UNIVERSIDAD
DE GRANADA

FACULTAD DE CIENCIAS

DEPARTAMENTO DE GENÉTICA

ESTUDIOS GENÓMICOS Y DINÁMICA EVOLUTIVA EN *Helicobacter pylori*

Memoria para optar al grado de

Doctor en Biología presentada por:

Jerson Alexander García Zea

Programa de Doctorado en Biología Fundamental y de Sistemas

Director: **José Carmelo Ruiz Rejón**

Editor: Universidad de Granada. Tesis Doctorales
Autor: Jerson Alexander García Zea
ISBN: 978-84-1117-364-3
URI: http://hdl.handle.net/10481/75451

Parte de los resultados que se presentan en esta memoria, han sido previamente publicados: García-Zea JA, de la Herrán Moreno R, Robles F, Navajas-Pérez R, Ruiz Rejón C. 2019. Detection and variability analyses of CRISPR-like loci in the *H. pylori* genome. Peer J 7:e6221 DOI: 10.7717/peerj.6221

Asimismo, parte de los resultados, han sido presentados en los siguientes congresos:

Alexander García-Zea, Roberto de la Herrán, Francisca Robles, Rafael Navajas-Pérez, Carmelo Ruiz Rejón. Estudios de genómica comparada en *Helicobacter pylori* y otros *Helicobacter* para la identificación de genes de interés. **V Congreso de Estudiantes de Investigación Biosanitaria 2019 (CEIBS).** Granada, 2019

Alexander García-Zea, Roberto de la Herrán, Francisca Robles, Rafael Navajas-Pérez, Carmelo Ruiz Rejón. Duplicated genes under concerted evolution in *Helicobacter pylori.* **XLII Congreso de la Sociedad Española de Genética.** On line, 2021

Índice general

Resumen

En esta tesis hemos llevado a cabo un análisis genómico, poblacional y evolutivo de la bacteria *H. pylori* junto a varias especies del mismo género. *H. pylori* es reconocida como uno de los patógenos obligados humanos más comunes que coloniza el estómago y el duodeno en la mitad de la población humana, causando inflamación que puede desarrollar úlceras y cáncer gástrico. Esta especie presenta un comportamiento poblacional panmíctico, con frecuente recombinación homóloga mutua y mostrando una gran diversidad en términos de estructura del genoma y composición de genes, así como una alta variación en secuencia de nucleótidos debido a la elevada tasa de mutación y recombinación.

Como hemos podido comprobar, el genoma de esta bacteria presenta una longitud media de 1.621.671+- 43.785 bp con un contenido medio de G/C del 40% por lo que se encuadraría dentro del grupo de genomas bacterianos pequeños de ~2 Mb. Esta reducción en el tamaño está en concordancia con otras bacterias patógenas de vida libre en ambientes extremos.

Igualmente, el contenido génico también es bajo, teniendo una media de 1.551+- 42 genes por genoma y un genoma central constituido por 802 genes. Ambos valores obtenidos en esta Tesis son más bajos que en otros estudios previos de esta misma especie. Estas diferencias podrían ser debidas a varias causas entre ellas: el diferente número de cepas incluido en cada análisis, el programa de cálculo y los respectivos filtros de significación del genoma central usado, así como la inclusión en alguno de estos estudios de genes parálogos y ortólogos conjuntamente. Sin embargo, podemos señalar que el tamaño de 802 genes para el genoma central no es extraordinario ya que, por ejemplo, las Actinobacteria de vida libre presentan un genoma de aproximadamente 800 genes.

Sin embargo, a pesar del reducido tamaño de su genoma en nucleótidos y en genes esta bacteria presenta una enorme variabilidad lo que se manifiesta en la gran cantidad de SNPs que presentan los genes del genoma central en el conjunto de las 53 cepas analizadas. Esta gran variabilidad permite clasificar en un número cada vez mayor las diferentes subpoblaciones geográficas al inicialmente propuesto en función del análisis para solo siete genes del genoma.

Esta observación de tan alta variabilidad no es excepcional para *H. pylori* ya que la secuenciación de un gran número de genomas procarióticos y la comparación de secuencias de especies estrechamente relacionadas han puesto de manifiesto una alta frecuencia en un gran repertorio de variaciones genómicas, que pueden ir desde variaciones de un solo nucleótido hasta eventos de inserción-deleción de grandes bloques cromosómicos.

El origen de esta gran cantidad de variabilidad puede tener varias explicaciones, ninguna de ellas mutuamente excluyentes. Entre estos podemos señalar: I) estilo de vida II) subproductos de recombinación ilegítima (recombinación homóloga) III) mecanismo de reparación no homólogo impreciso durante la replicación aberrante del ADN para reparar horquillas de replicación rotas y IV) presencia de elementos repetidos en su genoma.

Como resultado de los procesos erróneos de recombinación hemos podido comprobar que las inversiones son la mutación cromosómica más frecuente, existiendo inversiones características de cada región, existiendo subgrupos dentro de ellas que han podido producirse por procesos de deriva genética.

Esta misma asociación se produce cuando se realiza un análisis estructural y filogenético en base a los SNPs detectados, ya que podemos observar un patrón geográfico en el que es posible, en general, asignar cada cepa según su origen. También identificamos hasta seis cepas híbridas, cuatro de ellas por primera vez en este trabajo y cuyo genoma es el resultado de la recombinación de, al menos, tres genomas de origen geográfico diferente. En este sentido también hemos comprobado la gran influencia que el flujo génico tiene en la estructura poblacional de esta especie, produciendo una gran cantidad de mezcla entre distintas regiones geográficas, existiendo regiones donadoras y otras receptoras.

También hemos podido comprobar que, además de los mecanismos mencionados que contribuyen a la conformación del genoma, en *H. pylori* la evolución concertada y la selección juegan un papel importante en su dinámica evolutiva. Así, para siete genes duplicados, encontramos pruebas de evolución concertada es decir, de homogenización de secuencias intergénicas para aquellas regiones del gen en las que se pudo comprobar que ocurría recombinación.

Para poder tener una visión más amplia de la evolución de esta especie y su relación con otras especies del género *Helicobacter*, hemos llevado a cabo un análisis comparativo de *H. pylori* con siete especies del género. De estos análisis, podemos deducir que las cepas africanas de *H. pylori* son las más relacionadas con *H. acinonychis* especie también de ambiente gástrico.

Por último, hemos realizado un estudio sobre los distintos sistemas inmunes que presenta el género. Hemos comprobado que presentan una gran diversidad de sistemas en el conjunto de especies y que casi todas las especies presentan más de tres sistemas inmunitarios, lo que les permite una mejor defensa ante los ataques de fagos.

Summary

In this thesis we have carried out a genomic, population and evolutionary analysis of the *H. pylori* bacteria together with several species of the same genus. *H. pylori* is recognized as one of the most common human obligate pathogens that colonizes the stomach and duodenum in half the human population, causing inflammation that can lead to ulcers and gastric cancer. This species presents a panmictic population behavior, with frequent mutual homologous recombination and showing great diversity in terms of genome structure and gene composition, as well as high variation in nucleotide sequence due to the high rate of mutation and recombination.

As we have been able to verify, the genome of this bacterium has an average length of 1,621,671 + - 43,785 bp with an average G / C content of 40%, which is why it falls within the group of small bacterial genomes of ∼2 Mb. This reduction in size is in concordance with other free-living pathogenic bacteria in extreme environments.

Likewise, the gene content is also low, having an average of 1,551 + - 42 genes per genome and a central genome made up of 802 genes. Both values obtained in this Thesis are lower than in other previous studies of this same species. These differences could be due to several causes including: the different number of strains included in each analysis, the calculation program and the respective significance filters of the central genome used, as well as the inclusion in some of these studies of paralogue and orthologous genes. jointly. However, we can point out that the size of 802 genes for the central genome is not extraordinary since, for example, the free-living Actinobacteria have a genome of approximately 800 genes.

However, despite the small size of its genome in nucleotides and genes, this bacterium shows enormous variability, which is manifested in the large number of SNPs that the genes of the central genome present in the set of 53 strains analyzed. This great variability makes it possible to classify the different geographic subpopulations in an increasing number than the one initially proposed based on the analysis for only seven genes of the genome.

This observation of such high variability is not exceptional for *H. pylori* since the sequencing of a large number of prokaryotic genomes and the comparison of sequences of closely related species have revealed a high frequency in a large repertoire of genomic variations, which can ranging from single nucleotide variations to large chromosome block insertion-deletion events.

The origin of this large amount of variability can have several explanations, none of them mutually exclusive. Among these we can point out: I) lifestyle II) illegitimate recombination by-products

(homologous recombination) III) imprecise non-homologous repair mechanism during aberrant DNA replication to repair broken replication forks and IV) presence of repeated elements in its genome.

As a result of the erroneous recombination processes, we have been able to verify that inversions are the most frequent chromosomal mutation, existing inversions characteristic of each region, and there are subgroups within them that could have been produced by genetic drift processes.

This same association occurs when a structural and phylogenetic analysis is carried out based on the SNPs detected, since we can observe a geographical pattern in which it is possible, in general, to assign each strain according to its origin. We also identified up to six hybrid strains, four of them for the first time in this work and whose genome is the result of the recombination of at least three genomes of different geographical origin. In this sense, we have also verified the great influence that gene flow has on the population structure of this species, producing a large amount of mix between different geographic regions, with donor and other recipient regions.

We have also been able to verify that, in addition to the aforementioned mechanisms that contribute to the conformation of the genome, in *H. pylori* concerted evolution and selection play an important role in its evolutionary dynamics. Thus, for seven duplicated genes, we found evidence of concerted evolution, that is, of homogenization of intergenic sequences for those regions of the gene in which it was possible to verify that recombination occurred.

In order to have a broader vision of the evolution of this species and its relationship with other species of the genus *Helicobacter*, we have carried out a comparative analysis of *H. pylori* with seven species of the genus. From these analyzes, we can deduce that the African strains of *H. pylori* are the most related to H. acinonychis species also from the gastric environment.

Finally, we have carried out a study on the different immune systems that the genus presents. We have verified that they present a great diversity of systems in the set of species and that almost all species have more than three immune systems, which allows them a better defense against phage attacks.

Capítulo 1 Introducción

1.1. Antecedentes

En el año 1984, los científicos australianos Barry Marshall y Robin Warren **(Moodley, 2016)** fueron los primeros en cultivar el bacilo Gram negativo helicoidal *Helicobacter pylori* y demostrar su papel en la gastritis y la formación de úlceras peptídicas **(Marshall & Warren, 1984).** Este microorganismo se caracteriza por ser microaerofílico, curvo o recto, de crecimiento lento y con 2 a 6 flagelos unipolares que le confieren movilidad (Aznar, 2014). *H. pylori* pertenece al filo *Proteobacteria*, clase *Epsilonproteobacteria*, Orden *Campylobacterales* y Familia *Helicobacteraceae* (Tabla 1) **(Marshall & Goodwin, 1987)**.

Taxonomía

Dominio	**Bacteria**
Filo	Proteobacteria
Clase	Epsilonproteobacteria
Orden	Campylobacterales
Familia	*Helicobacter*aceae
Género	*Helicobacter*
Especie	*Helicobacter pylori*

Tabla 1. Clasificación *Helicobacter pylori*

El género *Helicobacter* comprende al menos 40 especies diferentes **(Amorim *et al.,* 2015)**, aunque su taxonomía aún sigue siendo confusa **(Josten *et al.,* 2016)**. Aunque algunas de estas especies han sido asociadas de manera ocasional con infecciones en los humanos, su huésped primario es de origen animal. El género puede subdividirse en especies gástricas y entero-hepáticas **(Solnick *et al.,* 2001)**. Las primeras (gástricas) habitan primordialmente el antro del estómago, como lo son *H. felis* (en felinos domesticos*), H. mustelae* (en hurones), *H. acynonichis* (en felinos salvajes), *H. cetorum* (en delfines y ballenas) o *H. heilmannii* (en gatos, primates, cerdos y mamíferos carnívoros). Las segundas (entero-hepáticas) se establecen en las criptas intestinales de su huésped, como, por ejemplo: *H. pullorum* (en aves de corral) y *H. cinaedi* (en gatos, perros, hámster, ratas, zorros y monos) **(Fox, 2002; Rossi & Hänninen, 2012; Smet *et al.,* 2011)**.

1.2. Características de *H. pylori*

H. pylori se caracteriza por colonizar de forma persistente el estómago humano por lo que es considerada uno de los patógenos humanos más exitosos **(Suerbaum & Josenhans, 2007**). Sin embargo, hay casos en que esta bacteria induce la inflamación de la mucosa gástrica conocida como gastritis superficial, además de ser un factor de riesgo para el desarrollo de la enfermedad úlcero péptica, el adenocarcinoma gástrico y el linfoma de tejido linfoide asociado a la mucosa (MALT) **(Gangwer *et al.,* 2010)**.

Esta bacteria, es una de las especies más diversas con un genoma altamente variable; sus cepas pueden variar entre sí, gracias a una gran dinámica genómica que consiste en mecanismos como las mutaciones, transferencia horizontal de genes, recombinación, presencia de secuencias repetidas e inversas y/o directas y elementos transponibles **(Zepeda-Gurrrola, 2012 ; Draper *et al.,* 2016). T**ambién, se caracteriza por exhibir panmixia y mosaicismo genético y una gran plasticidad genómica, que indica la capacidad de esta bacteria para adaptarse a nivel micro-evolutivo a su ambiente **(Suerbaum & Josenhans, 2007).** Excepcionalmente, las tasas de recombinación en esta especie son significativamente altas, lo que le permite a nivel genómico altos niveles de intercambio genético, tanto de grandes como pequeñas porciones de su genoma, lo cual actúa como una fuerza impulsadora dominante para la diversificación de la especie **(Suerbaum *et al.,* 1998; Bubendorfer *et al.,* 2016).**

En la última década la genómica comparada ha permitido encontrar genes específicos no solo entre cepas del mismo origen geográfico, sino también entre réplicas de la misma cepa, lo que pone de manifiesto el alto grado de diversidad genética que existe en esta especie **(Kabir, 2009**). Este aspecto es esencial para la adaptación de la bacteria al estómago del huésped e influye directamente en el resultado final de la infección **(Han *et al.,* 2003)**.

1.2.1 Epidemiología de *H. pylori*

Se estima que *H. pylori* está presente en más de la mitad de la población mundial y en su gran mayoría los individuos no exhiben síntomas **(Ghose *et al.,* 2005; Gangwer *et al.,* 2010; Salama *et al.,* 2013)**. Esta bacteria suele contraerse en la infancia y una vez adquirida, la colonización bacteriana suele durar toda la vida **(Moodley, 2016)**. En relación a su prevalencia, esta exhibe importantes diferencias en relación a la geografía, oscilando entre el 74-90 % en los países en vía de desarrollo y por debajo del 58% en los países desarrollados **(Aznar, 2014; Perez *et al.,* 2004**).

1.2.2 Nicho de *H. pylori*

H. pylori puede crecer a un pH de ~5,1 que es el pH del contenido gástrico al momento de la ingestión de alimentos, por lo tanto, esta bacteria se ha adaptado a través de un proceso de coevolución con su huésped humano para resistir a las condicioneees ácidas cuando se ingieren alimentos, lo que le permite colonizar el estómago del huésped de manera persistente **(Rherr *et al.*, 2014)**. Las cepas de laboratorio de *H. pylori* obtenidas de humano pueden adaptarse para colonizar otras especies, pero es muy raro encontrarlas de forma natural colonizando otros huéspedes no humanos **(Andersson *et al.*, 2008)**. También se ha podido demostrar que, la composición de la microbiota gástrica evoluciona junto con el nivel de colonización de *H. pylori* y la producción de ácido del estómago **(Wroblewski & Peek 2016)**

1.2.3 Colonización y Patogénesis

Como se ha comentado, la gran mayoría de casos de *H. pylori* son asintomáticos, sin embargo, entre aquellos sujetos que, si los manifiestan, el principal síntoma es la gastritis **(Peek & Blaser, 2002; Algood & Cover 2006; Blaser & Atherton 2004)**. La capacidad de esta bacteria para provocar síntomas en su huésped humano depende de un complejo sistema que involucra interacciones entre las bacterias, su huésped y el medio ambiente. En este sentido, se podría organizar la patogénesis e infección de *H. pylori* en los siguientes procesos durante la interacción con el huésped: **1)** Amortiguación del pH, **2)** Motilidad y quimiotaxis, **3)** Adhesión **4)** Factores de virulencia citotóxicos asociados a la patogenicidad y **5)** Evasión del sistema inmunológico **(Figura 1)**

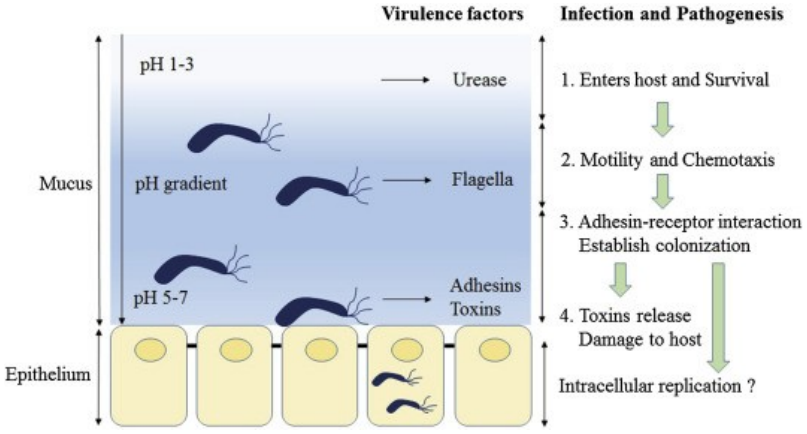


Figura 1. Diagrama esquemático de la patogénesis e infección por *H. pylori*. Tomada de **Kao, Sheu & Wu. (2015)**

1.2.3.1 Amortiguación del pH

La motilidad es facilitada por la capacidad de *H. pylori* de reducir la viscosidad del moco mediante la acción de la ureasa, que promueve el ajuste del pH periplásmico para asegurar la supervivencia al ambiente ácido **(Eaton & Krakowka, 1994).** En este proceso están implicados un grupo de siete genes de la ureasa incluidas las subunidades catalíticas (*ureA/, ureB*) y los genes accesorios (*ureI, ureE, ureF, ureH y ureG*) **(Mobley *et al.,* 1995).** Para permitir el ajuste rápido del pH periplásmico, los genes *arsR* y *arsS* controlan la transcripción del grupo de genes de la ureasa **(Pflock *et al.,* 2006).**

1.2.3.2 Motilidad

Existen evidencias que indican que la anatomía helicoidal de *H. pylori* favorece la motilidad y en consecuencia, permite a las bacterias penetrar en la capa de moco gástrico y colonizar el ambiente gástrico (**Eaton *et al.,* 1996).** La motilidad es facilitada por tres aspectos diferentes: **a)** los flagelos (generalmente de 4 a 6), que están compuestos por el cuerpo basal, el gancho... y el filamento flagelar **(Lertsethtakhan *et al.,* 2011)**. **b)** El filamento flagelar está formado por dos flagelinas y es codificado por los genes *flaA* y *flaB* y **c)** el gancho es codificado por el gen *flgE* **(Lertsethtakhan *et al.,* 2011)**.

Más de 40 genes participan en la biosíntesis y funcionamiento de los flagelos. Los genes relacionados con la síntesis de los flagelos se dividen en tres clases, dependiendo del factor que los regule. Así, el factor sigma σ^{80} de mantenimiento sintetizado por el gen *rpoD*, regula los genes de clase1 **Figura 2**. El factor sigma alternativo $\sigma^{54,}$ sintetizado por el gen *rpoN,* regula los genes de clase 2 y el factor σ^{28} sintetizado por el gen *fliA*, regula los genes de la clase 3 **(Kao, 2016)**. Los genes flagelares de clase 1 comprenden los principales genes reguladores (*rpoN, flgR, flgS, flhA*) y los genes estructurales (*motA y motB*) del sistema flagelar **Niehus *et al.,* 2004)**. Los genes flagelares clase 2 comprenden los *genes flaB, flgE, flgK, flgM, flgL* y el gen fliA que codifica el factor σ^{28} que es dependiente de *rpoN* **(Kao et al., 2104)**. Por último, los genes reguladores tardíos o de clase 3, que incluyen *flaA, Omp11 y EnvA* **Niehus *et al.,* 2004)**; y el *flhA* que son necesarios para la transcripción completa de los genes flagelares de clase 2 y 3 **(Kao, 2016).**

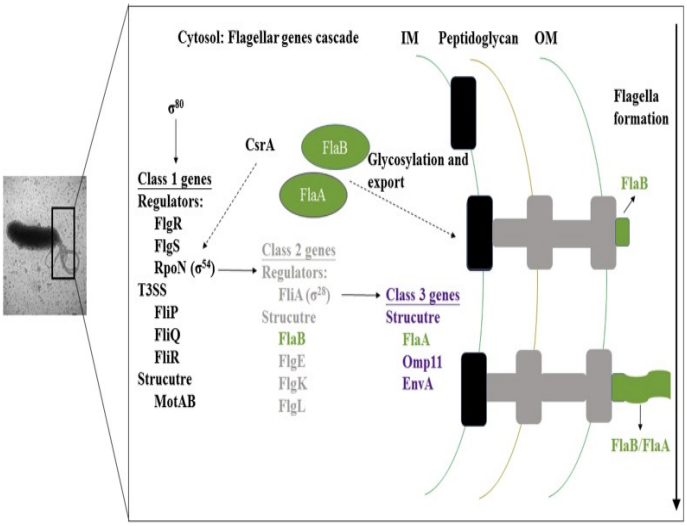


Figura 2. Modelo de la cascada reguladora de la transcripción flagelar para la biosisntesis flagelar de *H. pylori*. Tomado de **Kao. (2016)**

1.2.3.3 Adhesión y quimiotaxis

Una vez *H. pylori* ha penetrado en el moco, esta bacteria necesita adherirse a las células del huésped para establecer la infección **(Kao *et al.,* 2015)**. En esta etapa intervienen diversas adhesinas, las cuales pertenecen a la familia de proteínas de membrana externa de tipo 1 (OMP) **(Alm *et al.,* 2000)**. Todas las adhesinas identificadas hasta el momento son miembros de la familia *Hop* y los genes mejor caracterizados son los que codifican las proteínas de unión al antígeno del grupo sanguíneo (*babA*) (**Aspholm-Hurtig *et al.,* 2004)**, la proteína de unión al ácido siálico (*sabA*) **(Unemo *et al.,* 2005)**, la proteína inflamatoria externa A (*oipA*), y las lipoproteínas asociadas a la adherencia (*alpA, alpB*) **(Senkovich *et al.,* 2011; Prazkier *et al.,* 2016)**.

