | C_pneumoniae_AR39 | C_pneumoniae_CWL029 | C_pneumoniae_J138 | D_radiodurans_1 | D_radiodurans_2 | E_coli_K12 |
|---|---|---|---|---|---|---|---|---|
| Dist.en ds | 0.80434819 | 1.1432731 | 0.51006122 | 0.50461752 | 0.5112251 | 5.47894381 | 0.53829835 | 122.38937 |
| Valor real | 2 | 4 | 2 | 2 | 2 | 28 | 3 | 469 |
| Media | 4.23019802 | 5.6980198 | 5.76732673 | 5.79207921 | 5.79455446 | 9.26732673 | 6.11166253 | 3.98019802 |
| Des.estándar | 2.48648535 | 3.49872661 | 3.92109796 | 3.96339787 | 3.912171 | 5.11047402 | 5.5731176 | 3.83203213 |
| Varianza | 6.18260939 | 12.2410879 | 15.3750092 | 15.7085227 | 15.3050819 | 26.1169447 | 31.0596398 | 14.6844704 |
| Mínimo | 0 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Máximo | 14 | 22 | 29 | 29 | 29 | 40 | 42 | 29 |

| Organismo | E_coli_O157_H7_EDL933 | E_coli_O157_H7 | Haloarcula_sp | H_influenzae | H_pylori_26695 | H_pylori_J99 | L_lactis | M_genitalium |
|---|---|---|---|---|---|---|---|---|
| Dist.en ds | 81.4886131 | 76.9144225 | 1.9698707 | 0.46440405 | 0.82975302 | 0.48316318 | 0.44569795 | 0 |
| Valor real | 379 | 355 | 12 | 2 | 3 | 2 | 2 | 0 |
| Media | 6.93811881 | 6.71039604 | 8.43316832 | 6.41439206 | 5.68486352 | 6.16336634 | 6.51485149 | 1.99009901 |
| Des.estándar | 3.65095656 | 4.61551928 | 6.09177038 | 4.30659466 | 3.61553367 | 4.1393883 | 4.4873439 | 1.72987212 |
| Varianza | 21.6313969 | 21.3030182 | 37.1096664 | 18.5467575 | 13.0720837 | 17.1345355 | 20.1362553 | 2.99245756 |
| Mínimo | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 0 |
| Máximo | 41 | 44 | 56 | 23 | 24 | 25 | 28 | 10 |

| Organismo | M_jannaschii | M_leprae | M_leis | M_pneumoniae | M_pulmonis | M_thermautotrophicum | M_tuberculosis_CDC1551 | M_tuberculosis_H37Rv |
|---|---|---|---|---|---|---|---|---|
| Dist.en ds | 0.98600379 | 2.16457027 | 9.33946723 | 0 | 0.76416558 | 0.82246098 | 8.81940435 | 8.3293609 |
| Valor real | 3 | 11 | 73 | 0 | 3 | 5 | 52 | 50 |
| Media | 4.97277228 | 7.03465347 | 13.2277228 | 6.34405941 | 1.6980198 | 6.9529703 | 11.019802 | 11.019802 |
| Des.estándar | 3.04258465 | 5.08184011 | 7.81629168 | 3.94099651 | 3.92585071 | 6.07931579 | 3.8960858 | 6.00286151 |
| Varianza | 9.25732133 | 25.8250989 | 61.0944156 | 15.5314535 | 15.4123038 | 36.9580805 | 34.7638749 | 36.0343464 |
| Mínimo | 0 | 2 | 2 | 2 | 0 | 2 | 2 | 2 |
| Máximo | 19 | 54 | 60 | 22 | 67 | 39 | 39 | 39 |

| Organismo | N_meningitidis_MC58 | N_meningitidis_Z2491 | P_abyssi | P_aeruginosa | P_horikoshii | P_multocida | R_prowazekii | S_aureus_Mu50 |
|---|---|---|---|---|---|---|---|---|
| Dist.en ds | 3.86046578 | 4.79996326 | 0.3542639 | 5.80371969 | 1.11747886 | 1.81009006 | 0.56346792 | 0.77389366 |
| Valor real | 27 | 24 | 2 | 56 | 5 | 9 | 1 | 3 |
| Media | 6.9528536 | 6.46153846 | 7.22580645 | 15.9602978 | 7.00496278 | 7.39454094 | 2.87128713 | 6.11138614 |
| Des.estándar | 6.9939747 | 5.00003827 | 5.64550885 | 9.64898427 | 4.47435759 | 4.97212828 | 1.77472392 | 3.87650159 |
| Varianza | 48.9156821 | 25.0003827 | 31.8717702 | 93.1028974 | 20.0198758 | 24.7220597 | 3.14964499 | 15.0272646 |
| Mínimo | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 2 |
| Máximo | 52 | 36 | 34 | 84 | 36 | 30 | 15 | 20 |

| Organismo | S_aureus_N315 | S_aureus | S_meliloti | S_pyogenes | S_solfataricus | S_typhi | S_typhimurium_LT2 | Synechocystis_PCC6803 |
|---|---|---|---|---|---|---|---|---|
| Dist.en ds | 0.75054373 | 0.75090293 | 6.79520396 | 0.90632563 | 0.74252797 | 85.94139009 | 32.2681002 | 0.84890021 |
| Valor real | 3 | 3 | 39 | 4 | 4 | 411 | 303 | 5 |
| Media | 6.23514851 | 6.22524752 | 9.39851485 | 6.38613861 | 7.50742574 | 7.1975 | 7.31683168 | 6.08910891 |
| Des.estándar | 3.9971022 | 3.99519018 | 5.73934207 | 4.41342481 | 5.38700249 | 4.782328975 | 5.290990365 | 5.88997383 |
| Varianza | 15.976826 | 15.9615446 | 32.9400474 | 19.4783186 | 29.0197958 | 22.87067043 | 27.9936614 | 34.6917918 |
| Mínimo | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Máximo | 24 | 24 | 56 | 25 | 35 | 45 | 36 | 58 |

| Organismo | T_acidophilum | T_maritima | T_pallidum | T_volcanium | U_urealyticum | V_cholerae_1 | V_cholerae_2 | X_fastidiosa |
|---|---|---|---|---|---|---|---|---|
| Dist.en ds | 0.67088924 | 1.14159205 | 0.70531303 | 1.3415250 | 0 | 0.54168713 | 1.40948556 | 2.192216311 |
| Valor real | 4 | 7 | 5 | 5 | 0 | 4 | 7 | 11 |
| Media | 7.20544554 | 7.80940594 | 7.42326733 | 6.26237624 | 1.32009926 | 7.33995037 | 7.24813896 | 6.13647643 |
| Des.estándar | 5.962236 | 6.13178763 | 7.08905095 | 3.72710151 | 1.3908972 | 7.38433645 | 4.96635098 | 5.01775301 |
| Varianza | 35.5482581 | 37.5988195 | 50.2546434 | 13.8912857 | 1.93459501 | 54.5284249 | 24.6646421 | 25.1778453 |
| Mínimo | 2 | 2 | 2 | 0 | 0 | 2 | 2 | 2 |
| Máximo | 37 | 32 | 46 | 29 | 6 | 56 | 28 | 40 |

Datos de los valores reales y los valores de la distribución de los valores generados al azar. En sombreado, se muestran los organismos que tuvieron presencia significativa de elementos REP en su genoma.

Los resultados de estos análisis (Tabla 1) mostraron que, efectivamente, la presencia de elementos REP esta muy bien representada en las enterobacterias pero que, según parece, no son exclusivos de ese grupo. Otros tres organismos presentaron valores significativos en la presencia de elementos REP en su genoma, ellos son:

*Mycobacterium tuberculosis H37"RV"*, *Mycobacterium tuberculosis CDC1551* y *Mesorhizobium loti* (Números 6, 8 y 10 en Fig. 8, respectivamente; Tabla 1 y 2). Aunque hay organismos que presentaron datos similares a los de las tres bacterias anteriormente mencionadas, los análisis del universo de datos al azar permitieron averiguar que la presencia de los elementos era indistinguible de un fenómeno natural debido, seguramente, a la composición de bases del genoma.

Para las *Mycobacterium tuberculosis* los datos fueron similares, en ambas su valor real está a once desviaciones estándar de la media y rebasa la varianza muestral por diecisiete. Para *Mezorhizobium loti* su valor real está a 9.33 desviaciones de la media muestral y supera por doce la varianza de la distribución (Tabla 1). Aunque es importante señalar que los datos no son tan obvios como en las enterobacterias, en donde sus valores reales están entre cincuenta y siete (*Salmonella typhimurium LT2*) y ciento veintidós (*Escherichia coli K12*) desviaciones estándar de la media de la distribución de los valores al azar.

La presencia de elementos REP en bacterias no entéricas es un dato novedoso. Anteriormente, se había detectado la presencia de elementos REP sólo en enterobacterias a través de análisis por oligómeros. Ahora, con las herramientas bioinformáticas diseñadas en este trabajo se ha conseguido detección precisa y detallada de estos elementos en otros organismos.

La separación filogenética entre las bacterias que presentaron elementos REP es considerable. Las dos tuberculosis pertenecen a las actinobacterias dentro de los firmicutes. *Mezorhizobium loti* es una rhizobiacae de la subdivisión alfa perteneciente a

las proteobacterias. Las enterobactarias, aunque también son proteobacterias, pertenecen a la subdivisión gama. Este arreglo de los organismos con elementos REP en el árbol filogenético, junto con la presencia de bacterias sin éstos en ramas intermedias y las características que guardan los elementos REP en sus genomas, son indicio de que el origen de los elementos REP en organismos no entéricos se debe exclusivamente a eventos de transferencia horizontal.

**Tabla 2. Datos de la búsqueda de elementos REP por genoma según el porcentaje de similitud con respecto a la secuencia consenso de *Escherichia coli K12***

| Organismos | 90% | 83% | 77.70% | 74.50% | 72% | 70.89% | 70% | 69% | Total |
|---|---|---|---|---|---|---|---|---|---|
| [redacted] | | | | | | | | | |
| *Mesorhizobium loti* | 0 | 0 | 4 | 2 | 23 | 0 | 44 | 0 | 73 |
| *Mycobacterium tuberculosis CDC1551* | 0 | 0 | 1 | 8 | 19 | 0 | 24 | 0 | 52 |
| *Mycobacterium tuberculosis H37'Rv'"* | 0 | 0 | 1 | 6 | 19 | 0 | 24 | 0 | 50 |
| *Pseudomonas aeruginosa* | 0 | 1 | 1 | 10 | 17 | 0 | 27 | 0 | 56 |
| *Caulobacter crescentus* | 0 | 0 | 0 | 15 | 6 | 0 | 24 | 0 | 45 |
| *Sinorhizobium meliloti* | 0 | 0 | 3 | 0 | 11 | 0 | 25 | 0 | 39 |
| *Deinococcus radiodurans 1* | 0 | 0 | 0 | 7 | 4 | 0 | 18 | 0 | 29 |
| *Neisseria menengitidis MC58* | 0 | 0 | 1 | 3 | 7 | 0 | 16 | 0 | 27 |
| *Neisseria menengitidis Z2491* | 0 | 0 | 0 | 7 | 7 | 0 | 10 | 0 | 24 |
| *Halobacterium sp.* | 0 | 0 | 0 | 2 | 3 | 0 | 7 | 0 | 12 |
| *Bacillus subtilis* | 0 | 0 | 0 | 5 | 1 | 0 | 5 | 0 | 11 |
| *Mycobacterium leprae* | 0 | 0 | 0 | 7 | 1 | 0 | 3 | 0 | 11 |
| *Xylella fastidiosa* | 0 | 0 | | 6 | 1 | 0 | 4 | 0 | 11 |
| *Pasteurella multicida* | 0 | 0 | 0 | 2 | 2 | 4 | 1 | 0 | 9 |
| *Aeropyrum pernix* | 0 | 0 | 0 | 2 | 0 | 0 | 5 | 0 | 7 |
| *Thermotoga maritima* | 0 | 0 | 0 | 0 | 0 | 1 | 6 | 0 | 7 |
| *Vibrio cholerae 2* | 0 | 0 | 0 | 2 | 4 | 0 | 1 | 0 | 7 |
| *Synechocystis sp. PCC6803* | 0 | 0 | 0 | 0 | 1 | 3 | 1 | 0 | 5 |
| *Methanobacterium thermoautotrophicum* | 0 | 0 | 0 | 0 | 3 | 0 | 2 | 0 | 5 |
| *Treponema pallidum* | 0 | 0 | 1 | 0 | 2 | 0 | 2 | 0 | 5 |
| *Thermoplasma volcanium* | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 0 | 5 |
| *Pyrococcus horikoshii* | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 0 | 5 |
| *Archaeoglobus fulgidus* | 0 | 0 | 0 | 2 | 1 | 0 | 1 | 0 | 4 |
| *Streptococcus pyogenes* | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 0 | 4 |
| *Sulfolobus solfataricus* | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 4 |
| *Chlamydia muridarum* | 0 | 0 | 0 | 0 | 1 | 1 | 2 | 0 | 4 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Thermoplasma acidophilum* | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 4 |
| *Vibrio cholerae 1* | 0 | 0 | 0 | 3 | 0 | 0 | 1 | 0 | 4 |
| *Staphylococcus aureus Mu50* | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 3 |
| *Staphylococcus aureus N315* | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 3 |
| *Methanococcus jannaschii* | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 3 |
| *Mycoplasma pneumoniae* | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 3 |
| *Deinococcus radiodurans 2* | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 3 |
| *Helicobacter pylori 26695* | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 3 |
| *Aquifex aeolicus* | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 2 |
| *Helicobacter pylori J99* | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 2 |
| *Lactococcus lactis* | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 |
| *Bacillus halodurans* | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 2 |
| *Campylobacter jejuni* | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 2 |
| *Chlamydia pneumoniae AR39* | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 2 |
| *Chlamydia pneumoniae CWL029* | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 2 |
| *Chlamydia pneumoniae J138* | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 2 |
| *Pyrococcus abyssi* | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 2 |
| *Haemophilus influenzae* | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 2 |
| *Rickettsia prowazekii* | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| *Borrelia burgdorferi* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *Mycoplasma genitalium* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *Buchnera sp.* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *Mycoplasma pulmonis* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *Ureaplasma urealyticum* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

En negro se presentan las enterobacterias que tuvieron presencia significativa de elementos REP en su genoma. En gris oscuro se presentan las bacterias no entéricas en las cuales la presencia de secuencias REP fue significativa.

## 4.2 Organismos con Elementos REP

Es importante señalar que el número de elementos REP no fue constante en las cinco enterobacterias que los presentaron (Tabla 2). Hubo un decremento importante de noventa elementos (19%) entre los que tuvo *Escherichia coli K12* y los tuvo su cepa patógena *Escherichia coli O157H7*. Del mismo modo hubo un decremento de ciento catorce elementos (24.3%) entre la cepa *Escherichia coli K12* y la otra cepa patógena *Escherichia coli O157H7 EDL933* y de ciento sesenta y seis elementos (34%) con respecto a *Salmonella typhimurium LT2*. Entre las dos cepas patógenas hubo una

diferencia de veinticuatro elementos. La diferencia entre *Salmonella typhimurium LT2* y las *Escherichia coli* patógenas fue de setenta y seis y cincuenta y dos elementos con *Escherichia coli O157H7* y *Escherichia coli O157H7 EDL933* respectivamente. La otra *Salmonella* fue el segundo organismo con mayor número de elementos REP, presentó cincuenta y ocho elementos menos que *Escherichia coli K12*, treinta y dos, cincuenta y seis y ciento ocho elementos más que *Escherichia coli O157H7, Escherichia coli O157H7 EDL933 y Salmonella typhimurium LT2* respectivamente.

Con respecto al grado de conservación de los elementos se observaron diferencias importantes entre estos cinco organismos. Aparentemente, *Escherichia coli K12* presentó también el número más elevado de elementos conservados: 314. Las dos cepas patógenas de *Escherichia coli* tuvieron notorias diminuciones en el número de elementos conservados, 206 y 190 (Tabla 2), mientras que las dos *Salmonella* presentaron una asombrosa disminución, ya que *Salmonella typhimurium LT2* sólo mostró cuatro elementos altamente conservados (con un 90% o más similitud con la secuencia consenso) y *Salmonella typhi* sólo mostró dos. Eso sí, ambas con un enorme número de elementos que oscilaron en el intervalo de 74.5-70.89% de similitud (Tabla2).

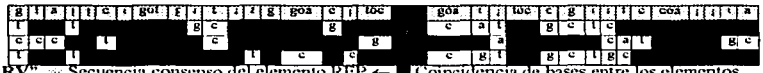
Existen dos posibles razones para explicar los datos encontrados en las *Salmonella*: i) La primera iría acorde con la pérdida de elementos en relación con la divergencia de los organismos en el tiempo, es decir, los elementos REP de estas bacterias presentan un deterioro en sus secuencias producto de la incapacidad de los mecanismos moleculares relacionados con las secuencias REP para conservarlas; ii) La segunda posibilidad sería que los elementos REP se adaptaran a las condiciones que dictara el organismo, es decir, que estos elementos estuvieran realmente conservados

pero que su secuencia consenso difiriera de la secuencia consenso de *Escherichia coli K12*.

Un examen detallado de las secuencias REP de ambas *Salmonella* que se hallaban en el intervalo de 74.5-70.89% de parecido con la matriz de peso (Tabla 3), mostró la presencia de posiciones conservadas entre la mayoría de estos elementos. Sin embargo, las bases de estas posiciones conservadas coincidían con la base reportada en el consenso de *Escherichia coli*. En las posiciones variables fue imposible determinar un patrón debido a que las frecuencias de las cuatro bases eran equivalentes. De este sencillo análisis se desprende que la alteración de los elementos REP en las *Salmonella* se debe más a una degradación de éstos por acumulación de mutaciones sin sentido, que a un cambio de forma para adaptarse a las condiciones internas de la bacteria.

Es importante señalar también que los cambios en las secuencias REP de las *Salmonella* no están en relación con la conservación del palíndromo interno. En la Tabla 3 se puede observar como, si dividimos al elemento REP en dos segmentos a partir de la región variable (**"RV"**), los cambios en el primero de los segmentos no se correlacionan con los cambios en el segundo.

**Tabla 3. Secuencias REP poco conservadas de *Salmonella***



**"RV"**. Secuencia consenso del elemento REP ←. Coincidencia de bases entre los elementos degradados de Salmonella typhimurium LT2 y la secuencia consenso. Bases que cambian con respecto al consenso.

Para los otros tres organismos cuyos análisis mostraron que la presencia de elementos REP en su genoma no era debida al azar, la historia es muy distinta. En las tres bacterias los elementos REP están muy mal conservados, no son tan abundantes como en las enterobacterias, cerca del 90% de ellos están inmersos total o parcialmente

en una región codificante y no se agrupan en unidades mayores a un elemento. Todas estas características reafirman la hipótesis de que estos elementos llegaron a las dos tuberculosis y a *Mezorhizobium loti* por transferencia horizontal. Dada la frecuencia de este fenómeno entre las bacterias, es posible que los elementos REP se hayan transferido desde las enterobacterias hasta otros procariontes, tal y como se demostró con los elementos REP encontrados en el "cluster" del gen de la fimbria en *Haemophilus influenzae* tipo b[118]. Lo que probablemente hizo la diferencia, es que en estos tres organismos las secuencias REP encontraron las herramientas moleculares propicias que los duplicaron y repartieron por el genoma. Posiblemente, las herramientas moleculares fueron deficientes e imperfectas, ocasionando que los elementos no se conservaran, y se distribuyeran sin aparente sentido dentro de los genomas (zonas codificantes).

El interés del trabajo es analizar la importancia de los elementos REP dentro de los genomas. Debido a que el posible origen de los elementos en las dos tuberculosis y en *Mesorhizobium loti* es la transferencia horizontal, se quebranta la línea que les dio origen dejando de lado la relevancia que pudieran desempeñar en los genomas donde en un inicio se seleccionaron y donde seguramente desempeñan su papel más primordial, objetivo perseguido en esta investigación. Además, la mala calidad que presentan es muestra de que en estos nuevos huéspedes no han podido desempeñar las mismas funciones, ya sea por que no existen los mecanismos moleculares necesarios o por que las bacterias no las necesitan. Por todo ello, los subsiguientes análisis se restringieron al

[118] Van Ham, S.M; van Alphen, L; Mooi, F.R; van Putten, O.M. (1994) *The fibrial gene cluster of haemophilus influenzae type b.* Molecular Microbiology. 13(4): 673-684.

Vasconcelos, A.T; Mattoso,M.A.G;de Almeida, D.F. (2000) *Short interrupted palindromes on the extrogenic DNA of Escherichia coli K-12, haemophilus influenzae and Neisseria meningitidis. Bioinformatics. 16(11):968-977.*

grupo de las enterobacterias, que es en donde estos elementos REP surgieron y en donde, dada la pulcritud de sus rasgos, realizan sus funciones más importantes.

## 4.3 Características de los Elementos REP Encontrados en las Enterobacterias

Es notoria la conservación de los elementos REP en las tres *Escherichia coli*. 70% de todos los elementos en *Escherichia coli K12*, el 53% en *Escherichia coli O157H7* y el 58% en *Escherichia coli O157H7 EDL933*, se parecieron en un 90% o más a la secuencia consenso, es decir, entre estos elementos la mayor divergencia posible es del 20% y esto ocurriría solamente cuando las bases que son distintas para una secuencia, con respecto al consenso, no son las mismas bases que son distintas para otra secuencia, también con respecto al consenso. Un elemento REP está compuesto por treinta y tres bases así que, aproximadamente, el cambio de tres bases equivale al 90% de similitud. Un cambio de 6 bases sería la máxima diferencia entre dos secuencias de este grupo.

Si se amplía el intervalo y se analizan las secuencias que comparten un 83% o más de similitud, en este caso con respecto a la matriz de peso formada por los elementos que guardan una similitud del 90% con la secuencia consenso, el número de elementos en las tres bacterias aumenta considerablemente. Ahora, *Escherichia coli K12* presentaría un 79% de sus elementos en este grupo, *Escherichia coli O157H7* sufriría un incremento importante y presentaría un 70% de sus elementos y finalmente, *Escherichia coli O157H7 EDL933* también aumentaría considerablemente el número de elementos hasta llegar a un 78% de éstos. Lo anterior muestra que, aunque en las

*Escherichia coli* patógenas hubo una caída en el número de elementos altamente conservados, es indiscutible el mantenimiento de las secuencias.

Un dato que es relevante en el estudio y que es un parámetro de la conservación de los elementos REP, es el cambio en la región variable **"RV"** del elemento (Fig. 1). Este cambio puede observarse tanto en la variación de las bases como en la variación del tamaño de esta región (Fig. 9).

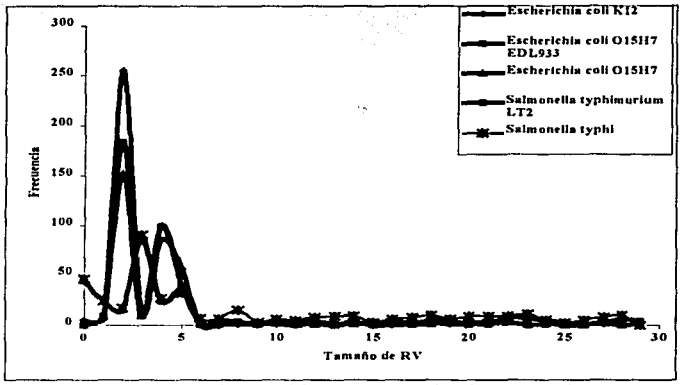



Fig. 9 El tamaño de las regiones variables ""**RV**"" en los elementos de las tres *Escherichia coli* y en las dos *Salmonella*.

Como se muestra en la figura 9, es muy marcada la preferencia por ciertos tamaños de la región variable en las tres *Escherichia coli*: 2, 4 y 5 bases, con un promedio entre las tres bacterias de 197, 95 y 45 elementos respectivamente. Aunque es posible que existan elementos con regiones que sobrepasen las 5 bases y que pudieran llegar a tener hasta 29 bases de largo, éstos siempre se mantienen en números reducidos (1 a 5 dependiendo del genoma). Aparentemente, el tamaño de esta región pareciera estar bajo una presión selectiva que le impide variar fuera de un cierto intervalo.

A diferencia de sus parientes, las dos *Salmonella* presentaron, como ya se había mencionado, un decremento importante en la conservación de sus elementos. En estas bacterias hay una clara preferencia por valores de **"RV"** distintos a los de sus parientes, siendo los tamaños de 0, 3 y 5 los predilectos, con frecuencias de 44, 84 y 36 respectivamente en *Salmonella typhimurium LT2* y de 46, 90 y 32 respectivamente en *Salmonella typhi*. *Salmonella typhimurium LT2* presentó un aumento considerable en cuanto a **"RV"** de gran tamaño se refiere, 77 de sus elementos sobrepasaron las 8 bases de tamaño. *Salmonella typhi*, por su parte, fue la que presentó las frecuencias más elevadas de **"RV"** de gran tamaño, 141 elementos superaron las 8 bases de tamaño. Este aumento desmedido en los tamaños de **"RV"** es una clara muestra de la degradación de los elementos, producto de una relajada presión de selección. Ya que esta región variable, en el elemento REP, es la que permite la formación del asa (Fig. 1 y 2), entre mayor sea ésta más energía es necesaria para la formación de la estructura secundaria. Aunado a esto, en las *Salmonella* también se encontró un elevado número de elementos (44 para *Salmonella typhimurium LT2* y 46 para *Salmonella typhi*) con **"RV"** de cero, característica que imposibilitaría la formación de una estructura de tallo y asa similar a la formada por las secuencias consenso de REP (la estructura que se llegaría a formar sería de menor tamaño ya que probablemente algunas bases del elemento pasarían a formar parte del asa).

La aparición de elementos REP que estaban inmersos total o parcialmente en una región codificante fue, para las tres *Escherichia coli*, de aproximadamente un 10% de los elementos. Lo interesante es que la presencia de estos elementos fue constante en los diferentes grupos de similitud formados por el programa. Es decir, el número de elementos REP en regiones codificantes en el grupo de elementos cuya similitud con la

secuencia consenso es de un 90% o más, era equivalente al número de elementos REP en regiones codificantes presentes en el grupo del 70.5-72% de similitud con la matriz de peso del turno previo. Sólo en el último grupo se percibe un aumento en estos elementos (20% de los elementos encontrados). Para *Salmonella typhymurium LT2* las cosas cambian, 95 de sus 303 elementos (equivalente al 31% de los elementos localizados) se encuentran total o parcialmente inmersos en una región codificante. Estos elementos se distribuyeron de la siguiente forma: en las primeras tres vueltas no hubieron, en la cuarta hubo 5, en la quinta 19, en la sexta 34, en la séptima 25 y en la octava 12. Como se podrá ver el número varía según el grupo de similitud. En los elementos muy cercanos (en secuencia) al consenso de *Escherichia coli K12* se mantienen en regiones extragénicas. Conforme los parámetros se hacen más laxos el número aumenta hasta un máximo que se obtiene en la sexta vuelta, para después volver a descender. Este descenso en el número de elementos REP en regiones codificantes en los últimos turnos permite inferir que la mayoría de éstos pudieran ser realmente elementos REP y no se tratan de falsos positivos. En la otra *Salmonella* sucede exactamente el mismo fenómeno, 116 de sus 411 elementos (30%) estuvieron en regiones codificantes con una distribución de 0, 0, 0, 1, 7, 68, 34 y 8 para las ocho vueltas que realizó el programa.

### 4.4 Características de las Unidades REP Encontradas

Sorprendentemente, el total de unidades no conserva la misma proporción que ocurrió con los elementos REP, pues ahora fueron las *Salmonella* las que poseyeron el mayor número. *Salmonella typhi* fue la que mayor número presentó con 368, seguida de *Salmonella typhimurium LT2* quien presentó 264. Finalmente, *Escherichia coli K12*

tuvo 254, *Escherichia coli O157H7* y *Escherichia coli O157H7 EDL933* presentaron 239 y 227 respectivamente.

Claramente se pueden distinguir una serie de características generales de las unidades presentes en las cepas de *Escherichia coli*. La más importante de todas es la asombrosa cantidad de unidades formadas por elementos solitarios, aproximadamente un 63% de todas las unidades de las tres *Escherichia coli* presentaron esta característica. En el restante 27% se encuentran las unidades con dos o más elementos (Fig. 10A, B, C). En *Salmonella typhimurium LT2* la proporción, elementos solitarios / elementos en unidades mayores a dos, es aún más grande, pues cerca del 92.5% de las unidades mostraron elementos solitarios (Fig. 10D). En el restante 7.5% se encuentran 21 unidades con dos elementos, una unidad con tres y dos unidades con cuatro. *Salmonella typhi* es similar a su congénere, el 89% de las unidades son solitarias, mientras que en el restante 11% hay cuarenta y dos unidades de dos elementos y dos unidades que tiene tres. En esta última bacteria el aumento en el número global de unidades es lo que mantiene equivalentes las proporciones con respecto a la otra *Salmonella* (Fig. 10E).

Con respecto a las unidades solitarias es relevante señalar que, la generalidad de las unidades idénticas que presentaron las tres *Escherichia coli* (113, aproximadamente 47% del total de unidades que presentaron) estaban formadas por elementos solitarios, es decir, alrededor del 62% de las unidades que se conservan totalmente están compuestas por elementos solitarios. Un restante 38% aproximadamente, estuvo compuesto por las unidades que podrían conformar estructuras más variadas de tallo y asa, esto gracias a la disposición de sus elementos (Fig.3 y Tabla 5).

Las unidades con dos y tres elementos estuvieron también bien representadas. Sin embargo, conforme el número de elementos por unidad se vio aumentado, la frecuencia de éstas decayó (Fig. 11). Fueron escasos los casos en donde el número de elementos por unidad superó la cifra de diez (Fig. 11 y Tabla 5).

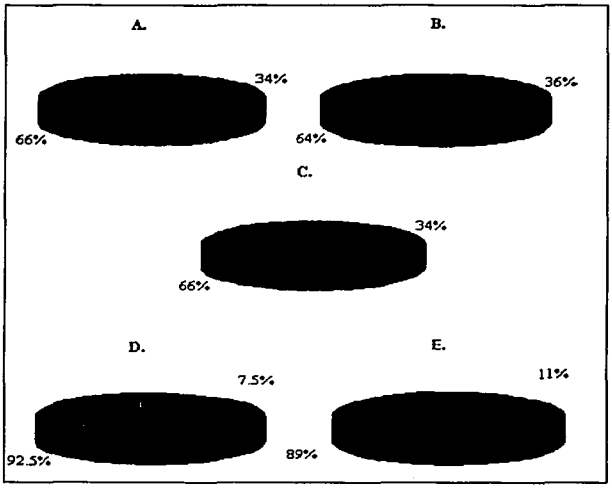



**Fig. 10** Proporción de unidades con elementos solitarios ■ o con dos o más elementos ■ **A.** Datos de *Escherichia coli K12* **B.** Datos de *Escherichia coli O157H7 EDL933* **C.** Datos de *Escherichia coli O157H7* **D.** Datos de *Salmonella typhimurium LT2*. **E.** Datos de *Salmonella typhi.*

Las unidades REP reportadas habían sido siempre relacionadas al extremo tres-prima de algún gen. Es por ello que fue una sorpresa encontrar elementos REP en regiones cinco-prima-cinco-prima. Tres unidades se localizaron en estas áreas en las cinco enterobacterias. Las tres *Escherichia coli* presentaron cuatro unidades compartidas con esta característica; *Escherichia coli K12* y la *Escherichia coli O157H7* tuvieron dos unidades extras en su genoma. Aunque son pocas las unidades que se

encontraron bajo esta situación, cuatro de ellas se hallaron altamente conservadas. Lo trascendental de este hallazgo es que las unidades REP, entre sus múltiples funciones asociadas, han figurado como terminadores de la transcripción[119] y como estabilizadores de mRNA[120]. Su presencia en regiones cinco-prima les impide realizar esas actividades.

La explicación más lógica a dicho hallazgo, dada la reducida cantidad de elementos encontrados, es que la maquinaria de generación de elementos REP esté, de algún modo, reconociendo las regiones tres prima de los genes e insertando elementos REP exclusivamente en dichas zonas, y que por algún fenómeno independiente a los elementos REP éstos aparecieron en regiones cinco-prima. Esto sucede raramente, pero cuando ocurre, las unidades pueden ser conservadas.

En las *Salmonella* hubo un aumento en el número de unidades REP en secciones cinco-prima de ambos genes, *Salmonella typhi* presentó 21 de estas unidades y *Salmonella typhimurium LT2* presentó 9. Este aumento va de acuerdo con la idea de que en estas bacterias la maquinaria molecular que genera los elementos REP está modificada y no es tan eficiente. Posiblemente, los fallos en el reconocimiento de la región tres-prima de algún gen conllevaron a la aparición de unidades en otras zonas, ya sean estas regiones codificantes o cinco-prima-cinco-prima. Aunque no es descartable la posibilidad de que las apariciones de unidades REP en regiones cinco-prima-cinco-

[119] Gilson, E; Rousset, JO; Clement, JM; Hofnung, M.(1986).*A subfamily of E.coli plindromic units implicated in transcription termination?*.Ann Inst Pasteur Microbiol. 137B(3):259-270.

Gilson, E; Saurin, W; Perrin, D; Bachellier, S; Hofnung, M. (1991) *The BIME family of bacterial highly repetitive sequences.* Research Microbiology.137B (2-3):217-222.

[120] Gilson, E; Saurin, W; Perrin, D; Bachellier, S; Hofnung, M. (1991) *The BIME family of bacterial highly repetitive sequences.* Research Microbiology.137B (2-3):217-222.

Newbury, S; Smith, N; Robinson, C; Hiles, I; Higgins, C. (1987) *Stabilization of translationally active mRNA by procariotic REP sequences.* Cell. 48: 297-310.

Merino, E; Becerril, B; Valle, F; Bolivar, F.(1987).*Deletion of a repetitive extragenic palindromic (REP) sequence dowstream from the structural gene of Escherichia coli glutamate dehydrogenase affects the stability of its mRNA.*Gene.58(2-3):305-309.

prima se deban exclusivamente al arreglo del material genético al interior del cromosoma.

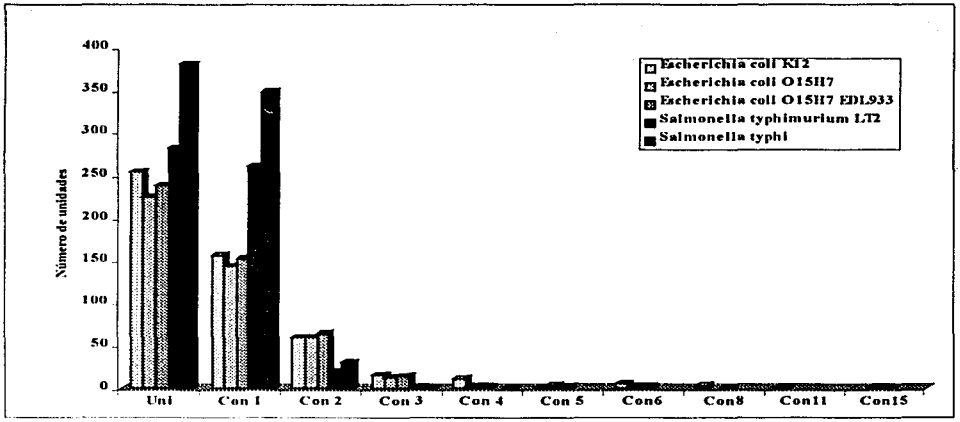



Fig. 11 Las diferentes unidades encontradas y su número de elementos.

La frecuencia de unidades con distintas orientaciones varía ampliamente (Fig.12). Las *Salmonella* son los dos organismos con mayor número de unidades solitarias. *Salmonella typhimurium LT2* es el organismo con mayor número de unidades que tiene la forma (←). Mientras que *Salmonella typhi* es la que obtuvo la mayor cantidad de unidades (→). Las unidades solitarias que se construyen de izquierda a derecha (→) fueron mayoritarias también en *Escherichia coli K12* y en *Escherichia coli*

*O157H7 EDL933*, y las que se construyen de derecha a izquierda (←) estuvieron presentes de manera importante en *Escherichia coli O157H7*. Es importante señalar que aunque uno de los elementos REP se encontró en mayor abundancia, las proporciones de ambas clases de elementos REP al interior del genoma fueron similares (Fig. 12).

