TÍTULO

IDENTIFICACIÓN DE REGIONES HIPERVARIABLES DE BACTERIAS PATÓGENAS MEDIANTE HERRAMIENTAS BIOINFORMÁTICAS

APLICACIÓN EN EL GENOTIPADO BACTERIANO

AUTORA

Mercedes Jiménez Bautista

Esta edición electrónica ha sido realizada en 2012

Director Enrique Viguera Mínguez

Curso Máster en Bioinformática

ISBN 978-84-7993-991-5

Universidad Internacional de Andalucía, 2012

Trabajo fin de Máster en Bioinformática

Identificación de regiones hipervariables de bacterias patógenas mediante herramientas bioinformáticas, aplicación en el genotipado bacteriano

Autor: Mercedes Jiménez Bautista
Dirigido por: Dr. Enrique Viguera Mínguez

Universidad Internacional de Andalucía

Septiembre 2012

Agradecimientos

A mi director de Proyecto Dr. Enrique Viguera Mínguez, a Dña. Rosario María Carmona Muñoz (Plataforma Andaluza de Bioinformática. Universidad de Málaga), a Carlos, a mi familia y a todos aquellos que de una forma u otra contribuyeron para la realización del presente trabajo, mi más sincera gratitud.

Índice General

Índice de Ilustraciones y Tablas

I. INTRODUCCIÓN GENERAL

1. Secuencias de DNA repetidas

La gran mayoría de los genomas de organismos procariotas analizados hasta la fecha poseen secuencias de DNA repetidas. Estas repeticiones se localizan tanto en regiones génicas como intergénicas y se clasifican en repeticiones dispersas y repeticiones en tándem. En cuanto a su organización física, estas repeticiones pueden constituir desde repeticiones en tándem de una secuencia de pocos nucleótidos a múltiples copias de genes completos, como en el caso del RNA ribosómico y del RNA transferente.

El **DNA repetido disperso** está repartido por todo el genoma. Existen distintas familias de elementos repetidos intercalados como podemos apreciar en la Tabla 1.

Repeticiones	Acrónimo	Características	Referencias
Palíndromos repetitivos intragénicos o unidades palindrómicas	REP or PU	Entre 21-65 bp Palíndromo imperfecto Secuencia intragénica	Higgins *et al*. (1982) Stem *et al*. (1984)
Elementos mosaico dispersos bacterianos	BIME	Entre 40-500 bp Combinación en mosaico de REP separadas por otros motivos	Gilson *et al*. (1991)
Repeticiones palindrómicas cortas regularmente dispuestas formando una matriz	CRISPR	Repeticiones no contiguas (entre 24-47 bp) separadas por trozos de secuencia del mismo tamaño (entre 26-72 bp)	Ishino *et al*. (1987) Mojica *et al*. (2000)

Elementos transponibles miniatura con repeticiones invertidas	MITE	Entre 100-400 bp No son autónomos (toda la longitud que puede transponerse es móvil *in trans* Probablemente deriven de IS por deleciones internas Flanqueados por repeticiones invertidas (entre 10-40 bp) Secuencia extra o intragénica	Correia *et al*. (1988) Delihas (2008)
Unidades repetidas intergénicas o regiones intergénicas repetitivas de enterobacterias	IRU or ERIC	Entre 69-127 bp Grandes secuencias palindrómicas	Sharples & Lloyd (1990) Hulton *et al*. (1991)
Secuencias de inserción	IS	Entre 0.7-3.5 kbp Elemento autónomo A menudo flanqueado por repeticiones invertidas (entre 10-40 bp) Codifica para una transposasa (lleva 1 o 2 ORFs) Transposición conservativa o replicativa	Mahillon & Chandler (1998)
Transposones	Tn	Elementos autónomos Codifica para transposasas y numerosos productos génicos (ej. Resistencia a antibióticos, virulencia, etc.) *Compuestos (Clase I)*: – Flanqueado por dos IS (indénticas o diferentes, directas o invertidas) – Transposición conservativa	

	No compuestos (Clase II): – Flanqueado por dos repeticiones invertidas – Transposición replicativa	
Elementos bacteriófagos	Fago Mu y elementos transponibles – Crecimiento lisogénico: transposición conservativa – Crecimiento lítico: transposición replicativa	Pato (1989)

Tabla 1: Familias de repeticiones intercaladas

El **DNA repetido en tándem** son repeticiones de secuencias idénticas o casi idénticas que se disponen una a continuación de otra. Se clasifican en función de la longitud de la unidad repetida y del número de repeticiones de dicha unidad:

- Satélites: son secuencias de entre 5 y varios cientos de nucleótidos que se repiten en tándem miles de veces generando regiones de entre 100 kb a varias megabases
- Minisatélites: se componen por una unidad básica de 6 a 25 nucleótidos repetidos en tándem que forman regiones de entre 100 y 20 000 pares de bases.
- Microsatélites: son secuencias repetidas de 1 a 6 nucleótidos.

En este proyecto nos centraremos en el estudio de los microsatélites.

Un **microsatélite** es una secuencia específica de bases de DNA que contienen mono-, di-, tri-, tetra-, penta- o hexanucleótidos repetidos en tándem. En la literatura anglosajona podemos encontrar distintas maneras de denominarlos:

Simple Sequence Repeats (SSRs), *Short Tandem Repeats* (STRs) o *Variable Number Tandem Repeats* (VNTRs). Estos microsatélites se caracterizan porque pueden tener una única unidad repetida o bien distintos tipos de unidades repetidas en tándem. Se distribuyen de manera variable, generalmente en regiones no codificantes del DNA. La composición y estructura de microsatélites varía enormemente entre grupos filogenéticos, e incluso de un individuo a otro de la misma especie.

En función del motivo repetido que presenten se clasifican en:

- SSR puro o perfecto: un único motivo repetido *n* veces en tándem. Ejemplo: $(CAA)_8$
- SSR interrumpido puro: un único motivo repetido *n* veces en el que se intercalan nucleótidos que no pertenecen al motivo repetido. Ejemplo: $(CAA)_3$ TG $(CAA)_{10}$
- SSR compuesto: al menos dos o más motivos repetidos en tándem. Ejemplo: $(CAA)_7$ $(TGG)_{13}$
- SSR interrumpido compuesto: en al menos uno de los motivos se intercalan nucleótidos. Ejemplo: $(CAA)_5$ TG $(CAA)_7$ AGC $(CTT)_{12}$
- SSR complejo: combinación de los anteriores. Ejemplo: $(CAA)_8$ TG $(AG)_{10}$ $(TGG)_4$ GC $(TGG)_8$

Las regiones flanqueantes al microsatélite suelen estar altamente conservadas y son utilizadas para el diseño de cebadores para su amplificación por PCR y caracterización del tamaño de la región repetida por electroforesis.

Los microsatélites se utilizan como marcadores moleculares por tener la más alta incidencia de polimorfismo en comparación con otros marcadores como cambios en los sitios de restricción (analizados por la técnica de RFLPs). Además, tienen unas determinadas características:

- son muy informativos: tienen herencia codominante y muchas repeticiones están presentes entre organismos estrechamente relacionados
- son muy abundantes y se encuentran uniformemente distribuidos por el genoma
- los loci están generalmente conservados entre especies relacionadas, e incluso entre algunos géneros
- son técnicamente simples y sensibles de analizar; por amplificación por PCR pueden utilizarse de forma fácil y rápida para automatizar su uso como marcadores

Los microsatélites pueden ser usados en otras aplicaciones como la elaboración de mapas de ligamiento, pruebas de paternidad, identificación del perfil genético de individuos, filogenias, estudios epidemiológicos, medicina forense y estudios demográficos (Chambers y MacAvoy, 2000; Metzgar *et al.,* 2000; Toth *et al.,* 2000; Ellegren, 2004; Buschiazzo y Gemmell, 2006). Distintos estudios han revelado que las variaciones en los microsatélites se asocian con un número de enfermedades humanas, siendo la más conocida los trastornos de repeticiones de trinucleótidos (Everett y Wood, 2004).

A día de hoy no se conoce muy bien la función biológica del DNA repetido, pero se cree que pueden desempeñar un papel en la regulación de la expresión génica (Subramanian *et al*., 2003). Se sabe que las repeticiones invertidas se encuentran en muchos orígenes de replicación de virus y bacterias y que son necesarias para poder comenzar la replicación (Lin y Kowalski, 1994). También se ha visto que las repeticiones de dinucleótidos y trinucleótidos en regiones codificantes de algunas proteínas son necesarias para desempeñar su función (Kashi *et al.*, 1997)

Como ya hemos dicho anteriormente, los microsatélites son altamente polimórficos en cuanto al número de repeticiones para un locus concreto. El origen de dicho

polimorfismo aún no está claro, pero diferentes estudios lo atribuyen a errores durante la replicación del DNA que está causado por un fenómeno llamado "slippage" o deslizamiento de la DNA polimerasa (Schlotterer y Tautz, 1992, Viguera et al., 2001; Zane et al., 2002). En este tipo de errores el extremo 3' de la hebra naciente se desaparea y se reaparea con otra de las repeticiones en el molde, formándose un bucle en una de las hebras. Dependiendo de que el bucle se forme en la hebra naciente o molde, se formarán respectivamente expansiones o deleciones en el DNA (Streisinger et al. 1966).

1.1 Conformaciones del DNA asociadas a la inestabilidad génica

El DNA puede tener diferentes tipos de conformaciones que generalmente están asociadas a puntos calientes de inestabilidad genómica (Wang y Vasquez 2006; Myers *et al*. 2008; Vasquez y Hanawalt 2009). Algunos procesos metabólicos como la replicación, la transcripción y la reparación del DNA necesitan que la doble hebra de DNA se desaparee (Mirkin 2006; Wang y Vasquez 2006; Mirkin 2007; Voineagu *et al*. 2009a). Si en esa zona hay repeticiones inversas en la secuencia, al desaparearse se pueden formar estructuras en forma de horquilla (Figura 2A) o estructuras cruciformes.

Cuando las repeticiones tienen un alto contenido en bases CG se pueden formar otro tipo de estructura llamada cuádruplex o tetrahélice (Figura 2B) a través de emparejamientos de Hoogsteen entre cuatro guaninas unidas por puentes de hidrógeno (Voineagu *et al.* 2009b). Estas estructuras se encuentran cerca de promotores de genes y en los telómeros.

Otra estructura que puede formarse es la ocasionada por deslizamiento de cadenas (Figura 2C) comentado anteriormente. Las cadenas pueden elongarse o acortarse en procesos metabólicos como la replicación. La probabilidad de la

formación de estructuras S-DNA se incrementa con la longitud de la repetición implicada y se reduce cuando las repeticiones no son perfectamente complementarias (Voineagu *et al.* 2009a).

Las repeticiones invertidas de fragmentos de DNA de polipurinas-polipirimidinas pueden formar estructuras en triplex (Figura 2D) o H-DNA (Mirkin 2006; Wang and Vasquez 2006; Mirkin 2007) En algunos casos, las moléculas de DNA con repeticiones de tipo GAA/TTC pueden formar una conformación denominada "*sticky* DNA" (Figura 2E), correspondiente a la formación de una estructura tríplex intramolecular entre dos regiones con este tipo de repeticiones en orientación directa (Wang y Vasquez 2006; Voineagu *et al.* 2009a).

La estructura en Z-DNA es una doble hélice levógira que forma un esqueleto en zig-zag (Wang y Vasquez 2006). Tiene una secuencia alternante de bases púricas y pirimidínicas y está favorecida por un alto contenido en C-G. Esta conformación tiene una gran inestabilidad en bacterias.

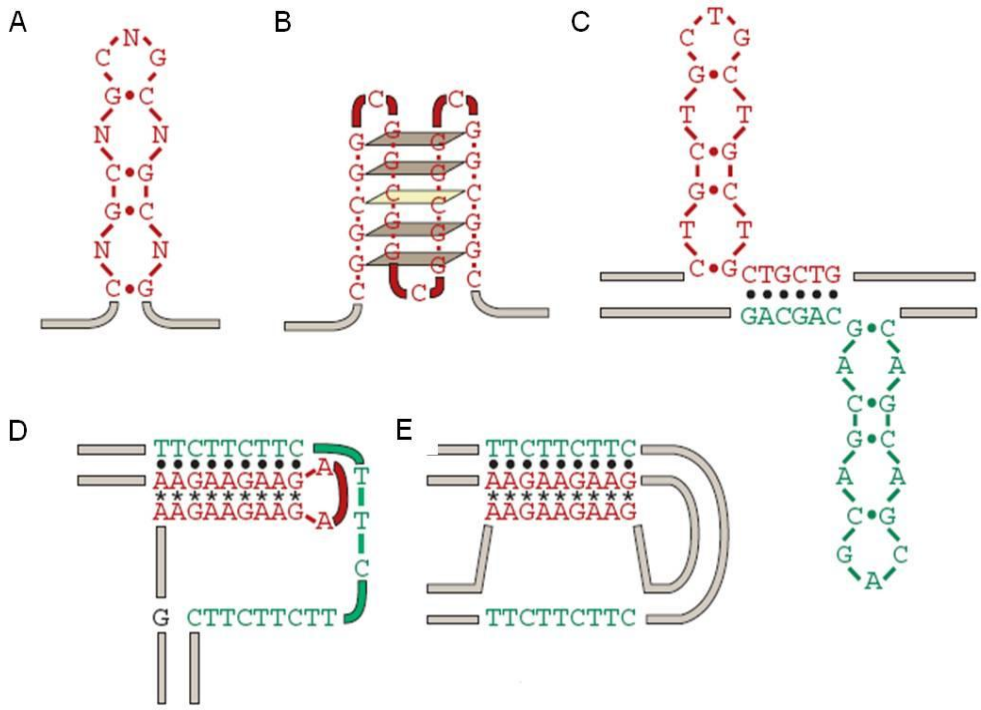


Figura 2: Estructuras de DNA inusuales formadas por secuencias repetidas.

A. Estructura en horquilla imperfecta, formada por repeticiones del tipo (CNG)n. B. Estructura en cuádruplex formada por repeticiones (CGG)n. C. Estructura S-DNA formada por repeticiones (CTG)n•(CAG)n. D. Estructura en tríplex formada por repeticiones (GAA)n. E. Estructura en tríplex intramolecular "sticky DNA". La hebra propensa a adoptar una estructura secundaria se muestra en rojo, su hebra complementaria en verde y el DNA flanqueante en gris. (Adaptada de Mirkin, 2004; Mirkin, 2007).

1.2 Inestabilidad de secuencias de DNA repetidas asociada a la patogénesis bacteriana

Las secuencias de DNA repetidas se caracterizan por su alta inestabilidad, lo que puede ser utilizada en beneficio de la propia célula tal y como se ha observado en determinadas especies de bacterias patógenas como *Haemophilus* o *Neisseria*.

En estas bacterias, mutaciones de cambio de fase de lectura en repeticiones de nucleótidos activan o inactivan determinados genes, llamados genes de

contingencia, de una forma reversible (Moxon *et al.,* 1994; Bayliss *et al.,* 2001). Algunas regiones de DNA repetidas parecen estar vinculadas con los mecanismos de adaptación, virulencia y patogenicidad de los microorganismos (Van Belkum *et al.,* 1998; Benson, 1999; Denoeud y Vergnaud, 2004). Así, en los promotores y en la región codificante de los genes de contingencia hay microsatélites (simple sequence contingency loci) que regulan el marco de lectura y el grado de expresión génica. La regulación de estos genes produce una rápida adaptación a entornos desfavorables, porque pueden desarrollar gran cantidad de posibles fenotipos con los que adaptarse a un ambiente hostil (Moxon *et al.,* 1994; Bayliss *et al.,* 2001) y aumenta la capacidad de virulencia de patógenos como *Haemophilus* o *Neisseria* entre otros.