La quimiotaxis de *H. pylori* depende de una serie de genes (*tlpA, tlpB tlpC* y *tlpD*) que codifican proteínas quimiorreceptoras fundamentales para la orientación espacial y control de la rotación flagelar dependiendo de las condiciones del entorno y adaptarlo para la supervivencia de la bacteria **(Aihara *et al.,* 2014)**. Las interacciones de estas proteínas son mediadas por los genes *cheA*, *cheW* y *cheY* **(Lowenthal *et al.,* 2009)** y la rotación flagelar es regulada por los genes *motA* y *motB* **(Aihara *et al.,* 2014)**.

1.2.4 Principales factores de virulencia citotóxicos implicados en la patogenicidad.

Una gran cantidad de factores de virulencia putativos han sido reportados y estos se clasifican en tres categorías: **1)** genes que comprenden estructuras y genotipos variables según la cepa como *vacA* y sus genes parálogos: *imaA*, *vlpC* y *faaA* **(Lu, 2007; Yamahoka, 2008); 2)** genes específicos de cepa (presentes solo en algunas) como por ejemplo la isla de patogenicidad cag PAI **(Roesler *et al.,* 2014); 3)** genes de la fase variable cuyo expresión puede cambiar durante el crecimiento de la bacteria o en condiciones ambientales específicas para adaptar la fisiología de *H. pylori* al medio y asegurar su supervivencia, como genes que codifican para proteínas de membrana (OMP: *oipA, sabA, sabB, babA, babC y hopZ*) **(de Vries *et al.,* 2002; Srikhanta *et al.,* 2011)**

1.2.4.1 Genes que comprenden estructuras y genotipos variables según la cepa

Uno de los determinantes de virulencia más importantes es la toxina secretada por el gen *vacA*. Esta es una toxina formadora de poros que provoca múltiples alteraciones en las células humanas, que incluyen vacuolización celular, despolarización del potencial de membrana, apoptosis, alteración de la permeabilidad de la membrana mitocondrial, inhibición de la presentación de antígenos y la inhibición de la activación y proliferación de células T **(Cover & Blaser, 1992; Yamaoka *et al.*, 2008)**. *VacA* se traduce en una prototoxina de 140 KDa que es escindida en N y C terminal durante el proceso de secreción para producir: una secuencia señal N-terminal, una toxina secretada madura de 88 KDa conocida como p88, un pequeño péptido secretado sin función conocida (péptido alfa); y un dominio beta barril C-terminal **(Gangwer et al., 2010; Posselt *et al.,* 2013**). El gen *vacA* es exclusivo de *H. pylori* y está presente en todas las cepas, pero con polimorfismos significativos **(Atherton *et al.,* 1995)** y se acumulan en tres regiones: **1)** en la región de la secuencia de señal, que se encuentra en dos variantes (s1, s2), **2)** en la región intermedia, con tres variantes (i1, i2, i3), y **3)** en la región media (m1, m2) **(Junaid *et al.,* 2016; M. J.Blaser & Atherton 2004).**

Además, en los genomas de *H. pylori* podemos encontrar tres genes similares a *vacA* que codifican proteínas de 260 a 348 Kda: **1)** el autotransportador inmunomodulador (*imaA*), **2)** el autotransportador A asociado a flagelos (*faaA*) y **3)** la proteína C similar a *vacA* (*vlpC*), que pueden tener funciones importantes en la colonización **(Tomb *et al.,* 1997; Sause *et al*., 2012;**

Radin *et al.,* 2013). Las regiones C-terminal de estas proteínas muestran homología con la región C-terminal de *vacA*, que es un dominio β-barrel necesario para la secreción de *vacA* a través de una vía de autotransporte de tipo V **(Praszkier *et al.,* 2016 book).**

1.2.4.2 Genes específicos de cepas. Isla de patogenicidad cag (cag PAI)

La isla de patogenicidad cag (cag PAI), es un segmento de ADN genómico de 40 kb que se adquiere mediante transferencia horizontal de ADN **(Akopyants et al., 1998)**. Dependiendo de las cepas, la isla cag PAI contiene entre 27 y 31 genes putativos, incluido el gen *cagA* (proteína del gen A asociado a citotoxinas). El gen *cagA* se localiza en un extremo de la isla de patogenicidad. La proteína de este gen varía en su tamaño entre 130 y 145 KDa, debido a un polimorfismo estructural en la región carboxiterminal (C-Terminal) **(Hatakeyama, 2017).** Esta región es la principal responsable del anclaje de *cagA* a la membrana.

1.2.4.3 Genes de variación de fase

La variación de fase (variación de la expresión de proteínas) es un mecanismo común utilizado por bacterias patógenas para generar diversidad intra-cepa para la adaptación a nichos y está estrechamente ligado con determinantes de la virulencia **(Salaün et al., 2004)**. En este sentido, *H. pylori* codifica un conjunto excepcionalmente grande de proteínas de membrana, que representa el 4% de los ORF predichos en el genoma y de las que, en gran parte se desconoce sus funciones. Aunque, algunas de estas han sido relacionados con funciones de adhesión **(Kim, 2016).**

Las adhesinas, son cruciales para la colonización e infección inicial estándo implicadas en la adherencia a la mucosa gástrica. Entre las mejor caracterizadas se encuentran: *babA,* proteínas de unión al antígeno del grupo sanguíneo y **(Pride & Blaser, 2001)**; *sabA,* adhesina de unión al ácido siálico **(Talarico *et al.,* 2012)***; oipA* de la cual hasta el momento no hay estudios sobre su estructura y sus receptores **(Yamahoka *et al.,* 2000; Horridge *et al.,* 2017; Farzi *et al.,* 2018)**; *hopQ* de la cual se sabe que posee dos alelos (*hopQI* y *hopQII*), **(Ohno *et al.,* 2009)** y *HopZ*, la cual posee dos variantes alélicas *hopZ-I* y *hopZ-II* **(Kenneman *et al.,* 2012)**.

1.2.4.5. Evasión del sistema inmune del huésped

H. pylori evade el sistema inmunológico innato mediante una variedad de mecanismos. Entre ellos, se encuentran la detección de receptores de reconocimiento de patrones (PPR), que son proteínas que reconocen patrones moleculares asociado a patógenos (PAMP) **(Taslima *et al.,* 2014)**. Los mecanismos para eludir los PPR incluyen: evitación del reconocimiento por receptores tipo Toll (TLR) e inhibición de la señalización mediada por la lectina de tipo C (DC-SIGN). En este sentido, la bacteria modula sus moléculas de superficie las cuales incluyen el lipopolisacárido (LPS) y flagelinas **(Cullen *et al.,* 2011; Brooks *et al.,* 2013)**.

La infección por *H. pylori* también promueve una respuesta inflamatoria en su huésped que conduce al reclutamiento de macrófagos, neutrófilos y linfocitos en el tejido gástrico **(Telford *et al.,* 1999)**. Esta posee fenotipos antifagocíticos que dependen del sistema de secreción de tipo IV, los cuales son codificados por la isla de patogenicidad del gen asociado a la citotoxina *cag* PAI **(Ramarao *et al.,* 2000; 2001)** y que, junto a una presencia positiva de *vacA* y *cag* PAI, da como resultado un mecanismo inusual para evitar la muerte por fagocitosis **(Allen *et al.,* 2000; Ritting *et al.,* 2003).**

H. pylori también ha desarrollado una serie de mecanismos para evitar la inmunidad adaptativa al interferir con la presentación de antígenos y la modulación de la respuesta por parte de las células T **(Taslima *et al.,* 2014).** Se ha demostrado que esta bacteria causa polarización de las células presentadoras de antígenos (APC), además de controlar las funciones de estas células de manera diferente **(Quinding *et al.,* 2010; Fehlings *et al.,* 2012)**.

H. pylori provoca el aumento de células T reguladoras lo que resulta en una disminución de la respuesta inmune **(Kao *et al.,* 2010).** Otra forma de evadir el sistema inmune es mediante factores de virulencia, como por ejemplo a través de la isla de patogenicidad Cag PAI **(Ramarao *et al.,* 2000; Ramarao & Meyer 2001).**

1.2.5. Genómica de *H. pylori*

El genoma de *H. pylori* varía entre 1.5 a 1.7 Mbp, con unos ±1.500 marcos de lectura abierto, con un contenido de G+C que oscila entre 38 al 39% **(Zawilak & Zakrzewska, 2017).** El primer genoma en ser secuenciado, fue la cepa 26695 desde pacientes con gastritis, cuyo genoma constaba de 1.667.867 pb, con 1590 secuencias codificantes y un contenido medio de G+C del 39% **(Tombs *et al.,* 1997)**. Posteriormente, la cepa J99 fue secuenciada a partir de un

aislamiento de un paciente con úlcera duodenal **(Alm *et al.,* 1999)**. La comparación de las cepas 26695 y J99, reveló grandes diferencias tanto a nivel genómico, de secuencias como de genes **(Kuipers *et al.,* 2000)**.

El número de genomas de *H. pylori* secuenciados ha aumento de manera significativa, gracias al rápido avance en las tecnologías de Nueva Generación de Secuenciación (NGS) **(Rouli *et al.,* 2015)**, resultando para mediados del año 2021 en 2016 genomas depositados en la base de datos NCBI (**Chromosome: 66**, Complete: 243, Contig: 1350, Scaffold: 357)**.** Una de las razones más importantes para secuenciar tantos genomas y específicamente secuenciar genomas replicados de *H. pylori,* ha sido la necesidad de estudiar la evolución cronológica dentro de un solo huésped **(Ahmed, 2009)**.

1.2.5.1 Pan-genoma

El pan-genoma se define como todo el repertorio genómico de un clado filogenético que codifica para todos los posibles estilos de vida de un organismo **(Varnikos *et al.,* 2015**). Este término fue usado por primera vez por **Tettelin et al. (2005)** y describe los genes o marcos de lectura abierto (ORF) que comparten los genomas **(Tettelin *et al.,* 2005)**. Para describir el pan-genoma, se utilizan los términos de genoma central (genes compartidos entre todos genomas) y genoma variable/flexible (genes presentes solo en algunos genomas) **(Varnikos *et al.,* 2015)**. Adicionalmente **Contreras-Moreira & Vinuesa. (2013)** desarrollaron el programa GET_HOMOLOGUES, que permite agrupar el pan-genoma en diferentes componentes: genoma central (genes presentes en el 100 %genomas), genoma soft-core (genes presentes en el 95 % de los genomas), genoma cloud (genes presentes solo en unos pocos genomas), genoma Shell (genes restantes, presentes en varios genomas.

Existe otra posibilidad para la definición del pan-genoma partiendo desde el alineamiento y no desde unidades funcionales (genes). De esta y mediante el método descrito por **Darling *et al.* (2004; 2009)**, Este método define el genoma central y el genoma flexible a partir de un tamaño mínimo de nucleótidos. Así, se considera genoma central si el alineamiento presenta al menos 50 columnas de nucleótidos libres de gaps, que no se encuentren intercalados por 50 o más gaps en cualquiera de las cepas, aunque no está establecido de manera unánime el tamaño mínimo de nucleótidos y existe discrepancia en cuanto a este tamaño) **(Darlin *et al.,* 2004; Didelot *et al.,* 2009)**.

El genoma central está constituido por todos aquellos genes responsables de procesos básicos de la biología de una especie y sus principales rasgos fenotípicos. Mientras que el genoma variable, contribuye a la diversidad de la especie y puede codificar vías bioquímicas suplementarias y funciones que no son fundamentales para el desarrollo bacteriano, pero que representan ventajas selectivas como adaptación a diferentes ambientes, resistencia a antibióticos o colonización de un nuevo huésped **(Medini *et al.,* 2005)**. Todos estos mecanismos, actúan como una fuerza impulsadora para la diversificación y adaptación de esta especie **(Suerbaum *et al.,* 1998; Suerbaum & Josenhans, 2007; Bubendorfer *et al.,* 2016).**

A través del análisis del pan-genoma se pueden encontrar múltiples proteínas con gran similitud, considerándose homólogos. Estos pueden ser de dos tipos: ortólogas (de genes que divergieron a través de eventos de especiación de un antepasado en común) y parálogas (de genes que divergieron a través de eventos duplicación génica) **(Kuzniar *et al.,* 2008; Fouts *et al.,* 2012)**.

Los estudios llevados a cabo para caracterizar el pan-genoma de *H. pylori* han mostrado resultados variables, debido quizás a la metodología y/o programas implementados o a los parámetros utilizados al momento de elegir la similitud entre las secuencias parálogas que se considerarían como genoma central.

Un resumen sobre estudios del pan-genoma llevados a cabo en *H. pylori* y otras especies del género *Helicobacter* pueden verse a continuación en la **Tabla 2**.

Compartimentos del pangenoma

Numero de genomas analizados	Genoma Central ó core	Genoma variable	Genoma Soft-core	Genoma cloud	Genoma Shell	Origen geográfico de las cepas	Estudio
15	1281	362				Diverso	**Salama *et al.* (2000)**
39	1139	1421				Diverso	**Ali *et al.* (2014)**
47	1059	No data				Diverso	**You et al. (2015)**
36	1266	727				Malasia	**Kumar *et al.* (2015)**
30	1248	991				Diverso	**Uchiyama *et al.* (2016)**
75	1173	3236				Diverso	**Cao *et al.* (2016)**
99	682	11382				*H. pylori* y otras especies de *Helicobacter*	**Cao *et al.* (2016)**
168	399	x	89	12888	1719	*H. pylori* y otras especies de *Helicobacter*	**Smet *et al.* (2018)**

Tabla 2. Estudios del pan-genoma de *H. pylori* y otras especies del género *Helicobacter.* de la fila 1 a la fila 7, estudios con la definición de los compartimentos del pan-genoma de forma clásica: genoma central y genoma variable. Fila 7, estudio de los compartimentos del pan-genoma mediante la metodología **Contreras-Moreira & Vinuesa. (2013)** genoma core, genoma soft-core, genoma shell, genoma cloud.

1.2.5.2 Variación genómica

Son numerosos los mecanismos que pueden contribuir a la alta diversidad genética observada en *H. pylori* la cual le confiere una gran dinámica genómica. Entre estos mecanismos podemos encontrar: mutaciones puntuales (translocaciones, inversiones, deleciones e inserciones), transferencia horizontal de genes (HGT), recombinación inter e intragenómica, reordenaciones cromosómicas, variación de fase de transcripción y traducción, presencia de secuencias repetidas inversas y/o directas, elementos transponibles, pseudogenes y genes de contingencia **(Cooke *et al.,* 2005; Kang & Blaser, 2006 Zepeda-Gurrola, 2012; Draper *et al.,* 2016).**

Además, esta bacteria también se caracteriza por exhibir una competencia natural para la captación de ADN a partir de su entorno **(Yeh *et al.,* 2002)** mediante "*Com apparatus*", que es un sistema de secreción tipo IV (T4SS) **(Baltrus & Guillemin, 2006; Dorer *et al.,* 2011)**. Por último *H. pylori* presenta panmixia y mosaicismo genético **(Suerbaum & Josenhans, 2007).**

1.2.5.3 Reordenaciones genómicas

Cada especie posee una estructura genómica especifica la cual varia lentamente a través del tiempo mediante reordenamientos del genoma, que incluyen inversiones, translocaciones, duplicaciones y transposiciones **(Noureen *et al*., 2019).** Estos reordenamientos genómicos cambian el orden de los genes y la orientación de los mismos para adaptarse a un entorno cambiante y así, dos genomas que pueden parecer funcionalmente idénticos sobre la base de su contenido, génico pueden ser muy diferentes estructuralmente debido al reordenamiento y orientación de sus genes (**Noureen *et al*., 2019)**.

Como se ha mencionado en párrafos anteriores, el reordenamiento del genoma y la alta tasa de mutación de *H. pylori* son características genéticas intrínsecas de esta especie. La capacidad de *H. pylori* de formar reordenamientos genómicos aberrantes e incorporar ADN no homólogo pueden mejorar la diversidad genética facilitando la adquisición y perdida de genes **(Schawartz *et al.,* 2008). Lara et al. (2011)** encontraron evidencias que mostraban que, la recombinación

homóloga contribuye a la aparición de inversiones en *H. pylori* y la recombinación ilegitima estuvo ligada a la generación de repeticiones invertidas (**Lara *et al.,* 2011).**

Por otro lado, se sabe que la replicación y la transcripción del ADN ocurren de forma simultánea en las bacterias, lo que puede conducir a la colisión entre el replisoma y las ARN polimerasas (RNAP), que puede derivar en un colapso de la horquilla de replicación, rompiendo las cadenas de ADN y aumentando la mutagénesis. Así mismo, se ha puesto de manifiesto que estos conflictos de replicación-transcripción pueden ocurrir en cualquiera de las dos orientaciones con consecuencias diferentes **(Merrikh & Merrikh, 2018).** Los conflictos de replicación-transcripción pueden clasificarse de dos formas: 1) Mientras que los encuentros codireccionales (CD) ocurren dentro de los genes codificados en la cadena principal cuando las bifurcaciones de replicación superan a los RNAP y son menos graves y 2) Los conflictos (HO) ocurren en los genes codificados en la cadena rezagada y son mucho más graves, sensibilizando las células ante agotamiento completo o parcial de proteínas de mitigación y aumentando las tasas de mutación local en la localización genómica del conflicto **(Thomason & Storz, 2010; Merrikh & Merrikh, 2018)**.

Podría producirse una reducción de los genes HO a través de la deleción o inversión de genes a la orientación CD. Entre estas dos posibilidades, solo los eventos de inversión son teóricamente capaces de prevenir conflictos de HO sin resultar en la pérdida de genes importantes. Como tal, los eventos de inversión genética son un medio óptimo por el cual la carga del conflicto celular podría reducirse a lo largo del tiempo evolutivo. Se encuentran disponibles al menos dos métodos para identificar genes / fragmentos invertidos: comparación del genoma completamente ensamblado entre cepas de la misma especie y análisis de sesgo de GC. Aunque la comparación del genoma completo es muy precisa, es de bajo rendimiento y requiere la comparación computacionalmente intensiva de genomas específicos. Además, este método solo identificaría eventos de inversión recientes que ocurrieron desde el momento en que los aislamientos en cuestión divergieron. Por el contrario, el análisis de sesgo de GC es más eficiente, ya que no requiere una comparación del genoma

En *H. pylori* se ha observado una transcripción anti-sentido generalizada que ocurre en todo genoma e independiente del contenido local de GC, sin observase un sesgo hacia genes del genoma central o variable **(Thomason & Storz, 2010).** Sin embargo un número importante de genes invertidos están presentes en los genomas de esta especie, cuya presencia es variable. Esto plantea preguntas sobre la historia evolutiva de la arquitectura del genoma, el impacto de

los conflictos de replicación-transcripción y si la presencia de inversiones está relacionada con la prevención de conflictos HO sin resultar en la perdida de genes importantes en esta especie.

1.2.5.4 Secuencias repetidas

Las secuencias repetidas son fundamentales para la dinámica genómica de *H. pylori*, ya que están asociadas con los mecanismos de plasticidad, mencionados anteriormente como la variación de fase, emparejamiento incorrecto de hebra deslizada y mayores tasas de recombinación, deleción e inserción **(Shak *et al.,* 2020).** La secuenciación genómica inicial de las cepas 26695 y J99 de *H. pylori* **(Tomb *et al.,* 1997; Alm *et al.,* 1999)** reveló la presencia de grandes cantidades de ADN repetitivo distribuido de forma no aleatoria en el genoma **(Achaz et al., 2002; Rocha & Viari, 1999). Aras et al. (2003) y Lovett. (2004)** demostraron que gran parte de estas repeticiones son superiores a 24 pb y se encuentran a una distancia inferior de 5 kb entre sí, dentro del rango factible para el deslazamiento de la replicación **(Aras *et al.* 2003; Lovett, 2004).** Debido a que muchos de los mecanismos de reparación por recombinación y reparación de errores comunes están ausentes o modificados en *H. pylori* **(Wang *et al.,* 1999; Pinto *et al.,* 2005)**, esta bacteria es especialmente susceptible a los efectos de diversificación del ADN repetitivo **(Shak *et al.,* 2020)**. De hecho, se ha demostrado que los loci en el genoma de *H. pylori* que contienen ADN repetitivo, muestran una amplia variación entre hospedadores (**Kang & Blaser, 2006; Kulick *et al.,* 2008)**.

Las secuencias repetidas e inversas están formadas por dos o más secuencias duplicadas, donde una o más de ellas se hallan en dirección opuesta **(Saunders *et al.,* 1998; Lovett, 2004; Zepeda-Gurrola, 2012)**. **Furuta et al. (2011)** propusieron que, en *H. pylori* un proceso por el cual ocurre la ganancia o pérdida de genes es un mecanismo de duplicación de ADN, denominado "duplicación del ADN asociado con inversiones" (DDAI). DDAI es similar al proceso de inversión replicativo que utilizan los transposones de ADN específicos **(Kawai *et al.,* 2006)** pero este se diferencia en dos aspectos fundamentales: la naturaleza del ADN duplicado y las secuencias en los puntos de corte **(Furuta *et al.,* 2011).**

Tom *et al.* (1997), Alm *et al.* (1999) demostraron la presencia de repeticiones simples en *H. pylori*, las cuales eran puntos calientas para el desajuste de la cadena de deslizamiento y por lo tanto, actuaban como loci de contingencia para la generación de poblaciones heterólogas **(Tom *et al.,* 1997; Alm *et al.,* 1999)**. Se ha demostrado que las SSR afectan la virulencia y la adaptación al huésped **(Appelmelk *et al.,* 1999; Solnick *et al.,* 2004)**. En este sentido, los SSR

intragénicos pueden afectar la traducción mediante la introducción de mutaciones de cambio de marco dentro de las regiones codificante que conllevan a la terminación prematura de la traducción o alteraciones en los extremos C terminales de las proteínas **(Josenhans *et al.,* 2000; Bayliss & Palmer, 2012; Zhou *et al.,* 2013; Aberg *et al.,* 2014)**. Los SSR intergénicos pueden influir en la transcripción cambiando el espacio de los elementos promotores o los sitios de unión del factor de transcripción **(Martin et *al.,* 2005)**. Mientras que, los mecanismos y funciones de los SSR en la regulación génica a nivel del ADN están establecidos, los efectos sobre la estabilidad del ARNm o el control postranscripcional son menos conocidos **(Saunders *et al,* 1998; Coenye & Vandamme, 2005; Zhou *et al.,* 2013; Aberg *et al.,* 2014).** En *H. pylori*, el impacto de los SSR ha sido observado principalmente en genes que codifican para proteínas de membrana externa, como: *alpA*, *alpB*, *babA*, *babB*, *sabA* y *sabB* **(Saunders *et al,* 1998; Coenye & Vandamme, 2005)**.