Para las unidades que presentaron más de dos elementos la organización más común para las cinco bacterias fue (→ ←), aún así, viéndose muy reducidas en las dos *Salmonella*. Conforme el número de elementos por unidad aumentó, la frecuencia de éstas en las tres *Escherichia coli* disminuyó (Fig. 12). Sólo las unidades con un número reducido de elementos fueron las que, al parecer, más se conservaron y las que más fácilmente han aparecido por todo el genoma.

Se encontró que las unidades muy grandes estaban hechas por la reiteración continua de una unidad inicial compuesta por uno o dos elementos, esto se pudo saber por una simple comparación de las secuencias de las unidades. Los cotejos mostraron que todas las secuencias de una unidad con un alto número de elementos (5-15) fueron prácticamente idénticas (incluso en las secuencias de la **"RV"** y de la región inter-REP, que son las que más cambios sufren), es decir, todas se formaron a partir de un mismo molde. Lo interesante es que estas unidades son muestra de la intensidad con la que la maquinaria que produce las duplicaciones puede actuar en sitios muy precisos dentro del genoma.

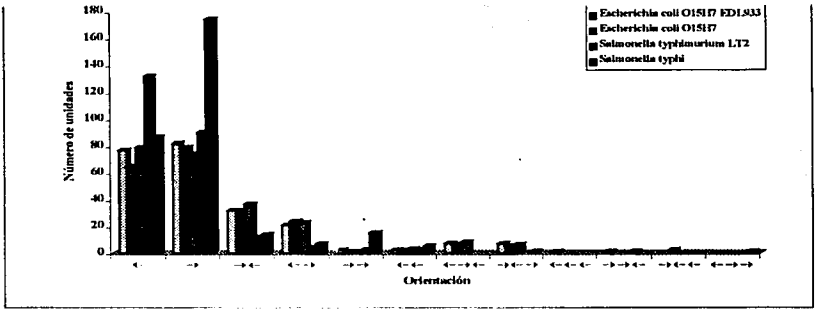



Fig. 12 Número de unidades y la orientación de los elementos. Se muestran las posibles conformaciones de las unidades con uno, dos y tres elementos.

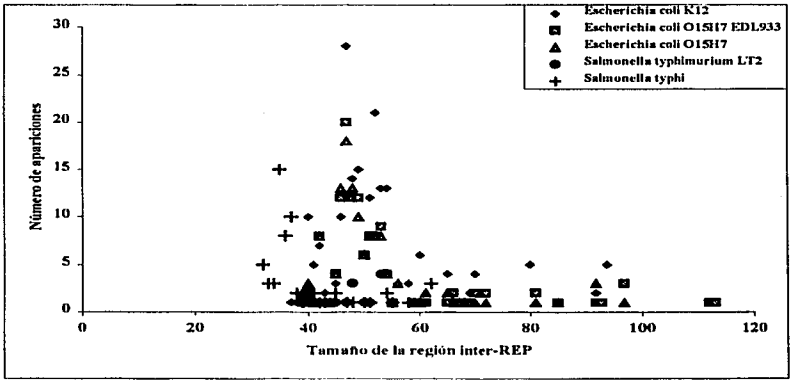



Fig. 13 Tamaño de las regiones inter-REP y su frecuencia en los distintos organismos. Obviamente, sólo las unidades con dos o más elementos fueron contempladas.

Una revisión de las distancias que existen entre los distintos elementos REP en aquellas unidades que poseen dos o más elementos (Fig. 13), mostró una aparente predilección por los tamaños que oscilan entre 39 y 60 bases. Tamaños menores a 39 no fueron encontrados, mientras que los tamaños mayores a 60 se hallaron en números muy reducidos. En ambas *Salmonella*, particularmente, los tamaños pueden llegar a ser de 35 bases, pero nunca superan las 60.

En lo referente a distancias entre unidades se encontró que éstas son poco constantes. Al igual que para las distancias inter-REP, pareciera que hay un tamaño mínimo. Dicho de otra forma, una unidad no se encuentra cerca de otra a menos que haya una cierta cantidad de bases (Tabla 4). Las distancias máximas entre unidades pueden considerar varias miles de bases (Tabla 4), sin embargo, el promedio de distancias entre *Escherichia coli K12 y Salmonella typhimurium LT2* es parecido (aproximadamente 17 000 bases) y menor a las distancias de las dos cepas patógenas (23 000 bases, Tabla 6). *Salmonella typhi*, por el contrario, presentó un decremento en todas las distancias, es decir, esta bacteria posee unidades más cercanas.

**Tabla 4. Distancias entre unidades**

| | Distancia mínima | Distancia máxima | Distancia Promedio |
|---|---|---|---|
| *Escherichia coli K12* | 641 | 175793 | 17800 |
| *Escherichia coli O157H7* | 648 | 199271 | 23000 |
| *Escherichia coli O157H7 EDL933* | 512 | 199303 | 23000 |
| *Salmonella typhimurium LT2* | 502 | 252279 | 16900 |
| *Salmonella typhi* | 218 | 121654 | 12591 |

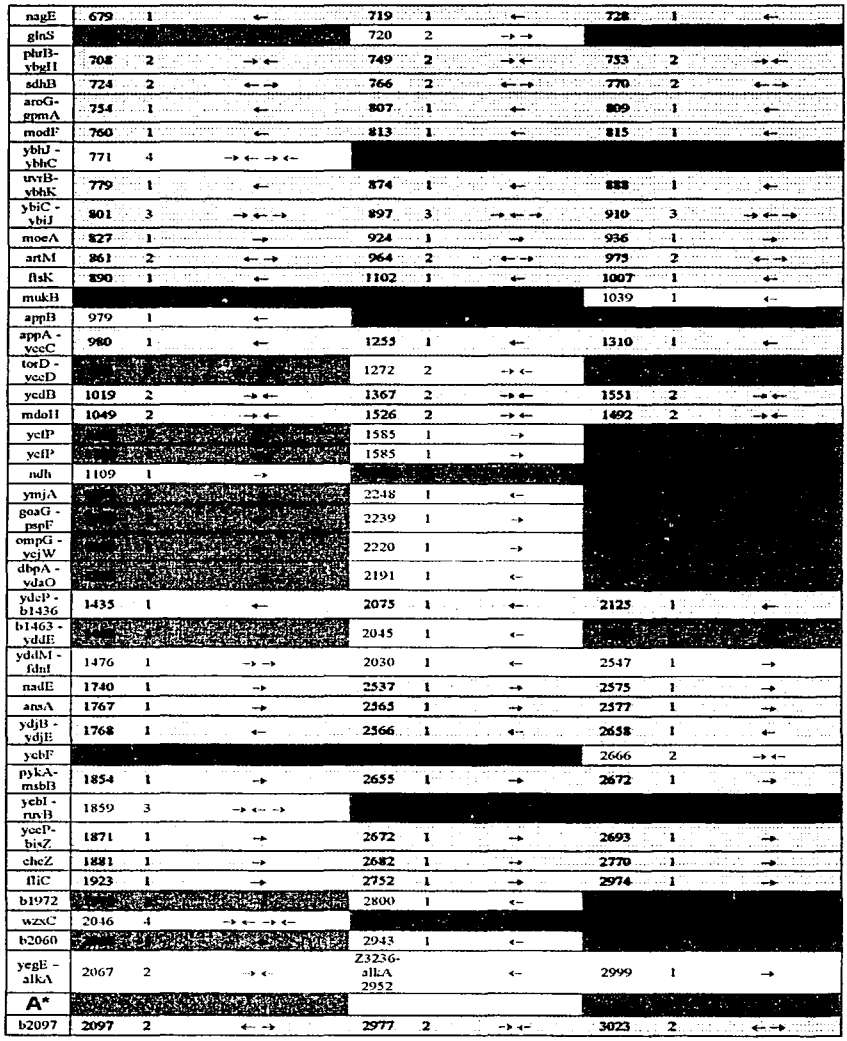
## 4.5 Análisis de los Grupos de Homología Formados Entre las Cinco Enterobacterias

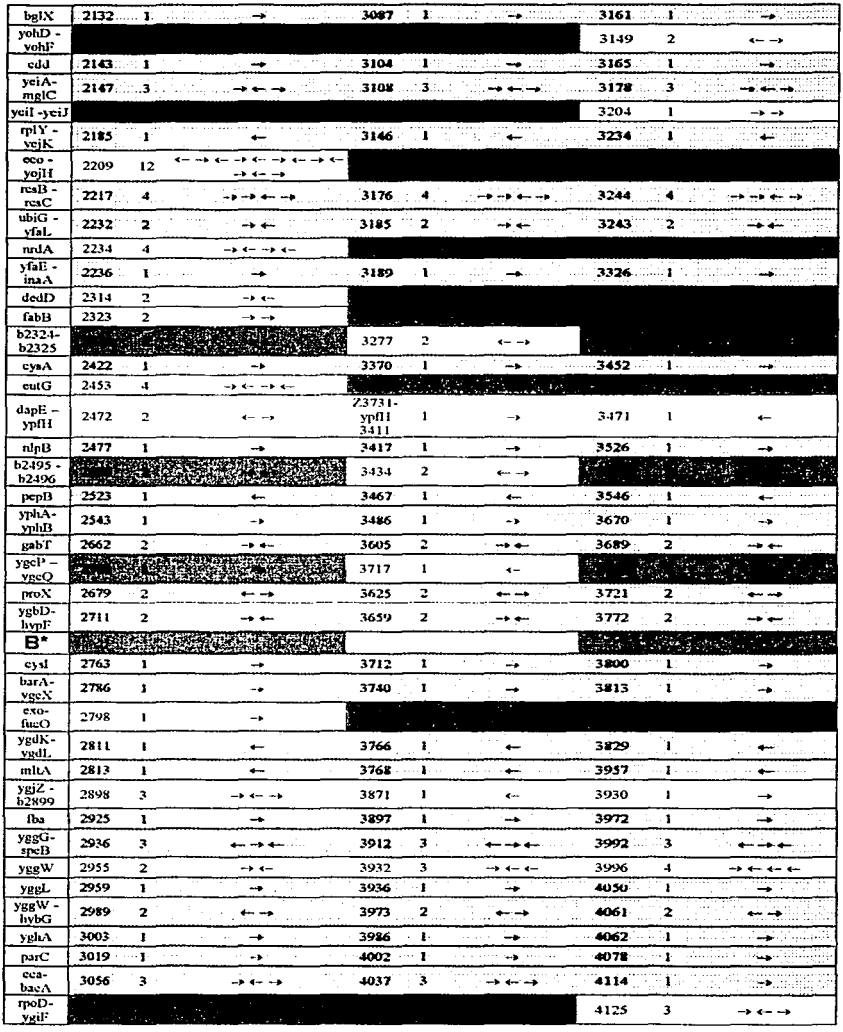
### Tabla 5. Grupos formados de acuerdo a la homología de los genes y a las unidades REP que tienen asociadas

### Tabla 5.1 Grupos formados para las *Escherichia coli*

**Tabla 5.1.1 Genes homólogos en donde las tres presentan unidades REP asociadas**

| *Escherichia coli K12* (180 unidades & 363 elementos) | | | *Escherichia coli O157H7 EDL933* (180 unidades & 299 elementos) | | | *Escherichia coli O157H7* (180 unidades & 308 elementos) | | |
|---|---|---|---|---|---|---|---|---|
| **Gen** | **Pos** | **REP** | **Orientación** | **Pos** | **REP** | **Orientación** | **Pos** | **REP** | **Orientación** |
| b0005 - yaaA | 6 | 2 | → ← | 6 | 2 | → ← | 6 | 2 | → ← |
| caiA | 39 | 2 | → ← | 42 | 2 | → ← | 42 | 2 | → ← |
| araA | 62 | 3 | ← ← ← | | | | | | |
| yabI - yabJ | 65 | 2 | → ← | | | | | | |
| yacK - gcd | 123 | 1 | ← | 127 | 1 | ← | 127 | 1 | ← |
| mesJ - yaeO | 188 | 1 | ← | 190 | 1 | ← | 190 | 1 | ← |
| cutF - yaeF | 192 | 2 | ← → | 194 | 2 | ← → | 194 | 2 | ← → |
| gmhA | 222 | 3 | ← → ← | | | | | | |
| yafL | 227 | 2 | ← → | 259 | 2 | ← → | 262 | 2 | ← → |
| yafM - fhiA | 228 | 4 | → ← → ← | | | | | | |
| yahG | 321 | 4 | → → → → | | | | | | |
| yahO - prpR | 329 | 1 | ← | 387 | 1 | ← | 392 | 1 | ← |
| prpE | 335 | 3 | → ← → | | | | | | |
| codA – cynR | 337 | 2 | → → | | | | | | |
| lacY | 343 | 1 | ← | 401 | 1 | ← | 405 | 1 | ← |
| mhpE | 352 | 9 | → ← → ← → ← → ← → | | | | | | |
| yaiL - yaiM | 354 | 4 | → ← → ← | 414 | | → | 418 | 2 | → ← |
| yaiW - b0379 | | | | 343 | 1 | → | | | |
| b0392 - yaiD | 392 | 1 | ← | 448 | 1 | ← | 452 | 1 | ← |
| yajD -tsx | 410 | 1 | ← | 467 | 1 | ← | 472 | 1 | ← |
| yajQ – yajR | 426 | 1 | → | 485 | 1 | → | 489 | 1 | → |
| gsk – ybaL | 477 | 2 | ← → | 536 | 2 | ← → | 540 | 2 | ← → |
| ybaQ - ybaR | 483 | 5 | ← → ← → ← | | | | | | |
| arcC-purK | 521 | 3 | → → ← | | | | | | |
| fepA | 584 | 1 | ← | 629 | 1 | ← | 635 | 1 | ← |
| cstA-ybdH | 598 | 1 | ← | 644 | 1 | ← | 649 | 1 | ← |


nagE 679 1 ← 719 1 ← 728 1 ←
glnS 720 2 → →
phrB-ybgH 708 2 → ← 749 2 → ← 753 2 → ←
sdhB 724 2 ← → 766 2 ← → 770 2 ← →
aroG-ppmA 754 1 ← 807 1 ← 809 1 ←
modF 760 1 ← 813 1 ← 815 1 ←
ybhJ - ybhC 771 4 → ← → ←
uvrB-ybbK 779 1 ← 874 1 ← 888 1 ←
ybiC - ybiJ 801 3 → ← → 897 3 → ← → 910 3 → ← →
moeA 827 1 → 924 1 → 936 1 →
artM 861 2 ← → 964 2 ← → 979 2 ← →
flsK 890 1 ← 1102 1 ← 1007 1 ←
mukB 1039 1 ←
appB 979 1 ←
appA - yceC 980 1 ← 1253 1 ← 1310 1 ←
torD - yceD 1272 2 → ←
ycdB 1019 2 → ← 1367 2 → ← 1551 2 → ←
mdoH 1049 2 → ← 1526 2 → ← 1492 2 → ←
ycfP 1585 1 →
ycfP 1585 1 →
ndh 1109 1 →
ymjA 2248 1 ←
gcaG - pspF 2239 1 →
ompG - ycjW 2220 1 →
dbpA - ydaO 2191 1 ←
ydeP - b1436 1435 1 ← 2075 1 ← 2125 1 ←
b1463 - ydjE 2045 1 ←
yddM - fdnI 1476 1 → → 2030 1 ← 2547 1 →
nadE 1740 1 → 2537 1 → 2575 1 →
ansA 1767 1 → 2565 1 → 2577 1 →
ydjB - ydjE 1768 1 ← 2566 1 ← 2658 1 ←
ycbF 2666 2 → ←
pykA-msbB 1854 1 → 2655 1 → 2672 1 →
ycbI - ruvB 1859 3 → ← →
yceP-bisZ 1871 1 → 2672 1 → 2693 1 →
cheZ 1881 1 → 2682 1 → 2770 1 →
fliC 1923 1 → 2752 1 → 2974 1 →
b1972 2800 1 ←
wzxC 2046 4 → ← → ←
b2060 2943 1 ←
yegE - alkA 2067 2 → ← Z3236-alkA 2952 ← 2999 1 →
A*
b2097 2097 2 ← → 2977 2 → ← 3023 2 ← →


bglX 2132 1 → 3087 1 → 3161 1 →
yohD - yohF 3149 2 ← →
eddJ 2143 1 → 3104 1 → 3165 1 →
yciA-mglC 2147 3 → ← → 3108 3 → ← → 3178 3 → ← →
yciI -yciJ 3204 1 → →
rplY - yejK 2185 1 ← 3146 1 ← 3234 1 ←
eco - yojH 2209 12 ← → ← → ← → ← → ← → ← →
rcsB - rcsC 2217 4 → → ← → 3176 4 → → ← → 3244 4 → → ← →
ubiG - yfaL 2232 2 → ← 3185 2 → ← 3243 2 → ←
nrdA 2234 4 → ← → ←
yfaE - inaA 2236 1 → 3189 1 → 3326 1 →
dedD 2314 2 → ←
fabB 2323 2 → →
b2324-b2325 3277 2 ← →
cysA 2422 1 → 3370 1 → 3452 1 →
eutG 2453 4 → ← → ←
dapE – ypfH 2472 2 ← → 23731-ypfH 3411 1 → 3471 1 ←
nlpB 2477 1 → 3417 1 → 3526 1 →
b2495 - b2496 3434 2 ← →
pepB 2523 1 ← 3467 1 ← 3546 1 ←
yphA-yphB 2543 1 → 3486 1 → 3670 1 →
gabT 2662 2 → ← 3605 2 → ← 3689 2 → ←
ygcI* – ygcQ 3717 1 ←
proX 2679 2 ← → 3625 2 ← → 3721 2 ← →
ygbD-hypF 2711 2 → ← 3659 2 → ← 3772 2 → ←
B*
cysI 2763 1 → 3712 1 → 3800 1 →
barA-ygcX 2786 1 → 3740 1 → 3815 1 →
exo-fucO 2798 1 →
ygdK-ygdL 2811 1 ← 3766 1 ← 3829 1 ←
mltA 2813 1 ← 3768 1 ← 3957 1 ←
ygjZ - b2899 2898 3 → ← → 3871 1 ← 3930 1 →
fba 2925 1 → 3897 1 → 3972 1 →
yggG-speB 2936 3 ← → ← 3912 3 ← → ← 3992 3 ← → ←
yggW 2955 2 → ← 3932 3 → ← ← 3996 4 → ← ← ←
yggL 2959 1 → 3936 1 → 4050 1 →
yggW - hybG 2989 2 ← → 3973 2 ← → 4061 2 ← →
yghA 3003 1 → 3986 1 → 4062 1 →
parC 3019 1 → 4002 1 → 4078 1 →
eca-bacA 3056 3 → ← → 4037 3 → ← → 4114 1 →
rpoD-ygiF 4125 3 → ← →

| C* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| yqjI-aer | 3071 | 2 | →← | 4052 | 2 | →← | 4129 | 2 | →← |
| ygjK | 3080 | 6 | →←→←→← | 4061 | 5 | →←→←→ | 4142 | 8 | →←→←→←→ |
| ygjO | 3084 | 2 | →← | 4065 | 2 | →← | 4160 | 2 | →← |
| yqjG | 3102 | 1 | → | 4083 | 1 | → | 4165 | 1 | → |
| yhaL - yhaM | 3107 | 2 | →← | 4088 | 2 | →← | 4211 | 2 | →← |
| yraP-yraQ | 3150 | 2 | →← | 4135 | 2 | →← | 4229 | 2 | →← |
| infB | 3168 | 2 | ←→ | 4153 | 2 | ←→ | 4415 | 2 | ←→ |
| yhdP | 3245 | 6 | →←←→←→ | 4224 | 1 | ← | 4300 | 1 | → |
| yhfR-yhfS | 3375 | 1 | ← | 4341 | 1 | ← | 4446 | 1 | ← |
| D* | | | | | | | | | |
| mreA-yrfE | | | | | | | 4428 | 1 | ← |
| yhgI | 3414 | 1 | → | 4373 | 1 | → | 4565 | 1 | → |
| arsC | 3503 | 1 | → | 4493 | 1 | → | 4592 | 1 | → |
| yhjD | 3522 | 2 | →← | 4521 | 2 | →← | 4594 | 2 | →← |
| yhjE - yhjG | 3523 | 1 | ← | 4522 | 1 | ← | 4594 | 1 | ← |
| yhjH | 3525 | 1 | ← | 4524 | 1 | ← | 4647 | 1 | ← |
| proK | 3545 | 2 | ←→ | 4547 | 2 | ←→ | 4615 | 2 | ←→ |
| aytA - viaI | 3572 | 1 | ← | 4580 | 1 | ← | 4654 | 1 | ← |
| aldB | 3588 | 2 | →← | 4590 | 2 | →← | 4830 | 2 | →← |
| gyrB | 3699 | 1 | → | 4756 | 1 | → | 4858 | 1 | → |
| pstA | 3726 | 1 | ← | 4782 | 1 | ← | 4913 | 1 | ← |
| rep-gppA | 3777 | 1 | ← | 4838 | 1 | ← | 4981 | 1 | ← |
| aslB - aslA | | | | 4862 | 1 | ← | | | |
| yigJ - yigK | 3824 | 1 | ← | 4889 | 1 | ← | 4961 | 1 | ← |
| ubiB - fadA | 3845 | 6 | →←→←→← | | | | | | |
| rffA - mobB | | | | | | | 4991 | 2 | →→ |
| hemN-glnG | 3868 | 2 | ←→ | | | | | | |
| yihN - yshA | 3875 | 1 | ← | 4939 | 1 | ← | 5010 | 1 | ← |
| yiiF-fdhE | 3891 | 1 | → | 4959 | 1 | → | 5030 | 1 | → |
| rhaA | 3904 | 8 | →←→←→←→← | 4973 | 4 | ←←→← | 5044 | 3 | ←→← |
| yiiM-cpxA | 3911 | 2 | ←→ | 4980 | 2 | ←→ | 5051 | 2 | ←→ |
| pfkA | 3917 | 1 | ← | 4985 | 1 | ← | 5056 | 1 | ← |
| metL | 3941 | 1 | ← | 5017 | 1 | ← | 5085 | 1 | ← |
| glnA | 3946 | 1 | → | 5022 | 1 | → | 5090 | 1 | → |
| pllD | 3952 | 1 | → | 5029 | 1 | → | 5097 | 1 | → |
| yijP | 3956 | 2 | →← | 5033 | 2 | →← | 5101 | 2 | →← |
| argE | 3958 | 6 | ←→←→←→ | | | | | | |
| aceA | 4016 | 3 | ←→← | 5095 | 3 | ←→← | 5163 | 3 | ←→← |
| yjbB-pepE | 4021 | 1 | ← | 5100 | 1 | ← | 5168 | 1 | ← |
| pgi | 4026 | 2 | ←→ | 5112 | 2 | ←→ | 5238 | 2 | ←→ |
| yjbH | 4030 | 1 | → | 5116 | 1 | → | 5242 | 1 | → |
| malE | 4035 | 3 | ←→← | 5121 | 3 | ←→← | 5247 | 3 | ←→← |
| lamB | 4037 | 3 | ←→← | | | | | | |
| ubiA - plsB | 4041 | 1 | ← | 5127 | 1 | ← | 5253 | 1 | ← |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| alr | 4054 | 1 | → | 5139 | 1 | → | 5265 | 1 | → |
| acs | 4070 | 2 | ← → | 5155 | 2 | ← → | 5281 | 2 | ← → |
| glpP - yjeO | 4078 | 11 | → ← → ← → ← → ← → ← → | | | | | | |
| phnK | 4098 | 2 | → ← | 5185 | 2 | → ← | 5310 | 2 | → ← |
| phnD | 4106 | 2 | ← → | 5192 | 2 | ← → | 5317 | 2 | ← → |
| phnA | 4109 | 12 | ← → ← → ← → ← → ← → ← → | | | | | | |
| yjdF | 4122 | 1 | ← | 5208 | 1 | ← | 5333 | 1 | ← |
| yjeI -yjeJ | 4145 | 1 | ← | 5232 | 1 | ← | 5356 | 1 | ← |
| vacB | 4180 | 1 | → | 5267 | 1 | → | 5389 | 1 | → |
| aidB - yjfN | 4188 | 1 | ← | 5275 | 1 | ← | 5397 | 1 | ← |
| ytfH-cpdB | 4213 | 1 | ← | 5302 | 1 | ← | 5425 | 1 | ← |
| msrA | 4220 | 1 | ← | 5309 | 1 | ← | 5432 | 1 | ← |
| ytfQ | 4228 | 2 | ← → | 5317 | 2 | ← → | 5440 | 2 | ← → |
| yjfG-yjgA | 4234 | 2 | ← → | 5322 | 2 | ← → | 5445 | 2 | ← → |
| nrdD | 4239 | 2 | → ← | 5327 | 2 | → ← | 5450 | 2 | → ← |
| mgtA-yjgF | 4243 | 1 | → | 5332 | 1 | → | 5454 | 1 | → |
| pepA | 4261 | 2 | → ← | | | | | | |
| yjgQ - yjgR | 4263 | 1 | ← | 5353 | 1 | ← | 5474 | 1 | ← |
| fimH - gntP | 4321 | 1 | ← | 5395 | 1 | ← | 5517 | 1 | ← |
| uxuB | 4324 | 1 | ← | 5398 | 1 | ← | 5520 | 1 | ← |
| yjjT | 4372 | 2 | → ← | | | | | | |
| yjjV - yjjW | 4379 | 8 | ← → ← → ← → ← → | | | | | | |
| nadR - yjjK | 4391 | 3 | ← → ← | 5464 | 3 | ← → ← | 5588 | 3 | ← → ← |
| trpR - yjjX | 4394 | 2 | → ← | 5467 | 2 | → ← | 5591 | 2 | → ← |

**Notas:**

A*: gsp 5'– 5' b2989 1 ← gsp 5'—5' Z4343 1 → ECs3872 5'—5' ECs3873 1 ←

B*: ivbL 5'—5 'emrD 1 → ivbL 5'—5 'emrD 1 ← ECs4613 5'—5' ECs4614 1 →

C*: yqjH 5'–5' yqjI 1 → yqjH 5'—5' yqjI 1 ← ECs3952 5'–5' ECs3953 1 →

D*: yjiR 5'–5' yjiS 1 ← yjiR 5'—5' yjiS 1→ ECs5302 5'–5' ECs5303 2 ← ←

**Tabla 5.1.2 Genes homólogos de las tres bacterias que presentan unidades REP asociadas sólo en *Escherichia coli K12* y *Escherichia coli O157H7 EDL933***

| *Escherichia coli K12* (6 unidades & 6 elementos) | *Escherichia coli O157H7 EDL933* (6 unidades & 6 elementos) | *Escherichia coli O157H7* |
|---|---|---|

| Gen | Pos | REP | Orientación | Pos | REP | Orientación | Pos | REP | Orientación |
|---|---|---|---|---|---|---|---|---|---|
| b1525 | | | | | | | | | |
| b2086 – gatR_1 | 2086 | 1 | → | 2969 | 1 | ← | | | |
| yeiL – yeiM | | | | | | | | | |
| glpD – yzgL | 3426 | 1 | ← | glpD-Z4787 4390 | 1 | → | | | |
| yicN | 3663 | 1 | → | Z5151-yicN 4721 | 1 | ← | | | |
| mdoB | | | | | | | | | |

**Tabla 5.1.3 Genes homólogos de las tres bacterias que presentan unidades REP asociadas sólo en *Escherichia coli K12* y *Escherichia coli O157H7***

| *Escherichia coli K12* (8 unidades & 12 elementos) | *Escherichia coli O157H7 EDL933* | *Escherichia coli O157H7* (8 unidades & 10 elementos) |
|---|---|---|

| Gen | Pos | REP | Orientación | Pos | REP | Orientación | Pos | REP | Orientación |
|---|---|---|---|---|---|---|---|---|---|
| lspA | 27 | 2 | → ← | | | | 30 | 1 | ← |
| mutT-b0100 | 99 | 4 | ← → ← → | | | | 103 | 2 | ← → |
| yleB | | | | | | | | | |
| yceC-yceF | | | | | | | | | |
| yqgF – yggR | | | | | | | | | |
| baeR | | | | | | | | | |
| nusA | | | | | | | | | |
| yicO | | | | | | | | | |

**Tabla 5.1.4 Genes homólogos de las tres bacterias que presentan unidades REP asociadas sólo en *Escherichia coli O157H7 EDL933* y *Escherichia coli O157H7***

| *Escherichia coli K12* | *Escherichia coli O157H7 EDL933* (16 unidades & 22 elementos) | *Escherichia coli O157H7* (16 unidades & 23 elementos) |
|---|---|---|

| Gen | Pos | REP | Orientación | Pos | REP | Orientación | Pos | REP | Orientación |
|---|---|---|---|---|---|---|---|---|---|
| caiF - caiE | | | | 37 | 1 | ← | 37 | 2 | → ← |
| prpB | | | | | | | | | |
| cynX | | | | | | | | | |
| yajF | | | | | | | | | |
| lysZ | | | | | | | | | |
| moeB | | | | | | | | | |
| cpsB | | | | | | | | | |
| hyfB | | | | | | | | | |
| purL | | | | | | | | | |
| b2332 | | | | 3285 | 2 | ← → | 3345 | 2 | → ← |
| phoU | | | | | | | | | |
| pstS | | | | | | | | | |
| glpK | | | | | | | | | |
| yjaB | | | | | | | | | |
| yjfM | | | | | | | | | |
| ytfT | | | | | | | | | |

**Tabla 5.1.5 Genes homólogos de las tres bacterias que presentan unidades REP asociadas sólo en *Escherichia coli K12***

| *Escherichia coli K12* (50 unidades & 71 elementos) | | | | *Escherichia coli O157H7 EDL933* | | | *Escherichia coli O157H7* | | |
|---|---|---|---|---|---|---|---|---|---|
| **Gen** | **Pos** | **REP** | **Orientación** | **Pos** | **REP** | **Orientación** | **Pos** | **REP** | **Orientación** |
| folA – apaH | 48 | 1 | → | | | | | | |
| lhuH - hemL | 153 | 2 | → ← | | | | | | |
| **E*** | | | | | | | | | |
| pgpA - yajO | 418 | 1 | → | | | | | | |
| actA – yhaM | 465 | 1 | → | | | | | | |
| nagD | 675 | 1 | → | | | | | | |
| pgm | 688 | 1 | ← | | | | | | |
| sucB | 727 | 4 | ← → ← → | | | | | | |
| glnQ | 809 | 1 | → | | | | | | |
| ybiT - ybiU | 820 | 1 | ← | | | | | | |
| ybjF - artJ | 859 | 1 | → | | | | | | |
| ycbL - aspC | 927 | 1 | → | | | | | | |
| ycdF - ycdG | 1005 | 1 | ← | | | | | | |
| putP | 1015 | 1 | ← | | | | | | |
| flgF | 1077 | 1 | ← | | | | | | |
| ydgA - uidC | 1614 | 1 | ← | | | | | | |
| sodB - b1657 | 1656 | 1 | ← | | | | | | |
| katE - ydjC | 1732 | 2 | → → | | | | | | |
| yegH - asmA | 2063 | 1 | → | | | | | | |
| narP - ecnH | 2193 | 1 | ← | | | | | | |
| nuoL | 2278 | 1 | → | | | | | | |
| b2339 | 2339 | 1 | → | | | | | | |
| b2430 - b2431 | 2430 | 1 | ← | | | | | | |
| b2532 | 2532 | 6 | → ← → ← → ← | | | | | | |
| yphH - glyA | 2550 | 1 | ← | | | | | | |
| srmB - yfiE | 2576 | 1 | ← | | | | | | |
| ascB - hycI | 2716 | 1 | → | | | | | | |
| yqeA - b2875 | 2874 | 1 | → | | | | | | |
| **F*** | | | | | | | | | |
| rnpB | 3123 | 1 | ← | | | | | | |
| dacB - yhbZ | 3182 | 1 | → | | | | | | |
| trpS | 3384 | 1 | ← | | | | | | |
| yhgF | 3407 | 2 | → ← | | | | | | |
| gntT - malQ | 3415 | 2 | → ← | | | | | | |
| livF | 3454 | 1 | → | | | | | | |
| yhhK - | 3459 | 1 | ← | | | | | | |

| | | | |
|---|---|---|---|
| livJ | | | |
| pilA-yhiO | 3493 | 1 | → |
| dppA | 3544 | 1 | → |
| xylR - hax | 3569 | 3 | → ← → |
| mtlA | 3599 | 1 | ← |
| yibL | 3602 | 1 | ← |
| yibK-eysE | 3606 | 1 | ← |
| uvrD-b3814 | 3814 | 2 | → ← |
| glnA | 3871 | 1 | ← |
| metJ | 3939 | 1 | → |
| yjeK | 4147 | 1 | → |
| treB | 4241 | 1 | ← |
| yjgD-b4256 | 4256 | 1 | ← |
| creA | 4398 | 1 | → |

**Notas:**

E*: b0332 389 5'—5' prpC 390 8 → ← → ← → ← → ← prpB 332 3'—5' prpC 333 prpB 394 3'—5' prpC 395

F*: ygjD 4045 5'—5' rpsU 4046 1 ← ygjD 3064 ygjD 4122

**Tabla 5.1.6 Genes homólogos de las tres bacterias que presentan unidades REP asociadas sólo en *Escherichia coli O157H7 EDL933***

| *Escherichia coli K12* | *Escherichia coli O157H7 EDL933* | *Escherichia coli O157H7* |
|---|---|---|
| | (5 unidades & 8 elementos) | |

| Gen | Pos | REP | Orientación | Pos | REP | Orientación | Pos | REP | Orientación |
|---|---|---|---|---|---|---|---|---|---|
| b0110 | | | | 104 | 2 | → ← | | | |
| ybbJ | | | | 555 | 3 | ← → ← | | | |
| glnX | | | | 707 | 1 | ← | | | |
| mtr | | | | 4146 | 1 | ← | | | |
| b2931 | | | | 5102 | 1 | → | | | |

**Tabla 5.1.7 Genes homólogos de las tres bacterias que presentan unidades REP asociadas sólo en *Escherichia coli O157H7***

| *Escherichia coli K12* | *Escherichia coli O157H7 EDL933* | *Escherichia coli O157H7* |
|---|---|---|
| | | (19 unidades & 21 elementos) |

| Gen | Pos | REP | Orientación | Pos | REP | Orientación | Pos | REP | Orientación |
|---|---|---|---|---|---|---|---|---|---|
| rhsD | | | | | | | 570 | 1 | ← |
| ybbT | | | | | | | 576 | 1 | ← |
| ybeV - ybeW | | | | | | | 700 | 1 | → |
| Int | | | | | | | 707 | 1 | ← |
| yefN | | | | | | | 1549 | 1 | → |
| prfA | | | | | | | 1786 | 1 | → |
| narG | | | | | | | 1799 | 1 | ← |
| yeaF | | | | | | | 2593 | 1 | ← |
| yeaX | | | | | | | 2614 | 1 | ← |
| G* | | | | | | | | | |
| b2390 | | | | | | | 3400 | 2 | → ← |

| | | | |
|---|---|---|---|
| tktB-b2466 | 3464 | 1 | → |
| yfgA | 3515 | 1 | → |
| sufI | 4076 | 1 | ← |
| agaB | 4195 | 1 | → |
| yhbU | 4219 | 1 | → |
| **H*** | | | |
| hemX | 4941 | 1 | → |
| xerC | 4949 | 1 | → |

**Notas:**
G*: b2016 2016 2898 b201 62490 5'–5' hisL 2491 1 ←
H*: yhjA 3518 yhjA 4517 yhjA 4588 5'–5' treF 4589 1 →

**Tabla 5.1.8 Gen exclusivo de *Escherichia coli K12* y *Escherichia coli O157H7 EDL933* que presenta una unidad REP asociada**

| *Escherichia coli K12* (1 unidad & 2 elementos) | *Escherichia coli O157H7 EDL933* (1 unidad & 2 elementos) | *Escherichia coli O157H7* |
|---|---|---|

| Gen | Pos | REP | Orientación | Pos | REP | Orientación | Pos | REP | Orientación |
|---|---|---|---|---|---|---|---|---|---|
| b2391-b2392 | | | | | | | | | |

**Tabla 5.1.9 Gen exclusivo de *Escherichia coli K12* y *Escherichia coli O157H7* que presenta una unidad REP asociada**

| *Escherichia coli K12* (1 unida & 2 elementos) | *Escherichia coli O157H7 EDL933* (1 unidad & 2 elementos) | *Escherichia coli O157H7* (1 unidades & 2 elementos) |
|---|---|---|

| Gen | Pos | REP | Orientación | Pos | REP | Orientación | Pos | REP | Orientación |
|---|---|---|---|---|---|---|---|---|---|
| agaA | | | | | | | | | |