2. Abundancia de repeticiones en genomas bacterianos

Los organismos procariotas tienen una gran variedad y riqueza de repeticiones en su genoma. Hay diferentes medidas para poder clasificar estas repeticiones que nos proporcionan diferentes perspectivas:

- Índice de repetición: mide el grado en el que las secuencias del genoma se repiten.
- RSF (Relative Simplicity Factor): mide la cantidad de SSR en el genoma
- Cobertura de la repetición: mide la longitud total de repeticiones no solapadas dividida entre el tamaño del genoma

Así, diferentes estudios concluyen que para repeticiones de más de 300 nucleótidos y considerando la cobertura de repetición, encontramos genomas de procariotas con un rango de repeticiones que van entre el 0%, es decir, no existen repeticiones de esa longitud, hasta un 42% del genoma. En el rango inferior se encuentran las bacterias endosimbiontes obligadas, como *Buchnera* spp., que

tiene pocas repeticiones en su genoma, mientras que las bacterias endoparásitas obligadas, como *Phytoplasma* spp. tienen gran densidad de repeticiones. A pesar de que ambos grupos han evolucionado por reducción del material genético (Shigenobu *et al*., 2000; Gil *et al*., 2003; Oshima *et al*., 2004), se piensa que patógenos como *Phytoplasma* spp. utilizan las repeticiones para generar variabilidad genética.

En general, los procariotas suelen tener SSRs de pequeño tamaño y dentro de ellos, los que tienen motivos SSR de mayor tamaño son los procariotas de vida libre. La mayoría de los genomas con grandes motivos SSR pertenecen a las Proteobacterias o a las Cianobacterias.

En el estudio realizado por Treangen *et al.,* 2009, utilizando la herramienta Repeatoire, se analizaron 720 cromosomas y encontraron 144.118 repeticiones clasificadas en 56.196 familias. En la Figura 3 a) podemos observar que aproximadamente el 99% de las repeticiones encontradas fueron menores de 10kb y más del 80% son de 2kb o incluso más pequeñas. En la Figura 3 b) se puede observar que hay grandes segmentos duplicados porque 267 familias de repeticiones abarcan 10 kb o más, variando su tamaño entre 10 y 143kb.

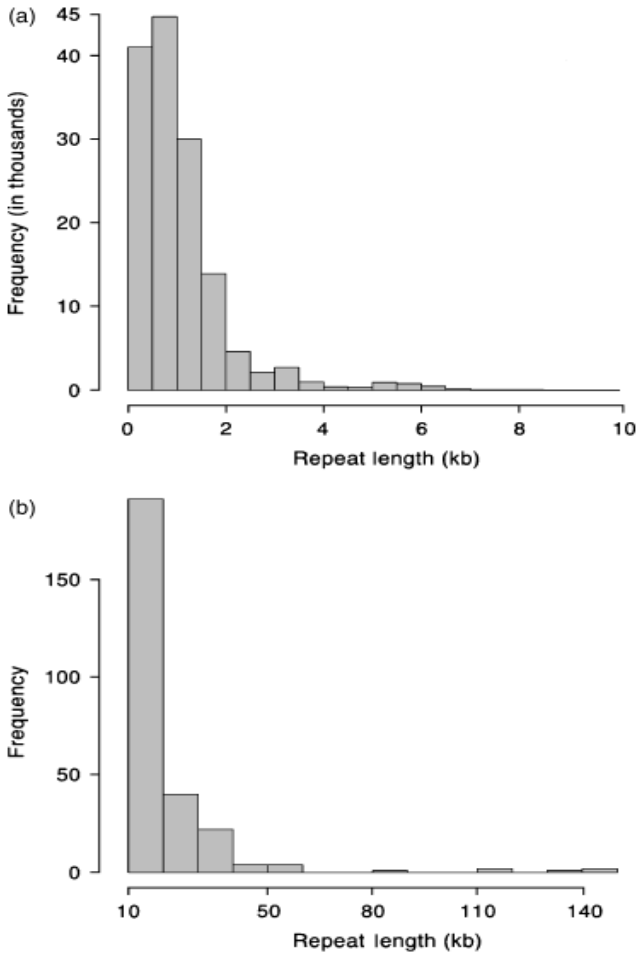


Figura 3: Distribución de repeticiones >300nt. Tomada de Treangen *et al.*, 2009

3. Técnica de MLVA (*Multiple Loci VNTR Analysis*)

Las secuencias repetidas descritas anteriormente son altamente inestables, pudiendo presentar variabilidad en el número de unidades de cada repetición, designándose entonces como número variable de repeticiones en tándem (VNTR) (Van Belkum *et al*., 1998; Fenollar y Raoult, 2004). Como ya se comentó anteriormente, el origen de la inestabilidad de los VNTRs puede ser ocasionado por el desapareamiento de la cadena de DNA durante su síntesis (Strand *et al*., 1993; Viguera *et al*., 2001), o en la reparación de la rotura de la doble cadena de DNA, entre otras causas (Ozenberger y Roeder, 1991). La variación puede ser

encontrada fácilmente cuando las bacterias contienen regiones de repetición de diversas longitudes en el mismo locus genómico (Weir, 1992; Smouse y Chevillon, 1998). Debido a su polimorfismo, los VNTR han sido utilizados como marcadores de DNA para la tipificación molecular de varias especies bacterianas (Chang *et al*., 2007).

La técnica de análisis de múltiples regiones de VNTR se designa como MLVA (Multiple Loci VNTR Analysis) y su objetivo es la identificación de loci de variabilidad en el genoma (Van Belkum, 2007). La utilización de cebadores para la PCR que delimiten la región de repetición, posibilita visualizar el polimorfismo en el número de unidades repetidas VNTR, mediante un análisis sencillo por electroforesis, siendo el número de repeticiones calculado en base al tamaño de las secuencias amplificadas (Van Belkum *et al*., 1998; Titze-de-Almeida *et al*., 2004).

Las regiones de DNA repetidas de tipo minisatélites fueron primero descubiertas en humanos por Wyman y White en 1980 (Wyman y White, 1980), y el primer ensayo MLVA fue también utilizado para genotipado humano (Balasingham, 2008). En estos casos estudiados, las regiones de DNA repetidas son muy frecuentes, comprendiendo el 10% o más del genoma (Benson, 1999). La característica polimórfica de los VNTRs llevó a su uso en áreas como las pruebas de paternidad y la medicina forense (Mayr, 1995).

Estas regiones están también presentes en los procariontes, dónde se incluyen los microorganismos de relevancia médica (Van Belkum *et al.,* 1998; Chang *et al.,* 2007), pudiendo utilizarse para el tipado de alta resolución de bacterias (O' Dushlaine *et al.,* 2005), con información significativa sobre sus relaciones genéticas y evolución.

El MLVA ha sido reconocido como una técnica novedosa, sencilla y flexible,

desarrollada esencialmente para el tipado de especies bacterianas de relevancia clínica, y pudiendo ser usada eficientemente para rastrear brotes u otras formas de diseminación bacteriana (Van Belkum, 2007). De esta forma, el análisis de variabilidad de las regiones conteniendo VNTRs contribuye a los estudios epidemiológicos (Van Belkum *et al.,* 1988; Doyle *et al.,* 2006) y filogenéticos de microorganismos de interés. Según Vergnaud y Pourcel (2009) la técnica MLVA podrá tornarse en el patrón oro para el tipado de muchos patógenos.

El tipado de microorganismos a través del análisis MLVA depende de la correcta selección de los marcadores (regiones de VNTR), y aunque éstos individualmente no aporten información relevante sobre los grupos de microorganismos, al presentar mucha variabilidad o un alto nivel de homoplasia, la combinación de regiones independientes, debidamente seleccionadas, puede ser altamente discriminatoria (Le Flèche *et al.,* 2001; Denoeud y Vergnaud, 2004; Cesar, 2008).

La técnica MLVA se ha venido aplicando con éxito creciente en especies bacterianas, tales como *Bacillus anthracis* (Keim *et al.,* 2000; Le Flèche *et al.,* 2001), *Brucella sp.* (Le Flèche *et al.,* 2006)*, Francisella tularensis* (Farlow *et al.,* 2001), *Legionella pneumophila* (Pourcel *et al.,* 2003), *Mycobacterium tuberculosis* (Le Flèche *et al.,* 2002), *Neisseria meningitidis* (Yazdankhah *et al.,* 2005; Liao *et al.,* 2006), *Pseudomonas aeruginosa* (Onteniente *et al.,* 2003; Vu-Thien *et al.,* 2007); *Salmonella enterica* (Lindstedt *et al.,* 2003; Liu *et al.,* 2003), *Staphylococcus aureus* (Sabat *et al.,* 2003; Hardy *et al.*, 2004), *Yersinia pestis* (Le Flèche *et al*., 2001; Pourcel *et al.,* 2004), y *Xylella fastidiosa* (Coletta-Filho *et al.,* 2001).

El tipado de microorganismos por MLVA ha hecho posible identificar aislados de forma sensible y específica, con rapidez, bajo costo, facilidades de implementación, uso e interpretación de los resultados (Cesar, 2008).

4. Características de las DNA polimerasas de translesión (TLS): dinB, polB y umuCD

La supervivencia de los organismos unicelulares depende de su eficiencia para replicar fielmente su información genética. En ocasiones, durante la replicación del DNA se producen errores inducidos por modificaciones de base en el DNA molde o cambios de estructura del DNA que hacen que la replicación se bloquee. Durante la evolución, las células han desarrollado un proceso conocido como síntesis de translesión (TLS) que permite que la replicación continúe y minimizan la muerte celular que se produce cuando se bloquea la replicación. En la síntesis de translesión hay implicadas DNA polimerasas propensas a errores, como la DNA polimerasa II (Familia B), la DNA polimerasa IV (Familia Y) y la DNA polimerasa V (Familia Y). La fidelidad reducida que presentan estas polimerasas de translesión se debe a diferencias estructurales con respecto a la DNA polimerasa replicativa, lo cual les permite adaptarse y replicarse a través de una lesión en el DNA (Schneider *et al*., 2009).

Al estudiar estas DNA polimerasas *in vitro* se observó que tenían una serie de características peculiares:

- cuando copian DNA no dañado producen una tasa mayor de errores que las DNA polimerasas replicativas
- carecen de actividad exonucleasa 3'-5' de corrección de errores
- copian el DNA de forma distributiva en contraposición a la alta procesividad observada para las DNA polimerasas de la replicación
- realizan la síntesis de translesión del DNA dañado

Las DNA polimerasas especializadas tienen un alto potencial mutagénico que la célula tiene que regular para poder mantener su estabilidad genómica. Se induce su expresión cuando hay DNA dañado (Friedberg *et al*., 2002).

- **DinB o DNA polimerasa IV:**

  La DNA polimerasa IV o DinB de *E. coli* está codificada por el gen *dinB* y se identificó inicialmente como un gen inducible por daño (*damage inducible gene*). Las mutaciones predominantes eran del tipo *frameshifts* y ocurrían en su mayoría en regiones de bases idénticas (Wood y Hutchinson, 1984). DinB es una DNA polimerasa con una longitud de 351 aminoácidos y su expresión se induce unas 10 veces cuando se activa la respuesta SOS de la bacteria.

  La fidelidad de síntesis de DNA de Pol IV es más baja que la de Pol II o Pol III (Nohmi, 2006). Pol IV interactúa con la subunidad β de la holoenzima Pol III, con su extremo C-terminal (Bunting *et al*., 2003), lo que es funcionalmente importante para sus efectos mutadores, así como para su actividad TLS (Lenne-Samuel *et al*., 2002). Esta polimerasa copia DNA no dañado cuando la horquilla de replicación se bloquea al encontrar alguna base dañada, en cuyo caso, la DNA polimerasa replicativa es reemplazada por la DNA polimerasa IV, pudiendo extender la extremidad 3' naciente. DinB no tiene actividad 3'-5' exonucleasa (de corrección) que previene este tipo de errores. La sobreexpresión de dinB provoca un aumento del número de mutaciones por cambio de marco de lectura. Esta mutagénesis asociada a DinB, es responsable de la producción de mutaciones adaptativas en la fase estacionaria, proporcionando a la bacteria mayor flexibilidad frente al estrés ambiental, aumento de la supervivencia a largo plazo y adaptación evolutiva.

  La estructura tridimensional de DinB presenta diferencias significativas con respecto a polimerasas de alta fidelidad como dominios de pulgar y palma más pequeños, un dominio adicional "meñique" o "*little finger*" altamente

flexible y un sitio activo amplio que posibilita que las polimerasas de la familia Y puedan acomodar aductos voluminosos en el DNA (Yang, 2003). La posible estructura tridimensional de la DNA polimerasa IV (ID Uniprot Q47155) la podemos observar en la Figura 4.

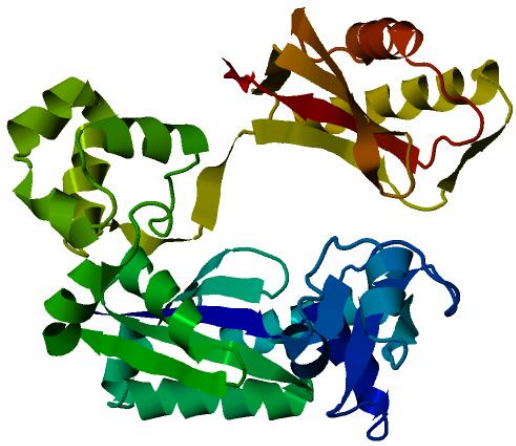


Figura 4: Estructura cristalina de DNA Pol IV o dinB (PDB ID: 1UNN)

- **PolB o DNA polimerasa II:**

  La DNA polimerasa II (ID Uniprot P21189) tiene actividad exonucleasa 3'-5', es codificada por el gen *polB* y es inducible por SOS. A pesar de esta actividad correctora, Pol II participa en la síntesis de translesión siendo una polimerasa propensa a errores de tipo de *frameshifts* -2 cuando replica a través de varios tipos de daños en el molde de DNA (Becherel y Fuchs, 2001). La principal función de la DNA polimerasa II es la reparación de DNA. Participa en la formación de mutaciones espontáneas en plásmidos F' en condiciones de estrés nutricional y colabora en el reinicio de la replicación en células con daños producidos por radiaciones ultravioleta (Nohmi, 2006). La estructura tridimensional de la polimerasa II la podemos observar en la Figura 5.

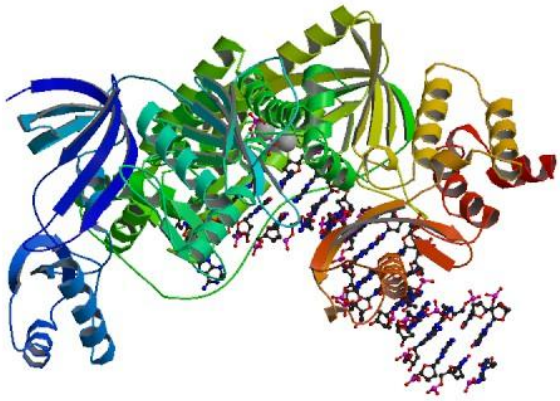


Figura 5: Estructura cristalina de DNA Pol II o polB (PDB ID: 3MAQ)

- **Proteína UmuCD o DNA polimerasa V:**

  La DNA polimerasa V está compuesta por dos subunidades umuC (ID Uniprot P0AG11) y umuD (ID Uniprot D6I6Z4), (Figura 6). La proteína umuCD está implicada en la protección de la célula frente a la radiación ultravioleta y es inducible por SOS.