También se ha descrito un número importante número de repeticiones directas dentro de los marcos de lectura abiertos de los genes *amiA*, *cagY* y *cagA* **(Aras *et al.,* 2003)**. Así, por ejemplo el gen *cagA* que es un importante factor de virulencia, destaca por la presencia de un número de motivos variables Glu-Pro-Ile-Try-Ala (EPIYA), los cuales se han clasificado en cuatro segmentos EPIYA (EPIYA-A-EPIYA-D) y que sirven como sitios de fosforilación de tirosina y los cuales se han indicado como fundamentales en la tumorogénesis *in vivo* **(Hatakeyama, 2014)**.

1.2.5.5 Secuencias de Inserción (IS)

Los IS clásicos son miembros del grupo de entidades genéticas denominados, elementos genéticos transponibles o móviles (TE o MGE) y presentan una longitud de entre 0,7 y 2,5 kb. Estos exhiben uno o dos marcos de lectura abierto que ocupan la totalidad del IS y terminan en secuencias de repetición terminal imperfecta (IR) que los flanquean **(Chandler *et al.,* 2015 -book mobiles elements).** En *H. pylori* se conocen hasta la fecha los siguientes IS: *IS200/IS605*, *IS606*, *IS607*, *IS608* e *IS609* **(Noureen *et al.,* 2021)**.

En el caso de *H. pylori* una vía de transposición, está asociada a las transposasas de tipo HUH, las cuales realizan un reconocimiento del extremo del transposón **(Chandler *et al.,* 2015)**. Las TPases HUH utilizan la tirosina como nucleófilo y generan un intermedio de transposición de ADN tirosina 5' covalente transitorio **(Chandler *et al.,* 2013)**. Estas transposasas son de ADN monocatenario de la familia *IS200/IS605* y solo movilizan sus transposones una vez estas son accesibles en su forma monocatenaria, por ejemplo en las hebras de retraso durante la

replicación o durante algunos tipos de reparación del ADN **(Chandler *et al.,* 2015; Ton-Hoang *et al.,* 2010; Mennecier *et al.,* 2006)**.

Esta familia *IS200/IS605* se divide en tres grupos según la presencia o ausencia de dos transposones: *tnpA* (codifica una TPase Y1) y *tnpB* (con función desconocida) **(Lam & Roth, 1983).** El grupo *IS200* comprende sólo *tnpA* y el grupo *IS1341* comprende *tnpB*, mientras que, el grupo *IS605*, comprende tanto *tnpA* como *tnpB* **(Ton-Hoang *et al.,* 2005; Murai *et al.,* 1998)**. En el caso de *H. pylori* se encuentran presentes *IS200* e *IS605*.

1.2.5.6 Evolución del contenido de GC en genomas bacterianos

La genómica comparada entre cepas bacterianas puede ayudar a revelar datos interesantes sobre la organización, evolución, patogenicidad y funcionamiento interno del genoma bacteriano **(Grigoriev, 2000; Lassalle *et al.,* 2015).** En este sentido, se han obtenido algunos datos generales como: a) la composición de bases de las secuencias genómicas varia ampliamente tanto entre especies como entre los cromosomas de una especie **(Lassalle *et al.,* 2015),** b) existe un sesgo en la composición de nucleótidos entre las cadenas principales y rezagadas **(Lobry, 1996; Casjens, 1998; Frank & Lobry, 1999; Bentley & Parkhill, 2004)**, y c) localización computacionalmente el origen de replicación y terminación, así como para las reordenaciones cromosómicas **(Kono *et al.,* 2018).**

1.2.5.7 Variación en contenido de GC

El contenido de nucleótidos genómicos varía mucho en las bacterias, con un contenido de GC que va desde el 13% al 75% entre especies. La variación en la composición de nucleótidos también puede ser sustancial entre cepas. Aunque se desconocen las causas específicas de estas variaciones de GC, tanto dentro como entre especies, se cree que son responsables una multitud de factores relacionados tanto con la historia evolutiva como con el medio **ambiente (Du *et al.,* 2018)**.

Entre los factores que muestran alguna asociación con la composición de bases genómicas en los procariotas se incluyen: el tamaño del genoma, la abundancia de oxígeno y nitrógeno, así como la capacidad de captación de ADN extraño mediante conjugación, transformación y transducción. La temperatura de crecimiento óptima también, puede influir en la composición del ADN genómico y, aunque este es un campo de debate, existe alguna evidencia de un papel de la

temperatura de crecimiento en la conformación del contenido de GC de genes individuales y ARN ribosómico. La riqueza de GC puede estar impulsada por la selección de un ADN más estable, ya que el apilamiento (y la rotura) de guanina y citosina normalmente requiere más energía que el de adenina y timina **(Merrikh & Merrikh, 2018).** Los genomas ricos en GC también pueden haber sido sometidos a selección para un uso de aminoácidos más favorable desde el punto de vista energético, ya que los codones ricos en GC codifican para aminoácidos que requieren menos energía que los codones ricos en AT. Además, muchas bacterias "silencian" secuencias de ADN ricas en AT extrañas, que a menudo se encuentran en los fagos. Por otro lado, se ha sugerido que la relajación de las presiones selectivas impulsa los genomas microbianos simbióticos hacia una mayor riqueza de AT debido al sesgo mutacional de AT y la pérdida de genes de reparación del ADN. Se ha descubierto que las partes no codificantes de los genomas procariotas son más ricas en AT que las partes codificantes y esto podría deberse a presiones selectivas relajadas en las regiones no codificantes en comparación con las regiones codificantes.

Asimismo, las regiones de codificante en el genoma central son significativamente más ricas en GC, tienen menos variación de contenido GC y una entropía relativa más alta, es decir, distribuciones de oligonucleótidos más sesgadas que las regiones de codificación del resto de los genomas correspondientes **(Merrikh *et al.,* 2012)**

1.2.5.8 Diferencia entre la cadena principal y rezagada

Está bien establecido que la guanina es más abundante que la citosina a lo largo de la cadena principal de cada hebra cromosómica, lo que da como resultado un sesgo de GC promedio positivo en ella. Los mecanismos evolutivos que influyen en la composición de nucleótidos incluye recombinación, mutaciones y presiones selectivas, con opiniones encontradas respecto al impacto que ella tiene sobre procesos como la replicación y la transcripción **(Francino *et al.,* 1996; Rocha *et al.,* 2006; Chen *et al.,* 2016),** En este sentido se puede decir que **Bhagwat *et al.* (2016)** demostraron experimentalmente que el sesgo de GC está influenciado por la replicación.

Algunas evidencias sugieren que el patrón de mutaciones espontáneas está sesgado hacia los nucleótidos AT, tanto en eucariotas como en procariotas Bajo tal sesgo, se espera que la presión selectiva favorezca generalmente una maquinaria de reparación de ADN sesgada por GC. La recombinación y la reparación son procesos estrechamente vinculados que utilizan muchas vías comunes. En la levadura, el análisis de la reparación en los productos meióticos indica que el

sesgo de conversión probablemente se deba a la maquinaria de reparación de desajustes (MMR). Los genes de los componentes de MMR implicados en la recombinación homóloga (MutS, MutL) generalmente se conservan entre bacterias y eucariotas. Sin embargo estos genes están ausentes en tres genomas bacterianos, *Campylobacter jejuni, H. pylori* y *Bifidobacterium longum,* como resultado de pérdidas ancestrales en *Delta-Proteobacteria* y *Actinobacteridae*.

1.2.5.9 Estructura cromosómica

La estructura cromosómica global es sugerida por la distribución no aleatoria de genes dentro de los genomas. Los orígenes de replicación y terminación, únicos, distribuyen típicamente los genes bacterianos casi simétricamente en dos réplicas de aproximadamente el mismo tamaño. Se sabe que la ubicación de algunos genes en relación con el origen de replicación es importante. Por ejemplo, la proximidad del gen spoIIR de *Bacillus subtilus* al origen de replicación permite su transcripción desde la espora recién formada **(Dworkin & Losick 2001).** Además, los genes *dnaA* están significativamente asociados con los orígenes de la replicación.

Fuera de estos casos especiales, se ha postulado poca importancia para las posiciones de otras unidades de transcripción en relación con el origen de replicación más allá del potencial de una mayor dosis de genes proximales al origen **(Liu & Sanderson 1995b, 1996).** Sin embargo, podemos inferir que el orden de los genes está determinado, ya que los mapas genéticos mantienen el orden frente a los mecanismos que pueden reorganizarlos. Más importante aún, los reordenamientos observados suelen ser simétricos con respecto a los orígenes y terminaciones de la replicación **(Eisen *et al.* 2000; Mackiewicz et al. 2001; Sanderson y Liu 1998; Suyama & Bork 2001; Tillier y Collins 2000),** lo que sugiere que las inversiones que reorganizan la estructura cromosómica (es decir, las que mueven los genes de cadenas que conducen a cadenas rezagadas) se contrarrestan.

1.2.6. Evolución y estructura poblacional

Hasta la fecha, el humano es el único huésped conocido para *H. pylori* y su evolución está estrechamente ligada a la historia evolutiva y migración de los primeros humanos modernos **(Falush *et al.,* 2003; Linz e*t al.,* 2007; Moodley *et al.,* 2009)**, ya que compartimos una estrecha relación co-evolutiva que abarca desde unos 60.000 a 100.000 años y posiblemente más **(Yamaoka, 2010; Moodley, 2014; Roesler *et al.,* 2014)**. La huella de dicha co-evolución se ve

reflejada en las secuencias de ADN, mostrando una alta heterogeneidad de las diferentes cepas según sus orígenes geográficos. Esta correlación ha posibilitado el estudio de las migraciones humanas desde fuera de África y es posible que esta larga e íntima relación **(Montano *et al.*, 2015)** haya permitido una amplia y diversa utilización de estrategias por parte de esta bacteria para la adaptación, colonización y el establecimiento en el nicho gástrico **(Suerbaum *et al.*, 2003, Blaser *et al.*, 2004).**

Como ya se ha comentado, a pesar de la alta diversidad genética de *H. pylori* y de no estar prácticamente relacionadas epidemiológicamente **(Mégraud *et al.*, 2016)**, esta bacteria exhibe un comportamiento genéticamente estructurado, el cual puede ser observado a través de patrones filogeográficos **(Kabamba *et al.*, 2018**). El primer estudio poblacional para comprender la alta variabilidad de *H. pylori*, implementó el uso de secuencias de ADN de un conjunto de fragmentos de genes constitutivos concatenados (*mutY, ppa, trcP, ureA y yphC*), además de dos genes asociados a la virulencia como son *vacA* y *cagA*, aislados desde diferentes regiones geográficas los cuales fueron la base para el tipaje de secuencias multilocus (MLST) **(Achtman *et al.*, 1999)**. En este estudio fue posible distinguir la genealogía a pesar de las recombinaciones frecuentes **(Achtman *et al.*, 1999)**. Estos datos fueron corroborados por **Falush et al. (2003)** al secuenciar 370 aislamientos de cepas *H. pylori* desde 27 regiones geográficas de diferentes grupos humanos **(Falus *et al.*, 2003)**.

Originalmente, **Falush et al. (2003),** definieron cuatro poblaciones principales hpAfrica1, hpAfrica2, hpEastAsia y hpEurope. Sin embargo, han ido surgiendo nuevas subestructuraciones: en África (hpNEAfrica, hpAfrica1 y hpAfrica2), en Europa (hpEurope) y en Asia (hpEasia, hpAsia2 y hpSahul) **(Falush *et al.*, 2003b; Linz *et al.*, 2007; Moodley & Linz, 2009; Moodley *et al.*, 2012)**. En los últimos el aumento del número de genomas estudiado ha permitido afinar más el número de subpoblaciones: hpEastAsia (hspAmerind, hspEAsia y hspMaori), hpNEAfrica (hspEastAfrica y hspCentralNEAfrica) y hpAfrica1 (hspSAfrica, hspWafrica y hspCAfrica) **(Tabla 3 y Figura 3) (Moodley *et al.*, 2012; Nell *et al.*, 2013).**

Population	Subpopulation	Geographic range
hpAfrica2	hspNorthSan	Namibia, Angola
	hspSouthSan	South Africa
hpAfrica1	hspWAfrica	Senegal, the Gambia, Burkina Faso, Morocco, Algeria, Nigeria, Cameroon, South Africa
hpNEAfrica	hspCAfrica	Cameroon, Namibia
	hspSAfrica	Namibia, Angola, South Africa
	hspENEAfrica	Sudan, Ethiopia, Somalia, Algeria
	hspCNEAfrica	Sudan, Cameroon, Nigeria, Algeria
hpSahul	hspAustralia	Australia
	hspNGuinea	New Guinea
hpAsia2	hspLadakh	India (Himalaya)
	hspIndia	India, Bangladesh, Malaysia, Thailand, the Philippines
hpEurope	Recombinant population	Europe as far east as Southeast Asia
hpEastAsia	hspEAsia	China, India, Malaysia, Singapore, Taiwan, Cambodia, Vietnam, Japan, Korea
	hspAmerind	Canada, the USA, Venezuela, Colombia, Peru
	hspMaori	Taiwan, the Philippines, Japan, Samoa, New Caledonia, Wallis and Futuna, New Zealand

Tabla 3. Resumen de las poblaciones y subpoblaciones modernas de *Helicobacter pylori*, incluida su distribución geográfica hasta donde se conoce**.** Adaptado desde **Moodley Chapter 1 *Helicobacter* pylori: Genetics, Recombination, Population Structure, and Human Migrations *Helicobacter pylori* Research (2016).**

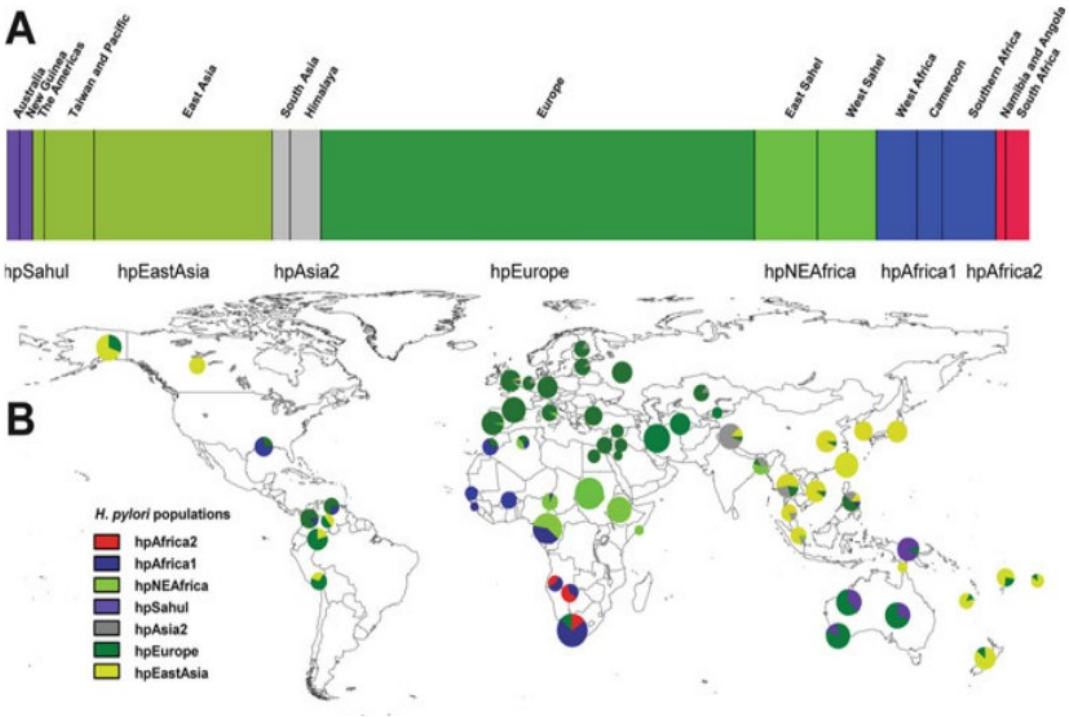


Figura 3. Panorama global de la estructura genética de *H. pylori*. **(a)** La estructura de la población global de *H. pylori* determinada mediante el análisis de agrupamiento bayesiano de las secuencias de genes de mantenimiento de 1716 cepas individuales. Hasta ahora se han descubierto siete poblaciones. Las ubicaciones indican la subestructura geográfica dentro de cada población. **(b)** Estructura geográfica mundial de siete poblaciones *de H.*

pylori. Las discontinuidades geográficas obvias en la distribución de hpEurope y hpAfrica1 reflejan los movimientos humanos recientes asociados con la expansión colonial europea y la trata de esclavos, respectivamente. Las distribuciones naturales estructuradas de las poblaciones permiten que se utilicen como marcadores para las migraciones humanas. Tomado de **Moodley Chapter 1 *Helicobacter* pylori: Genetics, Recombination, Population Structure, and Human Migrations *Helicobacter pylori* Research (2016).**

1.2.7 Evolución concertada

La evolución concertada escenifica la coevolución de secuencias de ADN dentro de miembros de una familia relacionada, de tal forma que conservan altos niveles de similitud entre sí, ya que estos divergen de los mismos miembros de la familia en otras especies o cepas **(Pride & Blaser, 2002).** Se han encontrado evidencias sobre la evolución concertada evidencias tanto en procariotas como eucariotas y se observa con mayor frecuencia en los genes de ARNr **(Wang & Chen, 2018).**

En el modelo de evolución concertada, **Brown et al. (1972),** propusieron que el proceso de homogeneización en el ADN requiere que las repeticiones individuales evolucionen de forma dependiente unas de otras **(Brown *et al.,* 1972)**. Así, la transmisión y acumulación de mutaciones que ocurren en la región de repetición se homogeneiza, es decir, las mutaciones se extienden por la matriz de repeticiones **(Naidoo *et al.,* 2013)**.

De esta forma por ejemplo los genes parálogos de una especie muestran una mayor similitud de secuencia entre sí que con los ortólogos de otras especies y frecuentemente, se agrupan de forma monofílica en el árbol distancia **(Wang & Chen, 2018).** Sin embargo, dicho patrón puede también surgir de la duplicación de genes específicos del linaje y para discernir entre estos escenarios, es esencial tener en cuenta la sintenia genética para dar luz a la ortología y la paralogía del gen en cuestión **(Nei & Rooney, 2005; Mansai & Innan, 2010).**

La conversión de genes es un mecanismo que favorece la evolución concertada, ya que acarrea la sustitución de un miembro de la familia por otro **(Pride & Blaser, 2002).** La conversión genética se define como la transferencia no recíproca de información genética de un sitio entre dos secuencias de ADN homólogas, pero no idénticas **(Santoyo & Romero, 2005)**. La mitad de las veces, este intercambio está asociado con el intercambio de segmentos flanqueantes y cuando esto ocurre, un segmento de una secuencia se transfiere a su zona homologa **(Figura 4).**

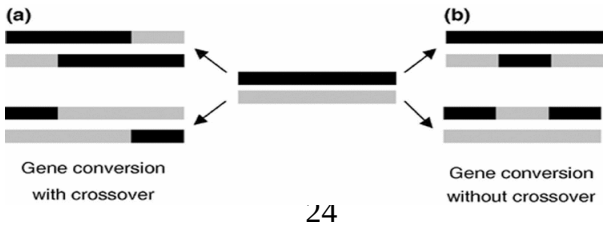


Figura 4. Asociación entre cruces y conversión génica. (a). La conversión génica se pude asociar la mitad de las veces a un cruce o un intercambio de regiones flanqueantes. **(b).** Para el resto de los casos, la conversión de genes ocurre sin una asociación de cruces. Tomado de **Santoyo & Romero. (2005)**.

La conversión de genes está presente en los tres dominios de la vida, arqueas, bacterias y eucariotas. En muchas bacterias patógenas, la conversión de genes es un mecanismo para mejorar la diversidad en un proceso que se denomina variación antigénica **(Santoyo & Romero, 2005; Vink *et al.,* 2012; Foley, 2015)**. Este proceso ha evolucionado de forma independiente en un grupo de bacterias filogenéticamente diferentes y se ha demostrado que, diferentes mecanismos moleculares operan en su evolución, entre ellos la conversión de genes **(Wasser *et al.,* 2021)**. La variación antigénica facilita a las especies patógenas eludir la respuesta inmune del huésped a través del cambio repetido de las estructuras de la superficie. Cuando el mecanismo molecular es la conversión genética, el genoma normalmente posee una copia funcional del gen respectivo pero además, puede contener en el genoma varias o muchas copias no expresadas, que sirven como sitios donantes para la conversión de genes.