**Tabla 5.1.10 Genes exclusivos de *Escherichia coli O157H7 EDL933* y *Escherichia coli O157H7* que sólo en *Escherichia coli O157H7 EDL933* presentan una unidad REP asociada**

| *Escherichia coli K12* | *Escherichia coli O157H7 EDL933* (16 unidades & 18 elementos) | *Escherichia coli O157H7* |
|---|---|---|

| Gen | Pos | REP | Orientación | Pos | REP | Orientación | Pos | REP | Orientación |
|---|---|---|---|---|---|---|---|---|---|
| afuA | | | | 420 | 1 | → | | | |
| Z1371 | | | | 1230 | 1 | → | | | |
| Z1812 | | | | 1645 | 1 | ← | | | |
| Z1910 | | | | 1733 | 1 | ← | | | |
| Z1918 | | | | 1740 | 1 | → | | | |
| Z1958 | | | | 1779 | 1 | ← | | | |
| Z2135 | | | | 1938 | 1 | ← | | | |
| Z2140 | | | | 1943 | 1 | ← | | | |
| Z2356 | | | | 2139 | 1 | → | | | |
| Z2360 | | | | 2143 | 1 | ← | | | |
| Z3320 | | | | 3031 | 1 | → | | | |

| | | | |
|---|---|---|---|
| Z4048-rpoS | 3688 | 1 | ← |
| Z4489 | 4116 | 2 | ← → |
| Z5002-yiaT | 4585 | 2 | → ← |
| Z5153 | 4722 | 1 | ← |
| Z5944 | 5420 | 1 | ← |

**Tabla 5.1.11 Genes exclusivos de *Escherichia coli O157H7 EDL933* y *Escherichia coli O157H7* que sólo en *Escherichia coli O157H7* presentan una unidad REP asociada**

| *Escherichia coli K12* | *Escherichia coli O157H7 EDL933* | *Escherichia coli O157H7* (11 unidades & 14 elementos) |
|---|---|---|

| Gen | Pos | REP | Orientación | Pos | REP | Orientación | Pos | REP | Orientación |
|---|---|---|---|---|---|---|---|---|---|
| ECs0415 | | | | | | | 424 | 1 | ← |
| ECs0549 | | | | | | | 559 | 3 | ← → ← |
| ECs1550 | | | | | | | 1615 | 1 | → |
| ECs1801 | | | | | | | 1877 | 1 | → |
| ECs1981 | | | | | | | 2067 | 1 | → |
| ECs2242 | | | | | | | 2337 | 1 | ← |
| ECs3594-Ecs3595 | | | | | | | 3748 | 1 | → |
| ECs4261 | | | | | | | 4451 | 1 | → |
| ECs4269-ECs4270 | | | | | | | 4459 | 1 | ← |
| ECs4459-ECs4460 | | | | | | | 4650 | 2 | ← → |
| ECs5305 | | | | | | | 5544 | 1 | → |

**Tabla 5.1.12 Genes exclusivos de *Escherichia coli K12* que presentan una unidad REP asociada**

| *Escherichia coli K12* (9 unidades & 13 elementos) | *Escherichia coli O157H7 EDL933* | *Escherichia coli O157H7* |
|---|---|---|

| Gen | Pos | REP | Orientación | Pos | REP | Orientación | Pos | REP | Orientación |
|---|---|---|---|---|---|---|---|---|---|
| b0395 - araJ | 395 | 4 | → ← → ← | | | | | | |
| yhgG | 732 | 1 | → | | | | | | |
| rhlE-ybiA | 797 | 1 | ← | | | | | | |
| cobB - b1121 | 1120 | 1 | ← | | | | | | |
| b1368 | 1368 | 1 | ← | | | | | | |
| b1372 | 1372 | 1 | ← | | | | | | |
| b2420-cysM | 2420 | 1 | → | | | | | | |
| b2973 | 2973 | 2 | → → | | | | | | |
| yjiQ-yjiR | 4340 | 1 | ← | | | | | | |

**Tabla 5.1.13 Genes exclusivos de *Escherichia coli O157H7 EDL933* que presentan una unidad REP asociada**

| *Escherichia coli K12* | *Escherichia coli O157H7 EDL933* (2 unidades & 2 elementos) | *Escherichia coli O157H7* |
|---|---|---|

| Gen | Pos | REP | Orientación | Pos | REP | Orientación | Pos | REP | Orientación |
|---|---|---|---|---|---|---|---|---|---|
| Z6036 | | | | 2330 | 1 | → | | | |
| Z3179 | | | | 2899 | 1 | → | | | |

**Tabla 5.1.14 Genes exclusivos de *Escherichia coli O157H7* que presentan una unidad REP asociada**

| *Escherichia coli K12* | *Escherichia coli O157H7 EDL933* | *Escherichia coli O157H7* (4 unidades & 4 elementos) |
|---|---|---|

| Gen | Pos | REP | Orientación | Pos | REP | Orientación | Pos | REP | Orientación |
|---|---|---|---|---|---|---|---|---|---|
| ECs1114 | | | | | | | 1152 | 1 | → |
| ECs4598-ECs4599 | | | | | | | 4794 | 1 | → |
| ECs4940 | | | | | | | 5170 | 1 | ← |
| ECs4973 | | | | | | | 5203 | 1 | → |

**Tabla 5.1.** "**Gen**" es el gen tres prima asociado a la unidad (véase metodología), excepto cuando se indica. El nombre del gen siempre proviene de *Escherichia coli K12* cuando está presente en el grupo. Cuando sólo son genes de las cepas patógenas, el nombre del gen de *Escherichia coli O157H7 EDL933* es el que se tomó como referencia. Sólo cuando el grupo está formado por genes de *Escherichia coli O157H7* es que aparecen los nombres reportados en esa bacteria. "**Pos**" es el número del gen en el genoma del organismo. "**REP**" es el número de elementos REP en esa unidad. "**Orientación**" es el arreglo de los elementos en la unidad. Cuando la unidad se encuentra entre dos regiones tres-prima ambos genes se señalan, las posiciones corresponden al primero de ellos. También se indican las unidades que se encuentran en regiones cinco-prima-cinco-prima (notas). Los colores de las asociaciones marcan cuando dos o más unidades son idénticas entre los organismos y son: muestra las unidades que comparten las tres bacterias que mantienen número y orientación constante de los elementos; muestra las unidades que conservan número y orientación sólo en *Escherichia coli K12* y *Escherichia coli O157 H7*. muestra las que han mantenido invariable su número y orientación sólo en las dos bacterias patógenas. muestra las unidades que han mantenido constante su orientación y su número en *Escherichia coli K12* y *Escherichia coli O157 H7 EDL933*. muestra las unidades que sí variaron su número y/o orientación, o que simplemente no tienen una unidad homóloga en algún otro organismo.

## Tabla 5.2 Grupos formados para las cinco enterobacterias

**Tabla 5.2.1 Genes homólogos en donde las cinco enterobacterias presentan unidades REP asociadas**

| S.typhi (41 uni & 56 ele) | | | | S.typhimurium LT2 (41 uni & 49 ele) | | | | Escherichia coli | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Gen | Pos | REP | Orientación | Gen* | Pos | REP | Orientación | Gen** | E.coli K12 (41 uni & 99 ele) | E.coli O157H7 (41 uni & 74 ele) | E.coli O157H7 EDL933 (41 uni & 81 ele) |
| thrC-yaaA | 4 | 1 | ← | Z0005-yaaA | 8 | 1 | ← | b0005-yaaA | → ← | → ← | → ← |
| caiA | 74 | 1 | → | | 113 | 1 | → | | → ← | → ← | → ← |
| araA | 106 | 2 | → → | | 155 | 1 | → | | ← ← ← | | |
| yabI-yabJ | 109 | 1 | → | yabJ | 160 | 1 | → | yabI-yabJ | → ← | | |
| gcd-cueO | 174 | 1 | ← | yacK-gcd | 174 | 1 | → ← | yacK – gcd | ← | ← | ← |
| gmhA | 326 | 1 | → | | 326 | 1 | → | | ← → ← | | |
| yajQ (STY0474) -yajR (STY0475) | 437 | 1 | ← | | 677 | 1 | ← | | → | → | → |
| ggrO (STY3237) -speI3 | 3008 | 1 | → | | 4371 | 1 | → | | ← → ← | ← → ← | ← → ← |
| yhgI (STY4285) | 3990 | 2 | ← ← | | 4962 | 2 | → ← | | → | → | → |
| nagE | [illegible] | 1 | [illegible] | | [illegible] | 1 | [illegible] | | [illegible] | [illegible] | [illegible] |
| sdhB | 718 | 1 | → → | | 1100 | 1 | → | | ← → | ← → | ← → |
| aroG-ppmA | 738 | 1 | → ← | | 1143 | 1 | → ← | | ← | ← | ← |
| artM | 848 | 1 | → | | 1335 | 1 | → | | ← → | ← → | ← → |
| ybhE-yblC | 1176 | 1 | → | | 754 | 1 | → | | → ← → ← | | |
| cdd | [illegible] | 1 | [illegible] | | [illegible] | 1 | [illegible] | | [illegible] | | |
| dedD | 2407 | 1 | → | | 3445 | 1 | → | | → ← | | |
| I* | | | | | | | | | | | |
| cutG | 2503 | 2 | ← → | | 3571 | 2 | → → | | → ← → ← | | |
| gabT | 2711 | 2 | → → | | 4000 | 1 | → | | → ← | → ← | → ← |
| proX | 2732 | 1 | ← | | 4031 | 1 | ← | | ← → | ← → | ← → |
| ygbD (STY2963) – hypF | 2758 | 1 | → → ← | | [illegible] | | | | [illegible] | [illegible] | [illegible] |
| exo-fucO | 2896 | | | | | | | | | | |
| STY3204-b2899 | 2977 | 1 | ← | b2899 | 4336 | 1 | ← | ygjZ – b2899 | → ← → | ← | → |
| parC | 3105 | 2 | ← → | | 4503 | 2 | ← → | | → | → | → |
| cca-lacA | 3134 | 1 | → ← | | 4546 | 1 | → ← | | → ← → | → ← → | → ← → |
| rpoD-mug | 3141 | 1 | → | rpoD-ygiF | 4557 | 1 | → | | | | → ← → |
| eco-yojI | 3314 | 1 | → | | 2304 | 1 | → | eco - yojII | ← → ← → ← → ← → ← → ← → ← → | ← → | ← → |
| yijP | 3500 | 1 | ← | | 5801 | 1 | → | | → ← | → ← | → ← |
| pfkA | 3551 | 1 | → | | [illegible] | | | | [illegible] | [illegible] | [illegible] |
| yhjD | 3904 | 1 | ← | | 5092 | 1 | → | | → ← | → ← | → ← |
| II* | | | | | | | | | | | |
| pepA (STY4816) | 4507 | 2 | → → | | 6321 | | → | | → ← | → ← → ← → ← | → ← → ← → ← |
| pgi | 4122 | 1 | ← | | 5937 | 1 | ← | | ← → | ← → | ← → |
| malE | 4130 | 2 | → ← | | 5950 | 1 | ← | | ← → ← | ← → ← | ← → ← |

## Tabla 5.2 Grupos formados para las cinco enterobacterias

**Tabla 5.2.1 Genes homólogos en donde las cinco enterobacterias presentan unidades REP asociadas**

| S.typhi (41 uni & 56 ele) | | | | S.typhimurium LT2 (41 uni & 49 ele) | | | | Escherichia coli | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Gen | Pos | REP | Orientación | Gen* | Pos | REP | Orientación | Gen** | E.coli K12 (41 uni & 99 ele) | E.coli O157H7 (41 uni & 74 ele) | E.coli O157H7 EDL933 (41 uni & 81 ele) |
| thrC-yaaA | 4 | 1 | ← | Z0005-yaaA | 8 | 1 | ← | b0005-yaaA | → ← | → ← | → ← |
| caiA | 74 | 1 | → | | 113 | 1 | → | | → ← | → ← | → ← |
| araA | 106 | 2 | → → | | 155 | 1 | → | | ← ← ← | | |
| yabI-yabJ | 109 | 1 | → | yabJ | 160 | 1 | → | yabI-yabJ | → ← | | |
| gcd-cueO | 174 | 1 | ← | yacK-gcd | 174 | 1 | → ← | yacK – gcd | ← | ← | ← |
| gmhA | 326 | 1 | → | | 326 | 1 | → | | ← → ← | | |
| yajQ (STY0474) -yajR (STY0475) | 437 | 1 | ← | | 677 | 1 | ← | | → | → | → |
| ygeG (STY3237) -speB | 3008 | 1 | → | | 4371 | 1 | → | | ← → ← | ← → ← | ← → ← |
| yhgI (STY4285) | 3990 | 2 | ← ← | | 4962 | 2 | → ← | | → | → | → |
| nagE | [illegible] | 1 | [illegible] | | [illegible] | 1 | [illegible] | | [illegible] | [illegible] | [illegible] |
| sdhB | 718 | 1 | → → | | 1100 | 1 | → | | ← → | ← → | ← → |
| aroG-gpmA | 738 | 1 | → ← | | 1143 | 1 | → ← | | ← | ← | ← |
| artM | 848 | 1 | → | | 1335 | 1 | → | | ← → | ← → | ← → |
| ybhE-yblC | 1176 | 1 | → | | 754 | 1 | → | | → ← → ← | | |
| cdd | [illegible] | 1 | [illegible] | | [illegible] | 1 | [illegible] | | [illegible] | [illegible] | [illegible] |
| dedD | 2407 | 1 | → | | 3445 | 1 | → | | → ← | | |
| I* | | | | | | | | | | | |
| eutG | 2503 | 2 | ← → | | 3571 | 2 | → → | | → ← → ← | | |
| gabT | 2711 | 2 | → → | | 4000 | 1 | → | | → ← | → ← | → ← |
| proX | 2732 | 1 | ← | | 4031 | 1 | ← | | ← → | ← → | ← → |
| ygbD (STY2963) – hypF | 2758 | 1 | → → ← | | | | | | | | |
| exo-fucO | 2896 | | | | | | | | | | |
| STY3204-b2899 | 2977 | 1 | ← | b2899 | 4336 | 1 | ← | ygiZ - b2899 | → ← → | ← | → |
| parC | 3105 | 2 | ← → | | 4503 | 2 | ← → | | → | → | → |
| cca-bacA | 3134 | 1 | → ← | | 4546 | 1 | → ← | | → ← → | → ← → | → ← → |
| rpoD-mug | 3141 | 1 | → | rpoD-ygiF | 4557 | 1 | → | | | | → ← → |
| cco-yojI | 3314 | 1 | → | | 2304 | 1 | → | cco - yojH | ← → ← → ← → ← → ← → ← → | ← → | ← → |
| yijP | 3500 | 1 | ← | | 5801 | 1 | → | | → ← | → ← | → ← |
| pfkA | 3551 | 1 | → | | [illegible] | | | | [illegible] | [illegible] | [illegible] |
| yhjD | 3904 | 1 | ← | | 5092 | 1 | → | | → ← | → ← | → ← |
| II* | | | | | | | | | | | |
| pepA (STY4816) | 4507 | 2 | → → | | 6321 | | → | | → ← | → ← → ← → ← | → ← → ← → ← |
| pgi | 4122 | 1 | ← | | 5937 | 1 | ← | | ← → | ← → | ← → |
| malE | 4130 | 2 | → ← | | 5950 | 1 | ← | | ← → ← | ← → ← | ← → ← |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| air | 4148 | 1 | ← | | 5980 | 1 | ← | | → | → | → |
| aes | 4176 | 1 | ← | | 6025 | 1 | ← | | ← → | ← → | ← → |
| glpP-yjcO | 4185 | 1 | ← | | 6038 | 1 | ← | | → ← → ← → ← → ← → ← → | → ← → | → ← → |
| III* | | | | | | | | | | | |
| nrdD | 4481 | 2 | ← → | | 6291 | 2 | ← → | | → ← | → ← | → ← |
| nadR-yiiK | 4610 | 1 | → | | 6466 | 1 | → | | ← → ← | ← → ← | ← → ← |
| trpR-yjjX | 4613 | 1 | ← | | 6470 | 1 | ← | | → ← | → ← | → ← |

Nota:
I*:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| fabB 5'-5' STY2610 | 2421 | 1 | ← | fabB 5'-5' STM2379 | 3463 | 1 | ← | | → → | | |

II*:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| mreA 5'-5' yrfD | 4007 | 2 | → ← | | | | | | | | |

III*:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| cpdB 5'-5' cysO | 4450 | 2 | → → | ytfII-cpdB | 6213 | 1 | → | ytfH-cpdB | ← | ← | ← |

**Tabla 5.2.2Genes homólogos de las cinco enterobacterias en donde *S.typhimurium LT2* y las tres *Escherichia coli* presentan REP**

| | S.typhimurium LT2 (9 uni & 11 ele) | | | Escherichia coli | | | | |
|---|---|---|---|---|---|---|---|---|
| Gen | Pos | REP | Orientación | Gen* | E.coli K12 (9 uni & 25 ele) | Gen** | E.coli O157H7 (9 uni & 14 ele) | E.coli O157H7 EDL933 (9 uni & 15 ele) |
| yajF-araJ | 617 | 1 | ← | araJ | → ← → ← | | | |
| pepB | | | | | | | | |
| yhaL-yhaN | 4600 | 1 | → | yhaL - yhaM | → ← | | → ← | → ← |
| yhdP | | | | | → ← ← → ← → | | ← | |
| rrfA - mobB | 5624 | 2 | ← → | | → | | → | → → |
| thaA | 5694 | 1 | → | | → ← → ← → ← → ← | | | |
| metL-STM4102 | 5775 | 1 | ← → | metL | ← | | ← | ← |
| gldA | 5786 | 1 | ← | | → | | → | → |
| pflD | | | | | | | | |

**Tabla 5.2.3 Genes homólogos de las cinco enterobacterias en donde *S.typhi* y las tres *Escherichia coli* presentan REP**

| | S.typhi (15 uni & 23 ele) | | | Escherichia coli | | | |
|---|---|---|---|---|---|---|---|
| Gen | Pos | REP | Orientación | Gen* | E.coli K12 (15 uni & 34 ele) | E.coli O157H7 (15 uni & 24 ele) | E.coli O157H7 EDL933 (15 uni & 25 ele) |
| ftsK | 881 | 2 | → → | | ← | ← | ← |
| mukB | 915 | 2 | ← ← | | → ← | → ← | → ← |
| aslB | 1685 | 1 | → | aslB -aslA | | ← | |
| ntdA | 2320 | 2 | ← ← | | → ← → ← | | |
| barA-yhaD | | | | | | | |
| argE | 3498 | 2 | → → | | ← → ← → ← → | | |
| aldB- | 3834 | 1 | ← | aldB | → ← | → ← | → ← |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| STY4117 | | | | | | | |
| yjbH | 4126 | 1 | ← | | → | → | → |
| aceA | 4107 | 2 | → → | | ← → ← | ← → ← | ← → ← |
| yjbB-pepE | 4111 | 2 | → → | | ← | ← | ← |
| lamB | 4132 | 2 | ← ← | | ← → ← | | |
| mpl - yjgA | 4465 | 1 | ← | yjfG- yjgA | ← → | ← → | ← → |
| arcC (STY4804) | 4495 | 1 | ← | arcC-parK | → → ← | → ← | → ← |
| STY4905-yjiT | 4589 | 1 | ← | yjiT | → ← | ← | ← |
| cstA | 4576 | 2 | → → | cstA-ybdH | ← | ← | ← |

**Tabla 5.2.4 Genes homólogos de las cinco enterobacterias en donde ambas *Salmonella* y alguna de las tres *Escherichia coli* presentan REP**

| | S.typhi (13 uni & 13 ele) | | | S.typhimurium LT2 (14 uni & 21 ele) | | | | Escherichia coli | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Gen | Pos | REP | Orientación | Gen* | Pos | REP | Orientación | Gen** | E.coli K12 (10 uni & 16 ele) | E.coli O157H7 (3 uni & 4 ele) | E.coli O157H7 EDL933 (2 uni & 3 ele) |
| lspA | 49 | 1 | → | | 74 | 1 | → | | → ← | ← | |
| folA-apaH | | | | | | | | | | | |
| fhuB-hemL | 202 | 1 | → | fhuB-STY0201-hemL | 29* | 3 | → → → | fhuB-hemL | → ← | | |
| | | | | | 80* | 1 | → | | | | |
| pqqA (STY0459)-yajO (STY0460) | | | | | | | | | | | |
| sucB | 720 | 1 | ← | | 1104 | 1 | ← | | ← → ← → | | |
| ybjF | 842 | 1 | ← | | 1326 | 1 | → | ybjF - artI | → | | |
| ung-artI | 6725 | 1 | → | | 6369 | 1 | → | | | | |
| dacB (STY3479) - STY3480 | | | | | | | | | | | |
| IV* | | | | | | | | | | | |
| yibK-STM3696 | 3818 | 1 | → | STY4190-yibK | 5217 | 1 | → | yibK-cysE | ← | | |
| uvrD | 3351 | 1 | → | | 5560 | 2 | → ← | uvrD-b3814 | → ← | | |
| rrlE-yjaB | | | | 5S_rRNA-yjaB | 5887 | 1 | → ← ← ← | | | | |
| pitA - uspB | 5067 | 1 | → | | 4013 | 1 | ← | pitA-yhiQ | → | | |

Nota:

IV*:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| yieH-STY3545-STY3546-yieJ | 3666 | 1 | → | yieH-phoU | 5423 | 1 | → | phoU | | | |

**Tabla 5.2.5 Genes homólogos de las cinco enterobacterias en donde *S.typhimurium LT2* y alguna de las tres *Escherichia coli* presentan REP**

| | S.typhimurium LT2 (1 uni & 1 ele) | | | Escherichia coli | | |
|---|---|---|---|---|---|---|
| Gen | Pos | REP | Orientación | Gen* | E.coli K12 (1 uni & 1 ele) | E.coli O157H7 EDL933 (1 uni & 1 ele) |
| STM4540 -mdoB | | | | mdoB | | |

**Tabla 5.2.6 Genes homólogos de las cinco enterobacterias en donde *S.typhi* y alguna de las tres *Escherichia coli* presentan REP**

| | S.typhi (18 uni & 19 ele) | | | Escherichia coli | | | |
|---|---|---|---|---|---|---|---|
| Gen | Pos | REP | Orientación | Gen* | E.coli K12 (11 uni & 14 ele) | E.coli O157H7 (4 uni & 5 ele) | E.coli O157H7 EDL933 (6 uni & 7 ele) |
| prpB | 369 | 1 | ← | | | | |
| lnt | 658 | 1 | → | | | | ← |
| yleB-asnB | 663 | 1 | → | | | | |
| rhlE | 788 | 1 | ← | rhlE-yhiA | ← | | |
| rluC-yceF (STY1228) | | | | | | | |
| cobB-potD | 1153 | 1 | → | cobB -b1121 | ← | | |
| narG | 1179 | 1 | → | | | | ← |
| yegH-asmA | 2152 | 1 | → | | → | | |
| b2431 (STY2683) | 2488 | 1 | ← | b2430 -b2431 | ← | | |
| glyA | 3508 | 1 | → | yphH -glyA | ← | | |
| purL | | | | | | | |
| STY3459-mtr | | | | | | | |
| nusA | | | | | | | |
| ccmH | 3701 | 1 | → | narP –ccmH | ← | | |
| STY4008-ascB | 3737 | 1 | ← | ascB- hycI | → | | |
| bax-xylR | 3852 | 1 | → | | → ← → | | |
| hlyF | 3957 | 2 | ← → | | → | | |
| gntT-malQ | 3989 | 1 | → | | → ← | | |

**Tabla 5.2.7 Genes homólogos de las *Salmonelas* con *Escherichia coli* K12 que sólo presentan REP en *S.typhimuriumLT2***
(68 uni & 71 ele)

| Gen | Pos | REP | Orientación |
|---|---|---|---|
| prpC | 574 | 1 | ← |
| ispA | 663 | 1 | ← |
| yajG | 690 | 1 | ← |
| ybaW-ybaX | 708 | 1 | ← |
| hemH | 756 | 1 | → |

**Tabla 5.2.8 Genes homólogos de las *Salmonellas* con *Escherichia coli* K12 que sólo presentan REP en *S.typhi***
(130 uni & 138 ele)

| Gen | Pos | REP | Orientación |
|---|---|---|---|
| fixX | 79 | 2 | ← → |
| murD | 130 | 1 | → |
| ftsW | 131 | 1 | → |
| secA | 140 | 1 | → |
| b2737 (ygbK) Φ | 167 | 1 | → |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| afbA | 785 | 1 | → | V* | | | |
| ybeJ | 832 | 1 | ← | hrpB | 197 | 1 | ← |
| kdpA | 1052 | 1 | ← | crl-phoE | 336 | 1 | → |
| nadA | 1137 | 1 | ← | yaiU (STY0405) | 374 | 1 | → |
| dinG | 1233 | 1 | ← | sbcC | 396 | 1 | → |
| ybjO | 1324 | 1 | → | b0441 (ppiD) | 457 | 1 | → |
| fnpA_1-STM0947 | 1424 | 1 | ← | adk | 494 | 1 | → |
| folA | 1443 | 1 | ← | tesA | 512 | 1 | → |
| dmsB | 1451 | 1 | → | ybbS (STY0562) | 520 | 1 | → |
| ycbB | 1502 | 1 | ← | VI* | | | |
| hpaE | 1665 | 1 | → | cysS-ybeI | 542 | 1 | → |
| hpaX | 1675 | 1 | → | fes | 583 | 1 | → |
| serX | 1717 | 1 | ← | entF | 585 | 1 | → |
| astD | 1976 | 1 | → | entC | 592 | 1 | → |
| pheT | 2031 | 1 | → | VII* | | | |
| ydiT | 2052 | 1 | ← | b1590 (STY0661)-ybdQ | 611 | 1 | → |
| ydiH-STM1368 | 2071 | 1 | ← | b0612 (citT) | 616 | 1 | → |
| orf242 | 2105 | 1 | ← | yhaG (STY0701) | 648 | 1 | → |
| ydgH | 2231 | 1 | ← | ybeZ | 661 | 1 | → |
| oppD | 2606 | 1 | ← | ybgF | 733 | 1 | ← |
| adhE-tdk | 2613 | 1 | → | b0829 (STY0887) | 815 | 1 | → |
| tyrT | 2625 | 1 | → | yliI-yliJ | 821 | 1 | ← |
| proQ | 2753 | 1 | ← | VIII* | | | |
| yebR | 2754 | 1 | ← | potI | 840 | 1 | → |
| ftnB-STM1933 | 2869 | 1 | → | ycaI | 904 | 1 | → |
| rtn | 3255 | 1 | → | IX* | | | |
| ccmG | 3300 | 1 | ← | ycbY (STY1082) | 995 | 1 | → |
| nrdB | 3335 | 1 | ← | b2245 (hpcH) | 1040 | 1 | → |
| yfcN-sixA | 3473 | 2 | → ← | b2582 (scsD) | 1049 | 1 | → |
| STM2453-eutR | 3563 | 2 | → ← | yjhC (STY1170) | 1066 | 1 | ← |
| eutM | 3575 | 1 | → | phoQ | 1162 | 1 | → |
| cadC | 3691 | 1 | ← | chaB-yehN (STY1283) | 1173 | 1 | → |
| purG | 3703 | 1 | → | oppF | 1198 | 1 | → |
| yfiM-kgtP | 3813 | 1 | ← | X* | | | |
| rrfG | 3817 | 1 | ← | yehW (STY1570) | 1437 | 1 | → |
| yghJ | 4171 | 1 | ← | ybgA (STY1731)-ydhZ | 1594 | 1 | → |
| ygcB | 4202 | 1 | ← | STY1755-b1685 (STY1756) | 1615 | 1 | → |
| dsbC | 4328 | 1 | ← | aroD-ppsA | 1620 | 1 | → |
| ubiI | 4344 | 2 | ← → | pheS | 1633 | 1 | ← |
| eafA | 4780 | 1 | → | b1745 (astB) | 1666 | 1 | → |
| XI* | | | | selD | 1675 | 1 | → |
| rrfB | 5825 | 1 | → | b1834 (STY1980) | 1820 | 1 | → |
| nrfE | 6034 | 1 | ← | purT-eda | 1922 | 1 | → |
| feoB | 4954 | 1 | ← | flhB | 1953 | 1 | → |
| yjbA | 5946 | 1 | ← | motA | 1962 | 1 | → |
| malM | 5955 | 1 | ← | yecI | 1971 | 1 | → |
| glgB | 4996 | 1 | ← | b1955 (STY2193)-STY2194 | 2022 | 1 | → |
| yjeE-STM4270 | 6017 | 1 | ← | sbcB-yecF | 2098 | 1 | ← |

| | | | |
|---|---|---|---|
| lpxO | 6042 | 1 | ← |
| yhhW | 5003 | 1 | → |
| XII* | | | |
| bcsC | 5104 | 1 | ← |
| rgbE-STM3678 | 5188 | 1 | → |
| ecmG | 5373 | 1 | ← |
| dgoK | 5390 | 1 | ← |
| yihX | 5668 | 1 | → |
| rhaR | 5698 | 1 | → |
| yjeB | 6153 | 1 | ← |
| treC | 6292 | 1 | ← |
| idnO | 6328 | 1 | ← |
| mrr | 6386 | 1 | → |
| STM4466 | 6307 | 1 | ← |
| yjiY | 6394 | 1 | → |
| rsmC | 6428 | 1 | ← |
| deoA | 6449 | 1 | ← |

**Tabla 5.2.9** Genes homólogos de las Salmonellas con Escherichia coli K12 que presentan REP

| | S.typhi (110 uni & 120 ele) | | | S.typhimurium LT2 (110 uni & 124 ele) | | | |
|---|---|---|---|---|---|---|---|
| Gen | Pos | REP | Ori | Gen* | Pos | REP | Ori |
| carB | 68 | 1 | → | | 105 | 1 | → |
| cstD | 71 | 1 | → | | 110 | 1 | → |
| b0055 (dsbA)-rluA | 97 | 1 | → | | 143 | 1 | → |
| ilvH | 121 | 1 | → | | 174 | 1 | → |
| hpt-yadF | 175 | 1 | → | | 261 | 1 | → |
| panB | 183 | 1 | ← | | 277 | 2 | → ← |
| yadB | 193 | 1 | ← | | 280 | 1 | ← |
| mreB | 198 | 2 | ← ← | | 286 | 1 | ← |
| yadR | 205 | 1 | → | | 309 | 1 | → |
| yeaG-yeaH cdaR-yaeH | 212 | 1 | ← | | 318 | 1 | ← |
| map | 216 | 2 | → ← | | 325 | 2 | → ← |
| dnaE | 232 | 1 | ← | | 353 | 1 | ← |
| b0189 (rof) | 239 | 1 | ← | | 363 | 1 | ← |
| yajK- phnV | 427 | 1 | → | tnrE-phnV | 665 | 1 | → |
| b3872 (phnR) | 432 | 1 | ← | | 670 | 1 | ← |
| phnX-thrL | 435 | 1 | ← | thrL | 675 | 1 | ← |
| | | | | phnX | 673 | 1 | ← |
| amtB-tesB | 469 | 1 | ← | | 725 | 1 | ← |
| htpG | 493 | 1 | → | | 753 | 4 | → → → → |
| ushA-ybaK | 500 | 1 | ← | | 764 | 1 | ← |
| ybbM - ybbN | 509 | 1 | → | | 774 | 1 | → |
| purE | 539 | 1 | → | | 824 | 1 | → |
| ybeL-ybeQ | 651 | 1 | ← | | 987 | 1 | ← |

| | | | |
|---|---|---|---|
| b2053 (rfbG) | 2124 | 1 | → |
| wcaB | 2147 | 1 | → |
| metG | 2204 | 1 | → |
| yeiB | 2244 | 1 | → |
| fruB | 2258 | 1 | → |
| spr (STY2450) | 2266 | 1 | → |
| b2195 (dsbE1) | 2290 | 1 | ← |
| napB | 2298 | 1 | → |
| b2249 (STY2523) | 2337 | 1 | → |
| b2291 (STY2562)-STY2563 | 2375 | 2 | → → |
| yfbT (STY2564) | 2377 | 1 | → |
| b2304 (STY2580)- hisP | 2393 | 2 | → → |
| STY2618-b2340 (STY2619) | 2429 | 1 | → |
| STY2690-yfeG | 2495 | 2 | → ← |
| b2458 (eutD) | 2507 | 1 | → |
| XIII* | | | |
| b2520 (STY2778) | 2576 | 1 | → |
| sseA | 2577 | 1 | → |
| b2523 (yfhI) | 2580 | 1 | ← |
| cadB | 2602 | 1 | → |
| XIV* | | | |
| ung- yfiF | 2637 | 1 | → |
| yfiP (STY2843) | 2640 | 1 | → |
| b3025 (tetD) | 2703 | 1 | → |
| oraA | 2744 | 1 | → |
| recA | 2745 | 1 | → |
| XV* | | | |
| hycD | 2767 | 1 | → |
| hypA | 2771 | 1 | ← |
| hypD | 2774 | 1 | → |
| iap-ygbF | 2852 | 1 | → |
| ygdH | 2893 | 1 | → |
| galR | 2935 | 1 | → |
| XVI* | | | |
| b2907 (visB) | 2985 | 2 | ← → |
| speA | 3010 | 1 | → |
| metK | 3013 | 1 | → |
| yqhA (STY3326) | 3084 | 1 | → |
| dnaG | 3140 | 1 | → |
| b1814 (ldcC) | 3169 | 1 | → |
| argG-secG | 3215 | 1 | → |
| nanT | 3261 | 1 | → |
| yheM (STY3526) | 3268 | 1 | → |
| yhdRP (STY3549) | 3290 | 1 | → |
| b3247 (rmG) | 3291 | 1 | → |
| yigB | 3352 | 1 | ← |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ybeY | 660 | 2 | → ← | | 1007 | 1 | ← |
| ybgK | 696 | 1 | ← | | 1063 | 1 | ← |
| hutI-ybhB | 760 | 1 | ← | | 1187 | 1 | ← |
| ybhS | 784 | 1 | → | | 1227 | 1 | → |
| ybjZ-caplD | 862 | 1 | ← | | 1418 | 1 | ← |
| XVII* | | | | | | | |
| hpaF (hpcD) | 3048 | 1 | ← | | 1669 | 1 | ← |
| mviM | 1101 | 1 | ← | | 1762 | 1 | ← |
| mviN-flgN | 1102 | 1 | → | | 1767 | 1 | → |
| yeiC | 1165 | 1 | ← | | 1869 | 1 | ← |
| pncA-STY1823 | 1221 | 2 | → → | yeaC-pncA | 1959 | 2 | ← ← |
| fumA | 1517 | 1 | → | | 2218 | 1 | ← |
| STY133-btuR | 1679 | 1 | ← | btuR-yciL | 2564 | 1 | → |
| yeaD-STY1828 | 1684 | 1 | → | STM1288-yeaD | 1951 | 1 | ← |
| mglB | 2242 | 1 | → | | 1222 | 1 | → |
| lysP | 2252 | 1 | → | | 1235 | 1 | → |
| STY2510-glpQ | 2324 | 1 | ← | STM2281-glpQ | 3441 | 1 | ← |
| pta | 2381 | 2 | → ← | | 3413 | 2 | → ← |
| hisJ | 2397 | 1 | → | | 3434 | 1 | → |
| glk | 2452 | 1 | ← | | 3494 | 1 | ← |
| lysV-xapR | 2462 | 1 | ← | | 3516 | 1 | ← |
| b2456 (eutN) | 2506 | 1 | → | | 3574 | 1 | → |
| purN | 2543 | 1 | → | | 3622 | 1 | → |
| suhB | 2590 | 1 | → | | 3673 | 2 | → → |
| yfhD-yfiC | 2610 | 1 | → | | 3707 | 1 | → |
| STY2822-acpS | 2619 | 1 | ← | yfhC-acpS | 3718 | 1 | ← |
| nadB-yfiC | 2631 | 1 | → | | 3795 | 1 | → |
| recN | 2671 | 1 | ← | | 3853 | 1 | ← |
| srlD | 2751 | 1 | ← | | 4063 | 1 | ← |
| hydN | 2760 | 1 | ← | | 4077 | 1 | ← |
| hypE | 2775 | 1 | ← | | 4096 | 1 | ← |
| sdaB | 2895 | 1 | ← | | 4234 | 1 | ← |
| argA-recD | 2915 | 1 | ← | | 4266 | 1 | ← |
| recB | 2916 | 1 | ← | | 4267 | 1 | ← |
| STY3151-ygeD | 2931 | 1 | ← | tas-ygeD | 4286 | 1 | ← |
| gcvP | 2981 | 1 | ← | | 4340 | 1 | ← |
| ictA-STY3221 | 2993 | 1 | → | ictA-ygeH | 4353 | 1 | → |
| gshB | 3018 | 1 | ← | | 4395 | 1 | ← |
| yggX | 3045 | 1 | ← | | 4419 | 1 | ← |
| nupG-speC | 3047 | 1 | ← | | 4424 | 1 | → |
| STY3330-exbD | 3087 | 1 | → | yghA-exbD | 4478 | 1 | → |
| STY3341-ygiR | 3097 | 1 | → | STM3316-ygiR | 4492 | 1 | → |
| STY3366-ygiD | 3119 | 1 | → | XIX* | | | |
| ygiF | 3132 | 2 | → → | | 4544 | 1 | → |

| | | | |
|---|---|---|---|
| yifL | 3356 | 1 | → |
| purD-hydG | 3454 | 1 | → |
| thiC | 3465 | 1 | → |
| metB | 3512 | 1 | → |
| STY3799 -cdh | 3548 | 1 | → |
| rhaS | 3565 | 1 | ← |
| yihW | 3590 | 1 | ← |
| yieG-STY3933 | 3667 | 1 | ← |
| thdF | 3672 | 1 | → |
| rpmH | 3675 | 1 | → |
| dsdX | 3711 | 1 | → |
| b4242 (mgtB) | 3750 | 1 | → |
| rfaD | 3803 | 1 | → |
| yibP (STY4090) | 3808 | 1 | → |
| STY4118-b3583 (yiaS) | 3835 | 1 | ← |
| b3581 (yiaQ) | 3838 | 1 | → |
| yiaF-yiaE | 3868 | 1 | → |
| yhjQ | 3892 | 1 | → |
| livG | 3956 | 2 | → → |
| glgA | 3980 | 1 | → |
| glgP-glgD | 3981 | 2 | → ← |
| nirD | 4025 | 1 | ← |
| yhfC | 4027 | 2 | → ← |
| aceB | 4106 | 1 | → |
| aceK-iclR | 4108 | 1 | ← |
| XVIII* | | | |
| STY4367 -oxyR (STY4468) | 4170 | 1 | ← |
| nrfF (STY4480) | 4183 | 1 | → |
| melA | 4200 | 1 | → |
| dmsC | 4209 | 1 | → |
| yjeM | 4394 | 1 | ← |
| yjeN -yjeP | 4395 | 1 | ← |
| pmbA | 4467 | 1 | → |
| ycdT (STY4904) | 4588 | 1 | → |
| deoC | 4601 | 1 | ← |
| deoB | 4603 | 1 | ← |