  El número de moléculas de DNA polimerasa V por célula es de aproximadamente 15. Tras la inducción de SOS, se aumenta la síntesis de la proteína umuCD hasta 200 moléculas por célula y se induce en 50 minutos tras someter a la célula a radiación ultravioleta (Nohmi, 2006).

  En respuesta a daños en el DNA, umuD_2C disminuye la tasa de replicación del DNA, proporcionando a la célula más tiempo para reparar el daño. umuD' podría inducir una transición de una reparación fiel a mutagénesis. Schlacher *et al.* propone que para la forma nativa de umuC y un patrón de

DNA lineal, el único requisito que necesita la DNA polimerasa V para activarse es RecA. Pol V y RecA interactúan de dos formas diferentes:

- vía umuC sin DNA y ATP
- vía umuD' en presencia de ATP y DNA

UmuCD, en su forma activa umu(D')$_2$C, se une junto a dos subunidades de la DNA polimerasa III, a la región del DNA donde se ha producido la lesión, formando un mutasoma. Este mutasoma facilita la replicación tendente a error, es decir, se introduce una mutación que hace que la replicación pueda continuar. La DNA polimerasa V es una de las responsables de la síntesis de translesión (Tang *et al.*, 1999; González y Woodgate, 2002).

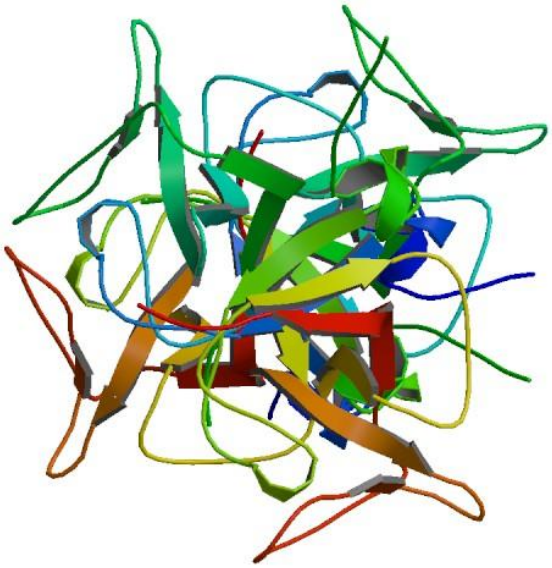


Figura 6: Estructura cristalina de umuCD (PDB ID: 1AY9)

II. OBJETIVOS

- Identificar y analizar las secuencias repetidas en los genomas de cepas de la bacteria patógena Haemophilus influenzae con fines de genotipado molecular.
- Identificar la presencia de genes codificantes para DNA polimerasas de translesión en los genomas de las bacterias en estudio.
- Estudiar una posible correlación entre la presencia de estas DNA polimerasas de translesión y la riqueza en secuencias repetidas en genomas de las bacterias analizadas.

III. METODOLOGÍA

1. Análisis informático de secuencias repetidas de DNA para su uso como candidatos VNTR

Se analizaron las secuencias de DNA genómico de las cepas de referencia de *Haemophilus influenzae* PittEE (referencia en la base de datos del NCBI NC_009566) y *Haemophilus influenzae* PittGG (con referencia NC_009567 en NCBI) en busca de secuencias repetidas de DNA. Se elige esta bacteria por ser de vida libre y no poseer DNA polimerasas de translesión (se puede ver el proceso de cómo se ha seleccionado en el apartado 2: Distribución de las DNA polimerasas de translesión en Enterobacteriaceae y Pasteurellaceae). Para ello se empleó el software de la base de datos de repeticiones en tándem desarrollada por Le Flèche *et al*. (2001) y Denoeud y Vergnaud (2004). Esta base de datos está disponible gratuitamente en http://minisatellites.u-psud.fr. En la Tabla 7 se muestran varias bases de datos públicas sobre elementos repetidos en genomas procariotas.

Database	Description	URL	References
ACLAME	Mobile genetic elements	http://aclame.ulb.ac.be	Leplae *et al*. (2004)
CBS Genome Atlas	Generic repeats	http://www.cbs.dtu.dk/services/GenomeAtlas	Hallin & Ussery (2004)
CRISPRdb	CRISPR repeats	http://crispr.u-psud.fr/crispr/CRISPRdatabase.php	Grissa *et al*. (2007)
IS-Finder	Insertion sequences	http://www-is.biotoul.fr	Siguier et al. (2006)
MICdb	Prokaryote microsatellites	http://www.cdfd.org.in/micas	Sreenu et al. (2003)
ProphageDB	Prophages	http://ispc.weizmann.ac.il/prophagedb	Srividhya et al. (2007)

Tandem Repeats DB	Tandem repeats	http://minisatellites.u-psud.fr	Le Flèche et al. (2001)

Tabla 7: Bases de datos de secuencias repetidas en procariotas (Treangen *et al.* 2009).

PROGRAM	Availability/URL	References
ADPLOT	Email: taneda@si.hirosaki-u.ac.jp	Taneda (2004)
CRISPRFINDER	http://crispr.u-psud.fr/Server/CRISPRfinder.php	Grissa et al. (2004)
CRT	http://www.room220.com/crt	Bland et al. (2007)
EULERALIGN	http://www.stat.psu.edu/~yuzhang	Zhang & Waterman (2003)
MREPATT	http://alggen.lsi.upc.es/recerca/search/mrepatt	Roset et al. (2003)
MREPS	http://bioinfo.lifl.fr/mreps	Kolpakov et al. (2003)
PATTERN LOCATOR	http://www.cmbl.uga.edu/software.html	Mrazek & Xie (2006)
PHOBOS	http://www.ruhr-uni-bochum.de/spezzoo/cm/cm_phobos.htm	NA
PILER	http://www.drive5.com/piler	Edgar & Myers (2005)
REAS	Email: ReAS@genomics.org.cn	Li et al. (2005)
RECON	http://selab.janelia.org/recon.html	Bao & Eddy (2002)
REPEATFINDER	http://www.cbcb.umd.edu/software/RepeatFinder	Volfovsky et al. (2001)
REPEATOIRE	http://wwwabi.snv.jussieu.fr/public/Repeatoire	Treangen et al. (2009)
REPEATSCOUT	http://bix.ucsd.edu/repeatscout	Price et al.(2005)
REPET	http://urgi.versailles.inra.fr/tools/REPET	NA
REPSEEK	http://wwwabi.snv.jussieu.fr/public/RepSeek	Achaz et al. (2007)
REPUTER	http://bibiserv.techfak.uni-bielefeld.de/reputer	Kurtz et al. (2001)
SPUTNIK	http://espressosoftware.com/sputnik	NA
SSRIT	http://finder.sourceforge.net	Temnykh et al. (2001)
STAR	http://atgc-montpellier.fr/star	Delgrange & Rivals (2004)
TRED	http://tandem.sci.brooklyn.cuny.edu/Tandem	Sokol et al. (2007)
TRF	http://tandem.bu.edu/trf/trf.html	Benson (1999)

Tabla 8: Programas para analizar repeticiones (Treangen *et al.* 2009).

Existen otros programas (ver Tabla 8) que permiten analizar repeticiones. Uno de ellos es el programa Tandem Repeat Finder (TRF) que se encuentra en la base de datos Tandem Repeats Data Base (TRBD) disponible gratuitamente en https://tandem.bu.edu/. Es un repositorio público de secuencias repetidas en

tándem en el DNA genómico. Contiene herramientas para el análisis de repeticiones, predicción de polimorfismos, selección de cebadores, visualización y descarga de los datos con distintos formatos.

En una primera fase, se han considerado diversos criterios para la selección de las posibles secuencias candidatos a VNTR. Su elección en la base de datos de repeticiones en tándem incluyó repeticiones iguales o superiores a 9 pb (U), repetidas por lo menos 3 veces (N) y una conservación interna igual o superior a 80% (Vergnaud y Pourcel, 2009). Se consideró igualmente la ausencia de similitud de la secuencia candidata y de sus regiones flanqueantes con secuencias pertenecientes a otros géneros taxonómicos (Monteil *et al*., 2007), además de la secuencia consenso, realizándose con ese fin un análisis BLASTN (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Las secuencias potenciales para ser VNTR se designaron como Hi (*H. influenzae* VNTRs) seguido por un número. La localización de cada Hi en el genoma de *H. influenzae* se designa como "locus". Un "alelo" corresponde a un número dado de unidades repetidas para un determinado Hi o locus.

2. Diseño de cebadores, amplificación, análisis y selección de los candidatos VNTR

El software desarrollado por Le Flèche *et al*. (2001) y Denoeud y Vergnaud (2004) permite obtener igualmente las secuencias flanqueantes de los candidatos para VNTR, 500 pb 5' y 500 pb 3' de la secuencia repetida (se puede seleccionar de 50, 100, 200 o 500 pb).

Los cebadores se diseñaron usando el programa informático, integrado en la base de datos de Le Flèche *et al*., *FastPCR* (Kalendar *et al.,* 2009). La selección de

éstos se realizó en las regiones flanqueantes del motivo repetido y alejados de éste al menos 40 pb, como sugieren Vergnaud y Pourcel (2009), por su especial importancia en la reacción de PCR multiplex. Los cebadores fueron analizados con el software Primer Blast (http://blast.ncbi.nlm.nih.gov/Blast.cgi) disponible, de forma gratuita, en el NCBI, para minimizar posibles productos de amplificación no deseados.

Los candidatos a VNTR obtenidos de este estudio se podrían ensayar en distintas cepas, siendo suficiente un número mínimo de diez (Vergnaud y Pourcel, 2006).

La amplificación de los loci de repeticiones en tándem se puede llevar a cabo mediante PCR multiplex. La separación de los productos de PCR se haría por electroforesis en gel de agarosa y visualización por tinción con bromuro de etidio o SyberSafe©, incluyéndose muestras procedentes de cepas de referencia, así como un marcador de tamaño molecular. Si se requiere una resolución más alta puede recurrirse a un gel de poliacrilamida o separación por secuenciador capilar marcando previamente los cebadores con un fluorocromo específico de cada fragmento.

La imagen del gel tras la electroforesis se puede analizar empleando un software específico como BioNumerics (Applied Maths, St. Martens Latem, Bélgica), que permite determinar el tamaño de las bandas. El número de repeticiones en los nuevos alelos se estimaría restando el tamaño de la región invariable flanqueante al tamaño del amplicón, y dividiendo por la longitud de la unidad de repetición, según lo determinado para la cepa de referencia *H. influenzae*.

La selección final de los candidatos a VNTR incluye diversos criterios descritos en Monteil *et al.*, 2007. Se excluirán los loci candidatos con menos de dos alelos diferentes; en caso de que dos VNTRs candidatos sean isomórficos, uno de ellos será excluido.

3. Validación del MLVA

Después de la selección de los candidatos sería necesario definir una colección apropiada de cepas del microorganismo a tipar. Este grupo de cepas debe haber sido caracterizado anteriormente y tipado usando los métodos de tipificación clásicos (Vergnaud y Pourcel, 2009), para que el MLVA pueda ser comparado en términos de eficacia. Frecuentemente son suficientes unas docenas de cepas para la fase de optimización del ensayo, la validez de éste se incrementa cuantas más cepas son genotipadas y comparando sus coeficientes de similitud (Vergnaud y Pourcel, 2006). Una vez validadas las parejas de cebadores específicos, esto nos permitiría su utilización para el tipaje de nuevas cepas.

4. Análisis de los datos de MLVA

Para evaluar los sistemas de tipificación epidemiológica microbiana, el Grupo de Estudio en Marcadores Epidemiológicos de la Sociedad Europea de Microbiología Clínica y Enfermedades Infecciosas recomienda que los criterios de eficacia estándar a utilizar en el nuevo esquema de MLVA, sean tipabilidad (T), reproducibilidad (R), estabilidad (S), concordancia epidemiológica (E), y poder discriminatorio (expresado como HGDI) (Struelens, 1996).

El poder discriminatorio es una característica clave de los sistemas de tipificación (Struelens, 1998). Para la evaluación del poder discriminatorio de los VNTRs seleccionados, se cuenta con el índice de diversidad de Hunter y Gaston (HGDI) (Hunter y Gaston, 1988). Este índice de discriminación para métodos de tipado se basa en la probabilidad de que dos cepas no relacionadas muestreadas de la población de prueba estén colocadas en diferentes grupos de tipado.

Se define como:

$$\mathbf{DI} = \frac{1}{N(N-1)} \sum_{j=1}^{s} nj(nj-1)$$

siendo:
N=número de aislados o cepas;
S=número total de alelos;
nj=número de aislados o cepas con el alelo j.

Según Hunter y Gaston (1988) en el desarrollo de un nuevo análisis de tipado, debe procurarse un índice discriminatorio tan elevado como sea posible, dependiendo del nivel aceptable de un número de factores, siendo conveniente un índice de más de 0,90 si se pretende que los resultados sean interpretados con confianza. Se considera un índice elevado cuando éste es superior a 0,95 (Struelens, 1996).

El índice de diversidad aplicado a los datos VNTR, es una medida de la variación del número de repeticiones en cada locus, y puede aplicarse a VNTRs individuales o combinados. Puede variar desde cero (sin diversidad) a uno (diversidad extrema). Es decir, loci con un número similar de repeticiones en cada muestra presentará un índice de diversidad inferior, mientras que cuando el número de repeticiones es diferente en casi todas las muestras tendrá un índice de diversidad muy alto (V-DIC, 2009). Los valores de índice de diversidad registrados se harán acompañar de intervalos de confianza.

La utilización del software BioNumerics permite catalogar y analizar para cada aislamiento el número de repeticiones en cada locus VNTR, y construir un dendrograma, en el que se pueden visualizar las similitudes de los elementos que lo integran, así como, los perfiles que encajan pueden ser identificados

rápidamente.

5. Gestión de datos

El resultado final del análisis consiste en los datos de tipado, expresados como el número de copias repetidas (Vergnaud y Pourcel, 2006). Estos datos se deben almacenar en bases de datos acompañados por toda la información sobre las cepas (características fenotípicas y bioquímicas, origen, información clínica o ambiental, etc.) (Vergnaud y Pourcel, 2009). Según Vergnaud y Pourcel (2006) en proyectos de pequeña dimensión, con unas pocas docenas de cepas y/o cuando la caracterización biológica de las cepas está limitada al tipado por MLVA, hay poca necesidad de un sistema de gestión de base de datos. Sin embargo, en proyectos de mayor envergadura se reclama un sistema de gestión. El software más usado para este fin es el BioNumerics (Applied Maths, Sint-Martens-Latem, Bélgica) descrito anteriormente. Es aconsejable poner los datos a disposición de los investigadores, como puede ser la base de datos del MLVAbank (http://minisatellites.upsud.fr) para el genotipado bacteriano (Vergnaud y Pourcel, 2009).

6. Colección de cepas de *Haemophilus influenzae*

Los aislados de *Haemophilus influenzae* a utilizar en el presente estudio son las *H. influenzae PittEE* y *H. influenzae PittGG* cuyos genomas completos son accesibles en la base de datos GenBank cuyos códigos de acceso son CP000671 y CP000672 respectivamente.