Moran et al. (2008) sugieren que la ventaja selectiva de la evolución concertada (EC) de genes puede relacionarse con un efecto de "dosis" para la expresión génica. Esta sugerencia es fuertemente respaldada por una correlación estadística positiva encontrada entre los niveles de expresión génica y la EC de genes duplicados **(Sugino e Innan, 2006).**

En relación a la conversión génica, **Pride & Blaser. (2002),** mediante la prueba de Sawyer proporcionaron evidencia de la evolución del segmento 3' de los genes *babA* y *babB* a través de la conversión de genes en *H. pylori* y pusieron de manifiesto que el foco más informativo para determinados estudios evolutivos pueden ser segmentos de genes y no genes completo. Además, propusieron varios mecanismos que podrían explicar la evolución concertada en estos genes como: entrecruzamiento desigual, deslizamiento de la replicación y la conversión de genes**. Jordan et al. (2001)** encontraron 18 genes de la familia Hop que codifican para proteínas de membrana en *H. pylori* con conversión génica. En ellos pudieron observar que, cuatro de seis eventos emparejados mostraban conversiones completas de la región codificante completa y en dos casos en la región flanqueanté**. Santoyo & Romero. (2004)** mediante el estudio de cepas merodiploides artificiales de *H. pylori*, detectaron conversión de genes, mediante la inserción de una copia silenciosa y truncada del gen *rpsL* (que codifica la proteína ribosómica S12) en una posición ectópica, revelando la homogeneización de la mutación entre las dos copias **(Santoyo & Romero, 2004).**

Todos mecanismos que conllevan una gran capacidad de evolución adaptativa, ha convertido a *H. pylori* en un organismo modelo para comprender la naturaleza de la evolución adaptativa y de la selección natural, permitiendo establecer que información se transfiere a siguientes generaciones **(Kobayashi, 2014)**.

1.2.7.1 Evolución concertada y selección positiva

En algunos casos, el proceso evolutivo de la evolución concertada, que promueve la homogeneidad entre genes estrechamente relacionados, aparentemente se podido combinar con la selección darwiniana positiva, que promueve la heterogeneidad entre genes estrechamente relacionados, de poblaciones de la misma especie favoreciendo distintas variantes génicas del mismo gen en ambientes distintos. Esta se ha explicado por la necesidad constante de nuevos compuestos en un entorno cambiante**. (Froy *et al.,* 1999; Duda & Palumbi 1999; Zhu *et al.,* 2004; Lynch, 2007)**

1.2.8 Sistemas CRISPR-Cas

Las repeticiones palindrómicas cortas agrupadas regularmente interpuestas (CRISPR) y las proteínas asociadas a estas Cas/Cmr o RAMP (proteínas misteriosas asociadas a repeticiones) son una forma de inmunidad adquirida en bacterias y arqueas, que abarcan complejos mecanismos para la integración de ácidos nucleicos extraños, principalmente de elementos móviles genéticos (MGE) **(Koonin & Makarova, 2019; Borges *et al.,* 2017; Cooper & Overstreet, 2014; Mojica *et al.,* 2005).** Se ha propuesto que estos sistemas están presentes en el 90% de las arqueas con un 70% de módulos Cmr **(Barrangou & Marraffini, 2014; Westra *et al.,* 2014; Karginov & Hannon, 2010)** y en cerca del 48% de las bacterias, con aproximadamente un 30% de módulos Cmr **(Makarova *et al.,* 2015; Barrangou & Marraffini, 2014; Westra *et al.,* 2014; Karginov & Hannon, 2010).**

Los sistemas CRISPR-Cas están constituidos por una matriz CRISPR que posee secuencias repetidas y espaciadores, una secuencia líder y proteínas Cas/Cmr asociadas a la matriz **(Karginov & Hannon, 2010; Kuzminov, 2014).** El número de matrices CRISPR-Cas/Cmr presentes en los genomas es variable, variando desde 1 a 15 y la longitud de cada locus CRISPR puede también variar, oscilando entre un solo espaciador entre dos repeticiones, hasta

un locus con 587 espaciadores **(Cooper & Overstreet, 2014; Kumar & Chen, 2012- desde Qiu 2016).**

Las matrices CRISPR utilizan proteínas Cas y Cmr que evolutivamente y funcionalmente son muy diversas y con objetivos (targets) diferentes. Es así que, las proteínas Cas van dirigidas a ADN exógeno y las proteínas Cmr o RAMP a ARN **(Hale *et al.,* 2009)**. Los sistemas CRISPR-Cas/Cmr funcionan mediante la inserción de fragmentos de ADN extraño en la matriz CRISPR, que posteriormente pasarán a ser denominadas espaciadores. Tras una infección, estos se transcribirán y los respectivos transcriptos serán procesados para generar ARN de CRISPR maduros que contendrán los espaciadores utilizados como guía para reconocer y escindir ADN o ARN específico de elementos genéticos extraños **(Faure *et al.,* 2019)**.

1.2.8.1 Clasificación de los sistemas CRISPR-Cas

Una característica fundamental de los sistemas CRISPR-Cas/Cmr es su versatilidad, que resulta en una gran diversificación de los sistemas CRISPR debido al contexto armamentístico en el que se desarrollan y evolucionan contra cualquier MGE o fago, derivando en nuevos tipos y subtipos de sistemas CRISPR-Cas/Cmr **(Koonin *et al.,* 2017; Burstein *et al.,* 2017; Makarova *et al.,* 2015, 2011).** En este sentido, las secuencias de las proteínas Cas y la organización genómica de los loci CRISPR-Cas muestran una diversidad sustancial **(Koonin & Makarova, 2019)**. Todos los sistemas CRISPR-Cas se dividen en dos clases distintas, en función de los módulos efectores. Así, los sistemas de Clase 1 poseen complejos efectores de múltiples subunidades que comprenden varias proteínas, mientras que los sistemas de Clase 2, el efector es una proteína grande, única y multidominio **(Figura 5) (Makarova *et al.,* 2015)**. La clasificación de los sistemas CRISPR no es una cuestión fácil, ya que no existen proteínas Cas universales que puedan usarse como marcadores filogenéticos e incluso la filogenia de la proteína más conservada evolutivamente, Cas1, no refleja de manera satisfactoria las relaciones entre los sistemas CRISPR-Cas debido a la evolución semi-independiente de los diferentes módulos **(Koonin & Makarova, 2019)**.

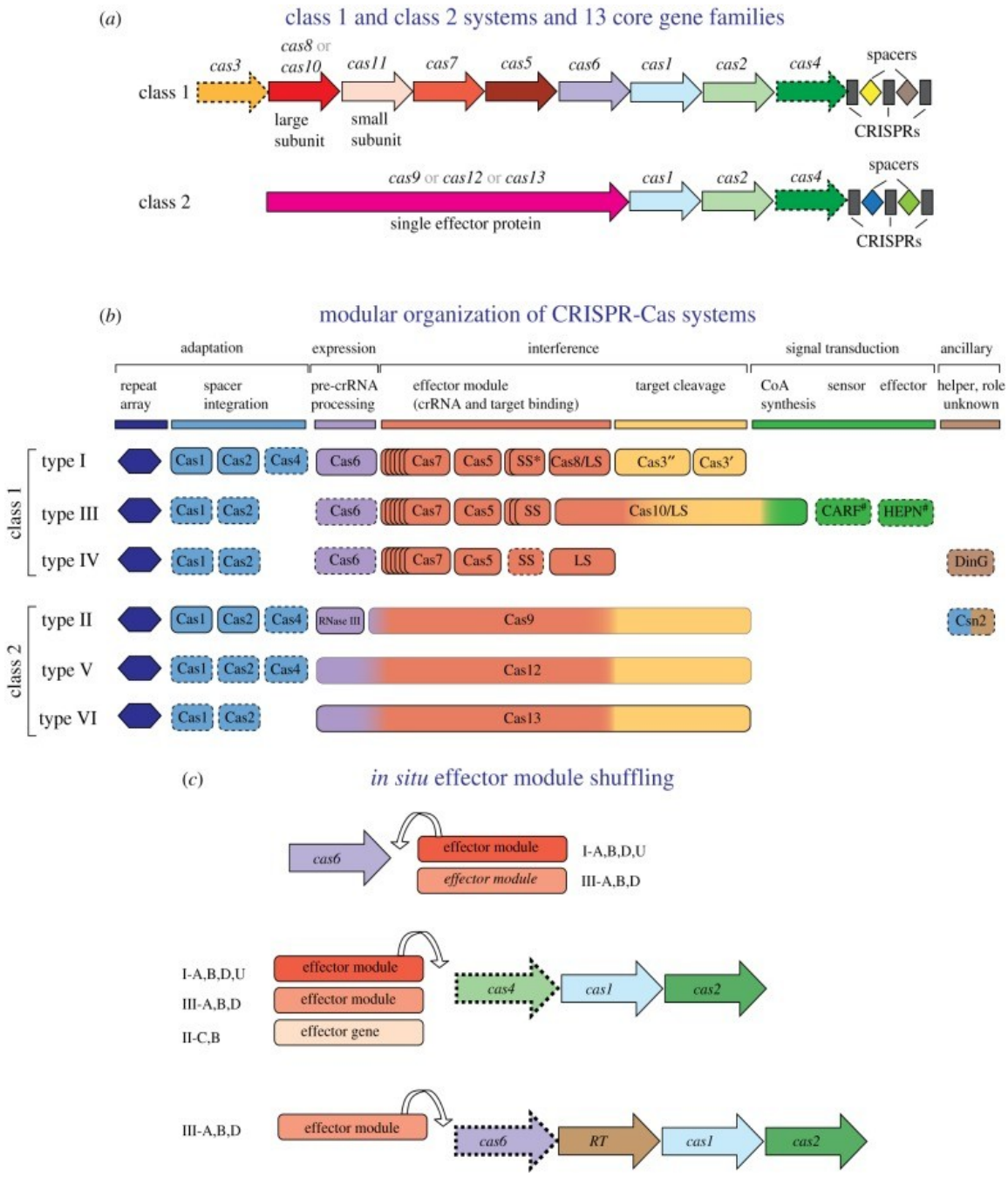


Figura 5. Sistemas CRISPR-Cas de clase 1 y clase 2: características clave, organización modular y reorganización de módulos. **(a)** Las arquitecturas generales de los sistemas CRISPR-Cas de clase 1 (complejos efectores multiproteicos) y clase 2 (complejos efectores de una sola proteína). Los genes se muestran como flechas; los genes homólogos se muestran con el mismo color. Los nombres de los genes siguen la nomenclatura y clasificación actuales. **(b)** Los principales componentes básicos de los tipos de sistemas CRISPR-Cas. Un asterisco indica la supuesta subunidad pequeña (SS) que podría fusionarse con la subunidad grande en varios subtipos de tipo I. El # junto a las etiquetas de los dominios CARF y HEPN indica que otros dominios efectores y sensores desconocidos pueden estar involucrados en la ruta de señalización. Los genes prescindibles se indican mediante un contorno punteado. El panel muestra esquemáticamente inferiores módulo arrastrando los pies en CRISPR- cas loci. Tomado de Koonin & Makarova, (2019).

1.2.8.2 Funciones alternativas de los sistemas CRISPR-Cas/Cmr

Hoy en día, se sabe que más allá de su rol en la inmunidad adaptativa de los sistemas CRISPR-Cas/Cmr y del conocimiento acerca de la matriz CRISPR en relación a su capacidad de cambiar de forma activa en respuesta a las infecciones virales **(Bozic *et al.,* 2019).** Así, también se ha propuesto que juegan un papel crítico en la regulación de la expresión génica endógena y en la regulación de importantes fenotipos bacterianos basados en el estilo de vida, como la patogenicidad **(Barrangou, 2015)**; ya que, como se sabe en muchos nichos los fagos y plásmidos evolucionan rápidamente haciendo obsoleta la función de defensa **(Makarova *et al.,* 2016)**, favoreciendo la adaptación y proveyendo beneficios evolutivos qué, incrementan la aptitud física como la resistencia a antibióticos y factores de virulencia **(Barrangou *et al.,* 2019)**.

Estos nuevos roles de los sistemas CRISPR-Cas/Cmr están relacionados con procesos de regulación transcripcional para regular la patogenicidad y la transferencia horizontal de genes, también en la reparación del ADN y como mecanismo de respuesta al estrés **(Westra et al., 2014; Barrangou, 2015; Koonin & Makarova, 2019).**

1.2.8.3 Matrices CRISPR huérfanas

En los últimos años, se ha encontrado matrices CRISPR huérfanas o no asociados a proteínas Cas/Cmr **(Makarova *et al.,* 2015)**. Generalmente, este tipo de matrices son consideradas vestigiales, aunque algunas de ellas podrían ser consideradas como funcionales **(Almendros *et al.,* 2016).** En este sentido, se ha observado que matrices CRISPR huérfanas del tipo IF encontradas en *Escherichia coli* controlan e interfieren con el material genético invasor, el cual puede albergar proteínas Cas funcionales, de esta manera evitan la adquisición de un sistema inmunitario especifico **(Almendros *et al.,* 2016).**

En relación al Género *Helicobacter,* los sistemas de tipo II están presentes en *H. cinaedi* y *H. mustelae*, mientras que el tipo III ha sido detectado en *H. cetorum* **(Burstein *et al.,* 2017).** En la especie *H. pylori,* se ha descrito la presencia de CRISPR-like por **García-Zea *et al.* (2019)**. Estos CRISPR-like fueron caracterizados en el gen *vlpC*, gen parálogo de *vacA* , cuya función no ha sido esclarecida en su totalidad, aunque se ha asociado a la resistencia de *H. pylori* a metronidazol **(Albert *et al.,* 2005).** La presencia de las matrices CRISPR podría tener relación con procesos de recombinación y el incremento de la variabilidad **(García-Zea *et al.,* 2019).**

1.2.9 Pan-inmune

Los virus se encuentran en cualquier entorno natural, se les considera las entidades biológicas más abundantes de la tierra y se he estimado que hay ~10^{31} fagos presente en la biosfera **(Fortier & Sekulovic, 2013).** En general, se cree que los virus superan en número a los huéspedes microbianos en una proporción de 10 a 1 o más **(Wommack & Colwell, 2000; Parikka, Wauters & Jacquet, 2017)**. Su abundancia se traduce en un 20 al 40% de la mortalidad bacteriana diaria. Esta abundancia, también implica que las bacterias estén bajo una constante presión evolutiva promovida por sus invasores virales **(Suttle, 2007)**.

Frente a la abundancia y diversidad de virus y la presión ejercida por estos sobre las bacterias y arqueas, estas han desarrollado diversas líneas de defensa que implican variados mecanismos inmunitarios, tanto innatos como adaptativos para hacer frente a esta presión, obstaculizando diferentes etapas del ciclo de vida del virus **(Dy *et al.,* 2014; van Houte, Buckling & Westra., 2016; Hampton, Watson & Fineran, 2020).** En su conjunto, estas líneas de defensa se han denominado el sistema inmunológico procariota o sistema "pan-inmune" **(Bernheim & Sorek., 2020)**.

Los genes asociados contra la defensa de los virus pueden llegar a constituir hasta el 10% del genoma microbiano **(Koonin, Makarova & Wolf, 2017)**. Sin embargo el aumento en la disponibilidad de datos genómicos y la aplicación de enfoques bioinformáticos han permitido expandir significativamente el estudio de estos sistemas, permitiendo la predicción de nuevos grupos de genes ligados contra la defensa de virus **(Isaev, Musharova & Severinov, 2021)**.

Los sistemas de defensa contra virus pueden dividirse en aquellos que se dirigen a ácidos nucleicos virales (restricción modificación RM y CRISPR-Cas), sistemas tipo Abi que llevan al huésped a la apoptosis una vez infectado **(Figura 6).**

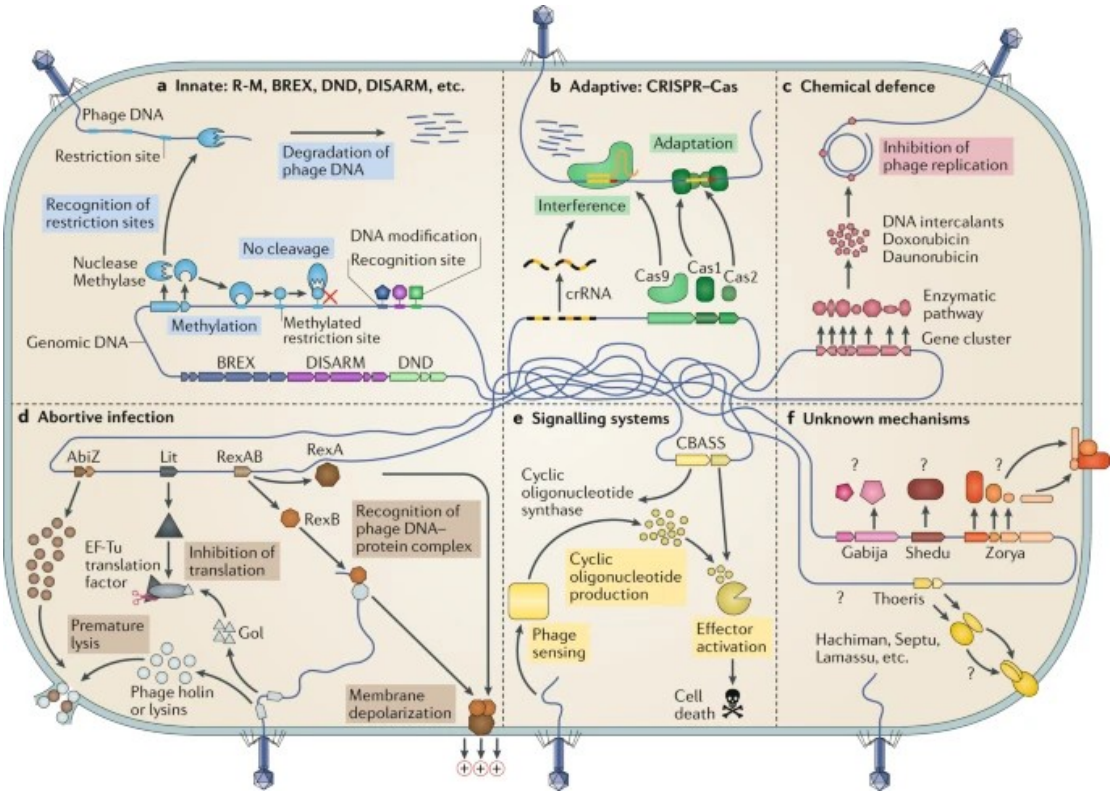


Figura 6. Los sistemas de defensa que se dirigen a los ácidos nucleicos abarcan la inmunidad innata como la adaptativa. **A).** Sistemas de restricción modificación (RM) y otros sistemas relacionados modifican motivos de secuencia específicos en el genoma del huésped y escinden o degradan el ADN extraño no modificado. **B).** Los sistemas CRISPR-Cas actúan en dos fase principales: adaptación e interferencia. **C).** Se ha descrito un sistema de defensa químico en que las bacterias producen una pequeña molécula anti-virica que se intercala en el ADN del fago e inhibe su replicación. **D).** Mecanismos abortivos que son diverso y junto con las holinas codificados por los fagos y las lisinas del fago aceleran la lisis antes de que se complete el ensamblaje del fago. **E).** CBASS (sistema de señalización anti-fagos basado en oligonucleótidos) detecta la presencia de fagos y genera una señal, que activa un efector que conduce a la muerte celular. **F).** Recientemente se ha demostrado que múltiples sistemas tienen funciones anti-fagos, pero sus mecanismos siguen siendo desconocidos. Abi (infección abortiva), BREX (exclusión de bacteriófagos), DISARM (sistema de islas de defensa asociado a RM). Tomado de **Bernheim &, Sorek., 2020**.

Un esfuerzo por mapear islas de defensa microbiana ha resultado en el descubrimiento de nuevos sistemas de defensa que se encuentran muy extendidos en los genomas bacterianos y de arqueas, cuyo mecanismo de acción molecular aún no han sido descifrado. Estos sistemas recibieron el nombre de deidades protectoras de la mitología mundial, como Hachiman, Thoeris, Zorya, Gabija y Shedu **(Doron *et al.*, 2018)**.

Para cada estrategia de defensa microbiana conocida, los virus desarrollan medios para la contra defensa **(Isaev, Musharova & Severinov, 2021)**, es por ello que las bacterias y arqueas

no pueden depender de una única línea de defensa y, por lo tanto, deben presentar varios sistemas de defensa como estrategia de supervivencia de cobertura de apuestas. Actualmente se sabe que los genomas procariotas pueden albergar simultáneamente diferentes sistemas de defensa **(Oliveira *et al.,* 2014; (Koonin, Makarova & Wolf, 2017; Koonin *et al.,* 2020)**. De hecho, una sola cepa puede llegar a codificar diversas líneas de defensa, que pueden incluir Abi, RM y CRISPR-Cas **(Bernheim & Sorek., 2020)**. También hay bacterias y arqueas que pueden codificar varios sistemas de defensa del mismo tipo, como por ejemplo la cepa de *H. pylori* F30, que codifica cuatro sistemas de restricción (RM) (3 del tipo RM I, 11 del tipo RM II, 1 tipo RM III y 1 tipo RM IV) **(Oliveira *et al.,* 2014)**.

Las bacterias y arqueas adquieren los sistemas de sistemas de defensa a través de la transferencia horizontal de genes (HGT) **(Koonin, Makarova & Wolf, 2017; van Houte, Buckling & Westra., 2016)**. Diversos estudios basados en análisis filogenéticos y genómica comparada, han confirmado la alta tasa de HGT de los sistemas de defensa **(Koonin, Makarova & Wolf, 2017; Oliveira *et al.,* 2014; Makarova., *et al.,* 2015; Makarova et al., 2013)**. Así por ejemplo, solo el ~4% de los sistemas RM se encuentran en los genomas centrales de especies procariotas, lo que sugiere eventos de transferencia recientes **(Oliveira *et al.,* 2014)**.

Dada la ventaja selectiva que confieren a las bacterias y arqueas los sistemas inmunes en su lucha contra los virus, se podría esperar que una vez adquiridos (ya sea por evolución o a través de la HGT), se acumulasen en los genomas procariotas y fuesen seleccionados **(Bernheim & Sorek., 2020)**. Sin embargo, se ha observado que los sistemas de defensa se pierden con frecuencia de los genomas microbianos en escalas de tiempo evolutivas cortas, lo que sugiere que pueden traer consigo desventajas selectivas en ausencia de presión de infección **(Koonin, Makarova & Wolf, 2017; van Houte, Buckling & Westra., 2016)**. Por ejemplo, una gran desventaja de los sistemas de defensa es la autouimunidad. En este sentido, los sistemas CRISPR-Cas pueden generar errores en el proceso de adquisición de espaciadores y adquirir espaciadores del cromosoma en lugar del elemento invasor **(Stern et al., 2010; Heussler, & O'Toole., 2016)**. Esto provoca que la maquinaria de interferencia CRISPR-Cas se dirija a atacar el cromosoma, lo que resulta en la muerte celular **(Stern et al., 2010; Heussler, & O'Toole., 2016)**. De manera similar, los sistemas de RM también alguna vez pueden apuntar al propio cromosoma bacteriano, cortando el ADN infligiendo un costo de aptitud **(Pleška *et al.,* 2015)**. Los sistemas Abi también, pueden generar una actividad no deseada provocando latencia o muerte celular **(Berngruber *et al.,* 2013)**. Además, a parte de la autouimunidad otro aspecto importante es que los sistemas de defensa pueden suponer una carga de energía a la celular. Así por

ejemplo, algunos sistemas RM requieren la hidrólisis de una molécula de ATP por par de bases para la translocación de la enzima de restricción a lo largo del ADN **(Seidel *et al.,* 2008)**.