Notes:

| | | | |
|---|---|---|---|
| V*: gcd 5'-5' hpt | 174 | 1 | ← |
| VI*: ppiB 5'-5' cysS | 541 | 1 | → |
| VII*: dsbG 5'-5'ahpC | 604 | 1 | → |
| VIII*: STY0903 5'-5' STY0904 | 830 | 1 | ← |
| IX*: STY1081 5'-5' STY1082 | 994 | 1 | → |
| X*: STY1458 5'-5' STY1459 | 1337 | 1 | → |
| XI*: udhA 5'-5'-yijC | 5812 | 1 | ← |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| yhbV | 3203 | 1 | → | | 4658 | 1 | → |
| yrbB | 3242 | 1 | → | | 4698 | 1 | → |
| STY3504 – mtgA | 3248 | 1 | ← | yrbL-mtgA | 4725 | 1 | ← |
| nanA | 3262 | 1 | → | | 4744 | 1 | → |
| sapA | 3265 | 1 | → | | 4747 | 1 | → |
| XX* | | | | | | | |
| STY3547-tldD | 3288 | 1 | ← | yhbS – tldD | 4778 | 1 | ← |
| menG-yiiU | 3389 | 1 | ← | | 5761 | 1 | → |
| yhfA | 3496 | 1 | ← | yhfA-prkB | 4898 | 1 | → |
| hofQ | 3523 | 1 | → | | 4929 | 1 | ← |
| yhgE-STY4298 yhgE-yrfI | 3528 | 1 | ← | | 4943 | 1 | → |
| spf | 3615 | 1 | → | spf-yihA | 5638 | 1 | ← |
| yhgG | 3673 | 1 | → | | 4955 | 1 | ← |
| yicI -yicH | 3764 | 1 | → | | 5291 | 1 | ← |
| glpG | 3884 | 1 | → | | 4979 | 1 | ← |
| yhiQ-yhiP | 3918 | 1 | ← | | 5072 | 2 | → ← |
| asd | 3921 | 2 | → ← | | 4997 | 2 | → ← |
| nikR -yhhJ | 3932 | 1 | → | yhhJ-nikR | 5060 | 1 | ← |
| yhhL-yhhM | 3943 | 1 | ← | yhhM-yhhL | 5042 | 1 | → |
| ugpB | 3949 | 1 | → | | 5022 | 1 | ← |
| rpoH | 3959 | 1 | → | | 5034 | 1 | ← |
| ppiC-ilvY | 3975 | 1 | ← | ppiC-ilvC | 5499 | 1 | → |
| fpr-yiiR | 3984 | 1 | ← | yiiR –fpr | 5755 | 1 | → |
| yidC | 3994 | 1 | ← | | 5408 | 1 | → |
| dppF- yhjV | 4002 | 1 | ← | | 5117 | 1 | → |
| argB | 4035 | 1 | → | | 5806 | 1 | ← |
| metH | 4110 | 3 | ← → → | | 5899 | 2 | → → |
| lysC | 4120 | 1 | ← | STM4219-lysC | 5936 | 1 | ← |
| lexA | 4138 | 1 | ← | | 5964 | 1 | ← |
| nrfA | 4178 | 1 | → | | 6027 | 2 | → → |
| fdhF | 4187 | 2 | → ← | | 6041 | 2 | → ← |
| melB- fumB | 4201 | 1 | ← | | 6057 | 1 | ← |
| STY4687 - yjeH | 4376 | 1 | → | fxsA-yjeH | 6094 | 1 | ← |
| fklB | 4443 | 1 | → | | 6206 | 1 | → |
| cycA-yfiL | 4446 | 1 | → | | 6209 | 1 | ← |
| fhp | 4464 | 1 | ← | | 6231 | 1 | ← |
| yjgN -valS | 4504 | 1 | → | | 6317 | 1 | → |
| yjgQ-STY4819 | 4509 | 1 | ← | yjgQ-idnR | 6324 | 1 | ← |
| hsdR | 4570 | 1 | ← | | 6385 | 1 | ← |
| deoD | 4604 | 1 | ← | | 6433 | 1 | ← |
| b4389 (radA) | 4609 | 1 | → | | 6464 | 1 | → |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| XII*: STM3595 yhiR | 5′-5′- | 5074 | 1 | → | | | |
| XIII*: maeB taiA | 5′-5′ | 2513 | 1 | → | | | |
| XIV*: STY2817 STY2818 | 5′-5′ | 2614 | 1 | → | | | |
| XV*: mltB srlA | 5′-5′ | 2747 | 1 | → | | | |
| XVI*: xerD fldB | 5′-5′ | 2972 | 1 | → | | | |
| XVII*: ycbC ahpC | 5′-5′ | 911 | 1 | → | 143 | 1 | → |
| XVIII*: qor dnaB | 5′-5′ | 4146 | 1 | → | | | |
| XIX*: ygiD ygiE | 5′-5′ | 4524 | 1 | ← | | | |
| XX*: mdh | 5′-5′argR | 3280 | 1 | → | 4766 | 1 | → |

**Tabla 5.2.10 Gene homólogo entre las *Salmonellas* que tienen REP en ambas**

| *S.typhi* (1 uni & 1 cle) | | | | *S.typhimurium LT2* (1 uni & 1 cle) | | |
|---|---|---|---|---|---|---|
| Gen | Pos | REP | Orientación | Pos | REP | Orientación |
| STM3940 (STY3619) | 3362 | 1 | ← | 3948 | 1 | ← |

**Tabla 5.2.11 Genes homólogos entre las *Salmonellas* que tienen REP sólo en *S.typhi***
(23 uni & 23 cle)

| Gen | Pos | REP | Orientación |
|---|---|---|---|
| STM0509 (STY0556) | 515 | 1 | → |
| hcp (STY0933) | 860 | 1 | → |
| STM1020 (STY1034) | 949 | 1 | → |
| STM0907 (STY1042) | 956 | 1 | → |
| XXII* | | | |
| cbiA | 2066 | 1 | → |
| STM2514 (STY2545) - nuoN | 2358 | 1 | → |
| ratB (ratC) | 2557 | 1 | → |
| STM4261 (STY2875) | 2676 | 1 | → |
| STM2788 (STY2908) | 2707 | 1 | → |
| invF | 2816 | 1 | → |
| STM3142 (STY3313) | 3073 | 1 | ← |
| tpiA - STM4080 (STY3790) | 3532 | 1 | ← |
| STM4051 (STY3821) | 3562 | 1 | → |
| yiiL-STM3042A (STY3832) | 3571 | 1 | ← |
| STM3791 (STY3991) | 3722 | 1 | → |
| STY4198 5'-5' STY4199 | 3907 | 1 | → |
| STM3548 (STY4263) | 3968 | 1 | → |
| STM4253 (STY4449) | 4154 | 1 | → |
| STM2905 (STY4518)-phoN | 4219 | 1 | → |
| STM4444 (STY4783) | 4473 | 1 | ← |
| STM4510 (STY4866)-STY4867 | 4554 | 1 | → |
| STM4529 (STY4887)-STY4888 | 4573 | 1 | → |

**Tabla 5.2.12 Genes exclusivos de *S.typhimurium LT2* con REP**
(9 uni & 10 cle)

| Gen | Pos | REP | Orientación |
|---|---|---|---|
| XXI* | | | |
| STM1049 | 1591 | 2 | ← ← |
| STM2589 | 3733 | 1 | → |
| STM2607 | 3752 | 1 | ← |
| STM2689 | 3861 | 1 | → |
| STM2751 | 3939 | 1 | ← |
| STM3773 | 5322 | 1 | ← |
| STM4196 | 5911 | 1 | ← |
| STM4425 | 6244 | 1 | → |

**Tabla 5.2.13 Genes exclusivos de *S.typhi* con REP**
(17 uni & 17 cle)

| Gen | Pos | REP | Orientación |
|---|---|---|---|
| STY3417 - STY3418 | 3166 | 1 | → |
| STM4313 | 3166 | 1 | → |
| STY0308 | 286 | 1 | → |
| clpA-STY0945 | 869 | 1 | ← |
| STY1006 | 923 | 1 | → |
| STY1295 | 1186 | 1 | ← |
| STY2039 | 1871 | 1 | → |
| STY2929 | 2725 | 1 | ← |
| STY3343 | 3099 | 1 | → |
| STY3645 | 3388 | 1 | ← |
| XXIII* | | | |
| STY4012 | 3740 | 1 | → |
| STY4573 | 4265 | 1 | → |
| STY4576 | 4268 | 1 | → |
| XXIV* | | | |
| hpcC | 1036 | 1 | → |
| xseA-shdA | 2555 | 1 | ← |

**Tabla 5.2.14 Genes homólogos entre las *Salmonellas* que tienen REP sólo en *S.typhimurium LT2***
(11 uni & 11 cle)

| Gen | Pos | REP | Orientación |
|---|---|---|---|
| STM0050 | 79 | 1 | ← |
| STM0571-STM0572 | 875 | 1 | ← |
| STM0926 | 1396 | 1 | ← |
| STM2720 | 3897 | 1 | → |
| XXV* | | | |
| STM2804 | 4017 | 1 | ← |
| STM3072 | 4363 | 1 | ← |
| STM3175 | 4504 | 1 | ← |
| STM3529 | 4985 | 1 | ← |
| STM4070-STM4071 | 5734 | 1 | ← |

| | | | |
|---|---|---|---|
| STM4207 | 5922 | 1 | ← |

**Notas:**

**XXI*:** STM1012 5'-5' STM1013 1523 1 ←
**XXII*:** STY1399) 5'-5' STY1399)1283 1 →
**XXIII*:** STY3671 5'-5' STY3672 3414 1 →
**XXIV*:** STY48205'-5' STY4821 4511 1 ←
**XXV*:** STM28035'-5' STM2803 4016 1 ←

**Tabla 5.2. "Gen"** es el gen tres prima asociado a la unidad (véase metodología), excepto cuando se indica. El nombre del gen siempre proviene de *Escherichia coli K12* cuando está presente en el grupo. Φ marca cuando los genes de ambas *Salmonella* son homólogos pero el nombre asignado varía, en dado caso, se señalan ambos. **"REP"** es el número de elementos REP en esa unidad. **"Orientación"** es el arreglo de los elementos en la unidad. Cuando la unidad se encuentra entre dos regiones tres-prima ambos genes se señalan, las posiciones corresponden al primero de ellos. También se indican las unidades que se encuentran en regiones cinco-prima–cinco-prima (notas). Cuando aparece **"Gen*"** o **"Gen**"** quiere decir que los genes asociados a la unidad variaron en los distintos organismos y se señalan los cambios ocurridos. Los colores de las asociaciones señalan cuando dos o más unidades son idénticas entre los organismos. Los colores de las asociaciones entre las tres *Escherichia coli* no varían a menos que se haya una *Salmonella* involucrada, en dado caso los colores son: A las que comparten ambas *Salmonella*. ■ las unidades que comparten las cinco bacterias que mantienen número y orientación constante de los elementos. ■ las que se mantienen iguales en *Salmonella* y *Escherichia coli K12*. ■ las que presentan *Salmonella typhimurium LT2* y la patógena *Escherichia coli O157H7*. ■ las que presentan *Salmonella typhi* y las *Escherichia coli* patógenas. A las que presentan *Salmonella typhi* y *Escherichia coli K12*. las que presentan *Salmonella typhimurium LT2* y *Escherichia coli K12*. ■ las que presentan *Salmonella typhimurium LT2 y* las *Escherichia coli K12 y OH157H7*. ■ las que presentan *Salmonella typhi* y *Escherichia coli K12 y OH157 H7*. ■ las que presentan *Salmonella* y las dos patógenas. ■ las que presentan *Salmonella typhimurium LT2* y las tres *Escherichia coli*. ■ las que presentan *Salmonella typhi* y *Escherichia coli H157H7 EDL933*. ■ las que presentan *Salmonella typhi* y *Escherichia coli K12 y OH157 H7K*. ■ las que presentan *Salmonella typhimurium LT2* y las tres *Escherichia coli*. □ muestra las unidades que sí variaron su número y/o orientación, o que simplemente no tienen una unidad homóloga en algún otro organismo. ■ presenta las unidades tres-prima-tres-prima que se fraccionaron.

La Tabla 5 muestra los diferentes grupos formados después de los análisis de homología entre los genes (véase metodología). Los grupos fueron formados de acuerdo a las asociaciones entre organismos, genes y unidades. Se hicieron dos comparaciones independientes: la primera fue exclusivamente entre genes homólogos de *Escherichia coli* y la segunda fue entre los genes homólogos de las cinco enterobacterias.

Del análisis de la Tabla 5 se desprenden aspectos importantes que pueden ayudar en el entendimiento de los elementos REP. Lo primero que resalta es la conservación de las unidades a través del tiempo. Por razones de parentesco, los organismos más cercanos filogenéticamente presentaron mayor número de unidades compartidas exactas (Tabla 5 y 6), es decir, las tres *Escherichia coli* compartieron más elementos entre sí

(180) que con las *Salmonella* (41), y viceversa (162 entre las *Salmonella*). Además, al interior de las *Escherichia coli*, las dos patógenas fueron más parecidas (196).

El origen de los elementos REP se remonta, por lo menos, al último ancestro que tuvieron en común las cinco bacterias analizadas. Hay 41 unidades que comparten todas ellas y que muy posiblemente heredaron de su ancestro. Aparentemente, con la divergencia de la rama *Escherichia coli* y la rama *Salmonella*, la importancia de las secuencias REP varió, en las *Escherichia coli* se hizo trascendental y en la otra decayó notoriamente. Sin embargo, en ambas se continuó la diseminación de dichas secuencias hasta llegar a números muy elevados.

Lo anterior forzosamente implica que hay entre un 87-92% de unidades que no se han mantenido constantes, esto es, que han aparecido o han desaparecido en las cinco diferentes bacterias. Entre *Escherichia coli* este porcentaje se reduce a entre un 25-20% y en las *Salmonella* a un 50%, dejando ver que en estas últimas ha habido un importante número de elementos que han aparecido de forma independiente. Del mismo modo se puede decir que, en las *Escherichia coli* la aparición de unidades REP ha estado más controlada o las bacterias han estado menos tiempo separadas.

**Tabla 6. Los diferentes grupos de bacterias formados de acuerdo a como comparten las unidades REP**

| Asociación | Unidades Exclusivas | Núm. y Ori. | Dos o más | Sol. | Unidades acumuladas | Núm y Ori. Acumulada | Dos o más Acum. | Sol. Acum. |
|---|---|---|---|---|---|---|---|---|
| CINCO | 41 | 2 | | 2 | 41 | 2 | | 2 (100%) |
| LKPQ | 9 | 4 | 1 | 3 | 50 | 6 (12%) | 1 (16.66%) | 5 (83.33%) |
| LTPQ | 2 | | | | 43 | 2 (4.65%) | | |
| LTKQ | 1 | | | | 42 | 2 (4.76%) | | |
| TKPQ | 15 | 1 | | 1 | 56 | 3 (5.35%) | | 3 (100%) |
| LPK | 1 | 1 | | 1 | 42 | 3 (7.142%) | | 3 (100%) |
| LTK | 9 | 4 | | 4 | 50 | 6 (12%) | | 6 (100%) |
| TKP | 2 | 1 | | 1 | 43 | 3 (6.97%) | | 3 (100%) |
| TKQ | 1 | 1 | | 1 | 42 | 3 (7.14%) | | 3 (100%) |
| PQK | 139 | 111 | 42 | 69 | 180 | 113 (62.77%) | 43 (38.05%) | 71 (62.83%) |
| TPQ | 2 | 1 | | 1 | 43 | 3 (6.97%) | | 3 (100%) |
| LP | | 2 | | 2 | 53 | 9 (16.98%) | | 12 (133.33%) |
| LK | 1 | 1 | | 1 | 42 | 23 (54.76%) | | 17 (73.91%) |
| LT | 109 | 102 | 4 | 98 | 162 | 116 (71.60%) | 4 (3.44%) | 100 (86.20%) |
| TP | 1 | 1 | | 1 | 63 | 22 (34.92%) | | 12 (54.54%) |
| TQ | 1 | | | | 63 | 23 (36.50%) | | |
| TK | 10 | 4 | | 4 | 79 | 34 (43.03%) | | 21 (61.76%) |
| KQ | 9 | 26 | 9 | 17 | 189 | 139 (73.54%) | 52 (37.41%) | 19 (13.65%) |
| KP | 7 | 11 | 3 | 8 | 187 | 124 (66.31%) | 46 (37.09%) | 10 (8.06%) |
| PQ | 16 | 45 | 24 | 21 | 196 | 158 (80.61%) | 67 (42.40%) | 23 (14.55%) |

En la tabla **K** es *Escherichia coli K12*. **P** es *Escherichia coli O157 h7 EDL933*. **Q** es *Escherichia coli O157 H7*. **L** es *Salmonella typhimurium LT2*. **T** es *Salmonella typhi*. La segunda columna contiene las unidades que son exclusivas de esa asociación. La tercera, cuarta y quinta son las características de dichas unidades, es decir, cuántas conservan número y orientación, y de éstas cuántas son solitarias y cuántas no son. Las últimas cuatro columnas es el mismo concepto pero para los valores acumulados de los genomas.

Haciendo un análisis comparativo entre genomas completos (Tabla 6) se observa que entre *Salmonella* y *Escherichia coli*, de las pocas unidades que comparten, una mínima proporción de éstas se conserva idéntica, es decir, manteniendo el número y la orientación de los elementos. Es importante señalar también que las unidades compartidas tienen una fuerte tendencia a ser solitarias. Al comparar genoma por genoma se encuentra que, las similitudes entre las unidades REP de las dos *Salmonella*

y las tres *Escherichia coli* son parecidas, lo que significa que las diferencias entre ellas son equivalentes.

De las unidades que comparten las tres *Escherichia coli* hay una enorme proporción en las que se conserva tanto el número como la orientación de los elementos que las componen (Tabla 6). A simple vista parecería que la conservación de las secuencias REP no sólo se da en la secuencia sino también en la relación gen-unidad-número y orientación de los elementos que la conforman en estos organismos. Cuando el mismo análisis se amplía a las cinco enterobacterias, se puede observar que esta asociación gen-unidad-número y orientación de los elementos desaparece pues de las 41 unidades que comparten todas, sólo dos han mantenido invariable su configuración (Tabla5).

En las dos *Salmonella* y en *Escherichia coli K12* ha habido recientes y numerosas apariciones de elementos REP. Lo anterior es posible saberlo dada la enorme cantidad de unidades exclusivas que presentan. La diferencia entre estos organismos es que en *Escherichia coli K12* los elementos se han mantenido con las características distintivas de los elementos REP. En las *Salmonella*, por el contrario, los elementos se encuentran muy deteriorados. Es por ello que lo anterior es un indicador importante de que es en *Escherichia coli K12* en donde los eventos que llevan a la formación y conservación de las secuencias REP están actuando de forma más intensa y precisa (Fig. 14, 15 y Tabla 5).

Por razones evolutivas evidentes hay un número muy elevado de genes con unidades REP que comparten las cepas patógenas de *Escherichia coli* (en muchas de éstas es apreciable la conservación tanto del número como la orientación de los elementos, Tabla 5). Aún así, pese a que muchas unidades se encuentran asociadas genes exclusivos de estas *Escherichia coli*, nunca están en el mismo gen exclusivo (Fig. 14, 15 y Tabla 5).

Otro aspecto interesante es que la gran mayoría de las diferencias entre unidades compartidas entre las dos *Salmonella* se deben a la inversión de grandes fragmentos de DNA cromosomal. Lo anterior se observa por la disposición de los elementos y por la inversión en el arreglo de los genes tres-prima-tres-prima.

Las unidades, junto con sus genes asociados, han variado de ubicación en el genoma, se han fragmentado, o se han repetido. Es interesante constatar como, en algunos organismos hay unidades agrupadas entre dos genes tres-prima que en los otros se encuentran separadas, producto de un movimiento del gen o por la inserción de fragmentos de DNA (Tabla 5). Este aparente movimiento de genes y elementos REP al interior de los genomas es el principal causante de que las unidades y sus asociaciones cambien tanto de un organismo a otro. Esto parecería opuesto a la idea de que las unidades tienen una asociación fuerte con el gen al que están asociadas. Pues si en un organismo las secuencias REP guardan una disposición con respecto a ciertos genes y en otros organismos esta disposición varía por la inserción, movimiento o pérdida de genes, las relaciones gen-REP se vuelven insignificantes. Estos cambios en las posiciones relativas de los genes-unidades son muestra de la impresionante movilidad

de los elementos que conforman los genomas de las bacterias (Tabla 5). Para mayores detalles véase Koonin *et al*[121].

La falta de constancia en la presencia de unidades REP con respecto a algunos genes homólogos pudiera deberse a alguna de las siguientes razones: i) La importancia de las unidades varía según el gen. Por ejemplo, no es igual de importante una unidad que está presente en los cinco genes homólogos de las bacterias aquí consideradas, a una unidad que sólo está presente en uno de los cinco genomas en cuestión. ii) La importancia de las unidades varía de acuerdo al organismo en el que están. En las cinco bacterias la presencia de las unidades asociadas a ciertos genes es trascendental, lo que cambia es la importancia de la relación gen-unidad. iii) La movilidad de las unidades es muestra de la poca importancia de la asociación específica gen-REP en el genoma de los organismos.

[121] Koonin, E; Aravind, L; Kondrashov, A (2000) *The impact of comparative genomics on our understanding of evolution.* Cell. 101: 573-576.

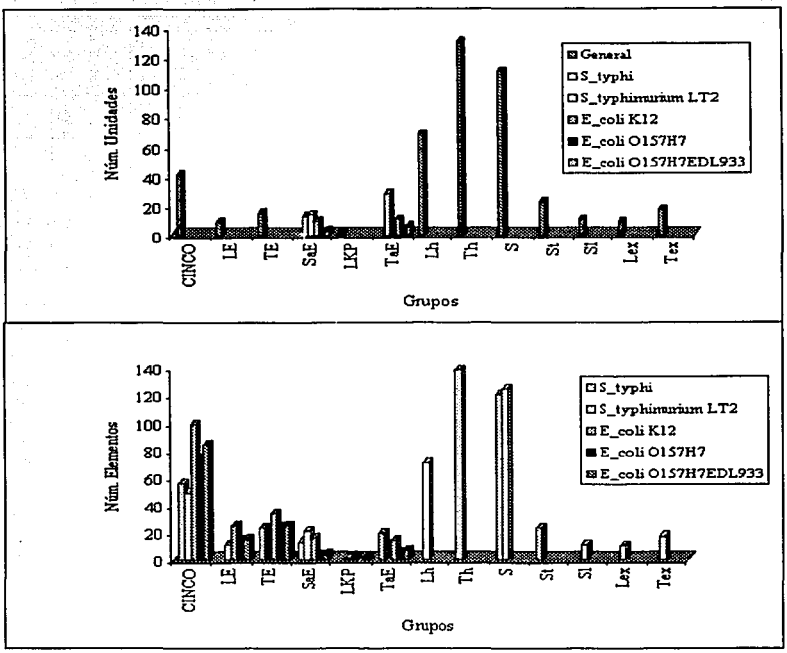



**Fig. 14** Características generales de los grupos de homología formados entre las cinco bacterias. La primera gráfica muestra las comparaciones por número de unidades. En la gráfica de abajo se muestran las comparaciones por número de elementos. Los grupos formados son: **CINCO**, unidad o elementos presentes en todas. **LE**, genes homólogos pero que sólo en la *Salmonella typhimurium LT2* y en las tres *Escherichia coli* está presente la unidad o los elementos. **TE**, genes homólogos pero que sólo en la *Salmonella typhi* y en las tres *Escherichia coli* está presente la unidad o los elementos. **SaE**, genes homólogos pero que sólo en las *Salmonella* y en alguna de las tres *Escherichia coli* está presente la unidad o los elementos. **LKP** genes homólogos pero que sólo en la *Salmonella typhimurium LT2* y en las *Escherichia* coli K12 *y O157H7 EDL933* está presente la unidad o los elementos. **TaE**, genes homólogos pero que sólo en la *Salmonella typhi* y en alguna de las tres *Escherichia coli* está presente la unidad o los elementos. **Lh**, genes homólogos pero que sólo en la *Salmonella typhimurium LT2* está presente la unidad o los elementos. **Th**, genes homólogos pero que sólo en la *Salmonella typhi* está presente la unidad o los elementos. **S**, genes homólogos pero que sólo en las *Salmonella* está presente la unidad o los elementos. **St**, genes homólogos de las *Salmonella* pero que sólo en la *Salmonella typhi* está presente la unidad o los elementos. **SI**, genes homólogos de las *Salmonella* pero que sólo en la *Salmonella typhimurium LT2* está presente la unidad o los elementos. **Lex**, genes exclusivos de *Salmonella typhimurium LT2* con unidad o elementos. **Tex**, genes exclusivos de *Salmonella typhimurium LT2* con unidad o elementos.

Un aspecto importante sobre la dinámica en el movimiento de las unidades REP, es su presencia en genes exclusivos. Éstos son muestra del reciente desplazamiento de unidades producto de una maquinaria molecular aún activa. La relación entre unidades REP y los genes exclusivos podría ser señal de que la asociación gen-unidad REP no está relacionada con el tipo de gen, pues muchos de ellos presentan funciones únicas. Por lo general, estas clases de unidades están formadas solamente por elementos solitarios (Fig. 14,15 y Tabla 5).

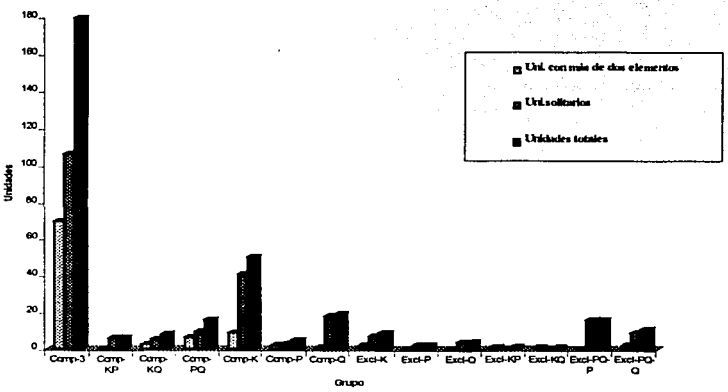



**Fig. 15** Características generales de los grupos de homología formados. **K:** *Escherichia coli K12*, **P:** *Escherichia coli O157H7 EDL933* y **Q:** *Escherichia coli O157H7*. **Comp-3**, unidad presente en las tres *Escherichia coli*. **Comp-KP**, genes homólogos pero que sólo en K y P está presente la unidad. **Comp-KQ**, genes homólogos pero que sólo en K y Q está presente la unidad. **Comp-PQ**, genes homólogos pero que sólo en P y Q está presente la unidad. **Comp-K**, genes homólogos pero que sólo en K. **Comp-P**, genes homólogos pero que sólo en P está presente. **Comp-Q**, genes homólogos pero que sólo en Q está presente. **Excl-K**, genes exclusivos de K con unidad asociada. **Excl-P**, genes exclusivos de P con unidad asociada. **Excl-Q**, genes exclusivos de Q con unidad asociada. **Excl-KP**, genes exclusivos de K y P con unidad asociada. **Excl-KQ**, genes exclusivos de K y Q con unidad asociada. **Excl-PQ-P**, genes exclusivos de P y Q con unidad asociada sólo en P. **Excl-PQ-Q**, genes exclusivos de P y Q con unidad asociada sólo en Q.

Sin embargo, pese a la aparente movilidad de algunos elementos REP dentro de los genomas, las propiedades de éstos no se han visto alteradas de forma considerable en *Escherichia coli*: su presencia se mantuvo preferentemente vinculada a la región tres prima de algún gen, los elementos se mantuvieron casi exclusivamente en regiones extragénicas y las secuencias se mantuvieron bastante bien conservadas. En *Salmonella* estas reglas se han visto modificadas considerablemente.

### 4.6 Conservación de los Palíndromos en las Unidades

La conservación de la forma palindrómica en los elementos solitarios es bastante constante para las tres *Escherichia coli*. Por lo general, sin importar su cercanía con la secuencia consenso, el cambio en una de las partes del palíndromo lleva al cambio en la base correspondiente de la segunda parte. Hay dos excepciones importantes que se encontraron en los elementos de estas tres bacterias: la primera es que siempre, en todos los elementos solitarios, se encontró una región en donde el palíndromo no coincide y que no forma parte de la región variable **""RV""**. Esta región es por lo general de dos bases de largo (puede alcanzar las tres) y su secuencia es constante, hay un par de timinas en un lado del palíndromo y una citocina y una guanina en el otro lado. Para la segunda forma de elemento REP las bases son exactamente las opuestas: dos adeninas y una guanina- citocina (Fig. 16A).

La segunda excepción que se encontró fue que en aquellos elementos donde la variación de la secuencia aumentaba, imposibilitando el mantenimiento del palíndromo, la región variable tenía un tamaño igual o superior a cuatro bases. Cabe señalar que no

todos los elementos que poseían una región variable de cuatro o más bases de largo presentaban un aumento en la variación de las mismas. Esta variación asociada a **"RV"** grandes se vio más frecuentemente en las cepas patógenas.

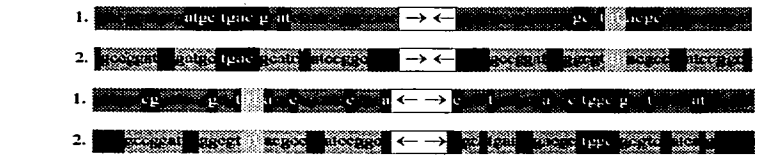
Para las unidades compuestas por dos o más elementos, las reglas del palíndromo son distintas ya que estos arreglos son doblemente palindrómicos (Fig. 2). En las tres *Escherichia coli* se encontró una tendencia a conservar el palíndromo interno de cada elemento sobre el palíndromo formado por los dos tipos de elemento REP. Un aspecto interesante que vale la pena resaltar, es que muchas veces la alteración en las bases del palíndromo formado entre los dos elementos REP se debía a una corrección en las bases del palíndromo interno (Fig. 16B). Estas características muestran que indudablemente las presiones de selección actúan sobre los elementos y la conservación del arreglo palindrómico interno.

Otro dato que llama la atención, es que en la mayoría de las unidades con dos elementos REP presentes en las tres *Escherichia coli*, siempre hubo un elemento con una región variable de cuatro o cinco bases, mientras que el otro siempre presentó una región variable de dos bases de largo (Fig. 16B). Por lo general el elemento REP que va en el sentido "→" es el que presentó la región variable más grande, sin importar el arreglo de los elementos en la unidad.

**A. Dos unidades con elementos solitarios en *Escherichia coli***


Regiones del elemento que no están involucradas en el palíndromo.
Regiones del elemento que son palindrómicas.
Regiones del elemento en donde el palíndromo no coincide.
Región variable.

**B. Dos unidades con dos elementos REP *Escherichia coli***



Regiones del elemento que no están involucradas en el palindromo.
Regiones del elemento que son palindrómicas.
Regiones del elemento en donde el palindromo no coincide.
Región variable de mayor tamaño asociada al elemento →.
Región variable de menor tamaño asociada al elemento ←.
Regiones de los elementos involucradas en el palindromo.
Regiones de los elementos en donde el palindromo no coincide.

**C. Tres elementos REP ← de *Salmonella typhymurium LT2***

Regiones del elemento que no están involucradas en el palindromo. Zonas del palindromo no conservadas. Zonas conservadas del palindromo. **"RV"**.

**Fig. 16 A.** Unidades compuestas por elementos solitarios, se muestran las dos posibilidades de elementos REP. Nótese que además de la región variable hay dos bases que no coinciden en el palíndromo. **B.** Unidades compuestas por dos elementos; se muestran repetidos dos de los arreglos que pueden tener. Nótese las diferencias que hay entre las bases que componen los posibles palíndromos, las bases que no coinciden en una de las posibilidades se deben a cambios en el otro palíndromo posible. **C.** Se muestran tres unidades solitarias de *Salmonella typhimurium LT2* pertenecientes al rango de 74.5-72%, que es donde los elementos REP son más numerosos (Tabla 2). Nótese el grado de deterioro de los elementos del palíndromo, por lo general el deterioro se da en las zonas cercanas a **"RV"**.

Para las *Salmonella* las cosas son aparentemente distintas. La mala conservación de muchos de sus elementos (Tabla 3) y la baja frecuencia de unidades con dos o más elementos (Fig. 10 y 11) podrían ser evidencia de la falta de importancia por el mantenimiento de cualquiera de los arreglos palindrómicos característicos de una secuencia REP. Como es apreciable en la Fig.16C, a diferencia de muchos elementos de las tres *Escherichia coli*, las acumulaciones de mutaciones en los elementos no son contrarrestadas con el fin de mantener el arreglo palindrómico, por el contrario, éstas

son acumuladas sin sentido, produciendo el deterioro notorio del arreglo de las bases en los elementos. Esta acumulación se da de manera más intensa en la periferia de la región variable **"RV"**.

### 4.7 Funciones de los Genes Asociados a Unidades REP

A continuación se enlistan las funciones de los genes de las *Escherichia coli* que presentaron una unidad REP asociada. Se incluyen dos tablas, una organizada con base en los organismos por separado (Tabla 7) y la otra con base en los grupos formados por la homología de los genes (Tabla 8).