7. Identificación de ortólogos de las DNA polimerasas de translesión

Para identificar homólogos de las DNA polimerasas de translesión nos hemos basado en la herramienta Blast del NCBI utilizando los parámetros establecidos por defecto para la búsqueda. Como *query* se introduce la secuencia aminoacídica de la DNA polimerasa TLS y no la secuencia de DNA para evitar que cambios sinónimos en su secuencia interfieran en la identificación de los ortólogos. Se han seleccionado las DNA polimerasas TLS de *Escherichia coli* puesto que en este organismo dichas DNA polimerasas han sido ampliamente estudiadas tanto desde el punto de vista genético como bioquímico. Se utiliza la secuencia proteica de los genes *dinB, polB y umuCD* para hacer la búsqueda de DNA polimerasas TLS en otras especies. Para el estudio que abordamos en este trabajo necesitamos identificar un microorganismo que no posea dichas DNA polimerasas de translesión y que además, sea de vida libre.

La primera búsqueda planteada fue encontrar este organismo dentro de las enterobacterias. Dado que estamos interesados en genomas completamente secuenciados, no realizamos una búsqueda global con Blast en todas las enterobacterias. Para identificar qué enterobacterias están totalmente secuenciadas, realizamos una consulta en la base de datos GOLD (www.genomesonline.org). Para identificar los ortólogos de polB, dinB y umuCD, se realizaron tres búsquedas. Por un lado, en los genomas seleccionados, se realizó una búsqueda directa en el listado de genes para seleccionar aquellos ortólogos ya identificados. Por otro lado, hemos consultado las fichas Uniprot (www.uniprot.org) de *dinB, polB* y *umuCD* en *E. coli* y hemos estudiado su distribución en otros genomas. Por último, hemos realizado un análisis Blast con los parámetros previamente indicados para identificar posibles candidatos no identificados hasta la fecha. Dado que los resultados de Blast obtenidos, coincidían con los descritos en las bases de datos específicas de cada organismo,

no realizamos análisis Blast recíproco. Todo este estudio, da como resultado un listado con las enterobacterias que poseen DNA polimerasas de translesión.

IV. RESULTADOS

1. Análisis informático de secuencias repetidas de DNA (candidatos VNTR) y diseño de cebadores

1.1. Análisis de secuencias candidatas a VNTR en las cepas de *Haemophilus influenzae* PittEE y *Haemophilus influenzae* PittGG

Se ha realizado un estudio de las secuencias candidatas a VNTR de las cepas *H. influenzae* PittEE (NC_009566.1) y *H. influenzae* PittGG (NC_009567.1). La búsqueda de VNTRs se realiza con el software Bionumerics a través de la base de datos de repeticiones en tándem disponible en la web: http://minisatellites.u-psud.fr/.

Anteriormente, en la Metodología, hemos explicado la manera de hallar los VNTRs. En nuestro trabajo hemos utilizado una utilidad del servidor http://minisatellites.u-psud.fr que suministra los cebadores que flanquean a un VNTR determinado en el genoma de *H. influenzae* (http://minisatellites.u-psud.fr/ASPSamp/base_ms/marqueurs.php). Si comparamos las cepas *H. influenzae* PittEE y *H. influenzae* PittGG nos proporciona una serie de VNTRs que comparten un mismo cebador (Tabla 9).

Nombre VNTR	Cebador *reverse*	Cebador *forward*
Hi4-10	GACAGATGAAAAGAAAAGAT	TATAATATGTTTTATTACAA
Hi4-11	TAAAAATGAATACAAAAATG	AAGTTTTAACAAATCCTACA
Hi4-12	AACGGCAAGTGTTGCTTATG	CTAGTTGTTCAGAAGCATTA
Hi4-2	ATTACCTGCAATAATGACAG	TATTCAATGAACGGTAGAAT
Hi4-3	CCTCTTATATTATGTAATAT	TTTAGTTTCTTTAATGCGTA
Hi4-4	CTAGTTGTTCAGAAACATTA	TAAATGCAAGCATAGCCTAT
Hi4-5	CTAGTTGTTCAGAAACATTA	GGCAGGTGTTGCTTATGCAG
Hi4-7	CTAATTGTTCAGAAACATTA	TAAATGCAAGCACAGTCTAT

Hi4-9	AAAATGAAAAGGATCTATAC	ACTACCGCAACGGTTTTATT
Hi5-2	GTGATTTTTATCGACAATCT	TACAGAGGGCATAATTTATG
Hi6-1	TCTACAATTTCTTGTTTTTC	ATGGTGTTGGAAGAACCTGC
Hi6-3	TGACATAATCTATCCTCTTG	TAGGTATAATACGAAAAGTT

Tabla 9: Cebadores comunes para VNTRs en *H. influenzae* PittEE y PittGG

Desarrollamos cada VNTR:

Hi4-10

Cebador *reverse*	GACAGATGAAAAGAAAAGAT
Cebador *forward*	TATAATATGTTTTATTACAA

	Cepa PittEE	Cepa PittGG
Secuencia consenso	TTGC	TTGC
Longitud (pb)	4	4
Nº copias	17	13
Posición en genoma	996670-996737	199854-199905

Hi4-11

Cebador *reverse*	TAAAAATGAATACAAAAATG
Cebador *forward*	AAGTTTTAACAAATCCTACA

	Cepa PittEE	Cepa PittGG
Secuencia consenso	AATC	TTGA
Longitud (pb)	4	4
Nº copias	29.5	56.5
Posición en genoma	1057609-1057728	44507-44734

Hi4-12

Cebador *reverse*	AACGGCAAGTGTTGCTTATG
Cebador *forward*	CTAGTTGTTCAGAAGCATTA

	Cepa PittEE	Cepa PittGG
Secuencia consenso	CAAC	TTGG
Longitud (pb)	4	4
Nº copias	19.8	30
Posición en genoma	1753955-1754033	1204392-1204511

Hi4-2

Cebador *reverse*	ATTACCTGCAATAATGACAG
Cebador *forward*	TATTCAATGAACGGTAGAAT

	Cepa PittEE	Cepa PittGG
Secuencia consenso	CAAT	TTGA
Longitud (pb)	4	4
Nº copias	12.8	24.8
Posición en genoma	1532091-1532141	1447451-1447549

Hi4-3

Cebador *reverse*	CCTCTTATATTATGTAATAT
Cebador *forward*	TTTAGTTTCTTTAATGCGTA

	Cepa PittEE	Cepa PittGG
Secuencia consenso	ATCA	TTGA
Longitud (pb)	4	4
Nº copias	23.3	25.3
Posición en genoma	43964-44056	1097031-1097131

Hi4-4

Cebador *reverse*	CTAGTTGTTCAGAAACATTA
Cebador *forward*	TAAATGCAAGCATAGCCTAT

	Cepa PittEE	Cepa PittGG
Secuencia consenso	CAA	TTGG
Longitud (pb)	3	4
Nº copias	29	16
Posición en genoma	1691431-1691526	1259127-1259190

Hi4-5

Cebador *reverse*	CTAGTTGTTCAGAAACATTA
Cebador *forward*	GGCAGGTGTTGCTTATGCAG

	Cepa PittEE	Cepa PittGG
Secuencia consenso	CAAC	TTGG
Longitud (pb)	4	4
Nº copias	19.8	30
Posición en genoma	1753955-1754033	1204392-1204511

Hi4-7

Cebador *reverse*	CTAATTGTTCAGAAACATTA
Cebador *forward*	TAAATGCAAGCACAGTCTAT

	Cepa PittEE	Cepa PittGG
Secuencia consenso	CAA	TTGG
Longitud (pb)	3	4
Nº copias	29	16
Posición en genoma	1691431-1691526	1259127-1259190

Hi4-9

Cebador *reverse*	AAAATGAAAAGGATCTATAC
Cebador *forward*	ACTACCGCAACGGTTTTATT

	Cepa PittEE	Cepa PittGG
Secuencia consenso	AACC	
Longitud (pb)	4	
Nº copias	11.5	
Posición en genoma	940723-940768	

Hi5-2

Cebador *reverse*	GTGATTTTTATCGACAATCT
Cebador *forward*	TACAGAGGGCATAATTTATG

	Cepa PittEE	Cepa PittGG
Secuencia consenso	TCGTC	
Longitud (pb)	5	
Nº copias	2.6	
Posición en genoma	867539-867551	

Hi6-1

Cebador *reverse*	TCTACAATTTCTTGTTTTTC
Cebador *forward*	ATGGTGTTGGAAGAACCTGC

	Cepa PittEE	Cepa PittGG
Secuencia consenso	AGAGCC	TTCTGGCTCTTTTTGTACATC
Longitud (pb)	6	21

Nº copias	5.5	2.5
Posición en genoma	340520-340552	751170-751221

Hi6-3

Cebador *reverse*	TGACATAATCTATCCTCTTG
Cebador *forward*	TAGGTATAATACGAAAAGTT

	Cepa PittEE	**Cepa PittGG**
Secuencia consenso	TTTTAA	AAATTA
Longitud (pb)	6	6
Nº copias	3.5	3.5
Posición en genoma	67739-67759	1073307-1073327

1.2. Análisis de secuencias candidatas a VNTR en las cepas de Escherichia coli K12 substr. MG1655 y Escherichia coli O157:H7 str. EC4115

Hacemos también este análisis comparativo (Tabla 10) entre las cepas de *Escherichia coli* K12 substr. MG1655 (NC_000913.2) y *Escherichia coli* O157:H7 str. EC4115 (NC_011353.1).

Nombre VNTR	**Cebador *reverse***	**Cebador *forward***
O157-1	GAGGGATTGTTACCTTGGTC TCAAAACAATGAAAGG	GTTCCAGCCCCTTCAACCTTA GCTTATTCTGGCTC
O157-11	GACCGGCAATCATCGGG CCAACCA	GATGCTGGAAAAACTGATGC AGACTCGCGT
O157-13	GCAGCAAACGCCACAGTACC CATGCC	GTAGGTCATCTGCCGTGG TTCGAGCGCT
O157-31	GCCGAAAAACGATGCAGCTG ACTTAGGCG	GACATTTCTGCCCGGGGGTTT GTTTATTTCTGC
O157-33	GTGAAGGATAAGCTGCATT TGTCAGTGATGTCCGAAG	GCCTGACGCTAAAGATAAAG AAGAAAGCGTCGCG
O157-4	GCCAGATAAACATCCAGCA GGTCGAACGTCC	GACTCTGCGGCAATATGGCG TCTTTAGTATCTCCTG

O157-57	GCGGCGCATTAGCGTCGT ATCAGGC	CAGTTTGGCCATGCGTCTGG GGTGAC
O157-58	GACTGAGGCTGTCATCTCG AAAGAGGGCATTCT	GCGCTGGGAGGTGTCGCTC AGATGG
O157-6	GTCTTCATATTGTTTGCGATG TCCCTGATGAACTTATTGA	GTCCAGACGCCAGTGCAGCTTATTCT CCACG
O157-63	GTTTGCTGTAGCCCAGGCC GTTGATCTTCTTC	GTTCCGGCGGCGAAAGTTTC CTCGTTAG
O157-64	GACTTACTCAGCGCCGCCAA CGAAGTCC	GCACCGCACGTTTCTGAAAAA GCGTCTACT
O157-7	GGGGCGATCCCACCCTCCAT CCTG	GAGCGGCAATTGTAATCCGGTGG CTTCC
O157-8	GCTGTTCCCGTTCTTTGGC TTTACCGCC	GCGTTACGCCGCAGAACCCA CCTGC
TR4/ O157-25	GGTGATGGCTTGATATTGA	GCCACACTGCGAGTATAGAG
TR7/ O157-19	CGCAGTGATCATTATTAGC	TGCTGAAACTGACGACCAGT
Vhec2/TR6	AACCGTTATGAAAGAAAGTCCT	TCGCCCAGTAAGTATGAAATC

Tabla 10: Cebadores comunes para VNTRS en *E. coli* K12 y O157:H7

O157-1

Cebador *reverse*	GAGGGATTGTTACCTTGGTCTCAAAACAATGAAAGG
Cebador *forward*	GTTCCAGCCCCTTCAACCTTAGCTTATTCTGGCTC

	Cepa K12	**Cepa O157:H7**
Secuencia consenso	TGCTACCCCGGACGG	GGCGTTGACATGAGAGAGGCTTACCTTCCC
Longitud (bp)	15	30
Nº copias	6.5	1.9
Posición en genoma	59040-59135	63902-63958

y también

	Cepa K12	Cepa O157:H7
Secuencia consenso	ATT	ATT
Longitud (bp)	3	3
Nº copias	4.3	4.3
Posición en genoma	59154-59166	64100-64112

O157-11

Cebador *reverse*	GACCGGCAATCATCGGGCCAACCA
Cebador *forward*	GATGCTGGAAAAACTGATGCAGACTCGCGT

	Cepa K12	Cepa O157:H7
Secuencia consenso	TGCAGG	TGCAGG
Longitud (bp)	6	6
Nº copias	3.2	6.2
Posición en genoma	3985945-3985962	4781562-4781597

O157-13

Cebador *reverse*	GCAGCAAACGCCACAGTACCCATGCC
Cebador *forward*	GTAGGTCATCTGCCGTGGTTCGAGCGCT

	Cepa K12	Cepa O157:H7
Secuencia consenso	CCGCCAGCA	CCGCCAGCA
Longitud (bp)	9	9
Nº copias	2	4
Posición en genoma	3688377-3688394	4432482-4432517

O157-31

Cebador *reverse*	GCCGAAAAACGATGCAGCTGACTTAGGCG
Cebador *forward*	GACATTTCTGCCCGGGGGTTTGTTTATTTCTGC

	Cepa K12	Cepa O157:H7
Secuencia consenso	GGCGGCATG	GGCGGCATG
Longitud (bp)	9	9
Nº copias	4	4
Posición en genoma	4370653-4370688	5227081-5227116

O157-33

Cebador *reverse*	GTGAAGGATAAGCTGCATTTGTCAGTGATGTCCGAAG
Cebador *forward*	GCCTGACGCTAAAGATAAAGAAGAAAGCGTCGCG

	Cepa K12	Cepa O157:H7
Secuencia consenso	AAAGTGCTATGCAGTAA	AAAGTGCTATGCAGTAA
Longitud (bp)	17	17
Nº copias	2.2	3.2
Posición en genoma	4437819-4437856	5295259-5295313

O157-4

Cebador *reverse*	GCCAGATAAACATCCAGCAGGTCGAACGTCC
Cebador *forward*	GACTCTGCGGCAATATGGCGTCTTTAGTATCTCCTG

	Cepa K12	Cepa O157:H7
Secuencia consenso	-	GCACCTCATTGTTGTCGGCGCTCTCTGTGTGGA
Longitud (bp)	-	33
Nº copias	-	2.3
Posición en genoma	1236376-1236673	1678733-1678808

O157-57

Cebador *reverse*	GCGGCGCATTAGCGTCGTATCAGGC
Cebador *forward*	CAGTTTGGCCATGCGTCTGGGGTGAC

	Cepa K12	Cepa O157:H7
Secuencia consenso	-	GAGCCG
Longitud (bp)	-	6
Nº copias	-	4.3
Posición en genoma	1039777-1040024	1219809-1219834

O157-58

Cebador *reverse*	GACTGAGGCTGTCATCTCGAAAGAGGGCATTCT
Cebador *forward*	GCGCTGGGAGGTGTCGCTCAGATGG

	Cepa K12	Cepa O157:H7
Secuencia consenso	GATTGCCG	GATTGCCG
Longitud (bp)	8	8
Nº copias	2.4	2.4
Posición en genoma	1058736-1058754	1237866-1237884

O157-6

Cebador *reverse*	GTCTTCATATTGTTTGCGATGTCCCTGATGAACTTATTGA
Cebador *forward*	GTCCAGACGCCAGTGCAGCTTATTCTCCACG