Dados los costes de aptitud que suponen los sistemas antivirales, es probable que ninguna cepa bacteriana o arquea pueda codificar, a largo plazo, todos los sistemas de defensa posibles sin sufrir serias desventajas competitivas. Por otro lado, el acceso a un conjunto diverso de mecanismos de defensa es fundamental para combatir la enorme diversidad genética y funcional de los virus **(Bernheim & Sorek., 2020).**

Objetivos

Dentro del género *Helicobacter* se encuadra la especie *H. pylori*, una bacteria patógena humana y con distribución global. Este organismo presenta un gran interés desde el punto de vista evolutivo ya que desarrolla estrategias adaptativas particulares a fin de adaptarse al ambiente extremo y específico en el que habita y además tiene una gran importancia desde el punto de vista epidemiologia y médico. Por lo tanto un estudio genómico y poblacional de esta especie es de interés tanto desde del punto vista básico como aplicado. Para llevar a cabo este estudio nos hemos planteado los siguientes objetivos

Objetivo general

Analizar los mecanismos moleculares que participan en la evolución de *H. pylori* mediante estudios de genómica comparada y estructuras poblacionales.

Objetivos específicos

1. Analizar el genoma completo de *H. pylori* a fin de determinar su composición en porcentaje de nucleótidos y número de genes y establecer el pan-genoma de esta especie.
2. Determinar el genoma central y genoma accesorio mediante comparación de diferentes cepas a fin de establecer los genes básicos para su adaptación a su modo de vida.
3. Estudiar la estructura poblacional a nivel global mediante análisis de SNPS de cepas de diferentes localizaciones `para identificar patrones de distribución geográfica.
4. Caracterizar los mecanismos que originan la variabilidad genética en esta bacteria a nivel cromosómico y nucleotídico y determinar su importancia en la organización genómica
5. Comparar el genoma de *H. pylori* con los de otras especies del género *Helicobacter* y determinar sus relaciones filogenéticas.
6. Establecer los mecanismos de inmunidad comunes del género *Helicobacter* mediante el establecimiento del “paninmune”

Capítulo 2 Pangenoma, estructura poblacional y evolución de *H. pylori*

2.1 Introducción

Para que la colonización y establecimiento de una bacteria en un huésped sea exitoso, es necesario que las bacterias promuevan una amplia variación en el contenido de sus elementos génicos, no solo a través de su "phyla", sino también entre los miembros de su misma especie, cepas, y hasta en las réplicas **(Lan & Reeves, 2000; Ochman *et al.,* 2000)**. En este sentido, las bacterias deben contener una gran variedad de genes que garanticen la activación de los mecanismos necesarios para la co-regulación de genes tanto imprescindibles como prescindibles para la adaptación al ambiente **(Martínez & Collado, 2003)**. Es así que, solo hasta hace poco comenzamos a comprender y descifrar el vínculo que existe entre la gran variedad de genotipos y los procesos tanto ecológicos como evolutivos que se ejercen sobre las poblaciones bacterianas **(Cordero & Polz, 2014)**.

En los últimos años, diversas investigaciones a través de diferentes estrategias han explorado las características que se encuentran conservadas y el contenido genómico compartido entre cepas y réplicas de *H. pylori* **(Tabla 1)**. Sin embargo, estos estudios se centraban en regiones geográficas específicas solo de *H. pylori* o los datos utilizados incluían genomas que no estaban completamente ensamblados. Sin duda, estos estudios han ayudado a fortalecer la visión en relación a la diversidad genética y estructura poblacional tanto de *H. pylori* como su vínculo con otras especies de *Helicobacter*.

Los mecanismos de infección y patogénesis de *H. pylori* son complejos, interviniendo interacciones entre cepas, el huésped y el nicho **(Kao, 2015)**. Se ha demostrado que una serie de genes son fundamentales durante la interacción con el huésped (Motilidad, amortiguación del pH, adhesión y quimiotaxis, factores citotóxicos y evasión del sistema inmunológico) para manipular y evadir las defensas del huésped y poder asegurar de esta manera su supervivencia en el ambiente gástrico **(Kumar *et al.,* 2016).** Así, para contrarrestar el pH del estómago, *H. pylori* utiliza la enzima ureasa **(Mobley, 1996);** posee un aparato flagelar y estructura helicoidal que le permite escapar a través del moco viscoso y llegar hasta el epitelio gástrico **(O'Toole, Lane & Porwollik, 2000)**; contiene una serie de proteínas de membrana externa (OMP) denominadas adhesinas para ejercer una fuerte adherencia a las células epiteliales **(Oleastro & Menard, 2013)**; y codifica diversos factores de virulencia como *cagA*, *vacA* e *iceA* que cumplen funciones importantes en la patogénesis **(Hatakeyama, 2004)**.

La comparación de genomas de diferentes especies y dentro de cepas de la misma especie ha permitido la caracterización de los mecanismos implicados en el cambio en el orden y estructura de los componentes genéticos a lo largo del genoma y, por consiguiente, que promueven la variabilidad **(Mira *et al.,* 2002; Abby & Daubin, 2007).** En este sentido, la identificación y caracterización del genoma central es una estrategia interesante para comprender los aspectos básicos para la vida celular de *H. pylori* y de otras especies del género *Helicobacter*. Es por ello, que se requieren estudios detallados de filogenómica y genómica comparada para comprender la asociación entre las diferentes cepas, réplicas y síntomas de la enfermedad. Además, es importante estudiar especies genéticamente distintas, pero ecológicamente similares, por el riesgo zoonótico que representan **(Dewhirst *et al.,* 2005; Haesebrouck *et al.,* 2009; Joosten *et al.,* 20015)**. Es así que, revelar cómo se constituye y funciona el genoma bacteriano y proporcionar una descripción más detallada de la diversidad de especies, cepas y réplicas es fundamental. Los análisis comparativos del patógeno humano *H. pylori*, junto con otras bacterias del género *Helicobacter*, ofrecen una oportunidad interesante para la identificación de regiones genómicas que podrían tener aplicaciones de diagnóstico y terapéuticas.

2.2 Material y métodos

2.2.1 Material

El material utilizado para elaboración de esta Tesis Doctoral está constituido por datos del Centro Nacional de Información Biotecnológica de E.U.A., NCBI **(National Center of Biotechnology Information).** Para alcanzar los objetivos de la presente estudio, se utilizaron genomas completos de 53 cepas de *H. pylori* cepas de diferentes *Helicobacter pylori* **(Tabla 1).** Los datos utilizados en este trabajo comenzaron a obtenerse en febrero del año 2017 y han ido actualizándose lo largo de periodo de investigación cuando era necesario**.**

NCBI	Cepa	Huesped/Asilamiento	Patología	Ng	Tg
NC_000921	J99	Human/Africa/USA	Duodenal ulcer	1541	1643831
NC_017357	908	Human/Africa/France	Duodenal ulcer	1479	1549666
NC_017361	S. Africa7	Human/South Africa	unknown	1580	1653913
NC_017371	Gambia94/24	Human/Africa/Gambia	unknown	1613	1709911
NC_017374	2017	Human/Africa/France	Duodenal ulcer	1484	1548238
NC_017381	2018	Human/Africa/France	Duodenal ulcer	1494	1562832
NC_022130	S. Africa20	Human/South Africa	unknown	1568	1622903
NC_000915	26695	Human/Europe/UK	Gastritis	1583	1667867
NC_008086	HPAG1	Human/Europe/Sweden	Atrophic gastritis	1571	1596366
NC_011333	G27	Human/Europe/Italy	unknown	1613	1652982
NC_011498	P12	Human/Europe/German	Duodenal ulcer	1587	1673813
NC_012973	B38	Human/Europe/France	MALT lymphoma	1523	1576758
NC_014256	B8	Human/unknown	Gastric ulcer	1583	1673997
NC_017362	Lithuania75	Human/Europe/Lithuania	unknown	1547	1624644
NC_018937	Rif1	Human/Europe/German	unknown	1640	1667883
NC_018938	Rif2	Human/Europe/German	unknown	1639	1667890
NC_018939	26695	Human/unknown	unknown	1590	1667892
NC_022886	BM012A	Human/Oceania/Australia	Asymptomatic-reinfection	1586	1660425
NC_022911	BM012S	Human/Oceania/Australia	Asymptomatic-reinfection	1585	1660469
NC_017354	52	Human/Asia/Korea	unknown	1511	1568826
NC_017360	35A	Human/Asia/Japan	unknown	1511	1566655
NC_017365	F30	Human/Asia/Japan	Duodenal ulcer	1509	1570564
NC_017366	F32	Human/Asia/Japan	Gastric cancer	1522	1578824
NC_017367	F57	Human/Asia/Japan	Duodenal ulcer	1537	1609006
NC_017368	F16	Human/Asia/Japan	Gastritis	1513	1575399
NC_017372	India7	Human/Asia/India	Peptic ulcer	1586	1675918
NC_017375	83	Human/unknown	unknown	1542	1617426
NC_017376	SNT49	Human/Asia/India	Asymptomatic	1525	1607577
NC_017382	51	Human/Asia/Korea	Duodenal ulcer	1525	1589954
NC_017926	XZ274	Human/Asia/China	Gastric cancer	1621	1634138
NC_020508	OK113	Human/Asia/Japan	unknown	1542	1616617
NC_020509	OK310	Human/Asia/Japan	unknown	1527	1591278
NC_021215	UM032	Human/Asia/Malasya	peptic ulcer	1518	1593537
NC_021216	UM299	Human/Asia/Malasya	unknown	1518	1594569
NC_021218	UM066	Human/Asia/Malasya	ulcer peptic disease	1565	1658047
NC_021882	UM298	Human/Asia/Malasya	unknown	1520	1594544
NC_010698	Shi470	Human/America/Peru	Gastritis	1535	1608548
NC_017355	v225d	Human/America/Venezuela	Gastritis	1558	1588278
NC_017358	Cuz20	Human/America/Peru	unknown	1565	1635449
NC_017359	Sat464	Human/America/Peru	unknown	1502	1560342
NC_017378	Puno120	Human/America/Peru	Gastritis	1530	1624979
NC_017379	Puno135	Human/America/Peru	Gastritis	1564	1646139
NC_017739	Shi417	Human/America/Peru	unknown	1566	1665719
NC_017740	Shi169	Human/America/Peru	unknown	1544	1616909
NC_017741	Shi112	Human/America/Peru	unknown	1586	1663456
NC_019560	Aklavik117	Human/America/Canada	Gastritis	1519	1614447
NC_019563	Aklavik86	Human/America/Canada	Gastritis	1430	1494183
NC_014555	PeCan4	Human/America/Peru	gastric cancer	1550	1629557
NC_014560	SJM180	Human/America/Peru	Gastritis	1559	1658051
NC_017063	ELS37	Human/America/El Salvador	Gastric cancer	1576	1664587
NC_017733	HUP-B14	Human/Europe/Spain	unknown	1515	1599280
NC_017742	PeCan18	Human/Africa/Peru	gastric cancer	1567	1660685
NC_021217	UM037	Human/Asia/Malasya	unknown	1618	1692794
Media				1551	1621671
Desviación estandar				42	43785
maximo				1640	1709911
Minimo				1430	1494183

Tabla 1. Resumen de los genomas descargados y utilizados en este estudio. Numero de acceso NCBI, cepa, huesped y origen del aislmeinto, patología asociada, número de genes (Hg) y tamaño del genoma (Tg). Media de genes calcualda, desviación estándar, minimo y máximo tanto de genes como del tamaño de los genomas analizados.

2.3 Métodos

2.3.1 Determinación del pan-genoma, genoma central y genoma variable

Para el establecimiento de los diferentes grupos de genes se utilizó el programa **GET HOMOLOGUES (Contreras & Vinuesa, 2013)**. Este programa aplica tres estrategias para la agrupación, BDBH (mejor acierto bidereccional), COG **(Kristensen *et al.,* 2010)** (de secuencias ortólogas) y OMCL agrupación (agrupación de Ortho Markov). OMCL **(Li *et al.,* 2003)** Como su nombre indica, utiliza el algoritmo de agrupación de markov para producir ortólogos y parálogos. En nuestro caso, las búsquedas de similitud y agrupamiento de las secuencias codificantes para los 53 genomas de *H. pylori*, así como de los 7 genomas restantes del género *Helicobacter*, se realizó utilizando OMCL, ya que es más sensible para encontrar ortólogos. Se impuso una cobertura mínima del 75% de la alineación por pares y un valor de corte de 1e-05, lo que dio como resultado el total de genes ortólogos detectados.

1) el script auxiliar auxiliar **compare clusters.pl** proporcionado por **GET HOMOLOGUES** se utilizó para formar la matriz pan-genómica en formato de texto a partir de los archivos de OMCL compuesto por genes ortologos.

2) Utilizando el script **parse_pangenome_matrix.pl** clasifica los genes contenidos en la matriz en: genoma central (core genome) (genes presentes en todos los genomas y soft-core genome su presencia varia del 95 al 99%, que constituyen el genoma central. Por otra parte, el sehll genoma su presencia varia del 15 al 94% y el genoma cloud su presencia es inferior o igual al 15%, que constituyen el genoma variable.

2.3.2 Clasificación funcional mediante análisis de Ontología génica del genoma central (términos GO)

Para determinar los términos GO de las secuencias que componen el genoma central en los dos conjuntos de datos (53 genomas de *H. pylori)* fue implementado el programa **Sma3s (Muñoz *et***

***al.,* 2014**). Este programa, escrito en lenguaje de programación Perl, está compuesto por tres módulos. Cada uno de estos realiza una búsqueda exhaustiva de **BLAST (Altschul *et al.,* 1997)** como punto de partida inicial y su diferencia en relación a otros programas radica en que su tercer módulo "**Sma3s**" que mejora la calidad de la anotación mediante el enriquecimiento de términos para identificar las anotaciones compartidas por grupos de secuencias similares. Para realizar las correspondientes anotaciones fue necesario descargar los dos siguientes archivos: "uniref90.annot.gz" y "uniref90.fasta.gz", desde el siguiente enlace: **http://www.bioinfocabd.upo.es/sma3s/db/.**

2.3.3 Realineamiento, concatenación y filtrado de los genes que conforman el genoma central.

Todos los genes correspondientes al genoma central obtenidos desde **GET_HOMOLOGUES** fueron realineados utilizando el programa para alineamientos múltiples **MAFFT (Katoh, 2013).** A continuación, cada uno de los alineamientos fue inspeccionado manualmente en búsqueda de secuencias duplicadas para elimarlas y así, proceder a la concatenación mediante el script: "**randomConcatenation**" del programa **core-genome-aligment** obtenido desde **https://github.com/tatumdmortimer/core-genome-alignment.** Una vez concatenado el siguiente paso, fue eliminar los sitios homoplásicos utilizando el programa **Noisy (Dress *et al.,* 2008),** ya que estos sitios no informativos pueden afectar el rendimiento de los algoritmos para la reconstrucción filogenética. A continuación, otro filtro fue aplicado a través del programa **Gblocks (Talavera & Castresana, 2007),** con el fin de eliminar las posiciones que pudieran estar mal alineadas y más divergentes del nuevo alineamiento del genoma central. Una vez aplicados estos filtros se procedió a la detección de polimorfismos.

2.3.4 Detección de Polimorfismos (SNP)

La búsqueda y detección de SNP se llevó a cabo con el programa SNP-Sites **https://github.com/sanger-pathogens/snp-sites (Page *et al.,* 2016).** Este programa genera archivos de salida con extensión fasta y VCF que fueron utilizados para posteriores análisis.

A continuación, el archivo con extensión VCF, que contiene los SNPs que forman parte del genoma central fueron analizados por el programa **VCFtools https://github.com/vcftools/vcftools (Danecek., *et al.,* 2011)**, con el fin de generar los ficheros

que el utiliza el programa **Plink v.1.9 (Purcell *et al.,* 2007)** para así recodificar estos archivos a los formatos correspondientes con extensión PED y MAP.

2.3.5 Análisis de Estructura Poblacional

La matrix de SNPs obtenidos tras los tratamientos con los programas **VCFtools** y **Plink**, se utilizaron para realizar las estimación de las poblaciones con los programas **ADMIXTURE (Alexander *et al.,* 2009)** y **FineSTRUCTURE (Yahara *et al.,* 2013),** a partir de los archivos con extensión PED y MAP, generados a través del programa PLINK.

El primer programa (**ADMIXTURE**) realiza una estimación de máxima verosimilitud de los ancestros dentro de un conjunto de datos (en nuestro caso SNPs) **(Pritchard *et al*, 2000; Falush *et al.,* 2003; Falush *et al.,* 2007**) mediante la utilización de un algoritmo de optimización numérico. Para calcular el número supuesto de diferentes componentes estructurales, ejecutamos, para ambos conjuntos de datos, una validación cruzada para K (poblaciones)= 4, 5, 6, 7, 8 con un "bootstrapping (B)" de 100.000. La interpretación será de la siguiente manera: si el valor K es alto en comparación con los otros valores de K, quiere decir que existe un error de validación cruzado, por lo tanto, la mejor K será representada por los valores más pequeños de esta. Para la visualización de los resultados de este análisis fue utilizado la plataforma **CLUMPAK (Kopelman *et al.,* 2015),** que fue desarrollada para su uso en programas tales como STRUCTURE y **ADMIXTURE** para obtener una representación gráfica de los resultados a partir de la matriz "Q" generada, en este caso por ADMIXTURE.

Para subsanar las limitaciones y mejorar la precisión y confiabilidad en relación a la identificación de subpoblaciones que nos proporciona **ADMIXTURE**, se han desarrollado programas tales como **FineSTRUCTURE** que revela estructuras poblacionales más sutiles **(Lawson *et al.,* 2012)**. Este programa infiere "fragmentos" derivados de la recombinación y reconstruye los haplotipos en el cromosoma receptor como una serie de fragmentos denominados "donantes" **(Yahara *et al.,* 2013).** Los resultados de FineSTRUCTURE son mostrados en una matriz de co-ascendencia, que contiene el número de eventos de recombinación de cada donante a cada individuo receptor **(Lawson *et al.,* 2012).**

Para poder implementar adecuadamente **FineSTRUCTURE**, los archivos con extensión PED y MAP obtenidos desde **Plink**, fueron previamente convertidos a los formatos que corresponden, con los scripts que proporciona la página del programa

https://people.maths.bris.ac.uk/~madjl/finestructure/toolsummary.html, (**plink2chromopainter.pl y makeuniformrecfile.pl**).

FineSTRUCTURE es un algoritmo hecho para la identificación de estructuras poblacionales, este utiliza la salida de programa ChromoPainter, que viene integrado en **FineSTRUCTURE**. **ChromoPainter** trabaja bajo un modelo de estadística de inferencia Bayesiana **(Lawson *et al.*, 2012)**, el cual se encarga de buscar haplotipos en las secuencias.

Los parámetros implementados para **FineSTRUCTURE** fueron los siguientes: para los SNPs se implementó el modelo ligado (Linkage Mode), ploidía= haploide, un número total de iteraciones para MCMC (Cadenas de Markov Monte Carlo) de 100.000, con un burn-in de 50.000. El número de maximización para la inferencia del árbol fue de 20.000. La visualización de la matriz de co-ancestría vinculada, como el mapa de color junto con el árbol fue inferido usando tanto la versión "GUI" de **FineSTRUCTURE**.

2.3.6 Análisis filogenéticos

Para reconstruir las filogenias fueron usados los datos del genoma central; específicamente se usaron los alineamientos de los genes ortólogos concatenados y filtrados como haplotipos en función de los SNPs. El programa utilizado fue **RAxML** (Randomized AxeleratedMaximum Likelihood), este programa analiza grandes conjuntos de datos bajo el concepto de máxima probabilidad y su principal fortaleza se basa en la rapidez de su algoritmo para la búsqueda de los árboles, devolviendo arboles con los mejores puntajes de probabilidad. Los resultados de **RAxML** fueron visualizados con el programa **SplitTree** aplicando los siguientes parámetros: 500 pseudorreplicados de arranque e implementando el modelo evolutivo GTR + GAMMA como modelo de sustitución (-s entrada -n salida -m GAMMAGTR - # 500 -p 123 -x 123 -fa).

2.4 Resultados

2.4.1 Características del genoma de *H. pylori*

2.4.1.1 Tamaño de genoma y contenido génico

Tras la comparación de los 53 genomas de *H. pylori*, se ha observado que los genomas tuvieron tamaños comprendidos entre 1.709.911 pb de máximo (Gambia94/24 cepa africana) y 1.494.183 pb como mínimo (Aklavik86 cepa amerindia), con una media de 1.621.671 ±43.785 **(Tabla 1).**

El análisis del conjunto de cepas mostró un contenido medio de GC del 40% ±0,002, con un mínimo del 39% y un máximo del 40%. El promedio de genes por genoma para *H. pylori* fue de 1.551 ± 42 y varió desde 1.430 (Aklavik86 cepa amerindia) a 1.640 (Rif1 cepa europea) **(Tabla 1).** Presentando todos ellos un sesgo de GC positivo.

El promedio de genes codificados en la cadena principal fue de 753 ±31 y en la cadena rezaga de 798 ± 37**).** El sesgo en GC en los genes de la cadena principal y retrasada es positivo **(Tabla 2).** Siendo el porcentaje de genes de la cadena rezagazada con sesgo positivo y de la cadena principal con sesgo positivo siempre mayor**.** El análisis de sesgo de GC para cada una de las cadenas mostró que en la cadena principal, de los 753 genes, hay un promedio de 343 con sesgo negativo (45%) y 410 con sesgo positivo (55%). Mientras que los 798 genes de la cadena rezagada mostraron un promedio de 281 con sesgo negativo (35%) y 517 con sesgo positivo (65%).