**Tabla 7. Organización de las funciones de los genes, que presentan una unidad REP en su región tres prima, por organismo**

| Funciones en *Escherichia coli K12* | | Funciones en *Escherichia coli O157H7 EDL933* | | Funciones en Escherichia coli O15H7 | |
|---|---|---|---|---|---|
| proteína hipotética | 77 | proteína hipotética | 63 | proteína hipotética | 67 |
| trasportadora | 22 | trasportadora | 20 | trasportadora | 20 |
| oxidoreductasa | 16 | oxidoreductasa | 15 | sintetasa | 17 |
| sintetasa | 16 | sintetasa | 14 | oxidoreductasa | 15 |
| deshidrogenasa | 13 | deshidrogenasa | 12 | profago | 13 |
| reguladora | 12 | isomerasa | 10 | deshidrogenasa | 11 |
| transferasa | 11 | reguladora | 10 | reguladora | 11 |
| isomerasa | 10 | transferasa | 10 | isomerasa | 10 |
| putativa | 9 | putativa | 7 | putativa | 9 |
| cinasa | 6 | cinasa | 6 | transferasa | 9 |
| estructural | 6 | profago | 5 | cinasa | 7 |
| peptidasa | 6 | proteína del ciclo celular | 5 | proteína del ciclo celular | 6 |
| proteína del ciclo celular | 5 | chaperona | 4 | no definida | 5 |
| profago | 4 | lipoproteína | 3 | peptidasa | 4 |
| glicosilasa | 3 | peptidasa | 3 | lipoproteína | 3 |
| helicasa | 3 | permeasa | 3 | permeasa | 3 |
| lipoproteína | 3 | respuesta al estrés | 3 | respuesta al estrés | 3 |
| permeasa | 3 | tRNA | 3 | acetilasa | 2 |
| respuesta al estrés | 3 | aldolasa | 2 | aldolasa | 2 |
| acetilasa | 2 | deaminasa | 2 | deaminasa | 2 |
| aldolasa | 2 | estructural | 2 | estructural | 2 |
| deaminasa | 2 | helicasa | 2 | glicosilasa | 2 |
| ligasa | 2 | ligasa | 2 | Helicasa | 2 |
| proteína alfa hélice | 2 | no definida | 2 | ligasa | 2 |

| | | | | | |
|---|---|---|---|---|---|
| sistema PTS | 2 | acetilasa | 1 | tRNA | 2 |
| activador transcripcional | 1 | activador transcripcional | 1 | activador transcripcional | 1 |
| carboxilasa | 1 | carboxilasa | 1 | carboxilasa | 1 |
| esterasa | 1 | exonucleasa | 1 | cofactor de síntesis | 1 |
| exonucleasa | 1 | Factor de iniciación de cadena proteica | 1 | colagenasa | 1 |
| factor de iniciación de cadena proteica | 1 | fosfoglucomutasa | 1 | esterasa | 1 |
| factor de transcripción | 1 | glicosilasa | 1 | exonucleasa | 1 |
| flavoproteína | 1 | nucleoproteína | 1 | factor de iniciación de cadena proteica | 1 |
| fosfatasa | 1 | peptido lider | 1 | factor de liberación de péptido | 1 |
| fosfoglucomutasa | 1 | receptor | 1 | factor de transcripción | 1 |
| hidrolasa | 1 | sulfatasa | 1 | flavoproteína | 1 |
| homólogo de factor de virulencia | 1 | | | fosfoglucomutasa | 1 |
| metilasa | 1 | | | metilasa | 1 |
| peptido lider | 1 | | | peptido lider | 1 |
| peroxidasa | 1 | | | peroxidasa | 1 |
| proteína de la subunidad 30S del ribosoma | 1 | | | profesa | 1 |
| proteína de membrana | 1 | | | proteína de membrana | 1 |
| receptor | 1 | | | proteína dnaK | 1 |
| RNAsa | 1 | | | proteína rhsD | 1 |
| tRNA | 1 | | | proteína sensora | 1 |
| | | | | receptor | 1 |
| | | | | sistema PTS | 1 |
| | | | | sulfatasa | 1 |
| | | | | trancetolasa | 1 |

**Tabla 8. Organización de las funciones de los genes por grupos de homología**

| Genes exclusivos de Q y P que sólo presentan en Q REPs | |
|---|---|
| proteína de profago | 9 |
| no definida | 5 |
| proteína hipotética | 3 |
| cinasa | 1 |
| proteínas de transporte | 1 |
| proteínas reguladoras | 1 |

| Genes exclusivos de P que presentan REPs | |
|---|---|
| no definida | 2 |
| proteína de profago | 1 |
| proteína hipotética | 1 |

| Genes que tienen las tres y presentan REPs en las tres | |
|---|---|
| proteína hipotetica | 50 |
| oxidoreductasa | 13 |
| trasportadora | 12 |
| sintetasa | 11 |
| deshidrogenasa | 10 |

| | |
|---|---|
| isomerasa | 10 |
| putativa | 7 |
| transferasa | 7 |
| cinasa | 5 |
| proteína del ciclo celular | 5 |
| reguladora | 5 |
| ptofago | 4 |
| lipoproteína | 3 |
| peptidasa | 3 |
| permeasa | 3 |
| respuesta al estrés | 3 |
| aldolasa | 2 |
| deaminasa | 2 |
| estructural | 2 |
| helicasa | 2 |
| ligasa | 2 |
| acetilasa | 1 |
| activador transcripcional | 1 |
| carboxilasa | 1 |
| exonucleasa | 1 |
| factor de iniciación de cadena proteica | 1 |
| glicosilasa | 1 |
| peptido lider | 1 |
| receptor | 1 |
| tRNA | 1 |

| Genes compartidos en donde sólo K y P presentan REP |
|---|
| tranferasa 2 |
| deshidrogenasa 1 |
| oxidoreductasa 1 |
| proteína hipotética 1 |
| reguladora 1 |
| transportadora 1 |

| Genes compartidos en donde sólo K y Q presentan REP |
|---|
| proteína hipotética 5 |
| esterasa 1 |
| factor de transcripción 1 |
| flavoproteína 1 |
| oxidoreductasa 1 |
| peptidasa 1 |
| reguladora 1 |
| sintetasa 1 |

| Genes compartidos en donde sólo P y Q presentan REP |
|---|
| transportadora 5 |
| proteína hipotética 3 |
| sintetasa 3 |
| reguladora 2 |
| cinasa 1 |
| deshidrogenasa 1 |
| fosfoglucomutasa 1 |
| sulfatasa 1 |
| transferasa 1 |
| tRNA 1 |

| Genes compartidos en donde sólo K presenta REP | |
|---|---|
| proteína hipotética | 16 |
| transportadora | 9 |
| reguladora | 5 |
| estructural | 4 |
| sintetasa | 4 |
| deshidrogenasa | 2 |
| glicosolasa | 2 |
| peptidasa | 2 |
| putativa | 2 |
| sistema PTS | 2 |
| transferasa | 2 |
| cinasa | 1 |
| fosfatasa | 1 |
| fosfoglucomutasa | 1 |
| helicasa | 1 |
| metilasa | 1 |
| oxidoreductasa | 1 |
| Peroxidasa | 1 |
| proteína alfa hélice | 1 |
| proteína de la subunidad 30S del ribosoma | 1 |
| RNAsa | 1 |

| Genes compartidos en donde sólo P presenta REP | |
|---|---|
| proteína hipotética | 2 |
| nucleoproteína | 1 |
| oxidoreductasa | 1 |
| reguladora | 1 |
| transportadora | 1 |
| tRNA | 1 |

| Genes compartidos en donde sólo Q presenta REP | |
|---|---|
| proteína hipotética | 5 |
| sintetasa | 2 |
| transportadora | 2 |
| cofactor de síntesis | 1 |
| colagenasa | 1 |
| factor de liberación de péptido | 1 |
| glicosilasa | 1 |
| metilasa | 1 |
| oxidoreductasa | 1 |
| proteína de membrana | 1 |
| proteína del ciclo celular | 1 |
| proteína dnaK | 1 |
| proteína rhsD | 1 |
| proteína sensora | 1 |
| proxidasa | 1 |
| putativa | 1 |
| reguladora | 1 |
| sistema PTS | 1 |
| trancetolasa | 1 |
| transferasa | 1 |

| Genes exclusivos de K y P' presentan REP' | |
|---|---|
| Proteína hipotética | 1 |

| Genes exclusivos de K y Q' presentan REP' | |
|---|---|
| acetilasa | 1 |

| Genes exclusivos de K que presentan REP' | |
|---|---|
| proteína hipotética | 5 |
| hidrolasa | 1 |
| homólogo de factor de virulencia | 1 |
| proteína alfa hélice | 1 |
| proteína de membrana | 1 |

| Genes exclusivos de P y Q' en donde sólo P' presenta REP' | |
|---|---|
| proteína hipotética | 5 |
| chaperona | 4 |
| reguladora | 1 |
| transportadora | 1 |

| Genes exclusivos de Q' presentan REP' | |
|---|---|
| proteasa | 1 |
| proteína hipotética | 1 |
| putativa | 1 |
| reguladora | 1 |

**Notas:**

1) La diferencia entre proteínas proteína hipotéticas y putativas es que las segundas presentan un gen homólogo con función conocida
2) En la agrupación de proteínas por función no se hizo ninguna distinción entre las proteínas putativas y no putativas.
Las únicas excepciones se hicieron en aquellos casos donde las anotaciones no contenían una función asociada
3) El grupo de transportadoras agrupa, además de las proteínas transportadoras, a las proteínas de unión a moléculas y las permeasas
4) El grupo de reguladoras agrupa a las proteínas reguladoras, las activadoras, las represoras y las de unión a DNA
5) El grupo de oxidoreductasas agrupa a las oxidoreductasas, las reductasas y las oxidasas
6) El grupo de sintetasas agrupa también a las proteínas de biosíntesis

**K:** *Escherichia coli K12*, **P:** *Escherichia coli O157H7 EDL933* y **Q:** *Escherichia coli O157H7.*

Hay tres aspectos importantes que se desprenden del análisis de las funciones de los genes que presentan una unidad REP asociada en las *Escherichia coli*: i) La falta de información es evidente, la gran mayoría de los genes que tienen una unidad REP asociada a ellos, carecen de función conocida. ii) Se deduce de las tablas que, aunque hay funciones muy bien representadas, no hay una clara relación entre la función que desempeña el producto del gen y la presencia de una unidad REP colindante. iii) Entre

más unidades presente el grupo, más variadas son las funciones de los genes, esto nuevamente enfatiza la carencia de una relación estrecha entre función y unidad REP.

Para las *Salmonella* son aplicables las mismas generalidades antes mencionadas, debido a la poca información brindada por este análisis, las tablas de datos no se incluyeron.

### 4.8 Asociación con las Secuencias IHF

En el estudio de las secuencias de reconocimiento de la proteína IHF (Integration Host Factor) se decidió examinar la periferia de las unidades REP para encontrar, de ser posible, este tipo de secuencias, tal y como habían sido reportadas previamente[122]. La tabla 9 resume los resultados encontrados para las tres *Escherichia coli*. En total se encontraron 44 de estas secuencias en la cercanía de las unidades REP. La inmensa mayoría de estas secuencias (41) se hallaron asociadas a unidades con elementos solitarios. Es conveniente destacar que siempre que se encontró una secuencia IHF cerca de un elemento REP ésta, invariablemente, se ubicaba al término del mismo, sin importar la orientación de éste, es decir, cuando el elemento REP presentaba una dirección →, la secuencia IHF le continuaba: [redacted] → secuencia IHF; con el otro elemento REP sucedió lo mismo: secuencia IHF ← [redacted] (Tabla 9). En los casos donde estas secuencias IHF fueron halladas entre dos elementos REP, siempre fue posible asociarlas al final de uno de ellos.

[122] Boccard, F; Prentki, P. (1993) Specific interaction of IHF with RIBs, a class of bacterial repetitive DNA elements located at the 3'end of transcription units. *EMBO Journal*. 12(13):5019-5027.

Clarkson, S; Bates, AD.(1996).*Action of DNA gyrase at RIP elements in E.coli*.Biochem Soc Trans. 24(3):420.

Oppenheim, AB; Rudd, KE; Mendelson, I; Teff, D. (1993) *Integration host factor binds to a unique class of complex repetitive extragenic DNA sequnces in Escherichia coli*. Molecular Microbiology. 10(1):113-122.

**Tabla 9. Genes cuyas unidades presentaron cercanía con una secuencia IHF exitosa**

| Compartidos por las tres | |
|---|---|
| Gen | Ubicación |
| yaaA | → HIF ← |
| gcd | IHF ← |
| yaeF | HIF ← → |
| tax | IHF ← |
| fepA | IHF ← |
| glnS | ← IHF ← |
| ybhK | IHF ← |
| ybiA | IHF ← |
| mdoH | → HIF ← |
| flicC | → IHF |
| pepB | IHF ← |
| mltA | IHF ← |
| yhjG | IHF ← |
| yiaI | IHF ← |
| yiiF | → IHF |
| pfkA | IHF ← |
| metL | IHF ← |
| plsB | IHF ← |
| msrA | IHF ← |
| yigR | IHF ← |
| yjjK | IHF ← |

| Exclusivos de K12 | |
|---|---|
| Gen | Ubicación |
| pgmI | IHF ← |
| ybiU | IHF ← |
| ycdG | IHF ← |
| fliF | IHF ← |
| b1121 | IHF ← |
| uidC | IHF ← |
| ccmH | IHF ← |
| b2431 | IHF ← |
| glyA | IHF ← |
| yfiE | IHF ← |
| mpB | IHF ← |
| dacB | → IHF |
| yzgL | IHF ← |
| mtlA | IHF ← |
| yjiQ | IHF ← |

| Exclusivos O15H7 | |
|---|---|
| Gen | Ubicación |
| b2016 | IHF ← |
| Ecs4270 | IHF ← |

| Exclusivos O16H7-EDL933 | |
|---|---|
| Gen | Ubicación |
| Z3179 | IHF ← |
| Z4787 | IHF ← |

| Exclusivos K 12-O15H7 | |
|---|---|
| Gen | Ubicación |
| b1436 | IHF ← |

| Exclusivos Patógenas | |
|---|---|
| Gen | Ubicación |
| ansC | → IHF |
| yjbI3 | → IHF |
| yhfR(yjiR) | → IHF |

Para los tres organismos la disposición más frecuente fue IHF ← (35). Es notorio que es *Escherichia coli K12* la que mayor número de relaciones presenta con 38, de las cuales 16 son únicas de ella.

Las *Salmonella* presentaron poca relación entre REP e IHF. *Salmonella typhimurium LT2* tuvo una sola coincidencia. La otra bacteria presentó sólo dos de estas concordancias.

Al mismo tiempo, se llevó a cabo una búsqueda de todas las secuencias IHF de los genomas de las tres *Escherichia coli*. Se buscaron todas las secuencias IHF que tuvieran mínimo, un 80% de parecido con la secuencia consenso (véase metodología). *Escherichia coli K12* presentó 135 secuencias, de las cuales 68 estaban en secciones codificantes. *Escherichia coli O157H7* obtuvo 154, con 87 en regiones codificantes.

Finalmente, *Escherichia coli O157H7 EDL933* presentó 159, con 90 de ellas contenidas en regiones codificantes.

Aunque es cierto que una infinidad de elementos REP no presentan una secuencia IHF asociada (y viceversa), cabe destacar que, de las aproximadamente 69 secuencias IHF que están en regiones extragénicas en las tres *Escherichia coli*, 40 están relacionadas con un elemento REP. Sólo una relación se dio en una zona codificante.

Para identificar si las relaciones encontradas eran o no producto del azar, se decidió variar 1000 veces todas las regiones IHF de los tres genomas para posteriormente realizar una búsqueda de relaciones REP-IHF. Los resultados fueron contundentes: de las mil vueltas que se dieron en aproximadamente 100 por genoma, se localizó una sola relación. Dicho de otra forma, las relaciones REP-IHF encontradas no son casuales en las *Escherichia coli*. Los resultados sugieren la existencia de una relación entre los mecanismos que dan origen a las secuencias REP y los que dan origen a las secuencias IHF en estas bacterias. Por el contrario, en las *Salmonella*, los resultados fueron indistinguibles de los obtenidos por cuestiones de azar, es por ello que se puede concluir que en este tipo de bacterias no existe la relación entre ambas secuencias.

Las asocaciones entre secuencias REP y secuencias IHF han sido denomiadas RIP por **R**epetitive **I**HF-binding **P**alindromic elements[123]. Hasta este trabajo habían sido identificadas veintidós de estas relaciones[124] solamente en *Escherichia coli K12*. El arreglo de las secuencias coresponde a la disposición encontrada, es decir, final del elemento REP-secuencia IHF.

[123]Oppenheim, AB; Rudd, KE; Mendelson, I; Teff, D. (1993) *Integration host factor binds to a unique class of complex repetitive extragenic DNA sequnces in Escherichia coli*. Molecular Microbiology. 10(1):113-122.

[124] Oppenheim, AB; Rudd, KE; Mendelson, I; Teff, D. (1993) *Integration host factor binds to a unique class of complex repetitive extragenic DNA sequnces in Escherichia coli*. Molecular Microbiology. 10(1):113-122.

## 4.9 Secuencias tipo-REP en Otros Organismos

Finalmente, para completar el análisis de los elementos REP como secuencias de aparente relevancia en el genoma de las enterobacterias, se decidió buscar en todos los genomas totalmente secuenciados de procariontes secuencias que cumplieran con las características que definen a una secuencia REP: extragénicas, abundantes, conservadas, doblemente palindrómicas, que su número no se disparara, que se agruparan en unidades mayores a un elemento, que tuvieran un alto contenido de GC y que fueran de un tamaño mayor a las diez bases.

Para cada organismo se buscó, en sus regiones extragénicas, la secuencia más común utilizando el algoritmo público conocido como MEME[125]. Posteriormente, partiendo de esa secuencia encontrada, se realizó una búsqueda de las características necesarias que definen a un elemento REP, utilizando una modificación del programa de búsqueda de elementos REP (véase metodología).

La figura 17 y la tabla 10 resumen los resultados obtenidos para cada uno de los organismos. En total se contemplaron trece características definitorias de un elemento REP. Los organismos se agruparon de acuerdo a la cantidad de características que presentaron.

[125] Bailey, L T; Gribskov, M. (1998). *Combining evidence using p-values: application to sequence homology searches.* Bioinformatics. 14: 48-54.

| | TOTAL | Palíndromo | Palíndromo Interno | Unidades + de uno | Extragenicidad NOR +50% | Abundancia NOR +50 | Conservación NOR +50% | Extragenicidad PAL +50% | Abundancia PAL +50 | Conservación PAL +50% | Crec. Disparado NOR | Crec. Disparado PAL | %GC +50% | Secuencia grande |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Escherichia coli | 13 | X | X | X | X | X | X | X | X | X | X | X | X | X |
| Escherichia coli O157:H7 | 13 | X | X | X | X | X | X | X | X | X | X | X | X | X |
| Escherichia coli O157:H7 EDL933 | 13 | X | X | X | X | X | X | X | X | X | X | X | X | X |
| Streptococcus pneumoniae TIGR4 | 12 | X | X | X | X | X | X | X | X | X | X | X | - | X |
| Sulfolobus solfataricus | 12 | X | - | X | X | X | X | X | X | X | X | X | X | X |
| Staphylococcus aureus Mu50 | 11 | X | X | X | X | - | X | X | X | - | X | X | X | X |
| Staphylococcus aureus N315 | 11 | X | x | X | X | X | - | X | X | - | X | X | X | X |
| Caulobacter crescentus | 10 | X | X | X | X | X | - | - | X | - | X | X | X | X |
| Deinococcus radiodurans 1 | 10 | X | X | X | - | - | X | X | X | - | X | X | X | X |
| Vibrio cholerae 1 | 10 | X | - | X | X | X | - | X | X | X | X | X | - | X |
| Mycobacterium leprae | 9 | X | - | - | X | - | X | X | - | X | X | X | X | X |
| Aeropyrum pernix | 9 | X | X | - | X | - | X | X | - | X | X | X | - | X |
| Agrobacterium tumefaciens 1 | 9 | X | X | X | X | - | - | - | X | - | X | X | X | X |
| Sinorhizobium meliloti | 9 | X | X | X | X | X | X | - | X +1000 | - | X | - | X | X |
| Deinococcus radiodurans 2 | 9 | X | - | X | X | - | X | X | - | X | X | X | - | X |
| Archaeoglobus fulgidus | 8 | X | X | X | - | - | - | X | - | - | X | X | X | X |
| Agrobacterium tumefaciens 2 | 8 | X | X | X | - | X | - | - | - | - | X | X | X | X |
| Methanococcus jannaschii | 8 | X | - | X | - | - | - | X | - | X | X | X | X | X |
| Pyrococcus abyssi | 8 | X | - | X | X | - | X | - | - | - | X | X | X | X |
| | | | | | | | | | | | | | | |
| Buchnera sp. APS | 7 | X | X | X | - | X | - | - | - | - | - | X | X | X |
| Mycobacterium tuberculosis H37Rv | 7 | X | - | X | X | X | - | - | X | - | - | - | X | X |
| Mesorhizobium loti | 7 | X | - | X | - | X | - | - | X | - | - | X | X | X |
| Neisseria meningitidis Z2491 (serogroup A) | 7 | X | - | X | X | X | - | X | X | - | - | - | - | X |
| Neisseria meningitidis MC58 ( serogroup B) | 7 | X | - | X | X | X | - | X | X | - | - | - | - | X |
| Thermoplasma acidophilum | 7 | X | - | - | X | - | X | - | - | - | X | X | X | X |
| Thermoplasma volcanium | 7 | X | - | - | X | - | X | - | X | - | X | X | - | X |
| Bacillus subtilis | 7 | X | - | - | - | X | - | X | - | X | X | X | - | X |
| Borrelia burgdorferi | 7 | - | - | X | X | - | X | - | - | - | X | X | X | X |
| Vibrio cholerae 2 | 7 | - | - | - | X | X | X | - | - | - | X | X | X | X |
| Aquifex aeolicus | 6 | X | - | - | X | - | X | - | - | - | X | - | X | X |
| Mycoplasma pneumoniae | 6 | X | - | X | - | X | - | X | - | - | - | X | - | X |
| Mycobacterium tuberculosis CDC1551 | 6 | X | - | X | - | X | - | - | X | - | - | - | X | X |
| Rickettsia prowazekii Madrid E | 6 | X | - | X | - | X | - | - | X | - | - | - | X | X |
| Mycoplasma pulmonis | 6 | - | - | - | X | - | X | - | - | - | X | X | X | X |
| Ureaplasma urealyticum | 6 | - | - | - | X | - | X | - | - | - | X | X | X | X |
| Treponema pallidum subsp. | 6 | - | - | - | X | - | X | - | - | - | X | X | X | X |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pallidum | | | | | | | | | | | | | | |
| Chlamydophila pneumoniae AR39 | 6 | - | - | - | X | - | X | - | - | - | X | X | X | X |
| Mycoplasma genitalium | 6 | - | - | - | X | - | X | - | - | - | X | X | X | X |
| Halobacterium sp. NRC-1 | 5 | X | X | X | - | X...+1000 | - | - | X...+1000 | - | - | - | X | X |
| Helicobacter pylori J99 | 5 | X | - | - | - | X | - | - | X | - | X | - | - | X |
| Chlamydophila pneumoniae CWL029 | 5 | - | - | - | X | - | - | - | - | - | X | X | X | X |
| Chlamydophila pneumoniae J138 | 5 | - | - | - | X | - | X | - | - | - | X | X | - | X |
| Xylella fastidiosa | 5 | - | - | - | - | X | - | - | - | - | X | X | X | X |
| Lactococcus lactis | 4 | X | X | X | - | X...+1000 | - | - | X...+1000 | - | - | - | - | X |
| Pyrococcus horikoshii | 4 | - | - | - | - | X | - | - | - | - | X | - | X | X |
| Chlamydia muridarum | 3 | X | X | - | - | X...+1000 | - | - | X...+1000 | - | - | - | - | X |
| Haemophilus influenzae Rd | 3 | X | X | - | - | X...+1000 | - | - | X...+1000 | - | - | - | - | X |
| Pasteurella multocida | 3 | X | X | - | - | X...+1000 | - | - | X...+1000 | - | - | - | - | X |
| Methanothermobacter thermautotrophicus | 3 | - | - | - | - | X | - | - | - | - | - | - | X | X |
| Helicobacter pylori 26695 | 2 | X | - | - | - | X...+1000 | - | - | X...+1000 | - | - | - | - | X |
| Clostridium acetobutylicum | 2 | X | - | - | - | X...+1000 | - | - | X...+1000 | - | - | - | X | - |
| Thermotoga maritima | 2 | X | - | - | - | X...+1000 | - | - | X...+1000 | - | - | - | X | - |
| Synechocystis PCC6803 | 2 | X | - | - | - | X...+1000 | - | - | X...+1000 | - | - | - | X | - |
| Campylobacter jejuni | 1 | X | - | - | - | X...+1000 | - | - | X...+1000 | - | - | - | - | - |

**Tabla 10.** Resultados de las secuencias extragénicas de cada uno de los organismos totalmente secuenciados. Se muestran en orden descendente de acuerdo a la cantidad de características que comparten con las secuencias REP. **TOTAL**, se refiere a la cantidad de características que comparte. **Palíndromo**, se refiere a si la secuencia buscada tiene un palíndromo de ella. **Palíndromo interno**, hace referencia a si la secuencia buscada poseía un palíndromo en su interior. **Unidades + de uno**, se refiere a si los elementos encontrados se agrupaban en unidades mayores a uno. **Extragenicidad NOR +50%, Abundancia NOR +50** y **Conservación NOR +50%**, se refieren a si los elementos encontrados son más del 50% extragénicos y conservados, además de tener un número superior a 50. **Extragenicidad PAL +50%, Abundancia PAL +50** y **Conservación PAL +50%**, es lo mismo que el anterior pero para el palíndromo de la secuencia buscada. **Crec. Disparado**, para ambas secuencia se refiere a si conforme se reducen los parámetros de semejanza con la secuencia de inicio, el número de elementos encontrados se incrementa exponencialmente. **%GC +50%**, se refiere a si la secuencia posee un 50% o más de contenido de GC. **Secuencia grande**, si la secuencia es mayor a las 10 bases. La franja que divide la tabla separa los organismos que pudieran tener secuencias tipo-REP en su genoma de los que es improbable que las posean. El símbolo **X...+1000**, quiere decir que el número de elementos encontrados superó los 1000.

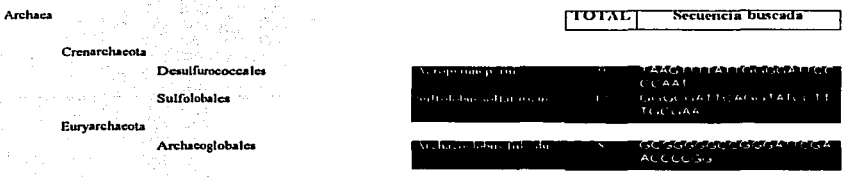
Es importante señalar que en las enterobacterias se encontró que la secuencia REP era la firma más abundante y significativa de las regiones extragénicas. Lo anterior prueba que el método utilizado para detectar secuencia tipo-REP era el adecuado. Los resultados obtenidos en el estudio de las secuencias más representativas de las regiones extragénicas de cada uno de los genomas, claramente muestran que es posible encontrar

toda la gama de posibilidades. Están presentes desde las secuencias que tienen todas las características necesarias, hasta las que apenas presentan una sola.

Además de las enterobacterias, ningún otro organismo presentó las trece características que definen a una secuencia REP. Es más, la gran mayoría de ellos presentaron menos de la mitad de éstas. El análisis de los resultados permitió decidir que a partir de las ocho características de semejanza, era presumible decir que las secuencias encontradas fueran en efecto secuencias tipo-REP.

En el grupo de organismos con posibles secuencias tipo-REP en sus genomas, es notorio cómo las características que más fallan son la conservación de los elementos, su abundancia y la presencia del palíndromo interno. En la mayoría de ellos, aparentemente, la presión de selección no es la suficiente como para contener la variación.

Se lograron detectar dos generalidades que comparten las secuencias tipo-REP: 1) Cuando se presenta una ausencia en el palíndromo interno, invariablemente los elementos se agrupan en unidades mayores a uno. Dicho de otra forma, la imposibilidad de generar una estructura con una sola secuencia se contrarresta juntando dos o más elementos por unidad. 2) Al contrario, cuando se presenta el palíndromo interno pero hay ausencia de unidades mayores a un elemento, existe una correlación en la abundancia, la conservación y la extragenicidad de ambos tipos de secuencias.



| | | | | |
|---|---|---|---|---|
| | Halobacteriales | Halobacterium sp. NRC-1 | 5 | |
| | Methanobacteriales | Methanothermobacter thermautotrophicus | 3 | |
| | Methanococcales | [illegible] | [illegible] | [illegible] |
| | Thermococcales | [illegible] | [illegible] | [illegible] |
| | Thermoplasmales | Thermoplasma acidophilum | 7 | |
| | | Thermoplasma volcanium | 7 | |
| Bacteria | | | | |
| Aquificales | | Aquifex aeolicus | 6 | |
| Firmicutes | | | | |
| | Bacillus/Clostridium group | | | |
| | Bacillaceae | Bacillus subtilis | 7 | |
| | Clostridiaceae | Clostridium acetobutylicum | 2 | |
| | Mycoplasmataceae | | | |
| | Mycoplasma | Mycoplasma genitalium | 6 | |
| | | Mycoplasma pneumoniae | 6 | |
| | | Mycoplasma pulmonis | 6 | |
| | Ureaplasma | Ureaplasma urealyticum | 6 | |
| | Staphylococcaceae | [illegible] | [illegible] | [illegible] |
| | Streptococcaceae | [illegible] | [illegible] | [illegible] |
| | | Lactococcus lactis | 4 | |
| | Actinobacteria | Mycobacterium tuberculosis H37Rv | 7 | |
| | | Mycobacterium tuberculosis CDC1551 | 6 | |
| | | [illegible] | [illegible] | [illegible] |
| Spirochaetales | | | | |
| | Spirochaetaceae | | | |
| | Borrelia | Borrelia burgdorferi | 7 | |
| | Treponema | Treponema pallidum subsp. pallidum | 6 | |
| Thermotogales | | Thermotoga maritima | 2 | |
| Thermus/Deinococcus group | | [illegible] | [illegible] | [illegible] |
| Planctomyces/Chlamydia/Verrucomicrobium group | | Chlamydia muridarum | 3 | |
| | | Chlamydophila pneumoniae CWL029 | 5 | |
| | | Chlamydophila pneumoniae AR39 | 6 | |
| | | Chlamydophila pneumoniae J138 | 5 | |
| Proteobacteria | | | | |
| | alpha subdivision | | | |
| | Caulobacter group | [illegible] | [illegible] | [illegible] |
| | Rhizobiaceae group | [illegible] | [illegible] | [illegible] |

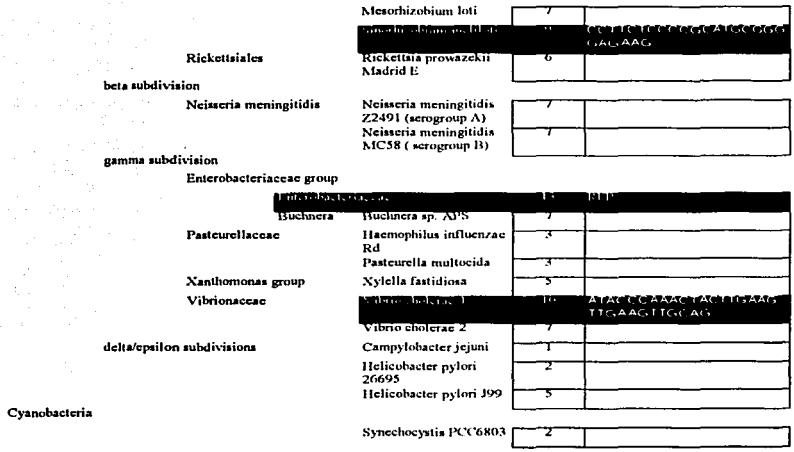



**Fig. 17** Arbol filogenético de los organismos procariontes en donde se representan los organismos de acuerdo a los resultados obtenidos en la búsqueda de secuencias tipo-REP en sus genomas. En sobreado negro se muestran los organismos con posibles secuencias tipo-REP en sus genomas. En blanco las que no las presentan.

A partir de la Fig. 17 es posible distinguir ciertos grupos compactos en donde es clara la presencia o ausencia de secuencias tipo-REP, por ejemplo en las crenarchaeota, las archeoglobales, las methanococcales, las thermoplasmales, las staphylococcaceae, las thermus/deinococcus y la subdivisión beta de las proteobacterias, presentan, en cada uno de sus miembros, las características suficientes para considerar que tienen secuencias tipo-REP. Del mismo modo, en los grupos halobacteriales, methanobacteriales, clostridiaceae, mycoplasmataceae, planctomyces/chlamydia/verrucomicrobium, delta/epsilon subdivisiones y cyanobacteria, claramente no hay secuencias tipo-REP.

De cualquier forma, el reparto general de estas secuencias en el árbol filogenético está disperso. Hay tanto en el dominio arquea como en el dominio bacteria.

Hay varios grupos donde algunos de sus miembros presentan las características necesarias, mientras otros no las exhiben; éstos son: thermococcales, bacillus, actinobacterias y las pertenecientes a las subdivisiones alfa y gamma de las proteobacterias.

Un último análisis que se realizó fue averiguar si los genes asociados a cada una de las secuencias tipo-REP, de aquellos organismos que tuvieron 8 o más características, eran homólogos en todas ellos. La Tabla 11 contiene los resultados correspondientes (véase metodología).

| | E_coli_K | S_pneum | S_solfat | S_aureu | S_aureu | M_lepr | C_cresc | D_radio | V_chol | A_perf | A_tum | S_melilot | D_radi | A_fulgi | A_tumi | M_jann | P_abyssi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E_coli | 526 - 263 | 138 | 20 | 20 | 20 | 20 | 45 | 26 | 127 | 10 | 56 | 415 | 50 | 29 | 612 | 28 | 18 |
| S_pne | 40 | 270 - 135 | 14 | 13 | 13 | 13 | 9 | 12 | 71 | 8 | 15 | 51 | 10 | 11 | 20 | 11 | 13 |
| S_solf | 14 | 12 | 212 - 106 | 5 | 5 | 5 | 2 | 7 | 6 | 2 | 1 | 16 | 2 | 3 | 3 | 1 | 13 |
| S_aure | 13 | 11 | 6 | 84 - 42 | 7 | 13 | 4 | 4 | 5 | 1 | 6 | 14 | 4 | 2 | 6 | 3 | 2 |
| S_aure | 16 | 20 | 7 | 20 | 96 - 48 | 5 | 4 | 4 | 13 | 0 | 4 | 18 | 3 | 3 | 5 | 3 | 3 |
| M_lep | 7 | 4 | 3 | 5 | 6 | 73 - 37 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 |
| C_cre | 27 | 11 | 3 | 3 | 4 | 4 | 156 - 78 | 0 | 12 | 0 | 13 | 0 | 0 | 2 | 9 | 1 | 1 |
| D_rad | 16 | 15 | 13 | 5 | 9 | 5 | 0 | 156 - 78 | 13 | 0 | 3 | 0 | 2 | 3 | 0 | 1 | 2 |
| V_cho | 67 | 24 | 9 | 8 | 11 | 8 | 22 | 19 | 69 - 18 | 1 | 19 | 0 | 0 | 4 | 12 | 3 | 3 |
| A_per | 1 | 3 | 2 | 1 | 1 | 1 | 0 | 0 | 1 | 18 - 9 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| A_tum | 20 | 21 | 1 | 1 | 1 | 1 | 21 | 2 | 33 | 1 | 54 - 27 | 0 | 0 | 2 | 8 | 2 | 1 |
| S_meli | 133 | 48 | 66 | 6 | 8 | 6 | 0 | 0 | 0 | 0 | 0 | 686 - 343 | 18 | 29 | 42 | 21 | 28 |
| D_rad | 15 | 8 | 4 | 7 | 7 | 7 | 0 | 3 | 0 | 0 | 0 | 50 | 48 - 24 | 3 | 5 | 4 | 3 |
| A_fulg | 10 | 11 | 5 | 3 | 3 | 3 | 3 | 3 | 8 | 1 | 2 | 33 | 6 | 94 - 47 | 9 | 1 | 1 |
| A_tum | 187 | 24 | 5 | 5 | 5 | 5 | 10 | 0 | 21 | 2 | 10 | 91 | 8 | 3 | 76 - 38 | 37 | 6 |
| M_jan | 5 | 10 | 2 | 1 | 1 | 1 | 3 | 2 | 11 | 2 | 3 | 46 | 7 | 2 | 68 | 58 - 29 | 4 |
| P_aby | 6 | 3 | 9 | 2 | 2 | 2 | 1 | 2 | 4 | 1 | 1 | 3 | 1 | 1 | 3 | 4 | 88 - 44 |

Tabla 11. Comparación de homología entre los genes que tienen asociada una secuencia tipo-REP. La diagonal interna de color amarillo señala el número de genes totales y el número de unidades totales para cada organismo. En ▒ se muestran las proteobacterias. En ▓ se muestran las firmicutes. En ▢ las deinococcales. En ■ se muestran las archeas.