	Cepa K12	Cepa O157:H7
Secuencia consenso	-	-
Longitud (bp)	-	-
Nº copias	-	-
Posición en genoma	3623107-3623342	664796-665031

O157-63

Cebador *reverse*	GTTTGCTGTAGCCCAGGCCGTTGATCTTCTTC
Cebador *forward*	GTTCCGGCGGCGAAAGTTTCCTCGTTAG

	Cepa K12	Cepa O157:H7
Secuencia consenso	AGCGCC	AGCGCC
Longitud (bp)	6	6
Nº copias	3.3	3.3
Posición en genoma	412078-412097	474838-474857

O157-64

Cebador *reverse*	GACTTACTCAGCGCCGCCAACGAAGTCC
Cebador *forward*	GCACCGCACGTTTCTGAAAAAGCGTCTACT

	Cepa K12	Cepa O157:H7
Secuencia consenso	CGACTT	CGACTT
Longitud (bp)	6	6
Nº copias	3.2	3.2
Posición en genoma	3449535-3449553	4186525-4186543

O157-7

Cebador *reverse*	GGGGCGATCCCACCCTCCATCCTG
Cebador *forward*	GAGCGGCAATTGTAATCCGGTGGCTTCC

	Cepa K12	Cepa O157:H7
Secuencia consenso	ACCACGCTGGCTA	ACTCGCTGGCAAGAACTCTGCCGTCTGGCAG CACCAGGAGTGGTGTAATGACCACGCGCCTG
Longitud (bp)	13	62
Nº copias	3.2	2.2
Posición en genoma	2014581-2014619	2640991-2641124

O157-8

Cebador *reverse*	GCTGTTCCCGTTCTTTGGCTTTACCGCC
Cebador *forward*	GCGTTACGCCGCAGAACCCACCTGC

	Cepa K12	Cepa O157:H7
Secuencia consenso	TGCCGGATGCTGAT	GCCGGATGCTGATC
Longitud (bp)	14	14
Nº copias	2.7	2.6
Posición en genoma	2579655-2579692	3297758-3297794

TR4 u O157-25

Cebador *reverse*	GGTGATGGCTTGATATTGA
Cebador *forward*	GCCACACTGCGAGTATAGAG

	Cepa K12	Cepa O157:H7
Secuencia consenso	-	TGCAAA
Longitud (bp)	-	6
Nº copias	-	4.3
Posición en genoma	1160727-1161059	1520733-1520758

TR7 u O157-19

Cebador *reverse*	CGCAGTGATCATTATTAGC
Cebador *forward*	TGCTGAAACTGACGACCAGT

O157-19:

	Cepa K12	Cepa O157:H7
Secuencia consenso	CATCATCACGATCACGAA	CACGAACATCAT
Longitud (bp)	18	12
Nº copias	3.2	6.8
Posición en genoma	2184320-2184376	2861992-2862069

y TR7:

	Cepa K12	Cepa O157:H7
Secuencia consenso	CAT	ACCACG
Longitud (bp)	3	6
Nº copias	21.7	11.7
Posición en genoma	2184305-2184369	2862005-2862074

Vhec2 o TR6 u O157-34

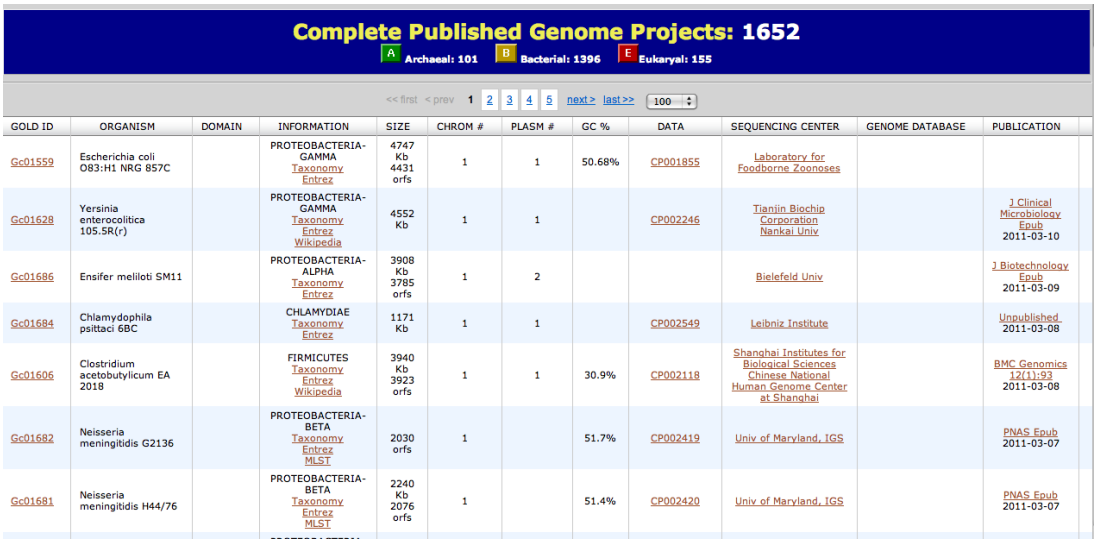
Cebador *reverse*	AACCGTTATGAAAGAAAGTCCT
Cebador *forward*	TCGCCCAGTAAGTATGAAATC

	Cepa K12	Cepa O157:H7
Secuencia consenso	TTAAATAATCCACAGGAG	TTAAATAATCTGCAGAAGTTAAATAAT ATACAGAAGTTAAATAATATACAGGAG
Longitud (bp)	18	54
Nº copias	12.4	3.5
Posición en genoma	4474015-4474238	5331530-5331718

2. Distribución de las DNA polimerasas de translesión en Enterobacteriaceae y Pasteurellaceae

Para identificar la presencia de genes homólogos a las DNA polimerasas de translesión *dinB*, *polB* y *umuCD* en enterobacterias hemos seleccionado, en primer lugar, los genomas secuenciados completamente a día de hoy. Esta información la hemos obtenido a partir de la base de datos GOLD (Figura 11).

La base de datos GOLD (Genomes OnLine Database) es un recurso de Internet que permite el acceso integral a la información sobre los proyectos de genomas completos y permanentes, así como metagenomas y metadatos, en todo el mundo.



Complete Published Genome Projects: 1652

A Archaeal: 101 B Bacterial: 1396 E Eukaryal: 155

<< first < prev 1 2 3 4 5 next > last >> 100

GOLD ID	ORGANISM	DOMAIN	INFORMATION	SIZE	CHROM #	PLASM #	GC %	DATA	SEQUENCING CENTER	GENOME DATABASE	PUBLICATION
Gc01559	Escherichia coli O83:H1 NRG 857C		PROTEOBACTERIA-GAMMA Taxonomy Entrez	4747 Kb 4431 orfs	1	1	50.68%	CP001855	Laboratory for Foodborne Zoonoses		
Gc01628	Yersinia enterocolitica 105.5R(r)		PROTEOBACTERIA-GAMMA Taxonomy Entrez Wikipedia	4552 Kb	1	1		CP002246	Tianjin Biochip Corporation Nankai Univ		J Clinical Microbiology Epub 2011-03-10
Gc01686	Ensifer meliloti SM11		PROTEOBACTERIA-ALPHA Taxonomy Entrez	3908 Kb 3785 orfs	1	2			Bielefeld Univ		J Biotechnology Epub 2011-03-09
Gc01684	Chlamydophila psittaci 6BC		CHLAMYDIAE Taxonomy Entrez	1171 Kb	1	1		CP002549	Leibniz Institute		Unpublished 2011-03-08
Gc01606	Clostridium acetobutylicum EA 2018		FIRMICUTES Taxonomy Entrez Wikipedia	3940 Kb 3923 orfs	1	1	30.9%	CP002118	Shanghai Institutes for Biological Sciences Chinese National Human Genome Center at Shanghai		BMC Genomics 12(1):93 2011-03-08
Gc01682	Neisseria meningitidis G2136		PROTEOBACTERIA-BETA Taxonomy Entrez MLST	2030 orfs	1		51.7%	CP002419	Univ of Maryland, IGS		PNAS Epub 2011-03-07
Gc01681	Neisseria meningitidis H44/76		PROTEOBACTERIA-BETA Taxonomy Entrez MLST	2240 Kb 2076 orfs	1		51.4%	CP002420	Univ of Maryland, IGS		PNAS Epub 2011-03-07

Figura 11: Base de datos GOLD

En función de este análisis, ofrecemos un listado de los genomas de enterobacterias que están secuenciados completamente:

- *Buchnera aphidicola*
- *Candidatus Riesia pediculicola*
- *Citrobacter rodentium, Citrobacter koseri*
- *Cronobacter turicensis, Cronobacter sakazakii*
- *Dickeya dadantii, Dickeya zea*
- *Edwardsiella tarda, Edwardsiella ictaluri*
- *Enterobacter cloacae, Enterobacter sp.*
- *Erwinia billingiae, Erwinia amylovora, Erwinia pyrifoliae, Erwinia tasmaniensis*

- *Escherichia coli, Escherichia fergusonii*
- *Klebsiella variicola, Klebsiella pneumoniae*
- *Pantoea sp., Pantoea vagans, Pantoea ananatis*
- *Pectobacterium wasabiae, Pectobacterium carotovorum, Pectobacterium atrosepticum*
- *Photorhabdus asymbiotica asymbiotica, Photorhabdus luminescens laumondii*
- *Proteus mirabilis*
- *Providencia alcalifaciens*
- *Rahnella sp.*
- *Salmonella enterica*
- *Serratia proteamaculans*
- *Shigella flexneri, Shigella boydii, Shigella sonnei, Shigella dysenteriae*
- *Sodalis glossinidius morsitans*
- *Wigglesworthia glossinidia*
- *Xenorhabdus nematophila, Xenorhabdus bovienii*
- *Yersinia enterocolitica, Yersinia pestis, Yersinia pseudotuberculosis*

Una vez que tenemos todas las entereobacterias con genomas secuenciados completamente, accedemos a sus secuencias genómicas en la base de datos de GenBank, accesible desde la web del NCBI (National Center for Biotechnology Information) a través del enlace www.ncbi.nlm.nih.gov/genbank/. El NCBI almacena y actualiza, constantemente, la información referente a secuencias genómicas en GenBank, gran cantidad de artículos científicos sobre biomedicina, biotecnología, bioquímica, genética y genómica en PubMed, una recopilación de enfermedades genéticas humanas en OMIM, y otros datos biotecnológicos de gran relevancia en diferentes bases de datos. También tiene herramientas

bioinformáticas para el análisis de secuencias de DNA, RNA y proteínas, como por ejemplo la herramienta BLAST utilizada en este trabajo.

En base a los genomas seleccionados en GOLD, obtenemos la información almacenada en GenBank a través del enlace www.ncbi.nlm.nih.gov/sites/genome En la Figura 12 se especifica, como ejemplo, la información obtenida para el genoma de *Buchnera aphidicola.*

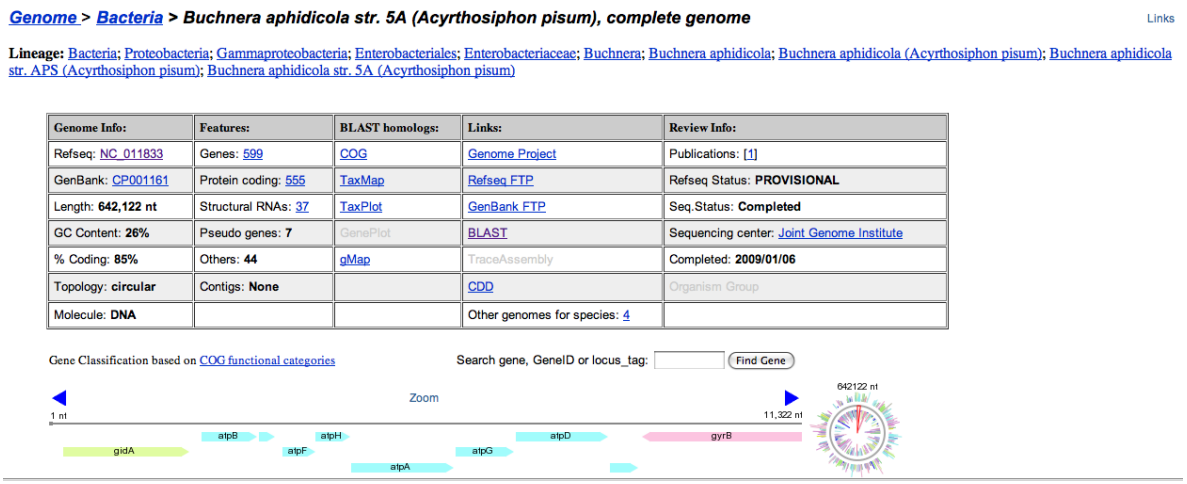


Figura 12: Tabla obtenida en la base de datos GenBank a través de *Genome* de *Buchnera aphidicola*

En esta tabla encontramos diversa información sobre cada enterobacteria, como los identificadores, longitud del genoma, contenido en C+G, topología, genes,... y también referencias a diferentes bases de datos. Como se indicó en el apartado de Metodología, consultamos en el listado de genes la presencia de ortólogos de los genes *polB*, *dinB* y *umuCD* de *E. coli* previamente identificados. Además, realizamos consultas BLAST dentro de la misma página para encontrar semejanzas lejanas no descritas entre la secuencia genómica y las secuencias de

las proteínas de translesión mencionadas. En conjunto, esta aproximación nos permite identificar qué enterobacterias poseen las DNA polimerasas II, IV y V.

Las DNA polimerasas de translesión están conservadas a lo largo de la evolución, lo que nos permite, al ser analizadas con esta herramienta, conocer su presencia en los diversos organismos y como consecuencia, identificar los organismos que las conservan y las que no.

Se ha elaborado la siguiente tabla en la que se muestra la distribución de las DNA polimerasas de translesión y sus identificadores en el genoma de cada enterobacteria, así como el identificador de cada especie.