		GENES CD y HO		Número de genes con sesgo positivo y negativo en las cadenas HO y CD				Proporcion de genes con sesgo positivo y negativo en las cadenas HO y CD			
Cepa	**GC**	**HO**	**CD**	**HO -**	**CD -**	**HO +**	**CD +**	**%HO -**	**%HO+**	**%CD -**	**%CD +**
J99	0,40	792	749	265	344	527	405	33	67	46	54
908	0,40	762	717	265	319	497	398	35	65	44	56
S. Africa7	0,39	792	788	284	354	508	434	36	64	45	55
Gambia94/24	0,40	851	762	298	354	553	408	35	65	46	54
2017	0,40	763	721	261	323	502	398	34	66	45	55
2018	0,40	773	721	264	323	509	398	34	66	45	55
S. Africa20	0,39	800	768	277	329	523	439	35	65	43	57
26695	0,40	829	754	307	305	522	449	37	63	40	60
HPAG1	0,40	806	765	302	332	504	433	37	63	43	57
G27	0,39	850	763	321	307	529	456	38	62	40	60
P12	0,40	833	754	321	294	512	460	39	61	39	61
B38	0,40	825	698	284	288	541	410	34	66	41	59
B8	0,40	741	842	274	311	467	531	37	63	37	63
Lithuania75	0,40	811	736	286	302	525	434	35	65	41	59
Rif1	0,39	860	780	320	316	540	464	37	63	41	59
Rif2	0,39	860	779	321	315	539	464	37	63	40	60
26695	0,40	828	762	309	307	519	455	37	63	40	60
BM012A	0,40	750	836	195	685	555	151	26	74	82	18
BM012S	0,40	749	836	195	685	554	151	26	74	82	18
52	0,40	764	747	272	308	492	439	36	64	41	59
35A	0,40	765	746	313	275	452	471	41	59	37	63
F30	0,40	749	760	253	324	496	436	34	66	43	57
F32	0,40	787	735	264	331	523	404	34	66	45	55
F57	0,40	766	771	285	302	481	469	37	63	39	61
F16	0,40	738	775	278	302	460	473	38	62	39	61
India7	0,40	822	764	285	359	537	405	35	65	47	53
83	0,39	755	787	287	308	468	479	38	62	39	61
SNT49	0,40	801	724	304	273	497	451	38	62	38	62
51	0,40	787	738	286	307	501	431	36	64	42	58
XZ274	0,39	827	794	274	337	553	457	33	67	42	58
OK113	0,40	783	759	289	310	494	449	37	63	41	59
OK310	0,40	753	774	289	307	464	467	38	62	40	60
UM032	0,40	762	756	217	610	545	146	28	72	81	19
UM299	0,40	763	755	217	609	546	146	28	72	81	19
UM066	0,39	786	779	341	289	445	490	43	57	37	63
UM298	0,40	763	757	217	610	546	147	28	72	81	19
Shi470	0,40	814	721	280	306	534	415	34	66	42	58
v225d	0,40	807	751	299	320	508	431	37	63	43	57
Cuz20	0,40	822	743	316	303	506	440	38	62	41	59
Sat464	0,40	780	722	274	303	506	419	35	65	42	58
Puno120	0,40	829	701	278	319	551	382	34	66	46	54
Puno135	0,40	829	735	296	292	533	443	36	64	40	60
Shi417	0,40	819	747	301	302	518	445	37	63	40	60
Shi169	0,40	802	742	310	283	492	459	39	61	38	62
Shi112	0,40	836	750	293	312	543	438	35	65	42	58
Aklavik117	0,40	817	702	286	271	531	431	35	65	39	61
Aklavik86	0,40	692	738	280	221	412	517	40	60	30	70
PeCan4	0,40	830	720	279	330	551	390	34	66	46	54
SJM180	0,40	814	745	264	302	550	443	32	68	41	59
ELS37	0,40	837	739	286	321	551	418	34	66	43	57
HUP-B14	0,40	822	693	280	286	542	407	34	66	41	59
PeCan18	0,40	822	745	281	338	541	407	34	66	45	55
UM037	0,40	865	753	253	407	612	346	29	71	54	46
Medias	0,39	798	753	281	343	517	410	35	65	45	55

Tabla 2. Resumen de de los principales datos obtenidos a partir del análisis de cada de cepa de *H. pylori*. Cepa, contenido de guanina-citosina (GC), genes de la cadena rezagada (HO), genes de la cadena principal (CD), número de genes de la cadena HO con sesgo negativo y positivo, número de genes de la CD con sesgo negativo y positivo y proporción de genes con sesgo posotivo y negativo para las cadenas HO y CD.

Implementamos un análisis de varianza de un factor para contrastar la hipótesis nula de que las medias de poblaciones en cuanto a número de genes y tamaño del genoma son iguales, frente a la hipótesis alternativa de que por lo menos una de las poblaciones difiere de las demás. Para ello, primero analizamos el tamaño del genoma a partir de siete grupos preestablecidos basados en origen geográfico en su conjunto. Los grupos establecidos fueron los siguientes: hpAfrica1, hpAfrica2, hpEAsia que fue subdivido en hpEAsia (solo asiáticas) y hspAmerind (solo amerindias), hpAsia2, hpEurope e híbridas.

El análisis de varianza de un factor para el conjunto de los grupos mostró que en al menos una de los grupos existe una diferencia significativa con una probabilidad p= 0,01.

Para identificar aquellos grupos que son significativas las diferencias se analizaron los genomas por tamaño pequeño contra genomas de tamaño grande, según las medias obtenidas en la **Tabla 3.**

Grupo	Media de genes	Media por tamaño del genoma
hpAfrica1	1522	1602896
hpEAsia	1532	1597292
hpsAmerind	1536	1610768
hpAsia2	1556	1641748
hpAfrica2	1574	1638408
Hybrids	1574	1661135
hpEurope	1581	1644150

Tabla 3. Medias para los tamaños de genomas y medias por número de genes por grupos.

Las diferencias estadísticamente significativas se obtuvieron para los siguientes grupos comparados: hpEAsia/hpEurope **(p= 0,0002)**, hpEAsia/hybrids **(p= 0,0002)** y hpEAsia/hpAfrica2 **(p= 0,04)**

Al igual que con el número de los genes, un análisis de varianza de un factor fue llevado a cabo para establecer si existen diferencias significativas. Primero realizamos un análisis entre todos los grupos, el cual mostró una probabilidad para **p= 0,008**. Demostrando que en al menos un grupo existe una diferencia significativa con respecto a los demás grupos.

Cuando se realizó un comparación entre grupos para el número de genes, se obtuvieron diferencias estadísticamente significativas entre los siguientes grupos: hpAfrica1/hpEurope **(p= 0,01)**, hpEAsia/hpEurope **(p= 0,0005)**, hspAmerind/hpEurope **(p= 0,01)** y hpEAsia/hybrids **(p= 0,010).**

2.4.1.2 Análisis del pangenoma *H. pylori*

El conjunto completo de genes en el pangenoma de los 53 genomas consistió en un total de 5.134 grupos de genes, que incluye tanto ortólogos como parálogos mediante la implementación de la estrategia OMCL. El genoma core o central, que representa un conjunto de genes cuya secuencia está altamente conservada y que están presentes en todos los genomas (100%) fue de 820 genes. El genoma *soft-core*, que representa los genes presentes desde el 95% al 99% de los genomas, fue de 1.162 genes. El genoma *shell*, que incluye genes presentes desde el 15% al 94% de los genomas, comprende 1.213 genes. Por último, el genoma *cloud*, definido como los grupos de genes cuya presencia es menor o igual al 15% de los genomas estaba compuesto de 2.759 genes. Los genes que conforman tanto el genoma *shell* como el genoma *cloud*, representan el genoma variable, y consistió en 3.972 genes.

2.4.1.3 Evaluación funcional de los grupos de genes del genoma central para *H. pylori*

De los 820 genes que conformaron el genoma central, se filtraron 18 genes que correspondían a genes parálogos. De los 802 restantes se determinó el perfil funcional y biológico mediante el programa Sma3s.

Este programa clasifica las anotaciones en tres categorías biológicas: componentes celulares, procesos metabólicos y procesos biológicos.

En la categoría de componentes celulares estuvieron involucrados 404 genes para 13 componentes celulares. Los más representados de estos fueron los componentes de membrana (188) y citoplasma (146).

En la categoría procesos metabólicos estuvieron involucrados 170 genes para 62 procesos del metabolismo siendo el más representado el formado por los genes vinculados a biosíntesis de amino ácidos (24) y biosíntesis de cofactores (20)

En la categoría procesos biológicos estuvieron involucrados 468 genes para 76 procesos biológicos. Los más representados fueron los procesos vinculados al transporte (60), biosíntesis de proteínas (35), biosíntesis de aminoácidos (30), ciclo celular (26) y metabolismo lipídico (21). Le siguieron en importancia los genes relacionados con la forma de la célula (15), reparación? del daño del ADN (15), transcripción (14), biogénesis/degradación de la pared celular (14), respuesta al estrés (14) y procesamiento de ARNt (11). **Figura 1.**

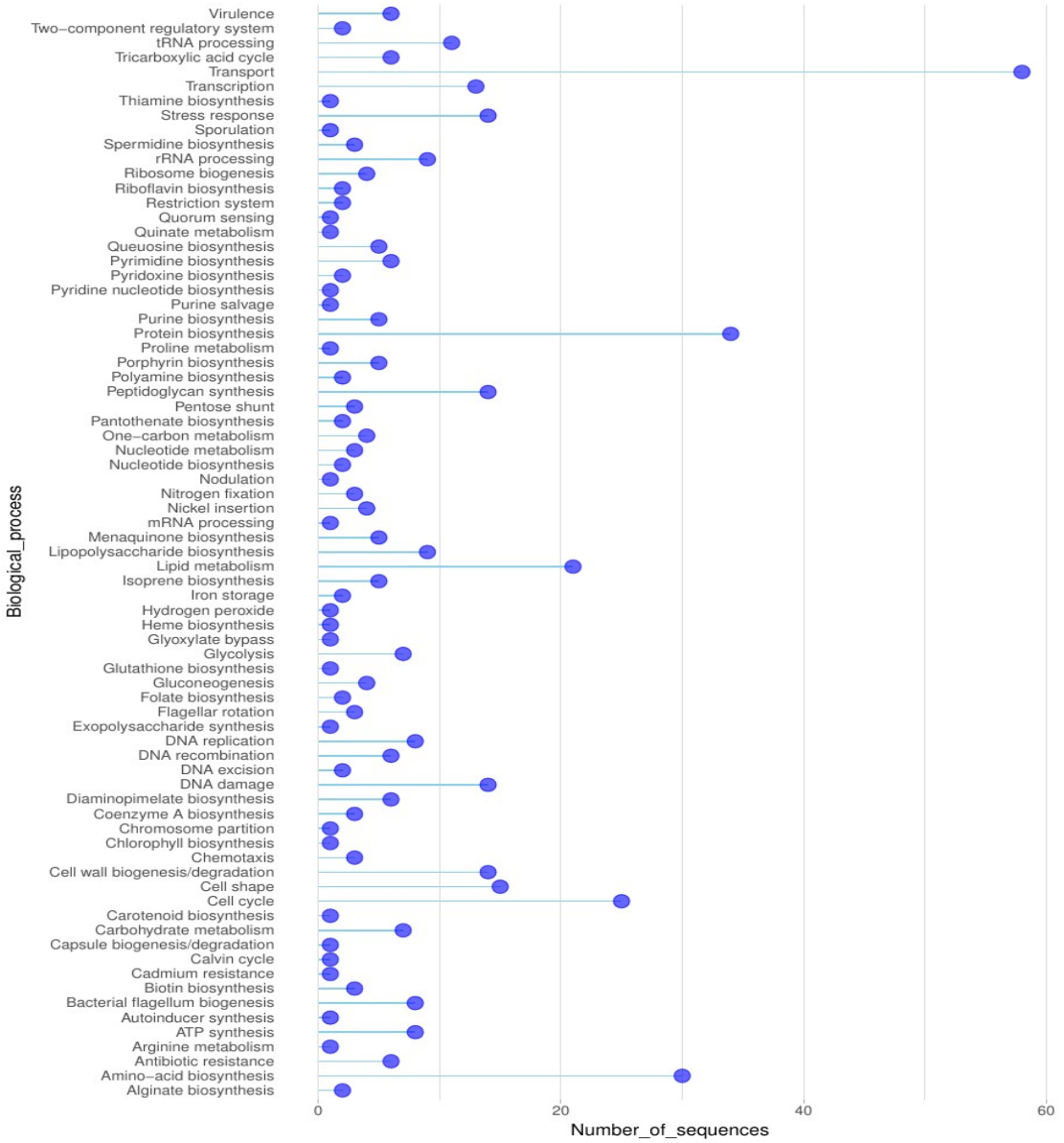


Figure 1. Genes clasificados mediante términos de GO para la categoría de procesos biológicos.

Otros hallazgos interesantes encontrados en el análisis de genoma central fueron genes relacionados con la respuesta al estrés (14 genes), resistencia a los antibióticos (6 genes) y un gen autoinductor (*quorum sensing*).

En cuanto a los genes para la colonización y patogénesis son numerosos los genes vinculados a estos procesos que forman parte del genoma central. Así, los genes de amortiguación del pH (genes de la ureasa que van desde *ureA* hasta *ureG*, que además también están implicados en la motilidad) son parte del genoma core o del soft-core **(Tabla 4**).

Por otro lado, para los procesos biológicos vinculados a la cascada de regulación de la transcripción y biosíntesis flagelar y asociados a la motilidad y quimiotaxis comprobamos que forman parte del genoma central o soft-core tanto los tres factores sigma 80, 54 y 28 (*rpoD*, *rpoN* y *fliA*) como los genes flagelares clase I (*flgR, fglS, flhA*), los genes estructurales del sistema flagelar (*motA*, *motB*), los genes flagelares clase II (*flaB, flgE, flgK, flgM, flgL*) y los genes facilitadores de la secreción de los genes flagelares clase II (*fliQ, fliR*). (**Tabla 4)**.

Además de estos genes, nuestro análisis del genoma central, también reveló otros 34 genes flagelares que sugieren ser esenciales en la cascada reguladora para la transcripción flagelar

Igualmente, en la **Tabla 4**, se puede apreciar que todos factores para la quimiotaxis (*tlpA, tlpB, tplC, tlpD, cheA, cheW* y *cheY*) y los genes de adhesión (*omp11, babA/babB, sabA, oipA*, y *alpA/alpB*) forman parte del genoma central o genoma *soft-core*, además de otras adhesinas (24 genes) detectados como parte del genoma core

	Gen	**Compartimento**	
Amortiguación del pH	*ureA*	core	ajuste del pH, factor de virulencia
	ureB	core	ajuste del pH, factor de virulencia
	ureI	core	ajuste del pH, factor de virulencia
	ureE	core	ajuste del pH, factor de virulencia
	ureF	soft core	ajuste del pH, factor de virulencia
	ureH	core	ajuste del pH, factor de virulencia
	ureG	core	ajuste del pH, factor de virulencia
	arsS	accesory	control de la transcripción
	arsR	core	control de la transcripción
Regulador de los genes clase I	*rpoD*	soft core	factor sigma σ^{80} de mantenimiento
Genes flagelares clase I, que comprenden los principales genes reguladores	*flgR*	soft core	
	flgS	core	
	flhA	core	
Genes estructurales del sistema flagelar	*motA*	core	rotación flagelar
	motB	core	rotación flagelar
Regulador de los genes clase II	*rpoN*	core	factores sigma alternativos σ^{54}
Genes flagelares clase II	*flaB*	core	Filamento flagelar
	flgE	soft core	
	flgK	core	
	flgM	core	
	flgL	soft core	
T3SS	*T3SS/fliP*	accesory	facilita la secreción ordenada de los genes clase 2 y proteinas hook
	T3SS/fliQ	core	facilita la secreción ordenada de los genes clase 2 y proteinas hook
	T3SS/fliR	core	facilita la secreción ordenada de los genes clase 2 y proteinas hook
Regulador de los genes clase III	*fliA*	core	factores sigma alternativo σ^{28}

	flaA	core	Filamento flagelar
Regulador de la expresión de σ 54	*csrA*	core	
adhesión	*omp11*	core	
	bab/babB	core	factor de virulencia
	sabA	core	factor de virulencia
	oipA	core	factor de virulencia
	alpA	core	factor de virulencia
	alpB	core	factor de virulencia
quimiotaxis	*tlpA*	core	
	tlpB	core	
	tplC	soft core	
	tplD	soft core	
	cheA	core	
	cheW	core	
	cheY	soft core	

Tabla 4. Resumen de los genes implicados en la colonización, patogénesis y cascada de la regulación flagelar.

También nuestros resultados ponen de manifiesto que en el genoma central aparecen genes vinculados a la patogenicidad como *vacA* y sus parálogos (*faaA, vlpC* e *imaA,* que para posteriores análisis fueron sacados del genoma central), así como otros, *cag23* (DNA trasnfer), *lpxE* (Lipid A 1-phosphatase), *dps* (DNA protection during starvation) y *murJ* (Putative lipid II flippase)

2.4.1.4 Genes no anotados del genoma core

Por último, un total de 124 (15%) genes fueron anotados por sma3s como hypothetical protein. Estos genes fueron analizados con el paquete HHblits para caracterizar en estas secuencias dominios y poder relacionarlos con una posible función. Por mencionar algunos hallazgos de interés, se encontraron genes con posible relación con: LPS export ABC transporter permease *LptG*, Multidrug export protein *EmrA*, Type IV secretion system protein B4, VirB2 type IV secretion protein, Urease-enhancing factor *Lpp*, Putative flagellar protein *FlgJ*, Putative motility protein chaperone *MotE*, Outer membrane protein *HorL*, Putative beta-lactamase *HcpC* y Chemotaxis protein.

2.4.2 Estructura poblacional *H. pylori*

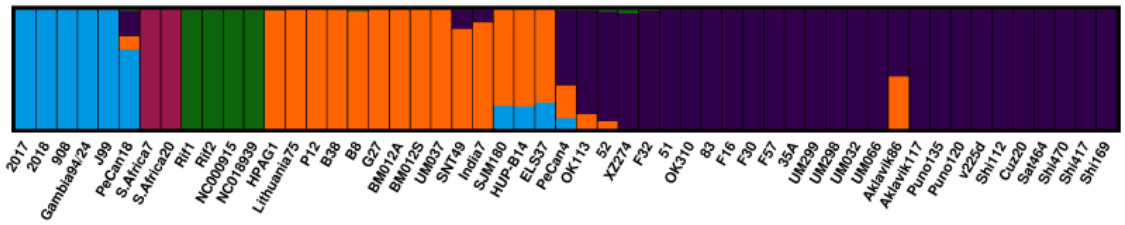
Para el estudio filogenético hemos aplicado tres metodologías diferentes y en todas ellas se utilizaron los 149.123 SNP obtenidos a partir de los 802 genes que conformaron el genoma central **(véase Material y Métodos).**

2.4.2.1 Componentes estructurales

A partir de los SNPs, el algoritmo ADMIXTURE fue implementado para calcular el número de componentes estructurales para una población utilizando el método validación cruzada (CV= Cross Validation) para "*K*" (número de poblaciones probadas. En nuestro caso estimamos los componentes para *K*= 4, 5, 6, 7, 8, con un bootstrapping "*B*" de 100.000 **(ver Materiales y Métodos).**

Los datos más ajustados para explicar la variación observado se obtienen en los niveles K=5 y K=6 **(Figura 2)**. Así, en el valor K=5 la variabilidad presente en las 53 especies se agrupa en cinco grupos principales que describimos de izquierda a derecha: África (azul), Sudáfrica (morado), Europa dividida en dos grupos (verde y naranja) y un gran grupo que comprende las cepas asiáticas y amerindias en color purpura.

K5: CV error (K=5): 0.68948



K6: CV error (K=6): 0.73218

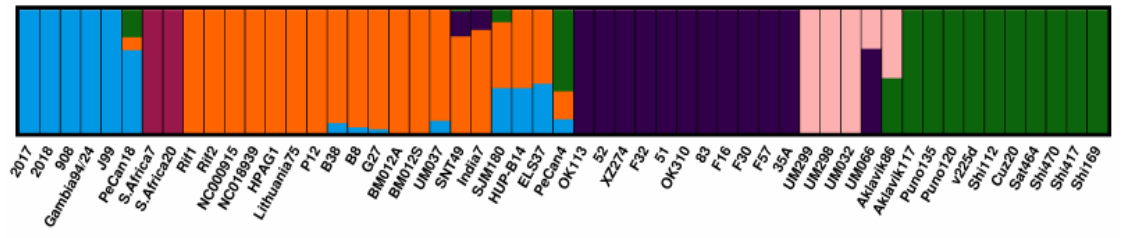


Figura 2. Estructura población para las 53 cepas de *H. pylori* determinada mediante la prueba de validación cruzada para *K* poblaciones mediante el programa.CV error (K=6): 0.73218. (CV=Cross validation).

Los resultados de validación cruzada obtenidos con *K=6* (0.73218) agruparon las cepas analizadas en seis grupos poblacionales distintos en parte coincidentes con el nivel de K=5. Así,

hay grupo un africano, formado por dos cepas sudafricanas (morado) y cinco cepas africanas. Aquí se incluye una cepa híbrida entre África, Europa y amerindia (PeCan18). Un segundo grupo es el constituido por las cepas europeas (14 cepas de color naranja), ademas también, se observan cuatro cepas híbridas (SJM180, ELS37, PeCan4 y UM037). Otro grupo es constituido por parte de las cepas asiáticas (11 cepas en color morado) ya que dentro de este grupo distinguimos 4 cepas malayas (color rosa). Y por último se agrupan las cepas de origen amerindio (11 cepas en color verde). ADMIXTURE identificó las cinco cepas híbridas previamente determinadas (con mezcla o ascendencia desde otras regiones geográficas). Así, la cepa PeCan4 (Perú) mostró tres componentes, amerindia en mayor proporción, seguido de Europa y África. PeCan18 (Perú), mostró mezcla principalmente de África, seguido de amerindia y Europa. SMJ180 (Perú), su genoma estuvo constituido principalmente por parte de Europa, seguido de África y amerindia. ELS37 (El Salvador), reveló mezcal en primer lugar de origen europeo seguido de África. Y por último UM037 (Malasia), mostró mezcla de componentes de Europa seguido de África **(Figura 2)**.