Es apreciable de la Tabla 11, que el número de genes homólogos que comparten los organismos analizados es muy reducido. Esto quiere decir que no hay una relación entre los genes que presentan una secuencia tipo-REP asociada. Al estudiar a los

organismos por separado, se encontró que tampoco existía una relación entre los genes que tienen asociada una secuencia tipo-REP, es decir, los genes presentaron funciones muy diversas y baja homología.

## 5. CONCLUSIONES

Los métodos bioinformáticos utilizados en este trabajo, han permitido complementar la información obtenida por las técnicas experimentales comunes. Es apreciable de la investigación realizada, como ha sido posible la detección exacta de los elementos REP, así como su localización precisa y la determinación detallada de todas sus características, aspectos que no habían sido esclarecidos completamente con anterioridad.

Aunque es cierto que los trabajos bioinformáticos son sumamente teóricos, muchas veces este tipo de aproximación a los problemas puede aportar datos novedosos que permiten la construcción de un escenario más apegado a la realidad.

Para llevar a cabo esta investigación, se construyeron diversos programas de cómputo, concebidos específicamente para resolver el conjunto de problemas planteados. El buen funcionamiento de cada uno de ellos permitió obtener datos sólidos y bien respaldados con la información contenida en el genoma de los distintos organismos procariontes. Fue precisamente el buen desempeño de los programas, lo que dejó ver una serie de incongruencias con la base de datos de secuencias REP reportadas en el genoma de *Escherichia coli K12*, en lo que a elementos REP poco conservados se refiere. Lo anterior fue el factor primordial para que en esta investigación se generaran completamente las bases de datos de elementos REP para todos los organismos, partiendo solamente de la secuencia consenso de los elementos REP de *Escherichia coli K12*[126].

[126]Merino, E; Bolivar, F. (1989) The ribonucleoside diphosphate reductase gene (nrdA) of *Escherichia coli* carries a repetitive extragenetic palindromic (REP) sequence in its 3'structural terminus. *Molecular Microbiology*. 3(6): 839-841.

Las características tan particulares que presentan los elementos REP al interior de *Escherichia coli K12* (extragenicidad, asociación a algún gene tres-prima cercano, abundancia, formación de unidades y su carácter de doble palíndromo) han sido materia de múltiples especulaciones. Fue tarea de este trabajo el aproximarse al problema desde un punto de vista novedoso para intentar dilucidar, hasta lo posible, la dinámica de los elementos REP en los genomas. Para ello, se realizó una búsqueda precisa de todos los elementos REP en los genomas totalmente secuenciados de procariontes.

Los resultados del proceso de localización, caracterización y validación estadística, arrojaron datos novedosos. Si bien las enterobacterias fueron los organismos con mayor número de elementos REP en sus genomas y donde estos conservan en mayor grado sus características distintivas, en otros tres organismos la presencia de elementos REP también fue cierta: *Mycobacterium tuberculosis H37"RV"*, *Mycobacterium tuberculosis CDC1551 y Mesorhizobium loti*.

Las distancias filogenéticas existentes entre todos los organismos con secuencias REP en sus genomas, son muy amplias; las dos tuberculosis, *Mesorhizobium loti* y las enterobacterias están muy lejanas entre sí. Junto con lo anterior, el mal estado de los elementos en las bacterias no entéricas son evidencia suficiente para aseverar que el origen de los elementos REP en éstas fue por transferencia horizontal. Dada la frecuencia de este tipo de fenómenos en las bacterias[1] y la ya probada transferencia horizontal de genes con unidades REP[127], se puede especular que en estas tres bacterias no entéricas, posiblemente se presentaron los mecanismos moleculares, aunque éstos no

[127] Van Ham, S.M; van Alphen, L; Mooi, F.R; van Putten, O.M. (1994) The fibrial gene cluster of *Haemophilus influenzae* type b. Molecular Microbiology. 13(4): 673-684.

Vasconselos, A.T; Mattoso,M.A.G;de Almeida, D.F. (2000) Short interrupted palindromes on the extragenic DNA of *Escherichia coli K-12, Haemophilus influenzae* and *Neisseria meningitidis*. *Bioinformatics*. 16(11):968-977.

fueron los ideales, para que algunos de los elementos REP se pudieran diseminar por el genoma del organismo.

En las enterobacterias analizadas (tres cepas de *Escherichia coli*, *Escherichia coli K12 Escherichia coli O157H7* y *Escherichia coli O157H7 EDL933* y dos de *Salmonella*, *Salmonella typhimurium LT2 y Salmonella typhi*) se encontraron los elementos REP más conservados y abundantes. Estos rasgos constituyen una evidencia indudable de que el origen de los elementos REP se dio en este grupo de bacterias. Existían dos posibles explicaciones a esto[128]. 1) los elementos REP son de origen reciente y se han diseminado rápidamente entre las enterobacterias por eventos de transposición y transferencia horizontal muy intensos. 2) El origen de los elementos REP data por lo menos del último ancestro en común de estas cinco bacterias.

Los datos encontrados en este trabajo confirman la segunda de estas hipótesis. Se encontraron cuarenta y un unidades que estuvieron en la región tres prima extragénica del mismo gene homólogo en las cinco enterobacterias. Esto implica que entre el 11% y el 18% de las unidades totales de estos organismos conservan las características de extragenicidad asociada al extremo tres-prima del mismo gene homólogo. La probabilidad de que esto suceda por eventos azarosos producto de transposiciones recientes y eventos de transferencia horizontal, dado el tamaño de los genomas, su número de genes y las múltiples secciones extragénicas que existen, es sumamente reducida.

De los atributos que presentaron los elementos REP en las cinco enterobacterias se pueden inferir una serie de características que debieron estar presentes en el último ancestro en común de todas ellas: 1) los elementos solitarios fueron los más abundantes

y los mejor seleccionados en el ancestro, esto debido a que en la actualidad lo siguen siendo y de forma muy marcada, además que existe una clara predilección por el palíndromo interno sobre cualquier otro arreglo, lo que demuestra que la selección se da al nivel de elementos y no de unidades. 2) La abundancia de los elementos que debieron de estar en este organismo pareciera ser baja dada la poca cantidad de unidades que se comparten. No obstante, es posible que haya pasado el tiempo suficiente como para que los rearreglos cromosomales y la acumulación de mutaciones movieran o desaparecieran a los elementos preexistentes. 3) Es posible asegurar que en esta bacteria los elementos se encontraban en regiones tres-prima extragénicas de algún gene, ya que la inmensa mayoría de los elementos en las cinco enterobacterias guardan estas características y es más probable que se hayan adquirido por línea vertical del ancestro, a suponer dos apariciones independientes en linajes tan cercanos.

Con la división de la línea *Salmonella* de la línea *Escherichia coli*, hubo cambios notables en cuanto a las secuencias REP se refiere. En la primera rama estos elementos sufrieron una explosiva diseminación, pero una pésima conservación, mientras que en el segundo linaje la abundancia de los elementos fue seguida de una perfecta conservación de los mismos. Ahora bien, para intentar entender qué fue lo que ocurrió, es necesario revisar las características de las secuencias REP y de sus unidades en cada uno de los organismos.

En ambas *Salmonella* hay una clara degeneración de los elementos REP con respecto a la secuencia consenso de *Escherichia coli K12*. La enorme mayoría de éstos se mantiene en el intervalo de 72%-70% de similitud (con las matrices previas respectivas). Se plantearon dos hipótesis para explicar dicho hallazgo: 1) los elementos

[128] Gilson,, E; Clément, J.M; Perrin, D; Hofnung, M. (1987) Palindromic units: a case os highly repetitive DNA sequences in

REP de estas bacterias presentan un deterioro en sus secuencias producto de la incapacidad de los mecanismos moleculares relacionados con las secuencias REP para conservarlas. 2) Los elementos REP han sufrido una adaptación en sus secuencias de acuerdo a las condiciones impuestas por el organismo, es decir, los elementos están conservados pero su secuencia consenso difiere de la secuencia consenso de *Escherichia coli K12*. Un detallado análisis de las secuencias REP encontradas en las dos *Salmonella* permitió averiguar que aunque había posiciones conservadas en la mayoría de ellas, éstas correspondían a las bases de la secuencia consenso de *Escherichia coli K12*, en el resto de los nucleótidos hubo enormes variaciones, y la frecuencia de las cuatro bases fue equivalente. Los resultados permiten decidirse por la primera de las hipótesis: los elementos en las *Salmonella* están realmente degenerados.

Otros datos corroboran lo anterior: el número de genes en regiones no extragénicas aumentó de un 10% en las *Escherichia coli* hasta un 30 % en las *Salmonella*. Asimismo, el número de unidades en regiones 5'-5' se incrementó considerablemente y las frecuencias de elementos con regiones variables "RV" de gran tamaño también se hicieron mayores. Estos datos señalan que hay evidentes fallas en los mecanismos moleculares que reparten y conservan a los elementos REP en estas bacterias.

Por el contrario, en la línea *Escherichia coli* la historia es muy distinta. *Escherichia coli K12* es la bacteria modelo en cuanto a secuencias REP se refiere ya que presentó una abundancia correlacionada perfectamente con la conservación de los elementos. Las dos *Escherichia coli* patógenas vieron desmejoradas ambas características.

---

bacteria. *Trends in Genetics*. 3(8): 225-230.

Es importante señalar que el 70% de los elementos en *Escherichia coli K12* y aproximadamente el 55% de éstos en las *Escherichia coli* patógenas, estuvieron altamente conservados y se mantuvieron extragénicos. Esto indica que tanto la secuencia como su localización han sido seleccionadas. Asimismo, un análisis minucioso de las secuencias permitió averiguar que los elementos que tuvieron un "RV" mayor a las cuatro bases, sufrían un aumento en la variación de sus secuencias, muestra de la importancia no sólo de la secuencia, sino también de la estructura secundaria que se pueda formar.

Una vez presentes las características de los elementos REP en los linajes de enterobacterias analizados, se puede intentar contestar la siguiente pregunta: ¿cuál fue el cambio que ocurrió en ambas ramas para que los elementos presentaran tan graves diferencias?

La aproximación al tema requiere examinar, una a una, las explicaciones históricas que se han dado. Para ir en orden cronológico es posible comenzar con las primeras dos funciones asignadas a los elementos REP: terminación de la transcripción y estabilización del mensajero.

La presencia de los elementos REP preferentemente en regiones tres-prima de algún gene sería una evidencia que apoyaría ambas posibilidades. Es indudable, de acuerdo a numerosos trabajos publicados[129], que las secuencias REP fungen como

[129] Newbury, S; Smith, N; Robinson, C; Hiles, I; Higgins, C. (1987) Stabilization of translationally active mRNA by procariotic REP sequences. *Cell*. 48: 297-310.

Merino, E; Becerril, B; Valle, F; Bolivar, F.(1987) Deletion of a repetitive extragenic palindromic (REP) sequence dowstream from the structural gene of *Escherichia coli* glutamate dehydrogenase affects the stability of its mRNA.*Gene*.58(2-3):305-309.

estabilizadores del mensajero. Sin embargo, y como se había señalado anteriormente[130], cualquier secuencia que pueda formar una estructura de tallo y asa al final del mensajero invariablemente lo estabiliza. Algunos análisis realizados han demostrado que no hay ninguna homología existente entre las distintas secuencias que estabilizan mensajeros. Entonces ¿por qué se conserva tanto la secuencia y la estructura de las secuencias REP? Esto claramente indica que hay algún otro mecanismo actuando en ese sentido.

Con respecto a los procesos de terminación, ya se había visto una falta de correlación entre la presencia de elementos REP y el producto génico[131]. Este trabajo permitió ahondar más en dicho aspecto. La falta de una conservación en la asociación entre los elementos REP y los genes aledaños en *Escherichia coli* y *Salmonella* permiten decir que, aunque no es descartable la posibilidad de que algunos de los elementos REP sirvan como terminadores de la transcripción, éste nuevamente no es el mecanismo que los conserva, de lo contrario la secuencia variaría y la relación gene-elemento REP se conservaría en los distintos linajes de bacterias entéricas.

En *Salmonella*, más específicamente, el asunto es aún más claro. En esta bacteria, el mal estado de sus elementos y la falta de conservación en cualquiera de sus palíndromos permiten suponer que en muchos casos la formación de una estructura

[130] Merino, E; Bolivar, F. (1989) The ribonucleoside diphosphate reductase gene (nrdA) of *Escherichia coli* carries a repetitive extragenetic palindromic (REP) sequence in its 3'structural terminus. *Molecular Microbiology*. 3(6): 839-841.

Yang, Y; Ames, G. (1988) DNA gyrase binds to a family of prokariotic repetitive palindromic sequences. *Proceedings of the National Academy of Sciences*. 85(23):8850-8854.

[131] Gilson, E; Rousset, JO; Clement, JM; Hofnung, M.(1986) A subfamily of *E.coli* plindromic units implicated in transcription termination?. *Annales d'Institut Pasteur Microbiology*. 137B(3):259-270.

Gilson, E; Bachellier, S; Perrin, S; Perrin, D; Grimont, P; Grimont, F; Hofnung, M. (1990) Palindromic units highly repetitive DNA sequences exhibit species specificity within Enterobacteriaceae in bacteria. *Research in Microbiology*. 141(9):1103-16.

Gilson,, E; Clément, J.M; Perrin, D; Hofnung, M. (1987) Palindromic units: a case os highly repetitive DNA sequences in bacteria. *Trends in Genetics*. 3(8): 225-230.

Goberdhan, D; Kenneth, R; Morgan, M; Bayat, H; Ames, G. (1992) Physical mapping of repetitive extragenic palindromic sequences in *Escherichia coli* and phylogenetic distribution among *Escherichia coli* strains and other enteric bactera. *Journal of Bacteriology*. 174(14):4583-4593.

secundaria estable no es posible y por lo tanto ambas funciones asignadas carecen de sentido.

El análisis de las funciones de los genes que tienen una unidad REP asociada es muy claro: no hay una asociación entre el elemento REP y el gene aledaño; entre más unidades por organismo, más variadas son las actividades de los genes asociados.

El movimiento evidente que ha habido en las unidades REP (inserciones, inversiones, deleciones y transposiciones, encontradas en este trabajo), detectado a partir de las comparaciones entre unidades de *Escherichia coli* y *Salmonella*, así como la presencia de unidades nuevas y exclusivas en genes únicos o en genes homólogos, son señales de que la importancia de los elementos REP no está relacionada con el tipo de genes asociados.

En las *Escherichia coli* los datos parecieran contradictorios con lo anteriormente expuesto, ya que la mayoría de las unidades conservan su asociación con el mismo gene homólogo en las tres bacterias, además de mantener en muchos casos la orientación y el número de éstos. Varios datos permiten vislumbrar que estas asociaciones no se deben a la importancia de la relación gene-unidad REP, sino al poco tiempo de divergencia. Desde la existencia del último ancestro en común de las tres bacterias, aproximadamente un 20% de las unidades ya han cambiado su asociación con el gene, ya sea por separaciones, deleciones o apariciones en nuevas zonas. Es un porcentaje bastante elevado si se toma en cuenta el poco tiempo que han estado separadas. Las asociaciones se conservan más entre las patógenas que con *Escherichia coli K12*, señal de que hay un cambio en las relaciones gene-unidad a medida que el tiempo pasa.

La predilección del palíndromo interno sobre cualquier otro arreglo de los elementos en las *Escherichia coli*, muestra que la selección actúa al nivel de elementos solitarios. Lo anterior, junto con la degradación de los elementos en *Salmonella* y la indistinta predilección por cualquiera de los dos tipos de elementos (→ y ←) visto en las abundancias relativas similares de ambos en todas las enterobacterias, explican por qué han variado tanto las orientaciones y el número de elementos en las unidades entre *Escherichia coli* y *Salmonella* (sólo dos unidades, 0.5%-0.8% de las unidades totales, se conservan idénticas en las cinco; ambas se componen de elementos solitarios). Esta tendencia a las modificaciones en los arreglos de las unidades ha producido que el porcentaje de unidades que ya han visto cambiado el número o la orientación de sus elementos es del 35% en *Escherichia coli*, señal de que los procesos de cambio en las asociaciones gene-unidad han empezado.

Lo anterior marca una fuerte contradicción con arreglos de elementos REP conocidos como BIME[132]. En ese planteamiento las unidades que deberían de estar siendo seleccionadas serían aquellas con dos o más elementos. Los datos expuestos claramente indican que esto no es posible y que son los elementos solitarios los que juegan el papel más importante.

Cabe destacar en esta sección a las unidades encontradas en regiones cinco-prima-cinco-prima, ya que su presencia en estas regiones es un dato novedoso. La

[132] Bachellier, S; Perrin, D; Hofnung, M; Gilson, E. (1993). Bacterial interspersed mosaic elements (BIMEs) are present in the genome of *Klebsiella*. *Molecular Microbiology*. 7(4):537-544.

Bachellier, S; Saurin,W; Perrin, D; Hofnung, M; Gilon, E. (1994) Structural and functional diversity among bacterial interspersed mosaic elements(BIMEs). *Molecular Microbiology*. 12(1):61-70.

Gilson, E; Saurin, W; Perrin, D; Bachellier, S; Hofnung, M. (1991) Palindromic units are part of a new bacterial interpersed mosaic element (BIME). *Nucleic Acids Research*. 19(7):1375-1383.

Gilson, E; Saurin, W; Perrin, D; Bachellier, S; Hofnung, M. (1991) The BIME family of bacterial highly repetitive sequences. *Research Microbiology*.137B (2-3):217-222.

estrecha relación entre las secuencias REP, la terminación de la transcripción[133], la estabilización del mRNA[134] y su relación bajo estas características con sus genes aledaños, resulta imposible, ya que su localización en las zonas cinco-prima-cinco-prima les impide realizar todas esas actividades.

La explicación más lógica a dicho hallazgo, dada la reducida cantidad de elementos encontrados, es que la maquinaria de generación de elementos REP esté, de algún modo, reconociendo las regiones tres prima extragénica de los genes e insertando elementos REP exclusivamente en dichas zonas y que por algún fenómeno independiente a los elementos REP (rearreglos cromosomales) éstos aparecieron en regiones cinco-prima. Esto sucede raramente en *Escherichia coli*, pero cuando ocurre, las unidades pueden ser conservadas.

En las *Salmonella* el aumento en el número de unidades REP en secciones cinco-prima de ambos genes puede entenderse como fallas en la maquinaria de reconocimiento de una región tres-prima extragénica o por la agudización de los movimientos de material genético. Los datos apuntan más hacia la segunda explicación ya que fueron notorias, durante la investigación, las repetidas inserciones de DNA, la separación de unidades y la inversión de fragmentos de material genético en estos organismos.

[133] Gilson, E; Rousset, JO; Clement, JM; Hofnung, M.(1986) A subfamily of *E.coli* plindromic units implicated in transcription termination?. *Annales d'Institut Pasteur Microbiology*. 137B(3):259-270.

Gilson, E; Saurin, W; Perrin, D; Bachellier, S; Hofnung, M. (1991) The BIME family of bacterial highly repetitive sequences. *Research Microbiology*. 137B (2-3):217-222.

[134] Gilson, E; Saurin, W; Perrin, D; Bachellier, S; Hofnung, M. (1991) The BIME family of bacterial highly repetitive sequences. *Research Microbiology*. 137B (2-3):217-222.

Newbury, S; Smith, N; Robinson, C; Hiles, I; Higgins, C. (1987) Stabilization of translationally active mRNA by procariotic REP sequences. *Cell*. 48: 297-310.

Merino, E; Becerril, B; Valle, F; Bolivar, F.(1987).Deletion of a repetitive extragenic palindromic (REP) sequence dowstream from the structural gene of *Escherichia coli* glutamate dehydrogenase affects the stability of its mRNA.*Gene*.58(2-3):305-309.

El origen de los elementos, como se demostró[135], es por eventos de duplicación. Aparentemente, estas duplicaciones están mediadas por la acción de la DNA polimerasa I que reconoce significativamente a las secuencias REP[136] y que además puede tener actividad de reversa transcriptasa[137]. Un modelo de aparición sería: la DNA polimerasa I reconoce a la secuencias REP, genera una copia del elemento y produce un híbrido DNA-RNA. Finalmente, la enzima transforma la secuencia de RNA a DNA y la integra al cromosoma en las cercanías del elemento original. Posteriormente, los mecanismos relacionados con la transposición, como podría ser IS1397[138], los podrían repartir por todo el genoma.

El número de elementos en las cinco bacterias es muy elevado, en especial en las *Salmonella*, el hecho de que el 90% de las unidades que no comparten con *Escherichia coli* y el 50% de las unidades que no son análogas entre ellas se deben a apariciones nuevas de elementos o a movimientos de unidades existentes previamente. Todo esto quiere decir que la aparición de los elementos REP y su diseminación por todo el genoma hacia zonas extragénicas tres-prima de algún gene, pueden darse sin la conservación estricta de los elementos. De otro modo no habría posibilidad de que aparecieran y/o se movieran en las *Salmonella*. Finalmente, se puede decir que los procesos de aparición y diseminación de los elementos no ejercen presión selectiva sobre el mantenimiento de la secuencia y/o la estructura de los elementos REP.

---

[135] Merino, E; Bolivar, F. (1989) The ribonucleoside diphosphate reductase gene (nrdA) of *Escherichia coli* carries a repetitive extragenetic palindromic (REP) sequence in its 3'structural terminus. *Molecular Microbiology*. 3(6): 839-841.

Shyamala, V; Schneider, E; Ames, G.F. (1990) Tandem chromosomal duplications: role os REP sequences in the recombination event at the join-point. *EMBO Journal*. 9:939-946.

[136] Gilson, E; Perrin, D; Hofnung, M. (1990) DNA polymerase I and a protein complex bind specifically to *Escherichia coli* palindromic unit highly repetitive DNA: implications for bacterial chromosome organization. *Nucleic Acids Research*. 18(13):3941-3952.

[137] Riechetti, M; Buc, H.(1993) *E.coli* DNA polymerase I as a reverse transcriptase. *EMBO Journal*. 12(2):387:396.

Del mismo modo, la extragenicidad de los elementos se entiende como un fenómeno relacionado exclusivamente con los mecanismos que generan y reparten los elementos en los genomas. De otra forma, no podríamos observar ninguno de los dos procesos en *Salmonella*.

Si las diferencias observadas entre *Escherichia coli* y *Salmonella* no tienen su origen en la aparición y dispersión de los elementos, ni en su asociación con el gene contiguo, ni con la terminación de la transcripción, ni con la estabilización del mensajero, entonces deben de existir otros mecanismos protéicos involucrados en su selección. Mecanismos que dadas las supuestas condiciones en el ancestro, se perdieron en las *Salmonella*.

La relación estrecha entre las secuencias IHF y las secuencias REP había sido señalada anteriormente[139] en *Escherichia coli K12*, ahora está perfectamente confirmada y se conocen con exactitud todas ellas en las cinco enterobacterias.

Si bien es importante señalar que la relación entre secuencias IHF-REP sólo existe en las *Escherichia coli*, lo cual de algún modo marcaría a estas relaciones como buenas candidatas para explicar lo ocurrido entre ambos linajes, sólo se encuentran en una proporción muy baja. Ni todas las secuencias REP tienen una secuencia IHF asociada, ni todas las secuencias IHF tienen una secuencia REP adyacente. No obstante,

---

[138] Clément, J.M; Wilde, C; Bachellier, S; Lambert, P; Hofnung, M. (1999) IS1297 is active for transposition into yhe chromosome of *Escherichia coli K-12* and inserts specifically into palindromic units of bacterial interspersed mosaic elements. *Journal of Bacteriology*. 181(22):6929-6936.

[139] Boccard, F; Prentki, P. (1993) Specific interaction of IHF with RIBs, a class of bacterial repetitive DNA elements located at the 3'end of transcription units. *EMBO Journal*. 12(13):5019-5027.

Clarkson, S; Bates, AD. (1996) Action of DNA gyrase at RIP elements in *E.coli*. *Biochemical Society Transactions*. 24(3):420.

Oppenheim, AB; Rudd, KE; Mendelson, I; Teff, D. (1993) Integration host factor binds to a unique class of complex repetitive extragenic DNA sequences in *Escherichia coli*. *Molecular Microbiology*. 10(1):113-122.

la relación entre IHF y REP en regiones exclusivamente extragénicas abarca a casi todas las secuencias IHF de estas áreas y está perfectamente corroborado, a partir de este trabajo, que estas asociaciones no son producto del azar, además que éstas siempre son las mismas, es decir, la secuencia IHF siempre está al final del elemento REP. Todo esto significa que, efectivamente, en estas bacterias las relaciones existen y son muy estrechas, pero no son suficientes para explicar todas las características de los elementos REP en las enterobacterias. Cabe señalar que, la gran mayoría de estas asociaciones exitosas se dieron con elementos REP solitarios, lo que nuevamente muestra como son éstos los verdaderamente importantes.

Se han identificado también las asociaciones con las proteínas HU y las girasas[140] en *Escherichia coli*. Las proteínas HU, que presentan una gran homología en su secuencia de aminoácidos con las dos subunidades de la proteína IHF[141], se unen a casi cualquier secuencia que pueda formar una estructura crucifrome en el DNA[142], la unión es precisamente para evitar esta distorsión. Lo anterior indica que la unión de estas proteínas tampoco actúa como presión de selección sobre los elementos REP. Lo que bien podría ser cierto es que, dada la gran cantidad de elementos REP en el genoma de las *Escherichia coli*, éstos podrían ser los que, de manera más significativa, estuvieran interaccionando con las proteínas HU. También podría ser cierto que las

---

[140] Clarkson, S; Bates, AD. (1996) Action of DNA gyrase at RIP elements in *E.coli*. *Biochemical Society Transactions*. 24(3):420.

Yang, Y; Ames, G. (1988) DNA gyrase binds to a family of prokariotic repetitive palindromic sequences. *Proceedings of the National Academy of Sciences*. 85(23):8850-8854.

[141] Freundlich, M; Ramani, N; Mathew, E; Sirko, A; Tsui, P. (1992) The role of integration host factor in gene expression in *Escherichia coli*. *Molecular Microbiology*. 6(18):2557-2563.

[142] Pentiggia, A; Negri, A; Beltrama, M; Blanchi, M.(1993) Protein HU binds specifically to kinked DNA. *Molecular Microbiology*. 7:343-350.

proteínas HU estén colaborando en la asociación de los elementos REP y otras proteínas, como podrían ser IHF y las girasas[143].

Se ha planteado que la organización y compactación del cromosoma bacteriano de estos organismos se debe precisamente a la unión de las girasas a las secuencias REP[144]. Aunque no se ha probado que esta unión realmente se lleve a cabo durante el superenrollamiento, sí se ha visto que la unión de las girasas es significativa sobre los elementos REP. Ahora bien, con los datos encontrados en esta investigación surgen nuevamente incongruencias en la explicación. Suponiendo que realmente la unión de las girasas a los elementos REP fuera el mecanismo de compactación del DNA, entonces:

1) Debería haber un número proporcional entre el tamaño del genoma y el número de elementos REP. Los datos encontrados señalan fallas en este aspecto, ya que las *Escherichia coli* patógenas presentan en su genoma un millón de bases extras que la cepa *Escherichia coli K12*, pero en cambio su número de unidades es menor.
2) Del mismo modo, debería de haber un cierto control en la aparición de la unidades y en el mantenimiento de su posición relativa en el genoma, ya que de lo contrario, la compactación del cromosoma cambiaría de conformación a cada momento, lo cual no tiene sentido biológico. Los datos muestran que hay una gran cantidad de

[143] Oppenheim, AB; Rudd, KE; Mendelson, I; Teff, D. (1993) Integration host factor binds to a unique class of complex repetitive extragenic DNA sequnces in *Escherichia coli*. *Molecular Microbiology*. 10(1):113-122.

[144] Clarkson, S; Bates, AD. (1996) Action of DNA gyrase at RIP elements in *E.coli*. *Biochemical Society Transactions*. 24(3):420.

Gilson,, E; Clément, J.M; Perrin, D; Hofnung, M. (1987) Palindromic units: a case os highly repetitive DNA sequences in bacteria. *Trends in Genetics*. 3(8): 225-230.

Yang, Y; Ames, G. (1988) DNA gyrase binds to a family of prokariotic repetitive palindromic sequences. *Proceedings of the National Academy of Sciences*. 85(23):8850-8854.

elementos que han cambiado de lugar o han aparecido en estas bacterias, lo que contradice el supuesto.

3) Si la relación entre las proteínas girasas y los elementos REP es trascendental para la bacteria y un fallo en ésta repercutiría en la adecuación del organismo, se debería encontrar que las secuencias REP se conservan notoriamente. Lo anterior sólo es aplicable a *Escherichia coli K12*. Las dos cepas patógenas mostraron un deterioro sensible en la calidad de sus elementos. En las *Salmonella* el problema es mayor. En estas bacterias no hay conservación de los elementos, ni en secuencia, ni en posición, ni en abundancia, ni en extragenicidad.

Todo lo anterior indicaría que si la relación existe, sólo está presente en *Escherichia coli K12*. En las dos *Escherichia coli* patógenas y en las *Salmonella* es menos eficiente (la deficiencia en las dos *Escherichia coli* podría deberse a los cambios en el DNA que sufren estos organismos por su carácter patógeno) o simplemente no existe. Sin embargo estas bacterias no han visto decaída su adecuación, pese a la supuesta importancia de la relación. Dicho de otra forma: ¿realmente es tan importante esta relación en *Escherichai coli K12*, como para que se seleccionen, de la forma con la que se hace, a los elementos REP, pese a que en sus cuatro parientes filogenéticamente más cercanos esta relación ha decaído sin que ellos se vieran afectados? La respuesta más lógica señalaría que en realidad, pese a que la relación existe en *Escherichia coli K12*, tampoco es una explicación suficiente.

Finalmente, se ha planteado que los elementos REP son secuencias de DNA egoístas[145], es decir, secuencias que utilizan los mecanismos moleculares de la célula para duplicarse, repartirse y conservarse, sin tener la menor incidencia en la adecuación de los organismos. Los datos del trabajo señalan aspectos importantes a este respecto. Claramente, el caso de las *Salmonella* muestra a una serie de elementos mal conservados, muchos de ellos lo suficientemente degradados como para no poder formar estructuras secundarias, necesarias para la enorme mayoría de las funciones anteriormente expuestas, y que, sin embargo, aparecen y se reparten por todo el genoma sin mayor problema. Lo anterior nos permite decir que la mala conservación de los elementos en las *Salmonella* es debido a la falla o la falta de los mecanismos de selección. Atributos definitorios de una secuencia egoísta.

En las *Escherichia coli* las cosas son distintas, ya que los elementos están muy bien conservados. Si fueran secuencias egoístas independientes, tendríamos un escenario como el de las *Salmonellas*. La ausencia de éste, quiere decir que en estos organismos hay mecanismos de selección producto de la importancia de los elementos REP en la adecuación.

Tomando en cuenta todo el conjunto de datos analizados hasta el momento, se puede decir que el marco más probable para entender las características de las secuencias REP en *Escherichia coli* y sus diferencias con *Salmonella*, es que los elementos REP continúen siendo secuencias egoístas en *Escherichia coli*, pero a diferencia de sus parientes, han mantenido la interacción con otros mecanismos moleculares (secuencias IHF o proteínas girasas y HU), los cuales indirecta o

[145] Gilson., E; Clément, J.M; Perrin, D; Hofnung, M. (1987) Palindromic units: a case os highly repetitive DNA sequences in bacteria. *Trends in* Genetics. 3(8): 225-230.

directamente han actuado sobre los elementos REP por separado y los han generado, repartido y conservado, tanto en un nivel de secuencia como de estructura.

A este respecto es importante mencionar los resultados obtenidos en el análisis de secuencias tipo-REP en los demás organismos. Pese a que ninguna de las secuencias encontradas cumplió con todas las características impuestas, indudablemente se encontraron organismos con presencia de este tipo de secuencias. El análisis filogenético de éstos mostró un falta de relación; los tres grupos con las mejores características no pudieron ser más distintos: enterobacterias dentro de las proteobacterias, streptococcaceas (*Streptococcus pneumoniae TIGR4*) dentro de los frimicutes y sulfolobales (*Sulfolobus solfataricus*) dentro de las archea. Aunque se encontraron grupos compactos donde la presencia o ausencia de este tipo de secuencia fue clara, la generalidad de la distribución de los organismos con secuencias tipo-REP en el árbol filogenético, es dispersa.

El análisis de homología de los genes que presentaron una secuencia tipo-REP asociada, fue como se esperaba dados los datos encontrados en las enterobacterias: no hubo relación homóloga alguna entre los distintos genes. Esto confirma también el carácter de secuencias tipo-REP en estos organismos.

Ahora bien, el desorden filogenético, la falta de homología y de relación con sus genes aledaños, la conservación de las características necesaria para ser consideradas como secuencias tipo-REP y lo datos provenientes de los elementos REP de las enterobacterias, permiten contemplar un panorama más amplio: los elementos REP y tipo-REP podrían ser secuencias de DNA egoístas que han aparecido y/o desaparecido innumerables veces de forma independiente en distintos puntos de la evolución de los organismos. En muchos casos la falta de asociaciones con los mecanismos moleculares

de los organismos conllevaron a que estos elementos no abundaran, se conservaran mal o desaparecieran. En muchos otros, la aparición de estas asociaciones permitió su selección y su permanencia íntegra en el genoma de los organismos.

Para terminar, es posible plantear una serie de trabajos que serían necesarios para ahondar más en el entendimiento del tema desarrollado en esta investigación:

1) Revisar mayor número de enterobacterias para poder ampliar, corroborar o cambiar las explicaciones expuestas sobre los elementos REP y sus características.
2) Comprobar si las girasas están unidas a los elementos REP durante el superenrollamiento del cromosoma bacteriano y analizar como se desarrolla este fenómeno en las otras enterobacterias, sobre todo en las *Salmonella*, donde los elementos están muy degradados.
3) Analizar qué tipo de reconocimiento de proteínas se pudieran estar dando en todas las secuencias tipo-REP propuestas.

## 6. BIBLIOGRAFÍA


1. Alm, R.A; Ling, L.S.L; Moir, D.T; King, B.L; Brown, E.D; Doig, P.C; Smith, D.R; Noonan, B; Guild, B.C; deJonge, B.L; Carmel, G; Tummino, P.J; Caruso, A; Uria-Nickelsen, M; Mills, D.M; Ives, C; Gibson, R; Merberg, D; Mills, S.D; Jiang, Q; Taylor, D.E; Vovis, G.F; Trust, T.J. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*. 397:176-180.

2. Altschul, S.F; Thomas, L; Madden, A.A; Schäffer, J.Z; Zheng, Z; Webb, M; Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 25:3389-3402.

3. Andersson, S.G; Zomorodipour, A; Andersson, J.O; Sicheritz-Ponten, T; Alsmark, U.C; Podowski, R.M; Naslund, A.K; Eriksson, A.S; Winkler, H.H; Kurland, C.G. (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*. 396: 133-140.

4. Bachellier, S; Perrin, D; Hofnung, M; Gilson, E. (1993). Bacterial interspersed mosaic elements (BIMEs) are present in the genome of *Klebsiella*. *Molecular Microbiology*. 7(4):537-544.

5. Bachellier, S; Saurin, W; Perrin, D; Hofnung, M; Gilon, E. (1994) Structural and functional diversity among bacterial interspersed mosaic elements(BIMEs). *Molecular Microbiology*. 12(1):61-70.

6. Bailey, L T; Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*. 14: 48-54.

7. Blattner, F.R; Plunkett, G. III, Bloch, C.A; Perna, N.T; Burland, V; Riley, M; Collado-Vides, J; Glasner, J.D; Rode, C.K; Mayhew, G.F; Gregor, J; Davis, N.W; Kirkpatrick, H.A; Goeden, M.A; Rose, D.J; Mau, B; Shao, Y. (1997) The complete genome sequence of *Escherichia coli K-12*. *Science*. 277:1453-1474.