Especie	Identificador	Identificadores de los TLS		
		dinB	polB	umuCD
***Escherichia coli* str. K12**	NC_000913	NP_414766	NP_414602	NP_415702
Escherichia fergusonii	NC_011740	YP_002383855	YP_002381299	YP_002383050
***Buchnera aphidicola* str. 5A**	NC_011833	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*
Candidatus Riesia pediculicola	NC_014109	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*
***Citrobacter rodentium* ICC168**	NC_013716	YP_003363971	YP_003363712	No encontrada secuencia con homología significativa*
***Citrobacter koseri* ATCC BAA-895**	NC_009792	YP_001454501	YP_001454841	YP_001453404
***Cronobacter turicensis* z3032**	NC_013282	YP_003209209	YP_003209054	YP_003210129
***Cronobacter sakazakii* ATCC BAA-894**	NC_009778	YP_001439181	YP_001439338	No encontrada secuencia con homología significativa*
***Dickeya dadantii* 3937**	NC_014500	YP_003884298	YP_003884640	No encontrada secuencia con

				homología significativa*
***Dickeya zea* Ech1591**	NC_012912	YP_003003252	YP_003002931	No encontrada secuencia con homología significativa*
Edwardsiella tarda	NC_013508	YP_003294844	YP_003294665	No encontrada secuencia con homología significativa*
***Edwardsiella ictaluri* 93-146**	NC_012779	YP_002932356	YP_002932160	No encontrada secuencia con homología significativa*
Enterobacter cloacae	NC_014618	YP_003943011	YP_003943190	YP_003941425 YP_003941656 ambas 100%
***Enterobacter* sp. 638**	NC_009436	YP_001175499	YP_001175345	YP_001176938
***Erwinia billingiae* Eb661**	NC_014306	YP_003740269	YP_003740089	YP_003741973 YP_003739321
***Erwinia amylovora* Ea273**	NC_013971	YP_003537982	YP_003537763	YP_003539119
***Erwinia pyrifoliae* Ep1/96**	NC_012214	YP_002649707	YP_002647748	YP_002648537
***Erwinia tasmaniensis* Et1/99**	NC_010694	YP_001908522	YP_001906675	YP_001907390
***Klebsiella variicola* At-22**	NC_013850	YP_003441049	YP_003441233	YP_003438756
***Klebsiella pneumoniae* NTUH-K2044**	NC_012731	YP_002917958	YP_002917769	YP_002918971
***Pantoea* sp. At-9b.**	NC_014837	YP_004114698	YP_004114521	YP_004115462
***Pantoea vagans* C9-1**	NC_014562	YP_003929936	YP_003929767	YP_003930853
***Pantoea ananatis* LMG 20103**	NC_013956	YP_003519175	YP_003518985	YP_003520059
***Pectobacterium wasabiae* WPP163**	NC_013421	YP_003260782	YP_003261160	No encontrada secuencia con homología significativa*
***Pectobacterium carotovorum* PC1**	NC_012917	YP_003018843	YP_003019179	No encontrada secuencia con homología significativa*
***Pectobacterium atrosepticum* SCRI1043**	NC_004547	YP_051558	YP_051940	No encontrada secuencia con homología significativa*

***Photorhabdus asymbiotica asymbiotica* ATCC 43949**	NC_012962	YP_003042052	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*
***Photorhabdus luminescens laumondii* TTO1**	NC_005126	NP_928550	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*
***Proteus mirabilis* HI4320**	NC_010554	YP_002150135	YP_002152045	YP_002152203
Providencia alcalifaciens	NZ_ABXW01000000	ZP_03317509	ZP_03319332	ZP_03319361
***Rahnella* sp. Y9602**	NC_015061	YP_004211708	YP_004214474	YP_004215593
Salmonella enterica* subsp. *enterica	NC_012125	YP_002635952	YP_002635738	YP_002637304
***Serratia proteamaculans* 568**	NC_009832	YP_001477196	YP_001476965	YP_001478835
***Shigella flexneri* 5 str. 8401**	NC_008258	YP_687878	YP_687644	YP_688703
***Shigella boydii* CDC 3083-94**	NC_010658	YP_001879027	YP_001878871	No encontrada secuencia con homología significativa*
***Shigella sonnei* Ss046**	NC_007384	YP_309293	YP_309096	No encontrada secuencia con homología significativa*
***Shigella dysenteriae* Sd197**	NC_007606	YP_402179	No encontrada secuencia con homología significativa*	YP_402864
Sodalis glossinidius morsitans	NC_007712	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*
Wigglesworthia glossinidia	NC_004344	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*
***Xenorhabdus nematophila* ATCC19061**	NC_014228	YP_003711526	YP_003714169	No encontrada secuencia con homología significativa*
***Xenorhabdus bovienii* SS-2004**	NC_013892	YP_003469184	YP_003467701	No encontrada

				secuencia con homología significativa*
Yersinia enterocolitica* subsp. *enterocolitica	NC_008800	YP_001007383	YP_001004989	YP_001006016
Yersinia pestis Angola	NC_010159	YP_001607653	YP_001607324	No encontrada secuencia con homología significativa*
***Yersinia pseudotuberculosis* YPIII**	NC_010465	YP_001722018	YP_001722266	No encontrada secuencia con homología significativa*

* Se descartan candidatos por no tener un mínimo de homología significativa al hacer Blast.

En los genomas de *Buchnera aphidicola*, *Candidatus Riesia pediculicola*, *Sodalis glossinidius morsitans* y *Wigglesworthia glossinidia* no existen ninguna de las tres polimerasas mutadoras, pero no podemos incluirlas en nuestro estudio dado que ninguna de ellas es de vida libre. Por lo tanto, continuamos la búsqueda en las Pasteurellaceae de un candidato que posea los criterios que buscamos, es decir, que no tengan DNA polimerasas de translesión y que sean de vida libre.

Hacemos lo mismo para Pasteurellaceae y vemos la distribución de las TLS:

Especie	Identificador	Identificadores de los TLS		
		Din B	Pol B	UmuCD
***Aggregatibacter actinomycetemcomitans* D11S-1**	NC_013416	YP_003255548	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*
***Aggregatibacter aphrophilus* NJ8700**	NC_012913	YP_003007457	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*
Actinobacillus pleuropneumoniae	NC_010939	YP_001968890	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*
Actinobacillus	NC_009655	YP_001344746	No encontrada	No encontrada

***succinogenes* 130Z**			secuencia con homología significativa*	secuencia con homología significativa*
***Haemophilus somnus* 2336**	NC_010519	YP_001784729	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*
***Mannheimia succiniciproducens* MBEL55E**	NC_006300	YP_088327	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*
***Pasteurella multocida* Pm70**	NC_002663	NP_245404	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*
***Haemophilus influenzae* F3031**	NC_014920	No encontrado gen con homología significativa*	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*
***Haemophilus parasuis* SH0165**	NC_011852	YP_002475392	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*
***Haemophilus ducreyi* 35000HP**	NC_002940	No encontrado gen con homología significativa*	No encontrada secuencia con homología significativa*	No encontrada secuencia con homología significativa*

* Se descartan candidatos por no tener un mínimo de homología significativa al hacer Blast

Encontramos que *Haempophilus influenzae* carece de DNA polimerasas de translesión y es un organismo de vida libre.

3. Distribución de secuencias repetidas en *E. coli* y *H. influenzae*

Para la comparación de la distribución de repeticiones entre *E. coli* y *H. influenzae* hemos accedido al estudio realizado por Treangen y colaboradores (Treangel *et al*., 2009) sobre repeticiones presentes en 659 genomas procariotas, accesible en la dirección http://wwwabi.snv.jussieu.fr/public/Repeatoire.

Este estudio fue efectuado con el programa procrastAligner disponible en la web http://alggen.lsi.upc.es/recerca/align/procrastination/. Este programa genera por defecto una salida en formato XFMA. Dado que nuestro interés era obtener un listado de distribución de repeticiones en formato XML, volvimos a ejecutar el programa generando una salida en formato XML con la siguiente línea de comandos:

```
./procrastAligner-linux --sequence=prueba.fasta --xml=salida
```

El fichero obtenido de esta forma es del tipo:

```
<procrastAlignment sequence="HInfluenzaPittEE.fasta">

<localAlignment id = "1" length = "23" multiplicity = "35" spscore="716336">

<component       id="1"       seqid="1"       leftend="90190"       length="23"
orientation="0">GCGGTAAATTTTTAATGTGTTTT      </component>

<component       id="2"       seqid="1"       leftend="156124"       length="23"
orientation="0">GCGGTTATTTTTTAGTGTGTTTT      </component>

<component       id="3"       seqid="1"       leftend="185230"       length="23"
orientation="0">GCGGTCAATTTTTACGGTCTTTT      </component>

...

</localAlignment>



<localAlignment id = "2" length = "23" multiplicity = "21" spscore="244201">

<component       id="1"       seqid="1"       leftend="13413"       length="23"
orientation="0">GCGGTAAATTTTAATCTAGTTTT      </component>

<component       id="2"       seqid="1"       leftend="65151"       length="23"
orientation="0">GCGGTCAAATTTAAATTAGTTTT      </component>

<component       id="3"       seqid="1"       leftend="114888"       length="23"
orientation="0">GCGGTAAATTTTTAGTTTGTTTT      </component>

...
```

donde:

- id es el identificador de la repetición (las repeticiones están ordenadas de la más a la menos frecuente)
- length es la longitud
- multiplicity es la frecuencia

A continuación se señalan una serie de líneas encabezadas por "component", que indica todos los sitios donde aparece esa repetición, así como la secuencia de nucleótidos de la repetición.

En este estudio hemos seleccionado las repeticiones entre 1 y 100 nucleótidos, repetidos con una frecuencia entre 5 y 100 veces. Para realizar este filtrado hemos utilizado la instrucción:

```
./procrastAligner-linux --sequence genome.fasta --l 1 --rmin
5 --rmax 100 --xml repeticiones.xml
```

A partir del archivo xml generado, extraemos las cabeceras, con el comando linux "grep":

```
>grep '<localAlignment' repeticiones.xml > cabeceras.txt
```

Dado que el programa no permite fijar un valor máximo de longitud de repetición, seleccionamos, en la tabla ordenada por longitud en Excel, aquellos que tengan valores igual o inferior a 100 nucleótidos.

Los resultados obtenidos se presentan en las Figura 13 a Figura 17.

La distribución de repeticiones en *Escherichia coli* K12 la podemos observar en la Figura 13.

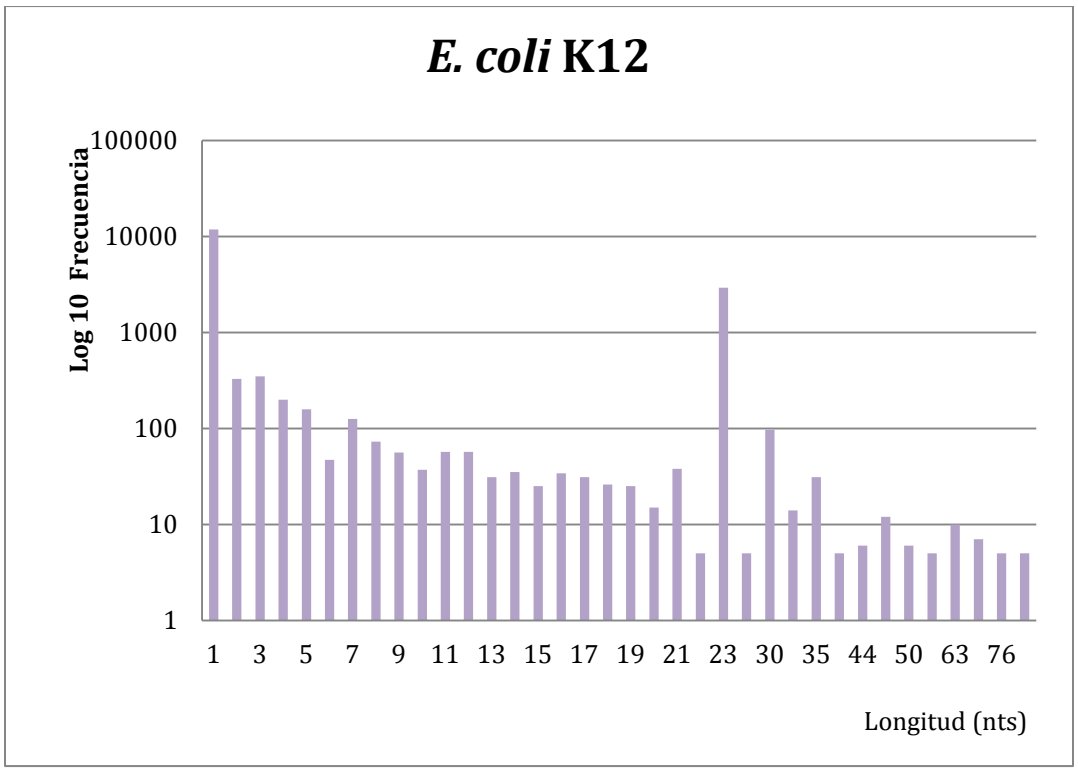


Figura 13: Distribución de repeticiones en *E. coli* K12

La distribución de las repeticiones en *Escherichia coli* O157:H7 la tenemos en la Figura 14.

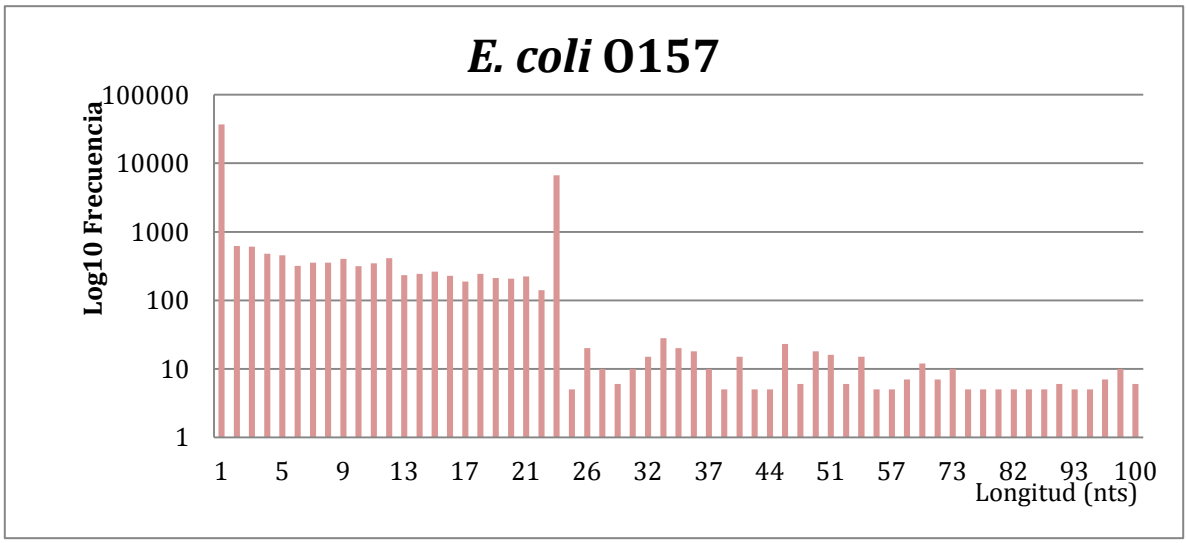


Figura 14: Distribución de repeticiones en *E. coli* O157

La distribución de repeticiones en *Haemophilus influenzae* PittEE es la que observamos en la Figura 15.

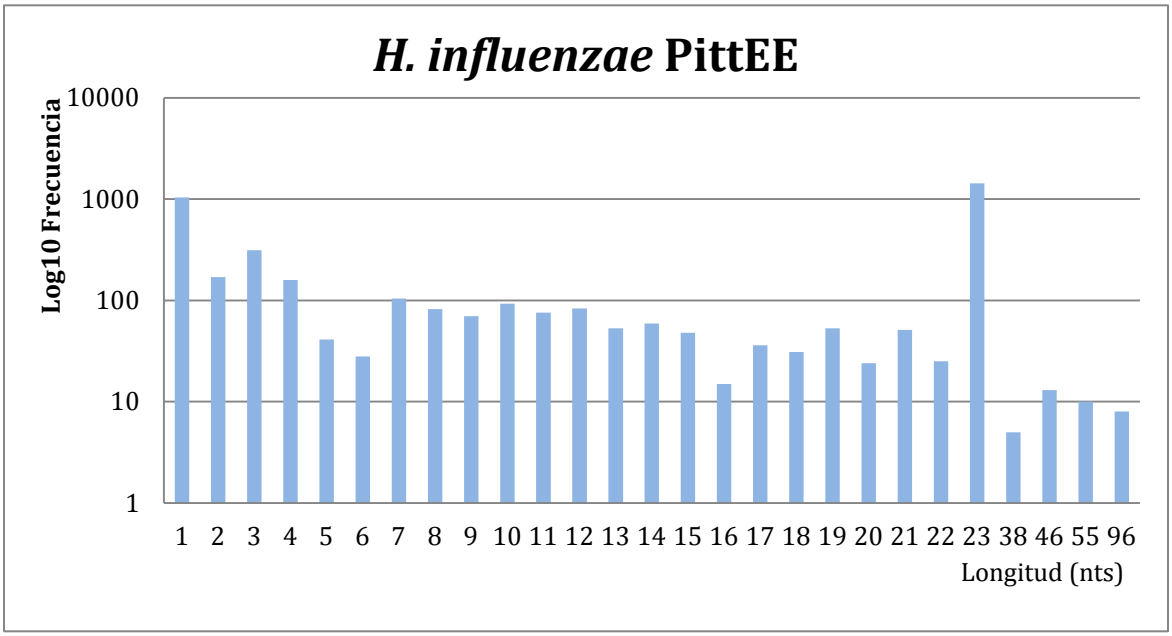


Figura 15: Distribución de repeticiones en *H. influenzae* PittEE

La distribución de repeticiones en *Haemophilus influenzae* PittGG es la que encontramos en la Figura 16.