También observamos que existe mezcla entre regiones geográficas, sin que ello derive en la caracterización de cepas híbridas, ya que estás cepas mantienen su posición geográfica de origen. Así, las cepas B38, B8 y G27 de origen europeo mostraron tener un componente ascendente a partir de África, pero este no es muy grande. Adicionalmente, determinamos que las India7 y SNT49 contienen un componente principal europeo seguido en menor medida del componente asiático. Curiosamente la cepa SNT49, exhibió un pequeño componente de ascendencia a partir de amerindia **(Figura 2)**.

2.4.2.2 Matriz de coascendencia y flujo génico de *H. pylori*

Un segundo análisis realizado a partir de los datos de SNPs es el llevado a cabo mediante el programa fineSTRUCTURE, que analiza los datos mediante un enfoque MCMC bayesiano (cadena de Markov Monte Carlo). Con este programa, se obtienen dos tipos de información. Por un lado obtenemos la filogenia entre las cepas que permite una mejor diferenciación entre grupos y dentro de cada grupo geográfico, agrupándolas en subpoblaciones que presentan mayor afinidad En segundo lugar, obtenemos una matriz de co-ascendencia que muestra el flujo génico entre las cepas de las distintas poblaciones distinguiendo entre población donadora y población receptora.

El análisis filogenético permitió una estructuración más fina entre regiones y dentro de regiones geográficas mostrando un conjunto de cinco subpoblaciones (hpAfrica1, hpAfrica2, hpEurope, hpEastAsia y hspAmerind). Dentro de cada grupo geográfico se distinguieron diferentes subpoblaciones:

Las cepas de la región geográfica de África se clasifican en dos grupos, hpAfrica1, hpAfrica2. En el primer grupo se incluyen dos subpoblaciones: hspWafrica_sg1 (2 cepas), hspWafrica_sg2 (3 cepas) y, un singleton híbrido (PeCan18). El segundo grupo hpAfrica2, estuvo constituido por las dos cepas sudafricanas (S. África7 y S. África20).

En la región geográfica de Europa, denominada hpEurope, fueron caracterizadas seis subpoblaciones: hpEurope_sg1 (4 cepas), hpEurope_sg2 (4 cepas), hpEurope_sg3 (dos cepas), hpEurope_sg4 (dos cepas), un singleton denominado hpEuropa_sg5 (HUP-B14). Además, en esta región aparecen dos singletons híbridos (UM037 y PeCan4) y un clado híbrido compuesto por ELS37 y SJM180. En esta región geográfica se incluyen dos cepas cuyo origen es el subcontinente asiático pero que su ascendencia es europea (India7, SNT49).

En la región geográfica de Asia, denominada hpEastAsia, fueron caracterizadas seis subpoblaciones: hspEAsia_sg (4 cepas), hspEAsia_sg1 (3 cepas), hpsEAsia_sg2 (3 cepas), hspEAsia_sg3 pertenecientes a la región malaya (3 cepas), hspEAsia_sg4 catalogada como singleton (UM066) y hspEAsia_sg5) catalogada como singleton.

En la región geográfica amerindia, denominada hpEastAsia, fueron caracterizadas seis subpoblaciones: hspAmerind_sg1 (2 cepas), hspAmerind_sg2 (2 cepas), hspAmerind_sg3 (2 cepas), hspAmerind_sg4 (2 cepas), hspAmerind_sg5 (2 cepas) y hspAmerind_sg6 catalogada como singleton (v225d)

Esta clasificación por regiones geográficas y la asignación de subpoblaciones puede verse resumida en la **Tabla 5**.

Cepa	fineStructure (Linkage Model)	Población
J99, Gambia94/24	hspWAfrica_sg1	hpAfrica1
2017, 2018, 908	hspWAfrica_sg2	hpAfrica1
PeCan18	Singleton (hybrid)	hpAfrica1
SouthAfrica7, SouthAfrica20	hpAfrica2	hpAfrica2
NC_000915, NC_018939, Rif1, Rif2	hpEurope_sg1	hpEurope
G27, B8, B38, P12	hpEurope_sg2	hpEurope
Lithuania75, HPAG1	hpEurope_sg3	hpEurope
BM012A, BM012S	hpEurope_sg4	hpEurope

UM037	Singleton (hybrid)	hpEurope
HUP-B14	Singleton (hpEurope_sg5)	hpEurope
ELS37, SJM180	(hybrid)	hpEurope
PeCan4	Singleton (hybrid)	hpEurope
India7, SNT49	hpAsia2	hpAsia2
F57, F30, 35A,F16	hspEAsia_sg	hpEastAsia
83, OK310, 51	hspEAsia_sg1	hpEastAsia
XZ274, F32, 52	hspEAsia_sg2	hpEastAsia
UM032, UM298, UM299	hspEAsia_sg3	hpEastAsia
UM066	Singleton (hspEAsia_sg4)	hpEastAsia
OK113	Singleton (hspEAsia_sg5)	hpEastAsia
Puno135, Puno120	hspAmerind_sg1	hpEastAsia
Shi470, Sat464	hspAmerind_sg2	hpEastAsia
Cuz20, Shi112	hspAmerind_sg3	hpEastAsia
Shi169, Shi417	hspAmerind_sg4	hpEastAsia
Aklavik86, Aklavik117	hspAmerind_sg5	hpEastAsia
v225d	Singleton (hspAmerind_sg6)	hpEastAsia

Tabla 5. Asignación de grupos geográficos y subpoblaciones mediante fineSTRUCTURE

Como hemos mencionado anteriormente fineSTRUCTURE permite averiguar el flujo génico entre genomas donantes y genomas receptores. En la **figura 3** se muestra este flujo génico. En la parte superior de la matriz de co-ascendencia se encuentran los genomas donantes y en la vertical la parte receptora. Esta matriz la hemos divido en cuatro zonas, cada una correspondiente a una región geográfica, según fuese el origen del genoma donante. (Amerindia/Asia, Europa/África). En esta figura la **zona 1** se considera como donantes a las cepas amerindias, en la **zona 2** las cepas asiáticas, en la **zona 3** las cepas europeas y en la **zona 4**, las cepas africanas.

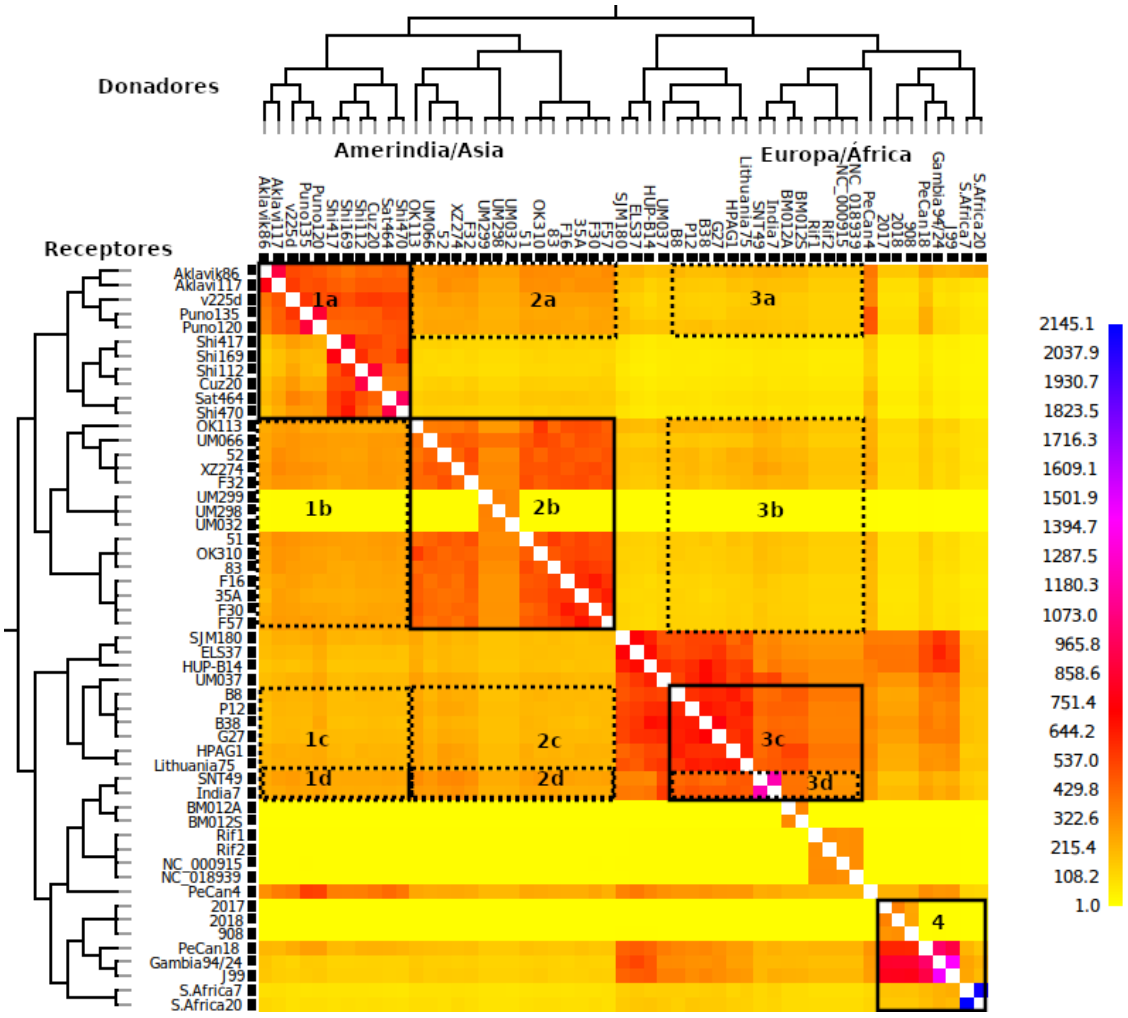


Figura 3. Matriz de co-ascendencia con estructura poblacional y flujo génico. Los rectángulos de líneas continuas indican el flujo genético entre subpoblaciones de la misma región geográfica, mientras que, los rectángulos de guiones indican el flujo génico con otras subpoblaciones de origen geográfico diferente. El color de cada celda de la matriz indica el número esperado de fragmentos importados de un genoma donante (Columna) a un genoma receptor (fila). El nombre de cada cepa se indica a la izquierda. E árbol de la izquierda muestra la agrupación para la asignación de las subpoblaciones que fue asignado en la **Tabla 5,** también reveló dos grandes clados (África/Europa y Asia/Amerindios)**.**

Cada zona esta subdivida en rectángulos continuos que indican los puntos más evidentes de mezcla entre las subpoblaciones de la misma región geográfica y los rectángulos de guiones que indican la mezcla entre subgrupos de diferentes regiones geográficas. La escala de color a la derecha indica el número de segmentos de ADN que son donados e importados entre cepas. Aquí, el término de "mezcla" hace referencia al flujo génico de fragmentos tanto entre grupos como entre subgrupos.

En la **zona 1**, el **rectángulo 1a**, indica un flujo génico importante interno dentro las cepas amerindias. El **rectángulo 1b** muestra el flujo génico moderado de estas con casi todas las

subpoblaciones asiáticas, excepto con la subpoblación hspEAsia_sg3 (cepas de Malasia) donde el flujo fue considerablemente bajo. El **rectángulo 1c**, muestra el flujo hacia las cepas europeas y el **rectángulo 1d,** hacia hpAsia2 (India7 y SNT49), en ambos casos el flujo génico es moderado.

En la **zona 2**, el **rectángulo 2a**, podemos observar un flujo moderado de las cepas asiáticas y las cepas amarindias incluidas en hspAmerind_sg1 (Puno120 y Puno135), hspAmerind_sg5 (Aklavi86 y Aklavi117) y el singleton hspAmerind_sg6 (v225d). El **rectángulo 2b**, muestra el flujo entre las cepas asiáticas observandose que este ha sido considerable entre casi todas las cepas de esta región, excepto con las cepas de la subpoblación hspEAsia_sg3 (Malasia), en cuyo caso, solo mantiene flujo entre ellas. El **rectángulo 2c**, muestra el flujo génico hacia la población de Europa que podemos considerar como bajo al igual que se puede observar en el **rectángulo 2d**, con el flujo hacia la población hpAsia2.

En la **zona 3**, el **rectángulo 3a**, se puede apreciar el flujo desde las cepas europeas hacia las subpoblaciones amerindias. Este flujo puede considerarse como moderado y principalmente con las subpoblaciones hspAmerind_sg1 (Puno120 y Puno135), hspAmerind_sg5 (Aklavi86 y Aklavi117) y con el singleton hspAmerind_sg6 (v225d). El **rectángulo 3b**, muestra el flujo génico hacia las subpoblaciones de Asia. Se puede inferir que flujo ha sido moderado con casi todas las subpoblaciones, excepto, con la subpoblación hspEAsia_sg3 (Malasia) donde el flujo es muy bajo. En el **rectángulo 3c**, se puede apreciar un valor alto del intercambio de material genético entre las cepas europeas siendo importante entre las subpoblaciones hpEurope_sg2 (G27, B8, B38, P12) y hpEurope_sg3 (Lithuania75, HPAG1). El **rectángulo 3d**, mostró que las subpoblaciones hpEurope_sg2 y hpEurope_sg3, también han tenido un flujo de genes importante con la población hpAsia2 (India7, SNT49).

En la **zona 4**, el **rectángulo 4a**, señala un flujo moderado entre la población africana y europea. En el c**uadro 4b** se puede apreciar un intercambio de material genético muy significativo entre las cepas de la subpoblación hspWAfrica_sg1 (J99, Gambia94/24). También se observó flujo intenso a partir de la subpoblación hspWAfrica_sg2 (2017, 2018, 908) hacia hspWAfrica_sg1. Y el flujo génico más alto se evidencio entre las cepas de la población hpAfrica2 (cepas sudafricanas).

En el caso de las cepas híbridas en la **Figura 4**, se puede observar que cinco cepas mostraron una clara mezcla de componentes genómicos desde al menos dos orígenes diferentes.

Los **rectángulos 1a** y **1b**, muestran que las cepas SMJ180 (Perú) y ELSE37 (Salvador) que se ubican en el clado europeo, son una mezcla principalmente de las subpoblaciones europeas, singleton hpEurope_sg5 (HUP-B14), hpEurope_sg2 (G27, B8, B38) y hpEurope_sg3 (Lithuania75, HPAG1), respectivamente, seguido de un componente moderado desde la subpoblación hspWAfrica_sg1 (J99 y Gambia94/24). En el **rectángulo 2**, se aprecia que la cepa UM037 (Malasia), presentó el mismo patrón de mezcla que las cepas SMJ180 y ELS37. Los **rectángulos 3a** y **3b**, revelaron que la cepa PeCan4 (Perú), es una mezcla a partir de las subpoblaciones europeas, singleton hpEurope_sg5 (HUP-B14), hpEurope_sg2 (G27, B8, B38) y hpEurope_sg3 (Lithuania75, HPAG1). Por último, en los **rectángulos 4a**, **4b** y **4c**, se puede apreciar que la cepa PeCan18 (Perú), es una mezcla principalmente de componentes la subpoblación hspWAfrica_sg1 (J99 y Gambia94/24), seguido de las subpoblaciones europeas singleton hpEurope_sg5 (HUP-B14), hpEurope_sg2 (G27, B8, B38) y hpEurope_sg3 (Lithuania75, HPAG1).

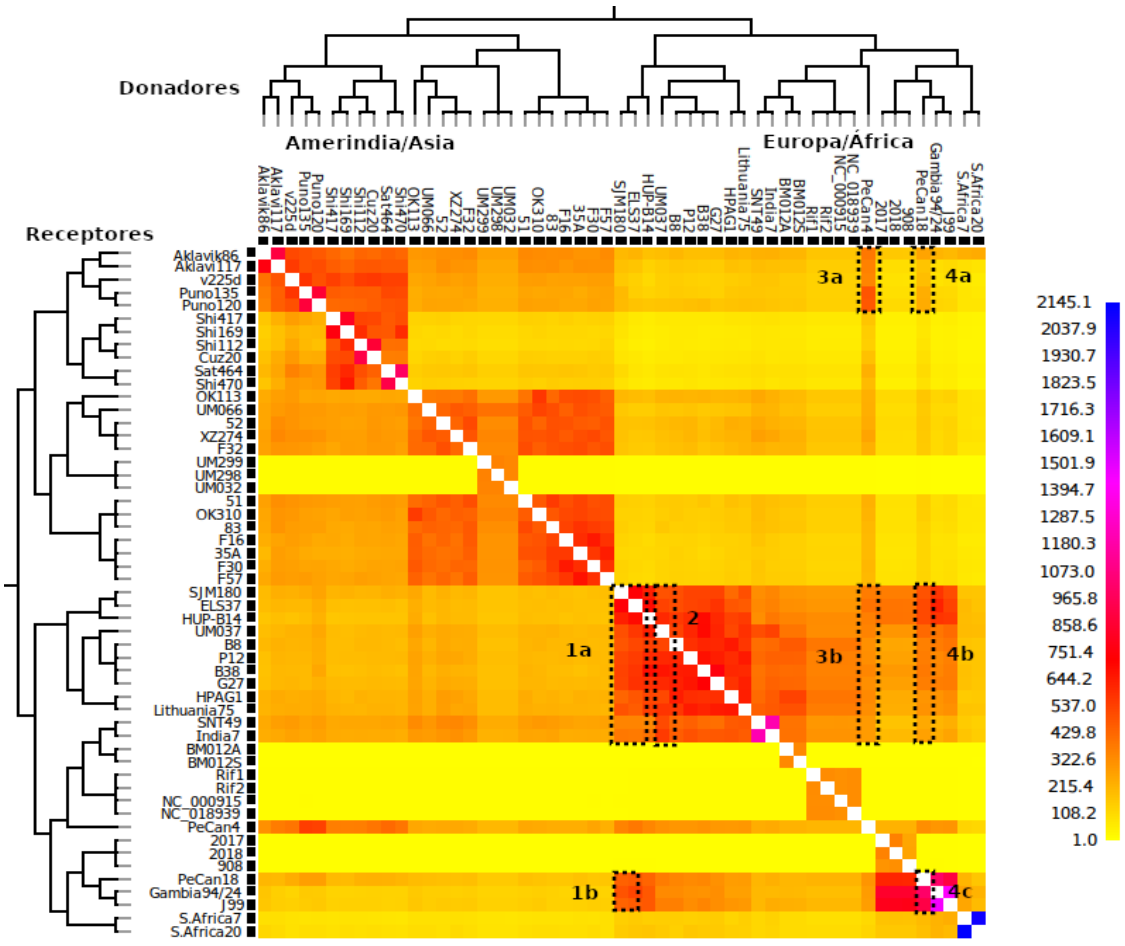


Figura 4. Matriz de co-ascendencia con estructura poblacional y flujo génico. Cepas híbridas y subgrupos mezclados. **1a-b.** la cepa SJM180 con signos de mezcla con las cepas europeas.y africana**. 2**. La cepa UM037

(Malasia) con signos de mezcla con las cepas europeas. **3a-b** PeCan4 singleton (hybrid) hpEuropa con signos de mezcla con las cepas europeas. **4a-c** La cepa PeCan18 singleton (hybrid) subpoblación hpAfrica1 con signos de mezcla con la subpoblación hspWAfrica_sg, europa y amerindia.

2.4.2.3 Análisis filogenético

Por último, el árbol filogenético basado en los SNPs detectados **(Figura 5)** reveló un agrupamiento en dos grandes clados principales formado, uno por las cepas de África/Europa y las de Asia/Amerindia por otro. Dentro de estos dos grandes clados, se pueden distinguir seis agrupaciones claramente asociadas con regiones geográficas **(Figura 5).** En el árbol filogenético también se pueden identificar cinco cepas híbridas, cuya posición no se correspondió con el origen de su aislamiento geográfico. Así, las cepas amerindias, PeCan18 (origen Perú) aparece ubicada en la región geográfica de África, PeCan4 (origen Perú), SMJ180 (origen Perú) y ELS37 (origen El Salvador) aparecen más próximas a la región geográfica de Europa. La cepa UM037 de origen malayo, también aparece ubicada en la región geográfica de Europa**.**

También es de destacar que: **1)** La rama formada por las dos cepas sudafricanas (S. Africa7, S. Africa20) aparece claramente separada como un clado aparte y a mayor distancia. **2)** Los aislamientos de origen australiano (BM012A y BM012S) están claramente ligadas al clado europeo.

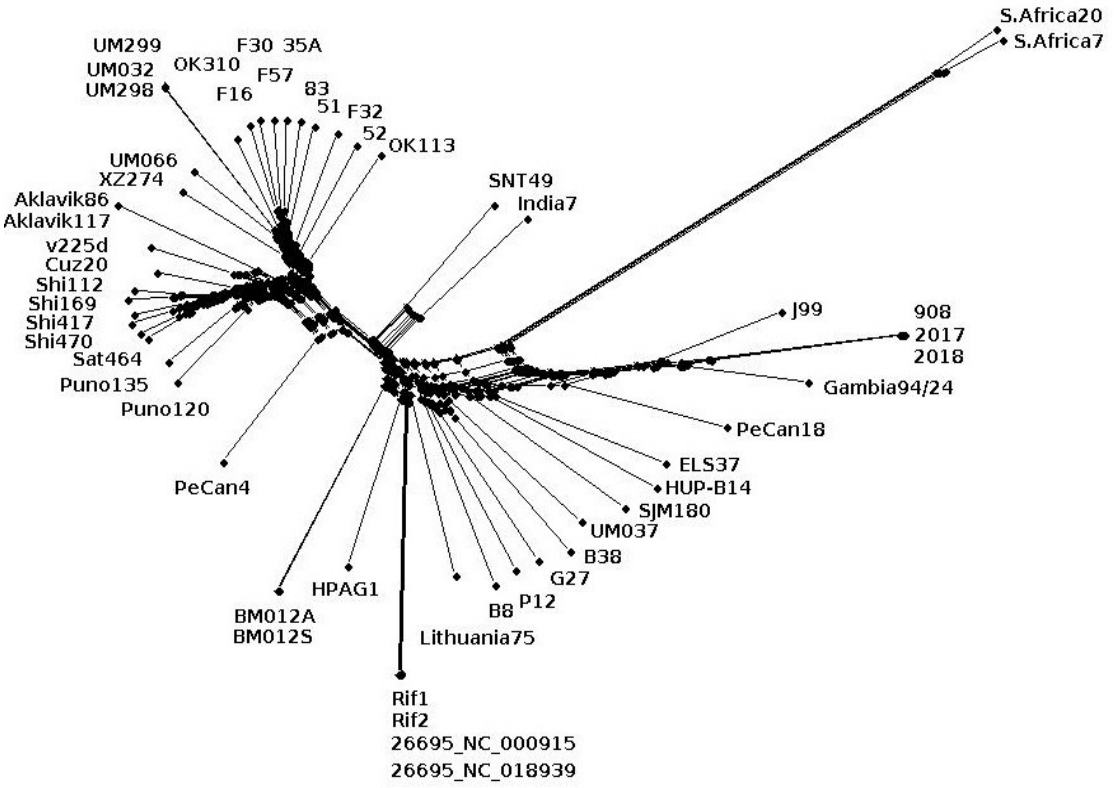


Figura 5. Análisis filogenético *H. pylori.* Árbol filogenético para 53 cepas de *H. pylori* (0.01) basado en 149.123 SNP. La topología del árbol revela dos grandes clados (África/Europa y Asia/Amerindios).