8. Boccard, F; Prentki, P. (1993) Specific interaction of IHF with RIBs, a class of bacterial repetitive DNA elements located at the 3'end of transcription units. *EMBO Journal*. 12(13):5019-5027.

9. Bolotin, A; Wincker, P; Mauger, S; Jaillon, O; Malarme, K; Weissenbach, J; Ehrlich, S.D; Sorokin, A. (2001) The complete genome sequence of the lactic acid bacterium *Lactococcus lactis ssp. lactis IL1403*. *Genome Research*. 11:731-753.

10. Bult, C.J; White, O; Olsen, G.J; Zhou, L; Fleischmann, R.D; Sutton, G.G; Blake, J.A; FitzGerald, L.M; Clayton, R.A; Gocayne, J.D; Kerlavage, A.R; Dougherty, B.A; Tomb, J.-F; Adams, M.D; Reich, C.I; Overbeek, R; Kirkness, E.F; Weinstock, K.G; Merrick, J.M; Glodek, A; Scott, J.D; Geoghagen, N.S; Weidman, J.F; Fuhrmann, J.L; Nguyen, D.T; Utterback, T; Kelley, J.M; Peterson, J.D; Sadow, P.W; Hanna, M.C; Cotton, M.D; Hurst, M.A; Roberts, K.M; Kaine, B.B; Borodovsky, M; Klenk, H.P; Fraser, C.M; Smith, H.O; Woese, C.R; Venter, J.C.

(1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science.* 273 (5278): 1058-1073.

11. Clarkson, S; Bates, AD. (1996) Action of DNA gyrase at RIP elements in *E.coli. Biochemical Society Transactions.* 24(3):420.

12. Clément, J.M; Wilde, C; Bachellier, S; Lambert, P; Hofnung, M. (1999) IS1297 is active for transposition into yhe chromosome of *Escherichia coli K-12* and inserts specifically into palindromic units of bacterial interspersed mosaic elements. *Journal of Bacteriology.* 181(22):6929-6936.

13. Cole, S.T; Brosch, R; Parkhill, J; Garnier, T; Churcher, C; Harris, D; Gordon, S.V; Eiglmeier, K; Gas, S; Barry III, C.E; Tekaia, F; Badcock, K; Basham, D; Brown, D; Chillingworth, T; Connor, R; Davies, R; Devlin, K; Feltwell, T; Gentles, S; Hamlin, N; Holroyd, S; Hornsby, T; Jagels, K; Krogh, A; McLean, J; Moule, S; Murphy, L; Oliver, S; Osborne, J; Quail, M.A; Rajandream, M.A; Rogers, J; Rutter, S; Seeger, K; Skelton, S; Squares, S; Sqares, R; Sulston, J.E; Taylor, K; Whitehead, S; Barrell, B.G. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature.* 393:537.

14. Cole, S.T; Eiglmeier, K; Parkhill, J; James, K.D; Thomson, N.R; Wheeler, P.R; Honore, N; Ganier, T; Churcher, C; Harris, D; Mungall, K; Basham, D; Brown, D; Chillingworth, T; Connor, R; Davies, R.M; Devlin, K; Duthoy, S; Feltwell, T; Fraser, A; Hamlin, N; Holroyd, S; Hornsby, T; Jagels, K; Lacroix, C; Maclean, J; Moule, S; Murphy, L; Oliver, Quail, M.A; Rajandream, M.-A; Rutherford, K.M; Rutter, S; Seeger, K; Simon, S; Simmonds, M; Skelton, J; Squares, R; Squares, S; Stevens, K; Taylor, K; Whitehead, S; Woodward, J.R; Barrell, B.G. (2001) Massive gene decay in the *leprosy bacillus. Nature.* 409:1007-1011.

15. Chambaud, I; Heilig, R; Ferris, S; Barbe, V; Samson, D; Galisson, F; Moszer, I; Dybvig, K; Wroblewski, H; Viari, A; Rocha, E.P.C; Blanchard, A. (2001) *The complete genome sequence of the murine respiratory pathogen Mycoplasma pulmonis. Nucleic Acids Research* 29:2145-2153.

16. de la Cruz, F; Davies, J. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends in Microbiology.* 8(3):128-129.

17. Deckert, G; Warren, P.V; Gaasterland, T; Young, W.G; Lenox, A.L; Graham, D.E; Overbeek, R; Snead, M.A; Keller, M; Aujay, M; Huber, R; Feldman, R.A; Short, J.M; Olson, G.J; Swanson, R.V. (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus. Nature.* 392:353.

18. Diekmann, S; Lilley, D. (1985) The anomalous gel migration of a stable cruciform: temperature and salt dependence. And some comparitions with curved DNA. *Nucleic Acids Research.*15:5765-5774.

19. Ferretti, J.J; McShan, W.M; Adjic, D; Savic, D; Savic, G; Lyon, K; Primeaux, C; Sezate, S.S; Surorov, A.N; Kenton, S; Lai, H; Lin, S; Qian, Y; Jia, H.G; Najar, F.Z; Ren, Q; Zhu, H; Song, L; White, J; Yuan, X; Clifton, S.W; Roe, B.A; McLaughlin,

R.E. (2001) Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proceedings of the National Academy of Sciences* 98: 4658-63.

20. Fleischmann, R.D; Adams, M.D; White, O; Clayton, R.A; Kirkness, E.F; Kerlavage, A.R; Bult, C.J; Tomb, J.-F; Dougherty, B.A; Merrick, J.M; McKenney, K; Sutton, G.G; FitzHugh, W; Fields, C.A; Gocayne, J.D; Scott, J.D; Shirley, R; Liu, L.I; Glodek, A; Kelley, J.M; Weidman, J.F; Phillips, C.A; Spriggs, T; Hedblom, E; Cotton, M.D; Utterback, T; Hanna, M.C; Nguyen, D.T; Saudek, D.M; Brandon, R.C; Fine, L.D; Fritchman, J.L; Fuhrmann, J.L; Geoghagen, N.S; Gnehm, C.L; McDonald, L.A; Small, K.V; Fraser, C.M; Smith, H.O; Venter, J.C. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. *Science*. 269:496-512.

21. Fleischmann, R.D; Alland, D; Eisen, J.A; Carpenter, L; White, O; Peterson, J; DeBoy, R; Dodson, R; Gwinn, M; Haft, D; Hickey, E; Kolonay, J.F; Nelson, W.C; Umayam, L.A; Ermolaeva, M; Salzberg, S.L; Delcher, A; Utterback, T; Weidman, J; Khouri, H; Gill, J; Mikula, A; Bishai, W. Whole genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. Aún sin publicar.

22. Fraser, C.M; Casjens, S; Huang, W.M; Sutton, G.G; Clayton, R.A; Lathigra, R; White, O; Ketchum, K.A; Dodson, R; Hickey, E.K; Gwinn, M; Dougherty, B; Tomb, J.-F; Fleischmann, R.D; Richardson, D; Peterson, J; Kerlavage, A.R; Quackenbush, J; Salzberg, S; Hanson, M; van-Vugt, R; Palmer, N; Adams, M.D; Gocayne, J.D; Weidman, J; Utterback, T; Watthey, L; McDonald, L; Artiach, P; Bowman, C; Garland, S; Fujii, C; Cotton, M.D; Horst, K; Roberts, K; Hatch, B; Smith, H.O; Venter, J.C. (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature*. 390: 580-586.

23. Fraser, C.M; Gocayne, J.D; White, O; Adams, M.D; Clayton, R.A; Fleischmann, R.D; Bult, C.J; Kerlavage, A.R; Sutton, G.G; Kelley, J.M; Fritchman, J.L; Weidman, J.F; Small, K.V; Sandusky, M; Fuhrmann, J.L; Nguyen, D.T; Utterback, T; Saudek, D.M; Phillips, C.A; Merrick, J.M; Tomb, J; Dougherty, B.A; Bott, K.F; Hu, P.C; Lucier, T.S; Peterson, S.N; Smith, H.O; Venter, J.C. (1995) The minimal gene complement of Mycoplasma genitalium. *Science*. 270:397-403.

24. Fraser, C.M; Norris, S.J; Weinstock, G.M; White, O; Sutton, G.G; Dodson, R; Gwinn, M; Hickey, E.K; Clayton, R; Ketchum, K.A; Sodergren, E; Hardham, J.M; McLeod, M.P; Salzberg, S; Peterson, J; Khalak, H; Richardson, D; Howell, J.K; Chidambaram, M; Utterback, T; McDonald, L; Artiach, P; Bowman, C; Cotton, M.D; Fujii, C; Garland, S; Hatch, B; Horst, K; Roberts, K; Watthey, L; Weidman, J; Smith, H.O; Venter, J.C. (1998) Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science*. 281: 375-388.

25. Freundlich, M; Ramani, N; Mathew, E; Sirko, A; Tsui, P. (1992) The role of integration host factor in gene expression in Escherichia coli. *Molecular Microbiology*. 6(18):2557-2563.

26. Galibert, F; Finan, T.M; Long, S.R; Puhler, A; Abola, P; Ampe, F; Barloy-Hubler, F; Barnett, M.J; Becker, A; Boistard, P; Bothe, G; Boutry, M; Bowser, L; Buhrmester, J; Cadieu, E; Capela, D; Chain, P; Cowie, A; Davis, R.W; Dreano, S;

Federspiel, N.A; Fisher, R.F; Gloux, S; Godrie, T; Goffeau, A; Golding, B; Gouzy, J; Gurjal, M; Hernandez-Lucas, I; Hong, A; Huizar, L; Hyman, R.W; Jones, T; Kahn, D; Kahn, M.L; Kalman, S; Keating, D.H; Kiss, E; Komp, C; Lelaure, V; Masuy, D; Palm, C; Peck, M.C; Pohl, T.M; Portetelle, D; Purnelle, B; Ramsperger, U; Surzycki, R; Thebault, P; Vandenbol, M; Vorholter, F.J; Weidner, S; Wells, D.H; Wong, K; Yeh, K.C; Batut, J. (2001) The composite genome of the legume symbiont *Sinorhizobium meliloti. Science.* 293: 668-572.

27. Gelfand, M; Koonin, E. (1997) Avoidance of palindromic words in bacterial and archeal genomes: a close connection with restriction enzymes. *Nucleic Acids Research.* 25(12): 2430-2439.

28. Gilson, E; Bachellier, S; Perrin, S; Perrin, D; Grimont, P; Grimont, F; Hofnung, M. (1990) Palindromic units highly repetitive DNA sequences exhibit species specificity within Enterobacteriaceae in bacteria. *Research in Microbiology.* 141(9):1103-16.

29. Gilson, E; Clément, J-M; Brutlag, D; Hofnung, M. (1984) A family of dispersed repetitive extragenic palindromic DNA sequences in *E. coli. EMBO Journal.* 3(6): 1417-1421.

30. Gilson, E; Bachellier, S; Perrin, S; Perrin, D; Grimont, P; Grimont, F; Hofnung, M. (1990) Palindromic units highly repetitive DNA sequences exhibit species specificity within Enterobacteriaceae in bacteria. *Research in Microbiology.* 141(9):1103-16.

31. Gilson, E; Perrin, D; Hofnung, M. (1990) DNA polymerase I and a protein complex bind specifically to *Escherichia coli* palindromic unit highly repetitive DNA: implications for bacterial chromosome organization. *Nucleic Acids Research.* 18(13):3941-3952.

32. Gilson, E; Perrin, D; Saurin, W; Hofnung, M.(1987) Species specificity of bacterial palindromic unit. *Journal of Molecular Evolution.* 25(4):371-373.

33. Gilson, E; Rousset, JO; Clement, JM; Hofnung, M.(1986) A subfamily of *E.coli* plindromic units implicated in transcription termination?. *Annales d'Institut Pasteur Microbiology.* 137B(3):259-270.

34. Gilson, E; Saurin, W; Perrin, D; Bacheller, S; Hofnung, M. (1991) Palindromic units are part of a new bacterial interpersed mosaic element (BIME). *Nucleic Acids Research.* 19(7):1375-1383.

35. Gilson, E; Saurin, W; Perrin, D; Bachellier, S; Hofnung, M. (1991) The BIME family of bacterial highly repetitive sequences. *Research in Microbiology.*137B (2-3):217-222.

36. Glass, J.I; Lefkowitz, E.J; Glass, J.S; Heiner, C.R; Chen, E.Y; Cassell, G.H. (2000) The complete sequence of the mucosal pathogen *Ureaplasma urealyticum. Nature.* 407: 757-762.

37. Goberdhan, D; Kenneth, R; Morgan, M; Bayat, H; Ames, G. (1992) Physical mapping of repetitive extragenic palindromic sequences in *Escherichia coli* and phylogenetic distribution among *Escherichia coli* strains and other enteric bactera. *Journal of Bacteriology*. 174(14):4583-4593.

38. Goodner, B; Hinkle, G; Gattung, S; Miller, N; Blanchard, M; Qurollo, B; Goldman, B.S; Cao, Y; Askenazi, M; Halling, C; Mullin, L; Houmiel, K; Gordon, J; Vaudin, M; Iartchouk, O; Epp, A; Liu, F; Wollam, C; Allinger, M; Doughty, D; Scott, C; Lappas, C; Markelz, B; Flanagan, C; Crowell, C; Gurson, J; Lomo, C; Sear, C; Strub, G; Cielo, C; Slater, S. (2001) Genome Sequence of the Plant Pathogen and Biotechnology Agent *Agrobacterium tumefaciens C58. Science*. 294 (5550), 2323-2328.

39. Gur-Arie, R; Cohen, C; Eitan, Y; Shelef, L; Hallerman, E; Kashi, Y. (2000) Simple sequence repeats in *Escherichia coli*: Abundance, distribution, composition, and polymorphism. *Genome Research*. 10:62-71.

40. Hayashi, T; Makino, K; Ohnishi, M; Kurokawa, K; Ishii, K; Yokoyama, K; Han, C.-G; 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; Shinagawa, H. (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli O157:H7* and genomic comparison with a laboratory strain *K-12*. *DNA Research*. 8:11-22.

41. Heidelberg, J.F; Eisen, J.A; Nelson, W.C; Clayton, R.A; Gwinn, M.L; Dodson, R.J; Haft, D.H; Hickey, E.K; Peterson, J.D; Umayam, L.A; Gill, S.R; Nelson, K.E; Read, T.D; Tettelin, H; Richardson, D; Ermolaeva, M.D; Vamathevan, J; Bass, S; Qin, H; Dragoi, I; Sellers, P; McDonald, L; Utterback, T; Fleishmann, R.D; Nierman, W.C; White, O; Salzberg, S.L; Smith, H.O; Colwell, R.R; Mekalanos, J.J; Venter, J.C; Fraser, C.M. (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae. Nature*. 405: 477-483.

42. Heilig, R. *Pyrococcus abyssi* genome sequence: insights into archaeal chromosome structure and evolution. Aún no publicado

43. Heringa, J. (1998) Detection of internal repeats: how common are they?. *Current Opinion in Structural Biology*. 8(3):338-45.

44. Higgins, C.F; Ames, G.F; Barnes, W; Clément, J-M; Hofnung, M. (1982) A novel intercistronic regulatory elment of procaryotic operons. *Nature*. 298:760-762.

45. Himmelreich, R; Hilbert, H; Plagens, H; Pirkl, E; Li, B.C; Herrmann, R. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae. Nucleic Acids Research* 24:4420-4449.

46. Kalman, S; Mitchell, W; Marathe, R; Lammel, C; Fan, J; Hyman, R.W; Olinger, L; Grimwood, J; Davis, R.W. Stephens, R.S. (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis. Nature* Genetics 21: 385-389.

47. Kaneko, T; Nakamura, Y; Sato, S; Asamizu, E; Kato, T; Sasamoto, S; Watanabe, A; Idesawa, K; Ishikawa, A; Kawashima, K; Kimura, T; Kishida, Y; Kiyokawa, C; Kohara, M; Matsumoto, M; Matsuno, A; Mochizuki, Y; Nakayama, S; Nakazaki, N; Shimpo, S; Sugimoto, M; Takeuchi, C; Yamada, M; Tabata, S. (2000) Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. *DNA Research*. 7:331-338.

48. Kaneko, T; Sato, S; Kotani, H; Tanaka, A; Asamizu, E; Nakamura, Y; Miyajima, N; Hirosawa, M; Sugiura, M; Sasamoto, S; Kimura, T; Hosouchi, T; Matsuno, A; Muraki, A; Nakazaki, N; Naruo, K; Okumura, S; Shimpo, S; Takeuchi, C; Wada, T; Watanabe, A; Yamada, M; Yasuda, M; Tabata, S. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis sp. strain PCC6803. II.* Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Research*. 3: 109-136.

49. Karlin, S; Burege, C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics*. 11(7):283-90. Review.

50. Kawarabayasi, Y; Hino, Y; Horikawa, H; Yamazaki, S; Haikawa, Y; Jin-no, K; Takahashi, M; Sekine, M; Baba, S; Ankai, A; Kosugi, H; Hosoyama, A; Fukui, S; Nagai, Y; Nishijima, K; Nakazawa, H; Takamiya, M; Masuda, S; Funahashi, T; Tanaka, T; Kudoh, Y; Yamazaki, J; Kushida, N; Oguchi, A; Kikuchi. H *et al.* (1999) Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Research*. 6: 83-101.

51. Kawarabayasi, Y; Sawada, M; Horikawa, H; Haikawa, Y; Hino, Y; Yamamoto, S; Sekine, M; Baba, S; Kosugi, H; Hosoyama, A; Nagai, Y; Sakai, M; Ogura, K; Otsuka, R; Nakazawa, H; Takamiya, M; Ohfuku, Y; Funahashi, T; Tanaka, T; Kudoh, Y; Yamazaki, J; Kushida, N; Oguchi, A; Aoki, K; Yoshizawa, T; Nakamura, Y; Robb, F.T; Horikoshi, K; Masuchi, Y; Shizuya, H; Kikuchi, H. (1998) Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii OT3* (supplement). *DNA Research*. 5: 55-76.

52. Kawashima, T; Yamamoto, Y; Aramaki, H; Nunoshiba, T; Kawamoto, T; Watanabe, K; Yamazaki, M; Kanehori, K; Amano, N; Ohya, Y; Makino, K; Suzuki, M. (2000) Determination of the complete genomic DNA sequence of *Thermoplasma volcanium GSS1*. *Proceedings of the National Academy of Sciences* 97: 14257-14262.

53. Klenk, H.P; Clayton, R.A; Tomb, J.-F; White, O; Nelson, K.E; Ketchum, K.A; Dodson, R.J; Gwinn, M; Hickey, E.K; Peterson, J.D; Richardson, D.L; Kerlavage, A.R; Graham, D.E; Kyrpides, N.C; Fleischmann, R.D; Quackenbush, J; Lee, N.H; Sutton, G.G; Gill, S; Kirkness, E.F; Dougherty, B.A; McKenney, K; Adams, M.D; Loftus, B; Peterson, S; Reich, C.I; McNeil, L.K; Badger, J.H; Glodek, A; Zhou, L; Overbeek, R; Gocayne, J.D; Weidman, J.F; McDonald, L; Utterback, T; Cotton, M.D; Spriggs, T; Artiach, P; Kaine, B.P; Sykes, S.M; Sadow, P.W; D'Andrea, K.P; Bowman, C; Fujii, C; Garland, S.A; Mason, T.M; Olsen, G.J; Fraser, C.M; Smith, H.O; Woese, C.R; Venter, J.C. (1997) The complete genome sequence of the

hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus. Nature.* 390:364-370.

54. Kohwi, Y; Kohwi-Shigematsu, T.(1988) Magnesium ion dependent triple-helix structure formed by homopurine-homopyrimidine sequences in supercoiled plasmid DNA. *Proceedings of the National Academy of Sciences.* 85:3781-3785

55. Koonin, E; Aravind, L; Kondrashov, A.(2000).The impact of comparative genomics on our understanding of evolution. *Cell.* 102:573-576.

56. Kunst, F; Ogasawara, N; Moszer, I; Albertini, A.M; Alloni, G; Azevedo, V; Bertero, M.G; Bessieres, P; Bolotin, A; Borchert, S; Borriss, R; Boursier, L; Brans, A; Braun, M; Brignell, S.C; Bron, S; Brouillet, S; Bruschi, C.V; Caldwell, B; Capuano, V; Carter, N.M; Choi, S.K; Codani, J.J; Connerton, I.F; Danchin, A. *et al.* (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis. Nature.* 390: 249-256.

57. Kuroda, M; Ohta, T; Uchiyama, I; Baba, T; Yuzawa, H; Kobayashi, I; Cui, L; Oguchi, A; Aoki, K; Nagai, Y; Lian, J; Ito, T; Kanamori, M; Matsumaru, H; Maruyama, A; Murakami, H; Hosoyama, A; Mizutani-Ui, Y; Kobayashi, N; Tanaka, T; Sawano, T; Inoue, R; Kaito, C; Sekimizu, K; Hirakawa, H; Kuhara, S; Goto, S; Yabuzaki, J; Kanehisa, M; Yamashita, A; Oshima, K; Furuya, K; Yoshino, C; Shiba, T; Hattori, M; Ogasawara, N; Hayashi, H; Hiramatsu, K. (2001) Whole genome sequencing of meticillin-resistant Staphylococcus aureus. *Lancet.* 357 (9264): 1225-1240.

58. Lawrence, J. Gene transfer, speciation, and the evolution of bacterial genome. *Current Opinion in Microbiology.* 2:519-523.

59. May, B.J; Zhang, Q; Li, L.L; Paustian, M.L; Whittam, T.S; Kapur, V. (2001) Complete genomic sequence of *Pasteurella multocida, Pm70. Proceedings of the National Academy of Sciences* 98: 3460-3465.

60. McClelland, M; Sanderson, K.E; Spieth, J; Clifton, S.W; Latreille, P; Courtney, L; Porwollik, S; Ali, J; Dante, M; Du, F; Hou, S; Layman, D; Leonard, S; Nguyen, C; Scott, K; Holmes, A; Grewal, N; Mulvaney, E; Ryan, E; Sun, H; Florea, L; Miller, W; Stoneking, T; Nhan, M; Waterston, R; Wilson, R.K. (2001) Complete genome sequence of *Salmonella enterica serovar Typhimurium LT2. Nature.* 413 :852-856.

61. Merino, E; Becerril, B; Valle, F; Bolivar, F.(1987) Deletion of a repetitive extragenic palindromic (REP) sequence dowstream from the structural gene of *Escherichia coli* glutamate dehydrogenase affects the stability of its mRNA.*Gene.*58(2-3):305-309.

62. Merino, E; Bolivar, F. (1989) The ribonucleoside diphosphate reductase gene (nrdA) of *Escherichia coli* carries a repetitive extragenetic palindromic (REP) sequence in its 3'structural terminus. *Molecular Microbiology.* 3(6): 839-841.

63. Mulligan, M; Hawley, D; Entriken, R; McClure, W.(1983). *Escherichia coli* promoter sequences predict in vitro RNA polymerase selectivity. *Nucleic Acids Research*. 12(1):789-800.

64. NCBI, National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/

65. Nelson, K.E; Clayton, R.A; Gill, S.R; Gwinn, M.L; Dodson, R.J; Haft, D.H; Hickey, E.K; Peterson, J.D; Nelson, W.C; Ketchum, K.A; McDonald, L; Utterback, T.R; Malek, J.A; Linher, K.D; Garrett, M.M; Stewart, A.M; Cotton, M.D; Pratt, M.S; Phillips, C.A; Richardson, D; Heidelberg, J; Sutton, G.G; Fleischmann, R.D; White, O; Salzberg, S.L; Smith, H.O; Venter, J.C; Fraser, C.M. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*. 399: 323-329.

66. Newbury, S; Smith, N; Robinson, C; Hiles, I; Higgins, C. (1987) Stabilization of translationally active mRNA by procariotic REP sequences. *Cell*. 48: 297-310.

67. Ng, W.V; Kennedy, S.P; Mahairas, G.G; Berquist, B; Pan, M; Shukla, H.D; Lasky, S.R; Baliga, N; Thorsson, V; Sbrogna, J; Swartzell, S; Weir, D; Hall, J; Dahl, T.A; Welti, R; Goo, Y.A; Leithauser, B; Keller, K; Cruz, R; Danson, M.J; Hough, D.W; Maddocks, D.G; Jablonski, P.E; Krebs, M.P; Angevine, C.M; Dale, H; Isenbarger, T.A; Peck, R.F; Pohlschrod, M; Spudich, J.L; Jung, K.-H; Alam, M; Freitas, T; Hou, S; Daniels, C.J; Dennis, P.P; Omer, A.D; Ebhardt, H; Lowe, T.M; Liang, P; Riley, M; Hood, L; DasSarma, S. (2000) Genome sequence of *Halobacterium species NRC-1*. *Proceedings of the National Academy of Sciences* 97:12176-12181.

68. Nierman, W.C; Feldblyum, T.V; Paulsen, I.T; Nelson, K.E; Eisen, J; Heidelberg, J.F; Alley, M; Ohta, N; Maddock, J.R; Potocka, I; Nelson, W.C; Newton, A; Stephens, C; Phadke, Nd; Ely, B; Laub, M.T; DeBoy, R.T; Dodson, R.J; Durkin, A.S; Gwinn, M.L; Haft, D.H; Kolonay, J.F; Smit, J; Craven, M; Khouri, H; Shetty, J; Berry, K; Utterback, T; Tran, K; Wolf, A; Vamathevan, J; Ermolaeva, M; White, O; Salzberg, S.L; Shapiro, L; Venter, J.C; Fraser, C.M. (2001) Complete Genome Sequence of *Caulobacter crescentus*. *Proceedings of the National Academy of Sciences* 98: 4136-4141.

69. Nolling, J; Breton, G; Omelchenko, M.V; Markarova, K.S; Zeng, Q; Gibson, R; Lee, H.M; Dubois, J; Qiu, D; Hitti, J; Wolf, Y.I; Tatusov, R.L; Sabathe, F; Doucette-Stamm, L; Soucaille, P; Daly, M.J; Bennett, G.N; Koonin, E.V; Smith, D.R. (2001) Genome Sequence and Comparative Analysis of the Solvent-Producing *Bacterium Clostridium acetobutylicum*. *Journal of Bacteriology*. 183: 4823-4838.

70. Ochman, H; Bergthorssan, U. (1995) Genome evolution in enteric bacteria. *Current Opinion in Genetic Development*. 5(6):734-738.

71. Ochman, H; Lawrence, J; Grolsman, E. Lateral gene transfer and the Nature of bacterial innovation. *Nature*. 405:299-304.

72. Ogata, H; Audic, S; Renesto-Audiffren, P; Fournier, P.-E; Barbe, V; Samson, D; Roux, V; Cossart, P; Weissenbach, J; Claverie, J.-M; Raoult, D. (2001) Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science*. 293:2093-2098.

73. Ogata, H; Audic, S; Barbe, V; Artiguenave, F; Fournier P.E; Raoult, D; Claverie J.M. (2000) Selfish DNA in protein-coding genes of *Rickettsia*. *Science*. 290: 347-350.

74. Oppenheim, AB; Rudd, KE; Mendelson, I; Teff, D. (1993) Integration host factor binds to a unique class of complex repetitive extragenic DNA sequnces in *Escherichia coli*. *Molecular Microbiology*. 10(1):113-122.

75. Parkhill, J; Achtman, M; James, K.D; Bentley, S.D; Churcher, C; Klee, S.R; Morelli, G; Basham, D; Brown, D; Chillingworth, T; Davies, R.M; Davis, P; Devlin, K; Feltwell, T; Hamlin, N; Holroyd, S; Jagels, K; Leather, S; Moule, S; Mungall, K; Quail, M.A; Rajandream, M.A; Rutherford, K.M; Simmonds, M; Skelton, J; Whitehead, S; Spratt, B.G; Barrell, B.G. (2000) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis Z2491*. *Nature*. 404: 502-506.

76. Parkhill, J; Dougan, G; James, K.D; Thomson, N.R; Pickard, D; Wain, J; Churcher, C; Mungall, K.L; Bentley, S.D; Holden, M.T.G; Sebaihia, M; Baker, S; Basham, D; Brooks, K; Chillingworth, T; Connerton, P; Cronin, A; Davis, P; Davies, R.M; Dowd, L; White, N; Farrar, J; Feltwell, T; Hamlin, N; Haque, A; Hien, T.T; Holroyd, S; Jagels, K; Krogh, A; Larsen, T.S; Leather, S; Moule, S; O'Gaora, P; Parry, C; Quail, M; Rutherford, K; Simmonds, M; Skelton, J; Stevens, K; Whitehead, S; Barrell, B.G. (2001) Complete genome sequence of a multiple drug resistant *Salmonella enterica serovar Typhi CT18*. *Nature* 413 (6858), 848-852.

77. Parkhill, J; Wren, B.W; Mungall, K; Ketley, J.M; Churcher, C; Basham, D; Chillingworth, T; Davies, R.M; Feltwell, T; Holroyd, S; Jagels, K; Karlyshev, A; Moule, S; Pallen, M.J; Penn, C.W; Quail, M; Rajandream, M.A; Rutherford, K.M; VanVliet, A; Whitehead, S; Barrell, B.G. *Nature* 403 (6770), 665-668 (2000) (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*. 403: 665-668.

78. Parkhill, J; Wren, B.W; Thomson, N.R; Titball, R.W; Holden, M.T.G; Prentice, M.B; Sebaihia, M; James, K.D; Churcher, C; Mungall, K.L; Baker, S; Basham, D; Bentley, S.D; Brooks, K; Cerdeno-Tarraga, A.M; Chillingworth, T; Cronin, A; Davies, R.M; Davis, P; Dougan, G; Feltwell, T; Hamlin, N; Holroyd, S; Jagels, K; Leather, S; Karlyshev, A.V; Moule, S; Oyston, P.C.F; Quail, M; Rutherford, K; Simmonds, M; Skelton, J; Stevens, K; Whitehead, S; Barrell, B.G. (2001) Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*. 413: 523-527.

79. Pentiggia, A; Negri, A; Beltrama, M; Blanchi, M.(1993) Protein HU binds specifically to kinked DNA. *Molecular Microbiology*. 7:343-350.

80. Perna, N.T; Plunkett, G. III, Burland, V; Mau, B; Glasner, J.D; Rose, D.J; Mayhew, G.F; Evans, P.S; Gregor, J; Kirkpatrick, H.A; Posfai, G; Hackett, J; Klink, S;

Boutin, A; Shao, Y; Miller, L; Grotbeck, E.J; Davis, N.W; Lim, A; Dimalanta, E; Potamousis, K; Apodaca, J; Anantharaman, T.S; Lin, J; Yen, G; Schwartz, D.C; Welch, R.A; Blattner, F.R. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli O157:H7. Nature.* 409:529-533.

81. Pettijohn, D.(1982) Structure and properties of the bacterial nucleoid. *Cell.* 30:667-6669.

82. Read, T.D; Brunham, R.C; Shen, C; Gill, S.R; Heidelberg, J.F; White, O; Hickey, E.K; Peterson, J; Umayam, L.A; Utterback, T; Berry, K; Bass, S; Linher, K; Weidman, J; Khouri, H; Craven, B; Bowman, C; Dodson, R; Gwinn, M; Nelson, W; DeBoy, R; Kolonay, J; McClarty, G; Salzberg, S.L; Eisen, J; Fraser, C.M. (2000) Genome sequences of *Chlamydia trachomatis MoPn* and *Chlamydia pneumoniae AR39. Nucleic Acids Research.* 28 (6): 1397-1406.

83. Ricchetti, M; Buc, H.(1993) *E.coli* DNA polymerase I as a reverse transcriptase. *EMBO Journal.* 12(2):387:396.

84. Saurin, W.(1987) Repetitive palindromic sequences in *Escherichia coli*.Detection and characterization with a new computer program. *Computer Applications in the BioSciences.* 3(2):121-127.

85. She, Q; Singh, R.K; Confalonieri, F; Zivanovic, Y; Allard, G; Awayez, M.J; Chan-Weiher, C.C; Clausen, I.G; Curtis, B.A; De Moors, A; Erauso, G; Fletcher, C; Gordon, P.M; Heikamp-de Jong, I; Jeffries, A.C; Kozera, C.J; Medina, N; Peng, X; Thi-Ngoc, H.P; Redder, P; Schenk, M.E; Theriault, C; Tolstrup, N; Charlebois, R.L; Doolittle, W.F; Duguet, M; Gaasterland, T; Garrett, R.A; Ragan, M.A; Sensen, C.W; Van der Oost, J. (2001) The complete genome of the crenarchaeon *Sulfolobus solfataricus P2. Proceedings of the National Academy of Sciences* 98: 7835-7840.

86. She, Q; Singh, R.K; Confalonieri, F; Zivanovic, Y; Allard, G; Awayez, M.J; Chan-Weiher, C.C; Clausen, I.G; Curtis, B.A; De Moors, A; Erauso, G; Fletcher, C; Gordon, P.M; Heikamp-de Jong, I; Jeffries, A.C; Kozera, C.J; Medina, N; Peng, X; Thi-Ngoc, H.P; Redder, P; Schenk, M.E; Theriault, C; Tolstrup, N; Charlebois, R.L; Doolittle, W.F; Duguet, M; Gaasterland, T; Garrett, R.A; Ragan, M.A; Sensen, C.W; Van der Oost, J. (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum. Nature.* 407: 508-513.

87. Shigenobu, S; Watanabe, H; Hattori, M; Sakaki, Y; Ishikawa, H. (2000) Genome sequence of the endoCellular bacterial symbiont of aphids *Buchnera sp. APS. Nature.* 407: 81-86.

88. Shirai, M; Hirakawa, H; Kimoto, M; Tabuchi, M; Kishi, F; Ouchi, K; Shiba, T; Ishii, K; Hattori, M; Kuhara, S; Nakazawa, T. (2000) Comparison of whole genome sequences of *Chlamydia pneumoniae J138* from Japan and *CWL029* from USA. *Nucleic Acids Research* 28:2311-2314.

89. Shyamala, V; Schneider, E; Ames, G.F. (1990) Tandem chromosomal duplications: role os REP sequences in the recombination event at the join-point. *EMBO Journal.* 9:939-946.

90. Simpson, A.J.G; Reinach, F.C; Arruda, P; Abreu, F.A; Acencio, M; Alvarenga, R; Alves, L.M.C; Araya, J.E; Baia, G.S; Baptista, C.S; Barros, M.H; Bonaccorsi, E.D; Bordin, S; Bove, J.M; Briones, M.R.S; Bueno, M.R.P; Camargo, A.A; Camargo, L.E.A; Carraro, D.M; Carrer, H; Colauto, N.B; Colombo, C; Costa, F.F; Costa, M.C.R; Costa-Neto, C.M; Coutinho, L.L; Cristofani, M; Dias-Neto, E; Docena, C; El-Dorry, H; Facincani, A.P; Ferreira, A.J.S; Ferreira, V.C.A; Ferro, J.A; Fraga, J.S; Franca, S.C; Franco, M.C; Frohme, M; Furlan, L.R; Garnier, M; Goldman, G.H; Goldman, M.H.S; Gomes, S.L; Gruber, A; Ho, P.L; Hoheisel, J.D; Junqueira, M.L; Kemper, E.L; Kitajima, J.P; Krieger, J.E; Kuramae, E.E; Laigret, F; Lambais, M.R; Leite, L.C.C; Lemos, E.G.M; Lemos, M.V.F; Lopes, S.A; Lopes, C.R; Machado, J.A; Machado, M.A; Madeira, A.M.B.N; Madeira, H.M.F; Marino, C.L; Marques, M.V; Martins, E.A.L; Martins, E.M.F; Matsukuma, A.Y; Menck, C.F.M; Miracca, E.C; Miyaki, C.Y; Monteiro-Vitorello, C.B; Moon, D.H; Nagai, M.A; Nascimento, A.L.T.O; Netto, L.E.S; Nhani Jr;A; Nobrega, F.G; Nunes, L.R; Oliveira, M.A; de Oliveira, M.C; de Oliveira, R.C; Palmieri, D.A; Paris, A; Peixoto, B.R; Pereira, G.A.G; Pereira Jr;H.A; Pesquero, J.B; Quaggio, R.B; Roberto, P.G; Rodrigues, V; de M. Rosa, A.J; de Rosa Jr;V.E; de Sa, R.G; Santelli, R.V; Sawasaki, H.E; da Silva, A.C.R; da Silva, F.R; da Silva, A.M; Silva Jr;W.A; da Silveira, J.F; Silvestri, M.L.Z; Siqueira, W.J; de Souza, A.A; de Souza, A.P; Terenzi, M.F; Truffi, D; Tsai, S.M; Tsuhako, M.H; Vallada, H; Van Sluys, M.A; Verjovski-Almeida, S; Vettore, A.L; Zago, M.A; Zatz, M; Meidanis, J; Setubal, J.C. (2000) The genome sequence of the plant pathogen *Xylella fastidiosa*. The *Xylella fastidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis. *Nature*. 406: 151-157.