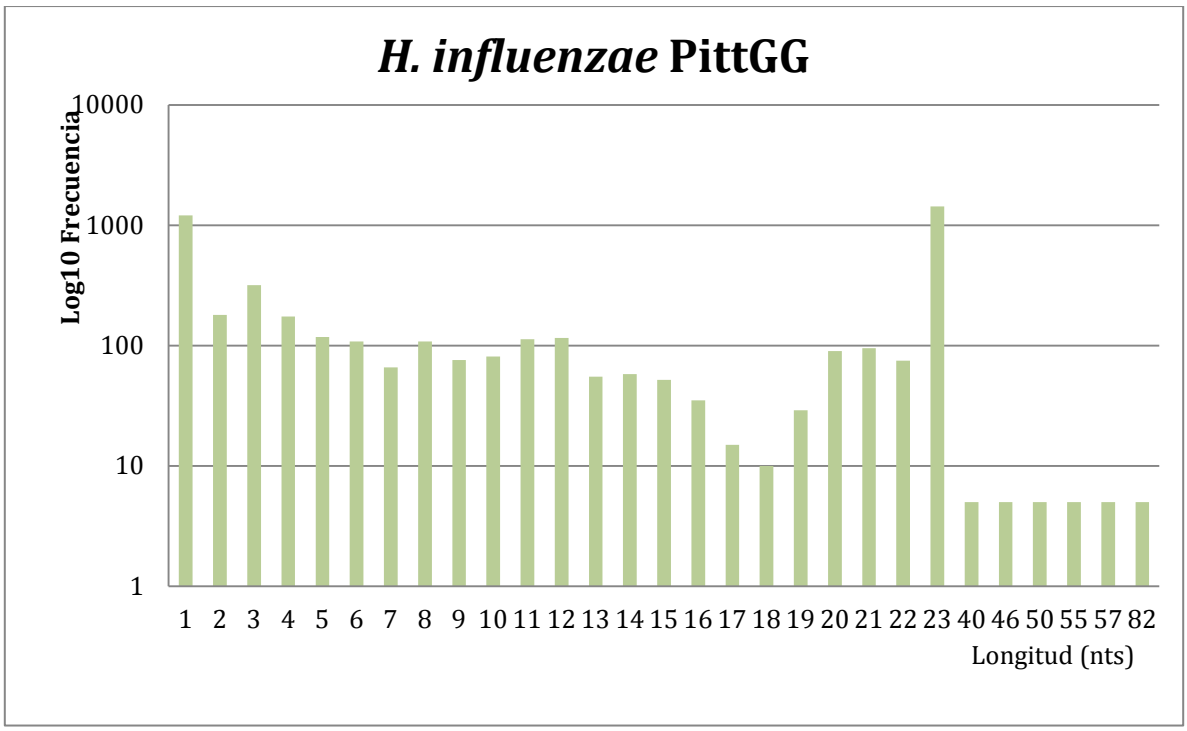


Figura 16: Distribución de repeticiones en *H. influenzae* PittGG

Si representamos los datos de la distribución de repeticiones (Figura 17) de las cuatro especies vemos que en *Escherichia coli* existe un mayor número de repeticiones de un nucleótido que con respecto a *Haemophilus influenzae*. Por otro lado, detectamos un aumento en el número de repeticiones cuya longitud es de 23 nucleótidos en los cuatro genomas, aunque *E. coli* tiene mayor cantidad de repeticiones que *H. influenzae*. Desconocemos las causas o consecuencias biológicas de este aumento. A partir de esta longitud se detecta una disminución en el número de repeticiones en los cuatro genomas seleccionados, siendo en general la frecuencia inferior a 10 (Figura 17).

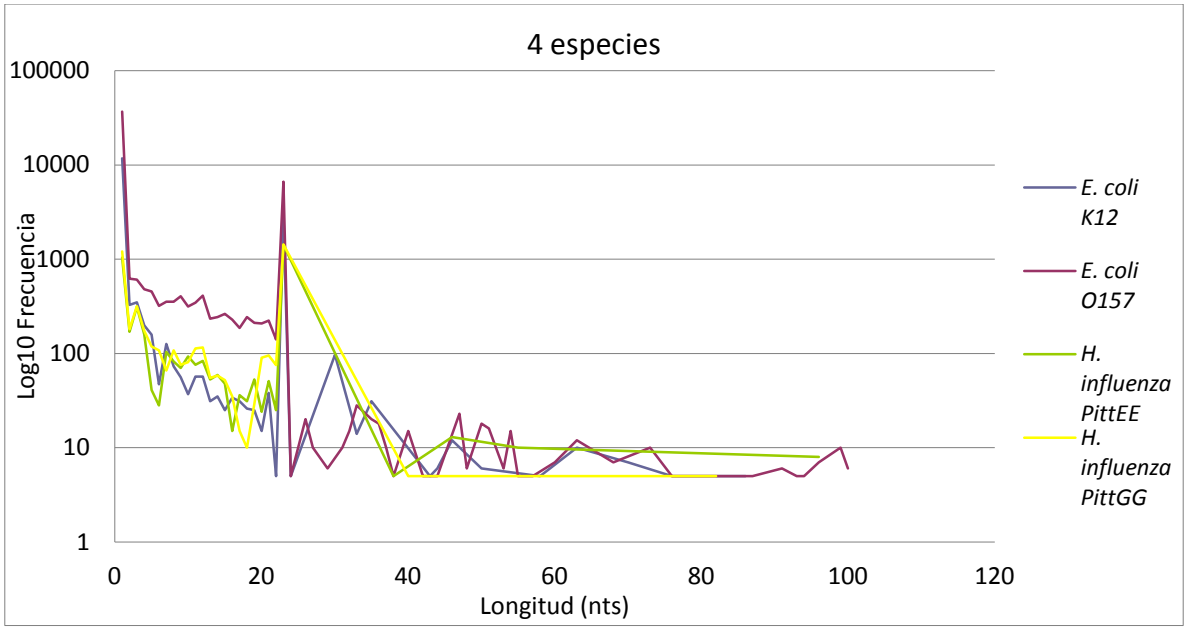


Figura 17: Distribución de repeticiones en *E. coli* K12, *E. coli* O157, *H. influenzae* PittEE y *H. influenzae* PittGG

V. DISCUSIÓN

Las DNA polimerasas de translesión actúan en la célula cuando hay un daño en el DNA que escapa al sistema de reparación (BER o MMR) y que causan el bloqueo de la DNA polimerasa replicativa. En estos casos, se produce un reemplazamiento de estas DNA polimerasas, por lo que una pequeña región del DNA será replicado por estas DNA polimerasas. La longitud de la región replicada por dichas DNA polimerasas de translesión viene determinada por la baja procesividad de dichas DNA polimerasas. Dado que estas DNA polimerasas causan pequeñas expansiones o deleciones de repeticiones en el DNA, cabe pensar que su actuación podría dejar una "huella" en el DNA en los organismos que las posean.

En este sentido, la finalidad de este estudio es identificar si hay una posible correlación entre la presencia de las DNA polimerasas de translesión y la riqueza en secuencias repetidas en los genomas de las bacterias que posean o carezcan de dichas DNA polimerasas en su genoma. Para ello hemos realizado un estudio encaminado a la identificación de genes ortólogos de las tres DNA polimerasas de translesión identificadas en *E. coli*, a saber, *polB*, *dinB* y *umuCD*. De esta forma, identificamos que *H. influenzae* carece de dichas DNA polimerasas y, por lo tanto, podría ser un buen ejemplo para realizar el análisis.

El primer bloque de este estudio está orientado al diseño de combinaciones de cebadores específicos de cara al tipado de cepas de *Haemophilus influenzae*.

En el segundo bloque, hemos estudiado la presencia o no de las DNA polimerasas de translesión en los genomas de Enterobacteriaceae y de Pasteurellaceae.

En el tercer bloque, hemos hecho un estudio de la distribución de las repeticiones de DNA en el genoma de *Haemophilus influenzae*. En varias cepas de *H. influenzae* se han identificado microsatelites asociados al promotor o a la región

codificante 5' de varios loci de contingencia que están relacionados con la virulencia. La repetición del dinucleótido TA, está en la región promotora del locus pilin (Van Ham *et al*., 1993). Las repeticiones de heptanucleótidos se asocian con genes de la adhesina HMW (Dawid *et al.*, 1999). Las repeticiones de tetranucleótidos se encuentran en los loci cuyas funciones son la adquisición de hierro, la biosíntesis de liposacáridos y un gen de restricción-modificación (Hood *et al*., 1996).

Los factores que afectan a la inestabilidad de microsatélites en *H. inluenzae* pueden ser el número de repeticiones y los mecanismos de reparación de errores (De Bolle *et al*., 2000; Bayliss *et al*., 2002, 2004, 2005). Por otra parte, en *E. coli*, es conocido que la inducción de la respuesta SOS afecta a la inestabilidad de secuencias repetidas (Morel et al, 1998). En *E. coli*, la activación de SOS aumenta la frecuencia de mutación ya que inducen las polimerasas de translesión: Pol II, Pol IV y Pol V, codificadas por los genes polB, dinB y umuCD respectivamente (Napolitano et al., 2000; Wagner et al., 2002). Sin embargo, al contrario de lo que sucede en *E. coli,* en el estudio de Notani y Setlow de 1980, vieron que en *H. influenzae* la inducción de SOS por exposición a la luz ultravioleta no aumentaba la frecuencia de mutación. En el estudio realizado en este trabajo hemos confirmado que el genoma de *H. influenzae* (Apartado 2, Distribución de las DNA polimerasas de translesión en Enterobacteriaceae y Pasteurellaceae) carece de las DNA polimerasas de translesión.

La ausencia de estas DNA polimerasas de translesión en el genoma de *H. influenzae* puede estar relacionada con una mayor estabilidad de las repeticiones durante la respuesta SOS (Sweetman, W. A., Moxon, E. R. y Bayliss, C. D., 2005).

El objetivo final que nos propusimos era estudiar una posible correlación entre la presencia de estas DNA polimerasas de translesión y la riqueza en secuencias repetidas en los genomas de las bacterias analizadas. En la consecución de este

proyecto hemos podido observar que efectivamente hay una mayor cantidad de repeticiones en el genoma de *Escherichia coli*, que posee DNA polimerasas de translesión, que en el de *Haemophilus influenzae*, que carece de estas DNA polimerasas de translesión.

Sin embargo, factores como el tipo de repetición, longitud o región en la que se localiza dicha repetición puede afectar a dicha inestabilidad, por lo que los datos obtenidos en este trabajo deberían ser contrastados empíricamente.

VI. BIBLIOGRAFÍA

- Balasingham, A., 2008. "Development of multi locus variable number tandem repeat analysis (MLVA) for the genotyping of *Legionella pneumophila* isolated from various habitats". Thesis for the Master's degree in Molecular Biosciences, Faculty of Mathematics and Natural Sciences, University of Oslo, Noruega.

- Bayliss, C. D., Field, D., Moxon, E. R., 2001. "The simple sequence contingency loci of *Haemophilus influenzae* and *Neisseria meningitidis*". J. Clin. Invest., 107(6): 657-662.

- Bayliss, C. D., Sweetman, W. A., Moxon, E. R., 2004. "Mutations in *Haemophilus influenzae* mismatch repair genes increase mutation rates of dinucleotide repeat tracts but not dinucleotide repeat-driven pilin phase variation rates". J. Bacteriol., 186(10): 2928-2935.

- Bayliss, C. D., Sweetman, W. A., Moxon, E. R., 2005. "Destabilization of tetranucleotide repeats in *Haemophilus influenzae* mutants lacking RnaseHI or the Klenow domain of PolI". Nucleic Acids Res., 33(1): 400-408.

- Bayliss, C. D., Van de Ven, T., Moxon, E. R., 2002. "Mutations in poll but not mutSLH destabilize *Haemophilus influenzae* tetranucleotide repeats". EMBO J., 21(6): 1465-1476.

- Becherel, O. J. and Fuchs, R. P., 2001. "Mechanism of DNA polymerase II-mediated frameshift mutagenesis". Proc. Natl. Acad. Sci. U S A, 98(15): 8566-8571.

- Benson, G., 1999. "Tandem repeats finder: a program to analyze DNA sequences". Nucleic Acids Res., 27(2): 573-580.

- Buntin, K. A., Roe, S. M., Pearl, L. H., 2003. "Structural basis for recruitment of traslesion DNA polymerase Pol IV/DinB to the beta-clamp". EMBO J., 22(21): 5883-5892.

- Buschiazzo, E., Gemmell, N. J., 2006. "The rise, fall and renaissance of microsatellites in eukaryotic genomes". Bioessays, 28(10): 1040-1050.

- Cesar, M. F. (2008). "Ocorrência de *Ehrlichia canis* em cães sintomáticos atendidos no hospital veterinário da Universidade de Brasília e análise de variabilidade em regiões genômicas de repetição". Dissertação de Mestrado em Saúde Animal, Universidade de Brasília.

- Chambers, G. K., MacAvoy, E. S., 2000. "Microsatellites: consensus and controversy". Comp. Biochem. Physiol. B. Biochem. Mol. Biol., 126(4): 455-476.

- Chang, C. H., Chang, Y. C., Underwood, A., Chiou, C. S., Kao, C. Y., 2007. "VNTRDB: a bacterial variable number tandem repeat locus database". Nucleic Acids. Res., 35(Database issue): D416-D421.

- Coletta-Filho, H. D., Takita. M. A., De Souza, A. A., Aguilar-Vildoso, C. I., Machado, M. A., 2001. "Differentiation of strains of *Xylella fastidiosa* by a variable number of tandem repeat analysis". Appl. Environ. Microbiol., 67(9): 4091-4095.

- Dawid, S., Barenkamp, S. J., St Geme, J. W., 1999. "Variation in expression of the *Haemophilus influenzae* HMW adhesins: a prokaryotic system reminiscent of eukaryotes". Proc. Natl. Acad. Sci. U S A, 96(3): 1077-1082.

- De Bolle, X., Bayliss, C. D., Field, D., Van de Ven, T., Saunders, N. J., Hood, D. W., Moxon, E. R., 2000. "The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gen with homology to type III DNA methyltransferases". Mol. Microbiol., 35(1): 211-222.

- Denoeud, F., Vergnaud, G., 2004. "Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a web-based resource". BMC Bioinformatics, 5:4.

- Doyle, C. K., Nethery, K. A., Popov, V. L., McBride, J. W., 2006. "Differentially expressed and secreted major immunoreactive protein orthologs of *Ehrlichia canis* and *E. chaffeensis* elicit early antibody responses to epitopes on glycosylated tandem repeats". Infect. Immun., 74(1): 711-720.