91. Smith, D.R; Doucette-Stamm, L.A; Deloughery, C; Lee, H.-M; Dubois, J; Aldredge, T; Bashirzadeh, R; Blakely, D; Cook, R; Gilbert, K; Harrison, D; Hoang, L; Keagle, P; Lumm, W; Pothier, B; Qiu, D; Spadafora, R; Vicare, R; Wang, Y; Wierzbowski, J; Gibson, R; Jiwani, N; Caruso, A; Bush, D; Safer, H; Patwell, D; Prabhakar, S; McDougall, S; Shimer, G; Goyal, A; Pietrovski, S; Church, G.M; Daniels, C.J; Mao, J.-i; Rice, P; Nolling, J; Reeve, J.N. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum deltaH*: functional analysis and comparative genomics. *Journal of Bacteriology*. 179:7135-7155.

92. Stern, M.J; Ames, G.F; Smith, N.H; Robinson, E.C. (1984) Repetitive extragenic palindromic sequences a mayor component of the bacterial genome. *Cell*. 37(3):1015-1026.

93. Stover, C.K; Pham, X.-Q.T; Erwin, A.L; Mizoguchi, S.D; Warrener, P; Hickey, M.J; Brinkman, F.S.L; Hufnagle, W.O; Kowalik, D.J; Lagrou, M; Garber, R.L; Goltry, L; Tolentino, E; Westbrook-Wadman, S; Yuan, Y; Brody, L.L; Coulter, S.N; Folger, K.R; Kas, A; Larbig, K; Lim, R.M; Smith, K.A; Spencer, D.H; Wong, G.K.-S; Wu, Z; Paulsen, I.T; Reizer, J; Saier, M.H; Hancock, R.E.W; Lory, S; Olson, M.V. (2000) Complete genome sequence of *Pseudomonas aeruginosa PA01*, an opportunistic pathogen. *Nature*. 406: 959-964.

94. Takami, H; Nakasone, K; Takaki, Y; Maeno, G; Sasaki, Y; Masui, N; Fuji, F; Hirama, C; Nakamura, Y; Ogasawara, N; Kuhara, S; Horikoshi, K. (2000) Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. *Nucleic Acids Research* 28: 4317-4331.

95. Tettelin, H; Nelson, K.E; Paulsen, I.T; Eisen, J.A; Read, T.D; Peterson, S; Heidelberg, J; DeBoy, R.T; Haft, D.H; Dodson, R.J; Durkin, A.S; Gwinn, M; Kolonay, J.F; Nelson, W.C; Peterson, J.D; Umayam, L.A; White, O; Salzberg, S.L; Lewis, M.R; Radune, D; Holtzapple, E; Khouri, H; Wolf, A.M; Utterback, T.R; Hansen, C.L; McDonald, L.A; Feldblyum, T.V; Angiuoli, S; Dickinson, T; Hickey, E.K; Holt, I.E; Loftus, B.J; Yang, F; Smith, H.O; Venter, J.C; Dougherty, B.A; Morrison, D.A; Hollingshead, S.K; Fraser, C.M. (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae. Science.* 293: 498-506.

96. Tettelin, H; Saunders, N.J; Heidelberg, J; Jeffries, A.C; Nelson, K.E; Eisen, J.A; Ketchum, K.A; Hood, D.W; Peden, J.F; Dodson, R.J; Nelson, W.C; Gwinn, M.L; DeBoy, R; Peterson, J.D; Hickey, E.K; Haft, D.H; Salzberg, S.L; White, O; Fleischmann, R.D; Dougherty, B.A; Mason, T; Ciecko, A; Parksey, D.S; Blair, E; Cittone, H; Clark, E.B; Cotton, M.D; Utterback, T.R; Khouri, H; Qin, H; Vamathevan, J; Gill, J; Scarlato, V; Masignani, V; Pizza, M; Grandi, G; Sun, L; Smith, H.O; Fraser, C.M; Moxon, E.R; Rappuoli, R; Venter, J.C. (2000) Complete genome sequence of *Neisseria meningitidis serogroup B strain MC58. Science.* 287: 1809-1815.

97. Thompson, E.O.P. (1955) *The insulin molecule.* Scientific American. 192(5): 36-41.

98. Tomb, J.-F; White, O; Kerlavage, A.R; Clayton, R.A; Sutton, G.G; Fleischmann, R.D; Ketchum, K.A; Klenk, H.P; Gill, S; Dougherty, B.A; Nelson, K; Quackenbush, J; Zhou, L; Kirkness, E.F; Peterson, S; Loftus, B; Richardson, D; Dodson, R; Khalak, H.G; Glodek, A; McKenney, K; Fitzegerald, L.M; Lee, N; Adams, M.D; Hickey, E.K; Berg, D.E; Gocayne, J.D; Utterback, T.R; Peterson, J.D; Kelley, J.M; Karp, P.D; Smith, H.O; Fraser, C.M; Venter, J.C. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori. Nature.* 388:539-547.

99. Van Ham, S.M; van Alphen, L; Mooi, F.R; van Putten, O.M. (1994) The fibrial gene cluster of *Haemophilus influenzae type b. Molecular Microbiology.* 13(4): 673-684.

100. Vasconselos, A.T; Mattoso, M.A.G;de Almeida, D.F. (2000) Short interrupted palindromes on the extragenic DNA of Escherichia coli K-12, Haemophilus influenzae and Neisseria meningitidis. Bioinformatics. 16(11):968-977.

101. Versalovic. (1991). *D*istribution of repetitive DNA sequences in eubacteria and applications to fingerprint of bacterial genomes. *Nucleic Acids Research.*19(24)6823:6831.

102. Watson, J.D; Crick, F.H.C. (1953) Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. *Nature.* 171: 964-967.

103. White, O; Eisen, J.A; Heidelberg, J.F; Hickey, E.K; Peterson, J.D; Dodson, R.J; Haft, D.H; Gwinn, M.L; Nelson, W.C; Richardson, D.L; Moffat, K.S; Qin, H; Jiang, L; Pamphile, W; Crosby, M; Shen, M; Vamathevan, J.J; Lam, P; McDonald, L; Utterback, T; Zalewski, C; Makarova, K.S; Aravind, L; Daly, M.J; Minton, K.W; Fleischmann, R.D; Ketchum, K.A; Nelson, K.E; Salzberg, S; Smith, H.O; Venter,

J.C; Fraser, C.M. (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans R1. Science.* 286: 1571-1577.

104. Yang, Y; Ames, G. (1988) DNA gyrase binds to a family of prokariotic repetitive palindromic sequences. *Proceedings of the National Academy of Sciences* 85(23):8850-8854.

# 7. ANEXO I. PROGRAMA DE BÚSQUEDA DE LOS ELEMENTOS REP

```
#!/usr/bin/perl
$continua=1;
while ($continua>0) {                    APERTURA DE LOS GENOMAS
$_='head -1 /murazaki/COMUN/PROGRAMAS/organismos`;
chomp;
$org =  $_;
print " org  $org\n";

$renglones=`cat -n/murazaki/COMUN/PROGRAMAS/organismos | tail -1 | cut-b 1-7 `;
chomp($renglones);
$renglones=$renglones-1;
print "renglones $renglones\n";
system("tail -$renglones/murazaki/COMUN/PROGRAMAS/organismos >
/murazaki/COMUN/PROGRAMAS/org");
system("cp /murazaki/COMUN/PROGRAMAS/org/murazaki/COMUN/PROGRAMAS/organismos");
system("cat /murazaki/COMUN/PROGRAMAS/organismos");

if ($org ne "final") {
 print " Analizo $org\n";
 $Archivo = "/murazaki/COMUN/RESULTADOS/condder".$org;
 $Archivo2 = "/murazaki/COMUN/RESULTADOS/condizq".$org;
 $Orgasec = "/murazaki/COMUN/FLAT/".$org.".flat";
 $Organum = "/murazaki/COMUN/NUMEROS/".$org.".num";
 $archiv = "/murazaki/COMUN/PROGRAMAS/datos";
 $sectrans = "/murazaki/COMUN/PROGRAMAS/sectrans";
 $archiv2 = "/murazaki/COMUN/PROGRAMAS/datos2";
 $erchrep ="/murazaki/COMUN/RESULTADOS/OTR/".$org.".rep";
 $erchfrec ="/murazaki/COMUN/RESULTADOS/".$org.".frec";
 $erchrep2 ="/murazaki/COMUN/RESULTADOS/OTR/".$org.".rep1";
 $erchfrec2 ="/murazaki/COMUN/RESULTADOS/".$org.".frec1";
 $avance = "/murazaki/COMUN/PROGRAMAS/avance";

 foreach $RRTT (0..5) {
        foreach $TTRR (0..18) {
                $acum1[$RRTT][$TTRR] =0;
                $acum1[$RRTT][$TTRR] =0;
                $acum2[$RRTT][$TTRR] =0;
                $acum2[$RRTT][$TTRR] =0;
                }
 }                                       INICIALIZACIÓN DE VARIABLES
 $MN =0;   @mn = "";   $sumCa3 = 0;  @Mn = "";  $sumCa = 0;  $porcentaje =90;  $Porcentaje =90;
 $valmax = 30;  $score =  27;  $scoreC1 = 12.15;   $scoreC2 =  14.85;  $Valmax = 30;  $Score = 27;
 $ScoreC2 = 12.15; $ScoreC1 = 14.85; # 90%  maximo  16.5
 system ("rm $sectrans");   system ("rm $archiv");  system ("rm $archiv2");
 open (OUT,">>$avance");
 print OUT "        $org\n";
 close (OUT);
 $J =0; $a=0; $b=0; $c=0; $uno=0; $j=0; $a1=0; $b1=0;
 $c1=0; $un =0;
 &inicio;          GENERACIÓN MATRIZ ELEMENTO -->
$J =0; $a=0; $b=0; $c=0; $uno=0; $j=0; $a1=0; $b1=0;
   $c1=0; $un =0;
 &iniciocom;              GENERACIÓN MATRIZ ELEMENTO <--
 foreach $jJ (1..8) {
        &uno;             BÚSQUEDA DE ELEMENTOS REP -->
        print "acabe uno\n";
```

```
        &dos;              BÚSQUEDA DE ELEMENTOS REP ←
        print "acabe dos\n";
  }
  open (OUT,">>$avance");
        print OUT "------->          @mn <------          @Mn\n";
  close (OUT);
}
else {print "ya termine\n"; $continua=0;
}
}                  FIN DEL PROGRAMA
##############################################################################
sub uno {                  BÚSQUEDA DE ELEMENTOS REP →
                  print "------------------>\n";
                  $NN = 14;    $NNN =17;    $MM = 15;      $MMM =18;

                  print "        Vuelta: $jJ\n\n\n";
                  @dv = 0; @Dv =0; $dv=0; $Dv=0;  $sumatc1=0;  $sumatc2=0;
                  print "loop1\n";
                  $vari =0;

                  system("./b.out");          LLAMADA SUBRUTINA EN C
```

**SUBRUTINA EN C**

```
 #include <stdio.h> #include <stdlib.h>
#include <malloc.h> #define beginprocedure {
#define endprocedure } #define begin {
#define end } #define total 7100000
char s[total];
int sec[total];
intm,tot,sec[total],caja1,caja2,sei,max,k,y,smay
or,separa,iter;
float score1[5][19];
float score2[5][16];   DEFINICIÓN DE
int frec1[5][19]; VARIABLES
int frec2[5][16];
int tra[255];
int i,j,k,l,tot,pos_return,reps;
char secuencia_in[100], matriz_out[100],
rep_out[100],
numero[10];
float filtro1,filtro2,filtro3;
char linea[111];
FILE *fdatos; FILE *fsecuencia_in;  FILE
*fmatriz_out;
FILE *frep_out; FILE *fsecuencia_trans;
/**********************/ main()
/**************************/
beginprocedure
fdatos=fopen("/murazaki/COMUN/PROGRA
MAS/datos2","r");
 fgets(secuencia_in,100,fdatos);
pos_return=strlen(secuencia_in);
secuencia_in[pos_return-1]=0;
 fgets(rep_out,100,fdatos);
pos_return=strlen(rep_out);   APRETURA
                              DE
rep_out[pos_return-1]=0;      ARCHIVOS
 fgets(matriz_out,100,fdatos);
pos_return=strlen(matriz_out);
```

```
matriz_out[pos_return-1]=0;
 for (i=1;i<=4;i++) {
    for (j=1;j<=18;j++) {
       fgets(numero,10,fdatos);
        score1[i][j]=atof(numero);
printf("%5.2f ",score1[i][j]);
  }
  printf("\n");
   }
   printf("\n");
 for (i=1;i<=4;i++) {
    for (j=1;j<=15;j++) {
       fgets(numero,10,fdatos);
        score2[i][j]=atof(numero);
printf("%5.2f ",score2[i][j]);
  }
   printf("\n");
  }
  printf("\n");
fgets(numero,10,fdatos);
       filtro3=atof(numero);
 fgets(numero,10,fdatos);
       filtro1=atof(numero);
fgets(numero,10,fdatos);
       filtro2=atof(numero);
printf("%5.2f %5.2f
%5.2f\n",filtro1,filtro2,filtro3);
fclose(fdatos);
for (i=1;i<=pos_return;i++) printf
("%c",secuencia_in[i]);
fsecuencia_in=fopen(secuencia_in,"r");
printf("Empiezo a leer secuencia
%s\n",secuencia_in);
fgets(s,total,fsecuencia_in);
fclose(fsecuencia_in);
printf("ya lei archivo \n");
tot=strlen(s);
printf("Total de bases leidas %d\n",tot);
fmatriz_out=fopen(matriz_out,"w");
frep_out=fopen(rep_out,"a");
fsecuencia_trans=fopen("/murazaki/COMUN/P
ROGRAMAS/sectrans","w");
 iter=tot-63;
for (i=0;i<256;i++) tra[i]=0;
for (i=0;i<tot+100;i++) sec[i]=0;
/** 97,65=a     99,67=c     103,71=g
116,84=t  **/
tra[97]=tra[65]=1; tra[99]=tra[67]=2;
tra[103]=tra[71]=3; tra[116]=tra[84]=4;
tra[42]=-100;
    for (i=0;i<tot+100;i++) sec[i]=tra[s[i]];
printf("Total de bases leidas %d\n",tot);
REPs();
fprintf(fsecuencia_trans,"%s\n",s);
fclose(fmatriz_out); fclose(frep_out);
fclose(fsecuencia_trans);
printf ("me salgo del programa compl.c\n");
endprocedure
/**********************/ REPs()
```

GENERACIÓN DE VENTANAS

```
/**********************/
{  COMPARACIÓN DE VALORES
float caja1,caja2,score_total,mayor;
int ii,ll,jj;
printf("empiezo a buscar REPs\n");
for (ii=1;ii<=4;ii++) {
  for (jj=1;jj<=18;jj++) {
   frec1[ii][jj]=0; } }
 for (ii=1;ii<=4;ii++) {
   for (jj=1;jj<=15;jj++) {
     frec2[ii][jj]=0;} }
reps=0;
for (i=0;i<=iter;i++)
  {
  caja1=0;
  for (j=1;j<=18;j++) {
caja1=caja1+score1[sec[i+j]][j]; }
    if (caja1>=filtro1)
      if (caja1<1000.0) {
     {
      for (l=0;l<30;l++)
        {
mayor=0.0; caja2=0;
        for (m=1;m<=15;m++) {
caja2=caja2+score2[sec[i+18+l+m]][m];
                if (mayor<caja2) {
mayor=caja2; ll=l; }
              }
        if (mayor>=filtro2)
  if (mayor<1000.0)
  {
   score_total=caja1+mayor;
   if (score_total>=filtro3) {
   reps=reps+1;
fprintf(frep_out,"%10d ..%8d\t loop:%2d
",i+2,i+35+ll,l);
fprintf(frep_out,"%5.2f  %5.2f  %5.2f
",caja1,
caja2, score_total);
                        for
(ii=0;ii<=17;ii++)
fprintf(frep_out,"%c",s[i+1+ii]);
fprintf(frep_out,"  ");
for (ii=0;ii<ll;ii++)
fprintf(frep_out,"%c",s[i+19+ii]);
                        fprintf(frep_out,"
");
                        for
(ii=0;ii<=14;ii++)
fprintf(frep_out,"%c",s[i+19+ll+ii]);
fprintf(frep_out,"\n");
for (ii=0;ii<=17;ii++)
frec1[sec[i+1+ii]][ii+1]=frec1[sec[i+1+ii]][ii+1
]+1;
for (ii=0;ii<=14;ii++)
frec2[sec[i+19+ll+ii]][ii+1]=frec2[sec[i+19+ll+
ii]][ii+1]+1;
for (ii=i+1;ii<=i+34+ll;ii++) s[ii]=42;
}
```

```
                                                    }
                                                  }
                                                 }
                                                 }
                                                }
                                               /*************************************
                                               **************/SALIDA DE LOS
                                               ELEMENTOS
                                               printf("Numero de REPs encontradas
                                               %d\n",reps);
                                               fprintf (fmatriz_out,"%d\n",reps);
                                               for (ii=1;ii<=4;ii++) {
                                               for (jj=1;jj<=18;jj++) {
                                                fprintf(fmatriz_out,"%d ",free1[ii][jj]);
                                                                        }
                                                fprintf(fmatriz_out,"\n");
                                                                        }
                                                fprintf(fmatriz_out,"\n");
                                                for (ii=1;ii<=4;ii++) {
                                               for (jj=1;jj<=15;jj++) {
                                                fprintf(fmatriz_out,"%d ",free2[ii][jj]);
                                                fprintf(fmatriz_out,"\n");
                                                                        }
                                                fprintf(fmatriz_out,"\n");
                                               fclose(fmatriz_out); fclose(frep_out);
                                               /*************************************
                                               *************/
                                               } FIN DE LA SUBRUTINA EN C


                print "acabe REPS\n";
                &sal2; &ca3; &nuevamat; &valmax; &matout; &hasta;
                LLAMADA A LAS OTRAS SUBRUTINAS DEL ELEMENTO REP -->

}

sub dos {          BÚSQUEDA DE ELEMENTOS REP <--

                print "\n\n\n<----------\n\n\n";
                $NN = 17;   $NNN =14;    $MM = 18;    $MMM =15;
                print "      Vuelta: $jJ\n\n\n";
                system("./a.out"); LLAMADO A LA SUBRUTINA EN C PARA
                print "acabe REPS\n";                EL ELEMENTO REP <--
                @dv1 = 0; @Dv1 =0; $dv1=0; $Dv1=0; $Sumatc1=0; $Sumatc2=0;
                print "loop1\n"; $variz =0;
                &sal21; &ca31; &nuevamat1; &valmax1; &matout1; &hasta1;
                LLAMADO A LAS OTRAS  SUBRUTINAS  PARA EL ELEMENTO REP <--
}

sub inicio { GENERACIÓN MATRIZ ELEMENTO -->
        open (SAL,">$archiv");
        print SAL "$Orgasec\n$erchrep\n$erchfree\n";
        close (SAL);
open (SAL,">>$archiv");

@vec1= qw(      0 0 0 0 0 0 0 0 1 0 0 0 0 0 0.5 0 0 0
                0 0 1 1 0 0 0 0 0 1 0 0 1 0 0.5
                0 1 0 0 0.5 1 0 0 1 0 1 0.5 0 1 0
                1 0 0 0 0.5 0 0 1 0 0 0 0 0 0 0.5);
```

```
@vec2= qw(      0 0.5 0 0 0 0 0 0 1 0 0 0.5 0 0 0 0 0 1 0
                0 1 0 0.5 1 0 0 0 0 1 0.5 0 0 1 1 0 0 1
                0.5 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0
                0 0 0 0.5 0 1 1 0 1 0 0 0 0 0 0 1 0 0);


        foreach $J (1..15) { $a=16 + $uno; $b=31+ $uno;$c=46+$uno; $uno +=1;
                $Mat1[1][$J] = $Vec1[$J];
                $Mat1[2][$J] = $Vec1[$a];
                $Mat1[3][$J] = $Vec1[$b];
                $Mat1[4][$J] = $Vec1[$c];
}
        foreach $j (1..18) { $a1=19 + $un; $b1= 37+ $un;
                $c1=55 +$un; $un +=1;
                $Mat2[1][$j] = $Vec2[$j];
                $Mat2[2][$j] = $Vec2[$a1];
                $Mat2[3][$j] = $Vec2[$b1];
                $Mat2[4][$j] = $Vec2[$c1];
}
        print "(.( ya acabe la matriz\n";
        foreach $Aa (1..4) {
                foreach $AaA (1..15) {
                        print SAL "$Mat1[$Aa][$AaA]\n";
}
}
        close (SAL);
        open (SAL,">>$archiv");
        foreach $Aa (1..4) {
                foreach $AaA (1..18) {
                        print SAL "$Mat2[$Aa][$AaA]\n";
        }
}
        close (SAL);
        open (SAL,">>$archiv");
        print SAL "$score\n$scoreC1\n$scoreC2";
        close (SAL);
}


sub sal2 { GENERACIÓN DE UNA MATRIZ TEMPORAL PARA EL ELEMENTO →

        print "empece sal2\n";
        open (SAL,"$erchfrec") || die "no pude abrir\n";
        print "         ###empece sal2\n";
        foreach  (1..11) {
                while (<SAL>) {
                        chomp;
                        $vari ++;    $_;   $db[$vari] = $_; $vari\n";
        }
}

        foreach $nucaja1 (2..5) {
                $dv2[$nucaja1] = $db[$nucaja1];
        }

foreach $nucaja1 (7..10) {
        $Dv2[$nucaja1] = $db[$nucaja1];
        }
```

```
        $dv = join(" ",@dv2);
        $Dv = join(" ",@Dv2);
        @dv = split(" ",$dv);
        @Dv = split(" ",$Dv);

        foreach $nucleo (1..4) {
                foreach $ZZ (1..15) {
                        $nunu =0 +$avan;  $avan ++;
                        $acum1[$nucleo][$ZZ] += $dv[$nunu];
        }
        foreach $ZZz (1..18) {
                $nunu2 =0+$avan2;  $avan2 ++;
                $acum2[$nucleo][$ZZz] += $Dv[$nunu2];
        }
}

        $avan=0;
        $avan2=0;
        $avante =0;
        $anute =0;


        foreach $Aa (1..4) {
                foreach $AaA (1..15) {
                        print  "$acum1[$Aa][$AaA] ";
                }
                        print "\n";
                }
                        print "\n";print "\n";
        foreach $Aa (1..4) {
                foreach $AaA (1..18) {
                        print  "$acum2[$Aa][$AaA] ";
                }
                        print "\n";
                }
}


sub ca3 {       CÁLCULO DE LA DESVIACIÓN ESTANDAR PARA EL ELEMENTO →
        chomp ($db[1]);
        $mn[$jJ] = $db[1];  $sumCa3 += $mn[$jJ];
        $ca3 = ($sumCa3/4)**(1/2);
                if ($sumCa3 < 1 ) {
                        print "********************************** FUE CERO\n";
                        $ca3 = 1;
                }

        print "\n\n     jJ   $jJ
        mn: @mn             sumCa3:   $sumCa3  \n\n";
}

sub nuevamat {  NUEVA MATRIZ PARA EL ELEMENTO →
        $espa = "";
        $punto = "#";
        print "ca3:  $ca3\n";
        foreach $LALA (1..$MM) {
                foreach $LULU (1..4) {
                        $Mat1[$LULU][$LALA] = ($acum1[$LULU][$LALA]/$ca3);
                        $_ = $Mat1[$LULU][$LALA];
                        s/\D/#/g; s/^\d.*#/$espa/g; @_ = split ("",$_); $r_ =
```

```
                        @_; $rr_ = $r_-2;
                                if ($r_ > 2) { foreach (0..$rr_-1) { chop
                                        ($Mat1[$LULU][$LALA]);
                                }
                        }
                }
        }
        foreach $LA (1..$MMM) {
                foreach $LU (1..4) {
                        $Mat2[$LU][$LA] = ($acum2[$LU][$LA]/$ca3);
                        $_ = $Mat2[$LU][$LA];
                        s/\D/#/g; s/^\d.*#/$espa/g; @r_ = split ("",$_); $r_ =
                        @r_; $rr_ = $r_-2;
                                if ($r_ > 2) { foreach (0..$rr_-1) { chop;
                                        ($Mat2[$LU][$LA]);
                                }
                        }
                }
        }
}

sub valmax { CÁLCULO DE NUEVOS VALORES PARA EL ELEMENTO →

        foreach $PP (1..$MM) {
                $mayor = 0;
                foreach $HH (1..4) {
                        if ($Mat1[$HH][$PP] > $mayor) { $mayor =
                                $Mat1[$HH][$PP];
                        }
                }
                $sumatc1 += $mayor;
        }
        foreach $PPP (1..$MMM) {
                $mayor1 = 0;
                foreach $HHH (1..4) {
                        if ($Mat2[$HHH][$PPP] > $mayor1) { $mayor1 =
                                $Mat2[$HHH][$PPP];
                                }
                }
                $sumatc2 += $mayor1;
        }

        $valmax = $sumatc1 + $sumatc2;
        $por = ($porcentaje-68)/3;
        $porcentaje = $porcentaje-$por;

                if ($jJ > 6) {
                        print " sub por\n";
                        $porcentaje = 70;
                }
                if ($sumCa3 < 1) {
                        $valmax = 30;
                        $sumatc1 = 12.15; #90%  maximo  13.5
                        $sumatc2 =  14.85;
                        print "************************************ FUE \n";

                }
                $score = ($valmax*$porcentaje)/100;
                $scoreC1 = ($sumatc1*$porcentaje)/100;
                $scoreC2 = ($sumatc2*$porcentaje)/100;
```

```
                $x_ = $score;
        foreach $xx_ (0..2) {
                $_ = $x_; s/\D/#/g; s/^\d.*#/$espa/g; @x_ = split
                ("",$_); $r_ = @x_; $rr_ = $r_-2;
                if ($r_ > 2) { foreach (0..$rr_-1) { chop ($x_);
                }
        }
        if ($xx_ == 0) { $score = $x_; $x_ = $scoreC1;}
        if ($xx_ == 1) { $scoreC1 = $x_; $x_ = $scoreC2;}
        if ($xx_ == 2) { $scoreC2 = $x_; }
        }
        print "\n\nvamax = $valmax  score= $score ... $scoreC1 ... $scoreC2 porcentaje:
        $porcentaje\n\n\n";
}


sub matout {      SALIDA DE LA NUEVA MATRIZ PARA EL ELEMENTO →
        open (MAT,">$archiv");
        if ($sumCa3 < 1) {
                $J =0; $a=0; $b=0; $c=0; $uno=0; $j=0; $a1=0; $b1=0;
                $c1=0; $un =0;
                &inicio;
        }
        else {
        foreach $aA6 (1..4) {
                foreach $aAa6 (1..15) {
                        print MAT "$Mat1[$aA6][$aAa6]\n";
                }
        }
        foreach $aA7 (1..4) {
                foreach $aAa7 (1..18) {
                        print MAT "$Mat2[$aA7][$aAa7]\n";
                }
        }
        print MAT "$score\n$scoreC1\n$scoreC2";
        close (MAT);
        }
}

sub hasta { INDICADOR PARA EL ELEMENTO →

        print  "van $jJ turnos
         #          elementos encontrados @mn
          #                en el ultimo turno $db[1]\n
        ##########################################################
        ##########################################################
        ##########################################################\n\n\n";
        open (OUT,">>$erchrep");
         print OUT  "\n\n van $jJ turnos
        ca3 =$ca3 vamax = $valmax  score= $score ... $scoreC1... $scoreC2 ...   % $porcentaje
         elementos encontrados @mn en el ultimo turno $db[1]\n
         ##########################################################
        ##########################################################
       ##########################################################\n\n\n";
        close (OUT);
}
```

# 8. ANEXO II. PROGRAMA DE FORMACIÓN DE UNIDADES Y ASIGNACIÓN DE GENES

```
#!/diego/bin/perl

&paso1;    AGRUPACIÓN DE LOS ELEMENTOS EN LAS  UNIDADES
        close (OUT);
        close (OUT2);
        close (OUT3);
        print "Elementos: $sumaE Unidades: $TtT\n";

&paso2;   CÁLCULO DEL TAMAÑO DE LAS UNIDADES
        close (OUT5);
        close (OUT4);
        print "Unidades = $GG\n";

&paso3;  ASIGNACIÓN DE LA UNIDAD A LOS GENES
        close (OU);
        close (OU2);
        close (OU3);
        close (OU4);
        print "Unidades: $FIN\n";

####################################


sub paso1 {

        print "paso1\n";
        system ("rm ULTIMO");
        open (OUT,">>ULTIMO");
        open (OUT2,"conceK")|| die "no pude abrir si";

                while (<OUT2>) {
                        chomp;
                        if (/^.\d/) {
                                $sub ++;
                                $as="$_ <--- ";
                                #$as=$_;
                                $espa = " ";
                                $esp="";
                                s/\D/$espa/g;
                                $nam = $_;
                                @nu2 = split (" ",$nam);
                                $nn{$nu2[0]} = $as;
                                $a[$sub] = $nu2[0];
                        }
                }
        open (OUT3,"conceK") || die "no pude abrir no";
                while (<OUT3>) {
                        chomp;
                        if (/^.(\d.)/) {
                                $aa="$_ ---> ";
                                $sub ++;
                                $espa = " ";$esp="";
                                s/\D/$espa/g;
```

```
                                        $naam = $_;
                                        @n2 = split (" ",$naam);
                                        $nn{$n2[0]} = $aa;
                                        $a[$sub] = $n2[0];
                                }
                        }
        $l = @a;

        @re =  sort {$a <=> $b}  @a;
        print OUT "$nn{$re[1]}          ";
        &resta
}

sub resta {

        foreach $eq (1..$l) {

                $tot =  $re[$eq+1] - $re[$eq];
                &grupo;

        }
}

sub grupo {

        if ($tot < 100) {
                $cm ++;
                print OUT "$nn{$re[$eq+1]}          ";
        }

        else {
                $sumaE += $cm;
                print OUT "\n# $cm";
                $TtT ++;
                $cm =1;
                print OUT "\n\n\n$nn{$re[$eq +1]}          ";
        }
}

#####################################

sub paso2 {

        print "paso2\n";
        system ("rm DESULTIMO");
        open (OUT5,">>DESULTIMO");
        print OUT5 "\n";
        &ultimo;
        open (OUT4,"ULTIMO");
                while (<OUT4>) {
                        chomp;
                        if (/^.\d/) {
                                        $GGf++;
                                        $w = 1 + $ww;
                                        $ww += 4;
                                        $espa = " ";
                                        $K = "#";
                                        s/score.{18}/$K/g;
                                        s/\D/$espa/g;
```

```
                                        $nn = $_;
                                        @nu = split (" ",$na);
                                        #print "@nu\n";
                                        $ll = @nu;
                                        $e =0;
                                        @nn =0;
                                        &mayor;
                                        &print;
                                }
                        }
}

sub ultimo {

        open (OUT4,"ULTIMO");
        @dd = OUT4>;
        #print "0 $d[0],1 $d[1],2 $d[2],3 $d[3],4 $d[4],5
        $d[5],6 $d[6],7 $d[7],8 $d[8],9 $d[9]\n";
}

sub mayor {

        foreach $ee (0..$ll) {
                if ($nu[$ee] > 1000 and $nu[$ee] < 50000000) {
                        $e ++;
                        $nn[$e] = $nu[$ee];
                }
        }

        $yy = @nn;
}

sub print {

        chomp($dd[$w-1]);
        $_ =$dd[$w-1];
        $espa="";
        $esp=" ";
        s/loop/9/g;
        #s/\s/$espa/g;
        s/:/$espa/g;
        s/\d/$espa/g;

        #print "$_\n";

        @OL =split(" ",$_);
        $WsD = join(" ",@OL);
        #print "0 $OL[0],1 $OL[1],2 $OL[2],3 $OL[3],4 $OL[4],5
        $OL[5],6 $OL[6],7 $OL[7],8 $OL[8],9 $OL[9]\n";
        #print "$WsD\n";

        print OUT5 "($nn[1]..$nn[$yy-1]) $WsD $dd[$w]";#$WsD
        $dd[$w] \n\n\n
}

########################################

sub paso3 {

        print "paso3\n";
```

```
        &posREP;
        &REP;
        print "termine con pos y REP\n";
        foreach (1..$rrr) {

                open (OU3, "ecoli.gbk") || die "no pude";
                $hh = 1 + $hhH;
                $hhH +=2;
                        while (<OU3>) {
                                if (/\/gene=".{1,6}/) {&nomb;}      #note="b.{1,6}
                                        if (/gene.{10,12}(\d).*/ ||
                                                /gene.{10,12}complement.*/) {&compl; &ti;
                                        }
                                }
                        }
}

sub posREP {

        system ("rm tmpK222");     # CAMBIO

        open (OU2,"DESULTIMO");
        print "posREP\n";
                while (<OU2>) {
                        chomp;
                        if (/^.(\d.)/) {
                                $DkkD ++;
                                $x +=2;
                                $espa = " ";
                                $espan = "";
                                s/score.*/$espan/g;
                                s/\D/$espa/g;
                                $sim = $_;
                                @sim = split (" ", $sim);
                                $sem[$x -1] = $sim[0];
                                $sem[$x] = $sim[1];
                                #print "                                $x, $xyx\n";
                                $rr =@sem;
                                $rrr = $rr/2;
                        }
                }

        #print "@sem\n";
        print " $DkkD\n";
}
sub compl {

    if (/gene.{10,12}(\d).*/) {
            $i = 1;
    } else {
            $i = 0;
    }
        $a [3] = $i;
        $a[1] = $a[2]; $a[2]=$a[3];
        #print"////////////////////////////////$a[1]////////////////////////////////\n";
}

sub ti {
    $esp = " ";
    s/\D/$esp/g;
```

```
    @te = split (" ",$_);
    $ti[3] = $te[0];
    $ti[1] = $ti[2]; $ti[2]=$ti[3];
    $tti[3] = $te[1];
    $tti[1] = $tti[2]; $tti[2]=$tti[3];

}
sub nomb {
        $espa = "";
        $espp = " ";
        s/$espp/#/g;
        s/#/$espa/g;
         $gene = "gene: ";
        $gana = "/gene=";
        s/$gana/$gene/g;
        $ta ="$_";

        $hash{$ti[2]} = $ta;
        chomp ($hash{$ti[2]});

                if ($sem[$hh] < $ti[2]) {
                        $W=0+$S; $S +=1; &be;
                        print "$sem[$hh] <= $ti[2]\n";
                        close (OU3);
                }
}

sub be {

        open  (OU,">>tmpK222");    # CAMBIO
        $idefix ++;
        print "!!!!!!!Andiamo... ($ti[1]-$tti[1])
        $hash{$ti[1]}          ##$idefix   ($ti[2]-$tti[2])
        $hash{$ti[2]}\n";

                if ($a[1] eq 1 and $a[2] eq 1) {
                        print OU  "$hash{$ti[1]} IDE:$idefix  $db[$W]";
                }
                if ($a[1] eq 0 and $a[2] eq 1) {
                        print  OU  "($ti[1]-$tti[1])   $hash{$ti[1]}**************5'  ($ti[2]-$tti[2])
IDE:$idefix  $hash{$ti[2]}**************$db[$W]";
                }
                if ($a[1] eq 0 and $a[2] eq 0) {
                        print  OU  "$hash{$ti[2]} IDE:$idefix  $db[$W]";
                }
                if ($a[1] eq 1 and $a[2] eq 0) {
                        print OU "$hash{$ti[2]} IDE:$idefix  $db[$W] - $hash{$ti[1]} IDE:$idefix
$db[$W]";

                }
                close (OU);
}


sub REP {

        print "REP\n";
        open (OU4,"DESULTIMO");
                while (<OU4>) {
```

```
                @db = <OU4>;
                chomp;
                #print "0 $db[0],1 $db[1],2 $db[2],3 $db[3],4 $db[4],5
                $db[5],6 $db[6],7 $db[7],8 $db[8],9 $db[9]\n";
        }
}
```