- Ellegren, H. (2004). "Microsatellites: simple sequences with complex evolution". Nat. Rev. Genet., 5(6): 435-445.

- Everett, C.M., Wood, N. W., 2004. "Trinucleotide repeats and neurodegenerative disease". Brain, 127: 2385-2405.

- Farlow, J., Smith, K. L., Wong, J., Abrams, M., Lytle, M., Keim, P., 2001. "*Francisella tularensis* strain typing using multiplelocus, variable-number tandem repeat analysis". J. Clin. Microbiol., 39(9): 3186-3192.

- Fenollar, F., Raoult, D., 2004. "Molecular genetic methods for the diagnosis of fastidious microorganisms". J. Clin. Microbiol., 42(11): 4919-4924.

- Friedberg, E. C., Wagner, R., Radman, M., 2002. "Specialized DNA polymerases, cellular survival and the genesis of mutations". Science, 296(5573): 1627-1630.

- Gonzalez, M., Woodgate, R., 2002. "The “tale” of UmuD and its role in SOS mutagenesis". Bioessays, 24(2): 141-148.

- Hardy, K. J., Ussery, D. W., Oppenheim, B. A., Hawkey, P. M., 2004. "Distribution and characterization of staphylococcal interspersed repeat units (SIRUs) and potential use for strain differentiation". Microbiology, 150(Pt 12): 4045-52.

- Hood, D. W., Deadman, M. E., Jennings, M. P., Bisercic, M., Fleischmann, R. D., Venter, J. C., Moxon, E. R., 1996. "DNA repeats identify novel virulence genes in *Haemophilus influenzae*". Proc. Natl. Acad. Sci. U S A, 93(20): 11121-11125.

- Hunter, P. R., Gaston, M. A., 1988. "Numerical index of the discriminatory ability of typing systems: an application of Simpson’s index of diversity". J. Clin. Microbiology, 26(11): 2465-2466.

- Kalendar R., Lee D., Schulman, A. H., 2009. "FastPCR Software for PCR Primer and Probe Design and Repeat Search". Genes, Genomes and Genomics, 3.

- Kashi, Y., King D., Soller, M., 1997. "Simple sequence repeats as a source of quantitative genetic variation." Trends Genet., 13(2): 74-78.

- Keim, P., Price, L. B., Klevytska, A. M., Smith, K. L., Schupp, J. M., Okinaka, R., Jackson, P. J. Hugh-Jones, M. E., 2000. "Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*". J. Bacteriol., 182(10): 2928-2936.

- Le Flèche P., Hauck, Y., Onteniente, L., Prieur, A., Denoeud, F., Ramisse, V., Sylvestre, P., Benson, G., Ramisse, F., Vergnaud, G., 2001. "A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*". BMC Microbiol., 1: 2.

- Le Flèche, P., Fabre, M., Denoeud, F., Koeck, J. L., Vergnaud, G., 2002. "High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing". BMC Microbiol., 2: 37.

- Le Flèche, P., Jacques, I., Grayon, M., Al Dahouk, S., Bouchon, P., Denoeud, F., Nöckler, K., Neubaur, H., Guilloteau, L. A., Vergnaud, G., 2006. "Evaluation and selection of tandem repeat loci for a *Brucella* MLVA typing assay". BMC Microbiology, 6: 9.

- Lenne-Samuel, N., Wagner, J., Etienne, H., Fuchs, R. P., 2002. "The processivity factor beta controls DNA polymerase IV traffic during spontaneous mutagenesis and translesion synthesis in vivo". EMBO Rep., 3(1): 45-49

- Liao, J. C., Li, C. C., Chiou, C, S., 2006. "Use of a multilocus variable-number tandem repeat analysis method for molecular subtyping and phylogenetic analysis of *Neisseria meningitidis* isolates". BMC Microbiology, 6:44.

- Lin, S., Kowalski, D., 1994. "DNA helical instability facilitates initiation at the SV40 replication origin." J. Mol. Biol., 235(2): 496-507.

- Lindstedt, B. A., Heir, E., Gjernes, E., Kapperud, G., 2003. "DNA fingerprinting of *Salmonella enterica* subsp. enterica serovar typhimurium with emphasis on phage type DT104 based on variable number of tandem repeat loci". J. Clin. Microbiol., 41(4): 1469-1479.

- Liu, Y., Lee, M. A., Ooi, E. E., Mavis, Y., Tan, A. L., Quek, H. H., 2003. "Molecular typing of *Salmonella enterica* serovar typhi isolates from various countries in Asia by a multiplex PCR assay on variable number tandem repeats". J. Clin. Microbiol., 41(9): 4388-4394.

- Mayr, W. R., 1995. "DNA markers in forensic medicine". Transfus. Clin. Biol., 2(4): 325-328.

- Metzgar, D., Bytof J., Wills, C., 2000. "Selection Against Frameshift Mutations Limits Microsatellite Expansion in Coding DNA." Genome Res., 10(1): 72-80.

- Mirkin, S., 2004. "Molecular models for repeat expansions". Chemtracts: Biochem. Mol. Biol., 17: 639-662.

- Mirkin, S. M., 2006. "DNA structures, repeat expansions and human hereditary disorders". Curr. Opin. Struct. Biol., 16(3): 351-358.

- Mirkin, S. M., 2007. "Expandable DNA repeats and human disease". Nature, 447(7147): 932-940.

- Monteil, M., Durand, B., Bouchouicha, R., Petit, E., Chomel, B., Arvand, M., Boulouis, H. J., Haddad, N., 2007. "Development of discriminatory multiple-locus variable number tandem repeat analysis for *Bartonella henselae*". Microbiology., 153(Pt 4): 1141-1148.

- Moxon, E. R., P. B. Rainey, M. A. Nowak, R. E. Lenski, 1994. "Adaptive evolution of highly mutable loci in pathogenic bacteria". Curr. Biol., 4(1): 24-33.

- Myers, S., Freeman, C., Auton, A., Donnelly, P., McVean, G., 2008. "A common sequence motif associated with recombination hot spots and genome instability in humans". Nat. Genet., 40(9): 1124-1129.

- Napolitano, R., Janel-Bintz, R., Wagner, J., Fuchs, R. P., 2000. "All three SOS-inducible DNA polymerases (Pol II, Pol IV and Pol V) are involved in induced mutagenesis". EMBO J., 19(22): 6259-6265.

- Nohmi, T., 2006. "Environmental stress and lesion-bypass DNA polymerases". Annu. Rev. Microbiol., 60: 231-253. Review.

- Nonati, N. K., Stelow, J. K., 1980. "Inducible repair system in *Haemophilus influenzae* unaccompanied by mutation". J Bacteriol., 143(1): 516-519.

- O'Dushlaine, C. T., Edwards, R. J., Park, S. D., Shields, D. C., 2005. "Tandem repeat copy-number variation in protein-coding regions of human genes". Genome Biol., 6(8): R69.

- Onteniente, L., Brisse, S., Tassios, P. T., Vergnaud, G., 2003. "Evaluation of the polymorphisms associated with tandem repeats for *Pseudomonas aeruginosa* Strain Typing". J. Clin. Microbiol., 41(11): 4991–4997.

- Ozenberger, B. A., Roeder, G. S., 1991. "A unique pathway of double-strand break repair operates in tandemly repeated genes". Mol. Cell. Biol., 11(3): 1222-1231.

- Pourcel, C., Andre-Mazeaud, Neubauer, H., Ramisse, F., Vergnaud, G., 2004. "Tandem repeats analysis for the high resolution phylogenetic analysis of *Yersinia pestis*". BMC Microbiology, 4: 22.

- Pourcel, C., Vidgop, Y., Ramisse, F., Vergnaud, G., Tram, C., 2003. "Characterization of a tandem repeat polymorphism in *Legionella pneumophila* and its use for genotyping". J. Clin. Microbiol., 41(5): 1819-1826.

- Power, P. M., Sweetman, W. A., Gallacher, N. J., Woodhall, M. R., Hood, D. W., 2009. "Simple sequence repeats in *Haemophilus influenzae*". Infect. Genet. Evol., 9(2): 216-228.

- Sabat, A., Krzyszton-Russjan, J., Strzalka, W., Filipek, R., Kosowska, K., Hryniewicz, W., Travis, J., Potempa, J., 2003. "New method for typing *Staphylococcus aureus* strains: multiple-locus variable-number tandem

repeat analysis of polymorphism and genetic relationships of clinical isolates". J. Clin. Microbiol., 41(4): 1801-1804.

- Schlötterer, C., Tautz, D. 1992. "Slippage synthesis of simple sequence DNA". Nucleic Acids Research, 20: 211-215.

- Schneider, S., Schorr, S., Carell, T. 2009. "Crystal structure analysis of DNA lesion repair and tolerance mechanisms". Curr. Opin. Struct. Biol., 19(1): 87-95.

- Smouse, P. E., Chevillon, C., 1998. "Analytical aspects of population-specific DNA fingerprinting for individuals". J. Hered., 89(2): 143-150.

- Strand, M., Prolla, T.A., Liskay, R.M., Petes, T.D., 1993. "Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair". Nature, 365(6443): 274-276.

- Streisinger, G., Y. Okada, J. Emrich, J. Newton, A. Tsugita, E. Terzaghi, M. Inouye, 1966. "Frameshift mutations and the genetic code. This paper is dedicated to Professor Theodosius Dobzhansky on the occasion of his 66th birthday". Cold Spring Harb. Symp. Quant. Biol., 31: 77-84.

- Struelens, M. J., 1996. "Consensus guidelines for appropriate use and evaluation of microbial epidemiologic typing systems". Clin. Microbiol. Infect., 2(1): 2-11.

- Struelens, M. J., 1998. "Molecular Epidemiologic Typing Systems of Bacterial Pathogens: Current Issues and Perpectives". Mem. Inst. Oswaldo Cruz, 93(5): 581-585.

- Subramanian, S., Mishra R. K., Singh, L., 2003. "Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions." Genome Biol., 4(2): R13.

- Sweetman W. A., Moxon E. R., Bayliss C. D., 2005. "Induction of the SOS regulon of *Haemophilus influenzae* does not affect phase variation rates at tetranucleotide or dinucleotide repeats". Microbiology, 151(Pt 8): 2751-2763.

- Tang, M., Shen, X., Frank, E. G., O'Donnell, M., Woodgate, R., Goodman, M. F., 1999. "UmuD'2C is an error-prone DNA polymerase, *Escherichia coli* pol V". Proc. Natl. Acad. Sci. U S A, 96(16): 8019-8924.

- Titze-de-Almeida, R., Willems, R. J., Top, J., Rodrigues, I. P., Ferreira II, R. F., Boelens, H., Brandileone, M. C., Zanella, R. C., Felipe, M. S., Van Belkum, A., 2004. "Multilocus variable-number tandem-repeat polymorphism among Brazilian *Enterococcus faecalis* strains". J. Clin. Microbiol., 42(10): 4879-4881.

- Toth, G., Gaspari, Z., Jurka, J., 2000. "Microsatellites in different eukaryotic genomes: survey and analysis". Genome Res., 10(7): 967-981.

- Treangen, T. J., Abraham, A. L., Touchon, M., Rocha, E. P. 2009. "Genesis, effects and fates of repeats in prokaryotic genomes". FEMS Microbiol. Rev., 33(3): 539-571.

- Van Belkum, A., 2007. "Tracing isolates of bacterial species by multilocus variable number of tandem repeat analysis (MLVA)". FEMS Immunol. Med. Microbiol., 49(1): 22-27.

- Van Belkum, A., Scherer, S., van Alphen, L., Verbrugh, H., 1998. "Short-sequence DNA repeats in prokaryotic genomes". Microbiol. Mol. Biol. Rev., 62(2): 275-293.

- Van Ham, S. M., Van Alphen, L., Mooi, F. R., Van Putten, J. P., 1993. "Phase variation of *Haemophilus influenzae* fimbriae: transcripcional control of two divergent genes through a variable combined promoter region". Cell., 73(6): 1187- 1196.

- Vasquez, K. M., Hanawalt, P. C., 2009. "Intrinsic genomic instability from naturally occurring DNA structures: An introduction to the special issue". Mol. Carcinog., 48(4): 271-272.

- V-DIC (VNTR Diversity and Confidence Extractor), Octubre, 2009. *http://www.hpabioinformatics.org.uk/cgi-bin/DICI/DICI.pl*.

- Vergnaud, G., Pourcel, C., 2009. "Multiple locus variable number of tandem repeats analysis". Methods Mol. Biol., 551: 141-158.

- Vergnaud, G., Pourcel, C., 2006. "Multiple Locus VNTR (Variable Number of Tandem Reapeat) Analysis". In: Stackebrandt, E. (Eds.), Molecular Identification, Systematics, and Populations Structure of Prokaryotes. Springer-Verlag, Berlin – Heidelberg, Alemania, pp, 83-104.

- Viguera, E., Canceill, D., Ehrlich, S. D., 2001. "In vitro replication slippage by DNA polymerases from thermophilic organisms." J. Mol. Biol., 312(2): 323-333.

- Viguera, E., Canceill, D., Ehrlich, S. D., 2001. "Replication slippage involves DNA polymerase pausing and dissociation". EMBO J., 20(10): 2587-2595.

- Voineagu, I., Freudenreich, C. H., Mirkin, S. M., 2009a. "Checkpoint responses to unusual structures formed by DNA repeats". Mol. Carcinog., 48(4): 309-318.

- Voineagu, I., Surka, C. F., Shishkin, A. A., Krasilnikova, M. M., Mirkin, S. M., 2009b. "Replisome stalling and stabilization at CGG repeats, which are responsible for chromosomal fragility". Nat. Struct. Mol. Biol., 16(2): 226-228.

- Vu-Thien, H., Corbineau, G., Hormigos, K., Fauroux, B., Corvol, H., Clément, A., Vergnaud, G., Pourcel, C., 2007. "Multiple-locus variable-number tandem-repeat analysis for longitudinal survey of sources of *Pseudomonas aeruginosa* infection in cystic fibrosis patients". J. Clin. Microbiol., 45(10): 3175-3183.

- Wagner, J., Etienne, H., Janel-Bintz, R., Fuchs, R. P., 2002. "Genetics of mutagenesis in *E. coli*: various combinations of translesion polymerases (Pol II, Pol IV and Pol V) deal with lesion/sequence context diversity". DNA repair (Amst), 1(2): 159-167.

- Wang, G., Vasquez, K. M., 2006. "Non-B DNA structure-induced genetic instability". Mutat. Res., 598(1-2): 103-119.

- Weir, B. S., 1992. "Population genetics in the forensic DNA debate". Proc. Natl. Acad. Sci. U S A, 89(24): 11654-11659.

- Wood, R. D., Hutchinson, F., 1984. "Non-targeted mutagenesis of unirradiated lambda phage in *Escherichia coli* host cells irradiated with ultraviolet light". J. Mol. Biol., 173(3): 293-305.

- Wyman, A. R., White, R., 1980. "A highly polymorphic locus in human DNA". Proc. Natl. Acad. Sci. U S A, 77(11): 6754-6758.

- Yang, W., 2003. "Damage repair DNA polimerases Y". Curr. Opin. Struct. Biol., 13(1): 23-30. Review.

- Yazdankhah, S. P., Lindstedt, B. A., Caugant, D. A., 2005. "Use of variable-number tandem repeats to examine genetic diversity of *Neisseria meningitidis*". J. Clin. Microbiol., 43(4): 1699-1705.

- Zane, L., Bargelloni, L., Patarnello, T., 2002. "Strategies for microsatellite isolation: a review. Mol Ecol., 11(1): 1-16 Review.