2.4.3 *H. pylori* con siete especies de *Helicobacter* no pylori (NHPH)

2.4.3.1 Aproximación a las características del genoma del género *Helicobacter*

El promedio de genes para el conjunto de los 60 genomas fue de 1.587, con un mínimo de 1.430 (Aklavik86) y máximo de 2.368 (*H. cinaedi*). El contenido medio de GC fue de un 40% para los 60 genomas y de 41% solo para las especies NHPH. El valor mínimo de GC en las especies NHPH fue 36% (*H. cetorum*) y el máximo 48% (*H. heilmannii*). En cinco de las siete especies NHPH (*H. acinonychis, H. cetorum, H. heilmannii, H. hepaticus y H. cinaedi*), los nucleótidos de citosina superaron a los nucleótidos de guanina en la cadena principal. En las especies *H. bizzozeronii* y *H. mustelae* los nucleótidos de guanina superaron a los nucleótidos de citosina en la cadena principal. Nuestros resultados revelaron que, el promedio de genes que son codificados en la cadena principal del brazo es de 924 genes y 940 para la cadena rezagada. Para los genes que son codificados en la cadena principal, se observó que el promedio de sesgo de GC negativo, como positivo, fue igual (25%). Mientras para los genes codificados en la cadena rezagada, se observó que el 18% posee un sesgo negativo y un 32% positivo.

2.4.3.2 Análisis del pangenoma de 60 genomas de *Helicobacter*

El análisis del pangenoma mostró un tamaño de 8.771 grupos de genes, con un genoma core de 505, genoma softcore 894, genoma shell 1.604 y genoma cloud de 6.171. La sumatoria del genoma shell y el genoma cloud, resulto en 7.775 grupos de genes que conforman el genoma variable.

2.4.3.3 Evaluación funcional de los grupos de genes de 60 genomas de *Helicobacter*

Con el fin de determinar el perfil funcional de los 466 grupos de genes ortólogos que comprenden el genoma core de las 53 cepas *H. pylori* y las 7 especies NHPH, los grupos consenso de genes fueron asignados a un término de ontología génica (GO) mediante el programa Sma3s.

Con este programa los genes del genoma central fueron clasificados en tres categorías: componentes celulares, procesos metabólicos y procesos biológicos,

Componentes celulares: En esta categoría estuvieron involucrados 234 grupos de genes para 11 para procesos celulares. El genoma central estuvo sobrerrepresentado principalmente por grupos de genes vinculados a componentes del citoplasma (100) y membrana (91). Todos los hallazgos se pueden ver en **Figura 6.**

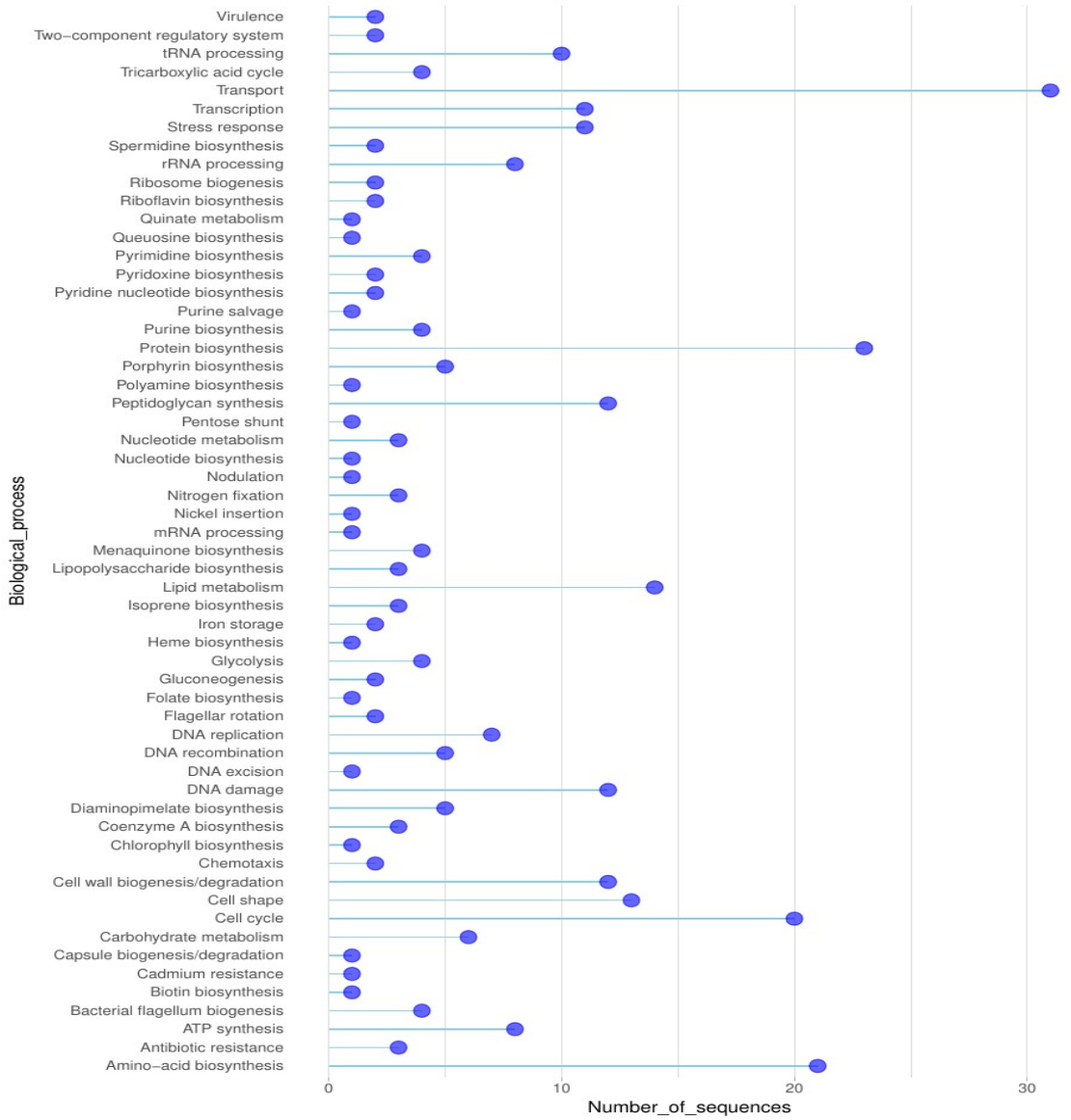


Figure 6. Clasificación de los genes del genoma central del género *Helicobacter.* Los genes fueron clasificados mediante términos de GO.

Procesos metabólicos: En este proceso estuvieron involucrados 108 grupos de genes para procesos metabólicos. El genoma central estuvo sobrerrepresentado principalmente por grupos de genes vinculados procesos de biosíntesis de amino ácidos (14) y cofactores de biosíntesis (14).

Procesos celulares: En este proceso estuvieron involucrados 321 grupos de genes para 60 para procesos biológicos. El genoma central estuvo sobrerrepresentado principalmente por grupos de genes vinculados a proteínas de transporte (32) biosíntesis de proteínas (24), biosíntesis de amino ácidos (21) y ciclo celular (21).

En cuanto a los genes para la colonización y patogénesis observamos que, para las especies NHPH los genes de amortiguación del pH y de motilidad (ureA hasta *ureG*) se encuentran presentes en casi todas estas especies, excepto en *H. cinaedi*. Los genes vinculados con la quimiotaxis (*tlpA, tlpB, tplC, tplD, cheA, cheW, cheW*) son parte de su genoma central en casi todas las especies NPHP, excepto en *H. cinaedi* y *H. hepaticus,* donde se encuentran ausentes los genes *tplB* y *tlpC*. Y los genes de adhesión *alpA/alpB*, solo estuvieron presentes en las especies *H. acinonychis* y *H. cetorum*. Los demás genes relacionados con la adhesión estuvieron ausentes en todas las especies NHPH. Otros genes de adhesión detectados como esenciales estuvieron presentes en las especies *H. acinonychis* y *H. cetorum*, que fueron quienes exhibieron un mayor número de estas adhesinas. Solo en la especie *H. cetorum* falto el gen *hofD***.**

Por otro lado, investigamos también los procesos biológicos vinculados a la cascada de regulación de la transcripción y biosíntesis flagelar. Así, en las especies NHPH pudimos constatar la presencia de los tres factores sigma 80, 54 y 28 (*rpoD*, *rpoN* y *fliA*), los genes flagelares clase I (*flgR, fglS, flhA*), los genes estructurales del sistema flagelar (*motA*, *motB*), los genes flagelares clase II (*flaB, flgE, flgK, flgM, flgL*), y los genes facilitadores de la secreción de los genes flagelares clase II (*fliQ, fliR*), son todos ellos parte del genoma core. Otros genes esenciales con función flagelar fueron detectados en las especies NHPH, entre ellos se encuentran: *fliW2, fliE, flgB, flgN, fliN, fliW1, flhF, fliM, fliY* entre otros.

En el caso de los genes en respuesta al estrés, casi todos los genes hicieron parte del genoma central para las especies NHPH. Solo el *htpX* estuvo ausente en especie *H. heilmanni.*

La presencia de otros genes de virulencia, fue establecida en su totalidad para las especies *H. acinonychis* y *H. cetorum*, En las demás especies, la presencia de estos otros factores de virulencia fue variable.

2.4.3.4 Filogenia 60 genomas de *Helicobacter*

De los 505 genes que hicieron parte del genoma core, se utilizaron los 466 genes, ya que 39 grupos de genes se correspondieron a parálogos. De este conjunto de datos genómicos, se obtuvieron 52.104 SNP mediante la aplicación de una serie de filtros implementados previamente (**ver Material y Métodos).** El árbol filogenético reveló un agrupamiento en dos grandes clados principales (África/Europa y Asia/Amerindios). En la **Figura 7** se muestra el árbol filogenético en el cual se puede observar que la especie *H. pylori,* más concretamente la subpoblación hpAfrica2

(S. Africa7, S. Africa20) se encuentra más cercana a las especies también gástricas *H. acinonychis* y *H. cetorum*. Más lejanas se encuentran otras dos especies de NHPH gástricas *H. heilmannii* y *H. bizzozeronii*, las cuales han divergido claramente de las especies enterohepáticas *H. hepaticus* y *H. cinaedi*. Y, por último, más cercana a estas últimas, se encuentra *H. mustelae*, cuyo nicho ecológico, es el gastrointestinal **(Figura 7)**.

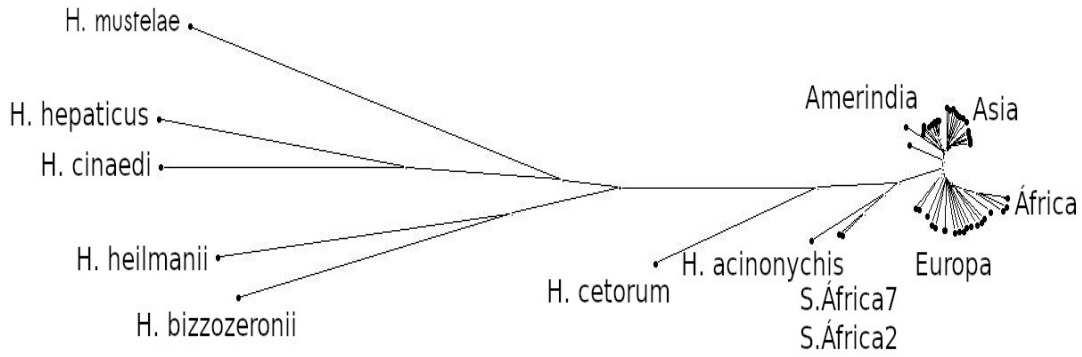
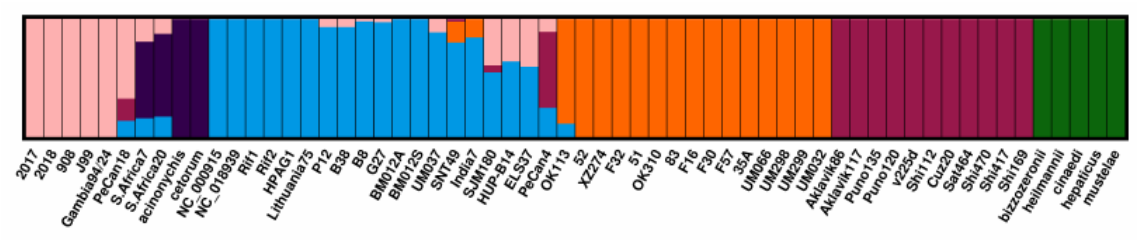


Figura 7. Análisis filogenético *H. pylori.* Árbol filogenético para 60 genomas de *Helicobacter* (0.01) basado en 52.104 SNP. La topología del árbol revela dos grandes clados (África/Europa y Asia/Amerindios).

2.4.3.5 Estructura poblacional de los 60 genomas de *Helicobacter*

Para las 60 cepas estimamos los componentes para poblaciones *K*= 4, 5, 6, 7, 8, con un bootstrapping "*B*" de 100.000 **(ver Materiales y Métodos).** Los resultados de ADMIXTURE revelaron que, los dos mejores valores para la prueba de validación cruzada fueron, *K*= 6 (0.48776), seguido de *K*=7 (0.48798). Los dos resultados obtenidos se ajustaron a lo esperado. En ambos casos las especies NHPH son estructuradas correctamente, sin embargo, interpretamos que *K*= *7* estructura mejor los componentes de las subpoblaciones europeas de *H. pylori* por lo tanto se adapta mejor **(Figura 8)**

A). *Helicobacter pylori* y 7 especies NHPH *K=6*



B). *Helicobacter pylori y* especies NHPH *K=7*

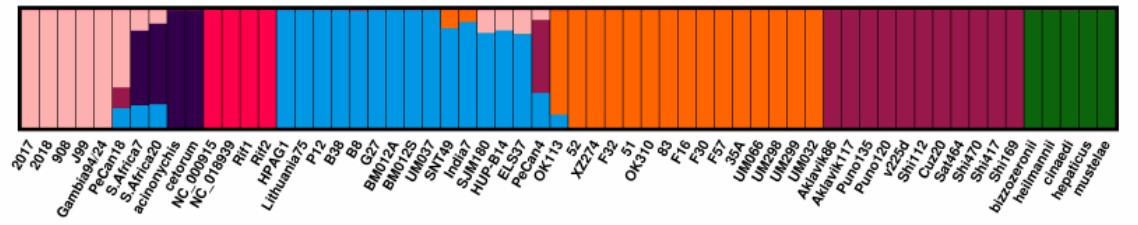


Figura 8. Estructura población de las 53 *H. pylori* y 7 especies de especies NHPH determinada mediante la prueba de validación cruzada mediante el programa ADMIXTURE para *K* poblaciones. **(A).** CV error K=6 (0.48776). **(B).** CV error K=7 (0.48798)**.**

2.4.3.6 Matriz de coascendencia con estructura poblacional y flujo génico de los 60 genomas de *Helicobacter*

A continuación, se aplicó el algoritmo fineSTRUCTURE para inferir el intercambio genético de segmentos de ADN entre las cepas de *H. pylori* y las especies de NHPH y dentro de las especies de NHPH. Lo primero que observamos, es que la matriz de co-ascendencia reveló muchos eventos de intercambio genético entre las especies NHPH, los más destacados se dieron entre *H. hepaticus/H. Mustelae (enterohepáticas) H. bizzozeronii/H. Heilmannii (gástricas)* y viceversa, seguidos de *H. cinaedi/H. Mustelae (enterohepática/gastrointestinal)*, *H. bizzozeronii/H. Mustelae (gástrica/gastrointestinal)* **(Figura 9, rectángulo 1)**.

La revisión del flujo por nichos ecológicos, mostró que en las cepas enterohepáticas *H. cinaedi, H. hepaticus,* las cuales están estrechamente relacionadas, el flujo fue considerablemente alto entre ellas. Así mismo, observamos un flujo génico importante desde *H. mustelae* (nicho ecológico gastrointestinal) hacia *H. hepaticus* y *H. cinaedi*, con las cuales esta emparentada. Este flujo también fue evidenciado desde estas dos últimas cepas hacia *H. mustelae.*

Para las especies gástricas (*H. heilmannii, H. bizzozeronii*), el flujo génico fue alto entre ellas. Cuando revisamos este flujo con las especies gástricas más cercanas al humano (*H. acinonychis* y *H. cetorum*), no observamos evidencia de flujo con *H. acinonychis* pero si con *H. cetorum*, en cuyo caso fue moderado. Por último, con *H. mustelae* (gastrointestinal) el flujo fue moderado y con las especies enterohepáticas fue bajo.

Por otro lado, evidenciamos que las especies *H. aninonychis* y *H. cetorum*, son estructuradas junto con la subpoblación sudafricana hpAfrica2 (S.Africa 7 y S.Africa20). Así mismo, un importante flujo génico desde *H. aninonychis* es evidente hacia esta subpoblación sudafricana, sin embargo, esto no fue observado desde *H. cetorum* hacia esta subpoblación **(Figura 9, rectángulo 2)**.

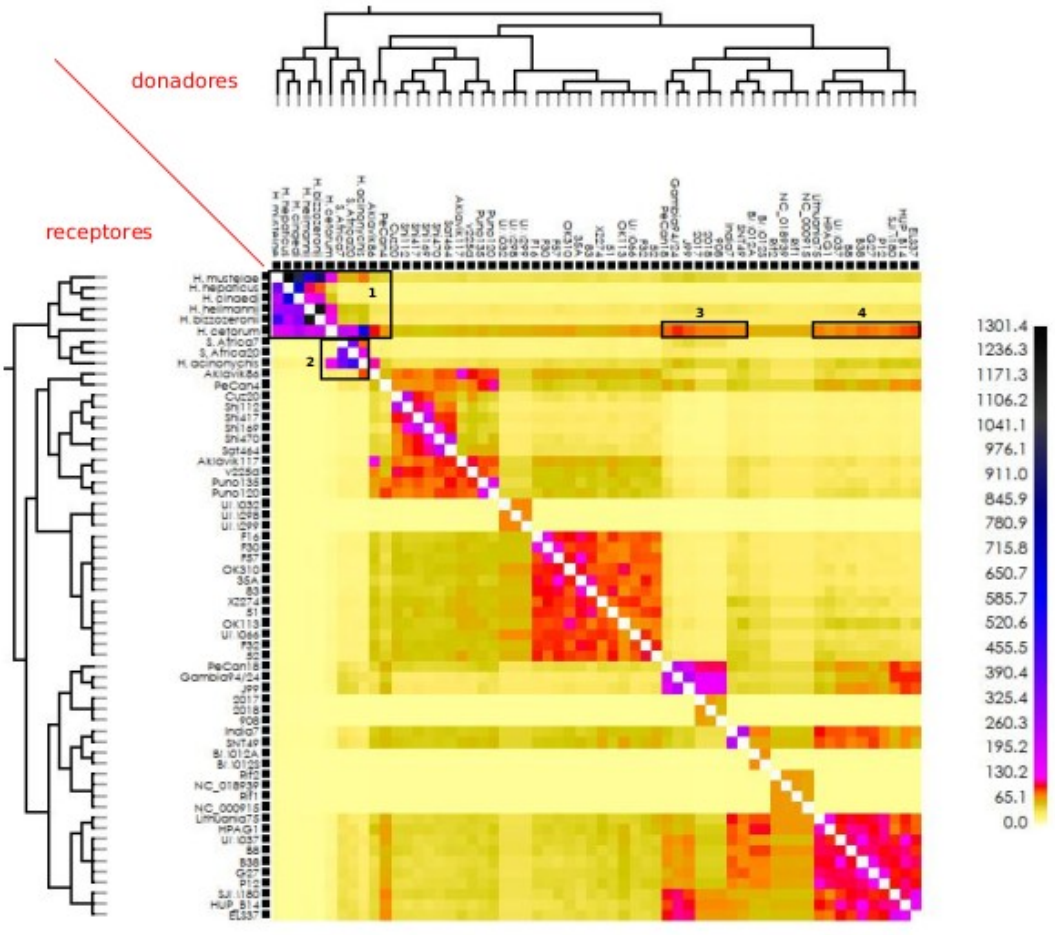


Figura 9. Matriz de co-ascendencia con estructura poblacional y flujo génico entre especie de *H. pylori* y otras especies de *Helicobacter*. Los recuadros **1** y **2** representan el flujo entre otras especies de *Helicobacter* y los rectángulos **3** y **4** el flujo génico observado de especies de *H. pylori* a *H. cetorum* principalmente.

2.5 Discusión

En esta tesis se han estudiado las características, genómicas de la bacteria patógena humana *H. pylori*. Los mecanismos de infección y patogénesis de *H. pylori* y las otras especies de NHPH son complejos, en estos intervienen interacciones entre cepas, el huésped y el nicho **(Kao, 2015)**. Se ha demostrado que una serie factores genéticos (genes) son fundamentales durante la interacción con el huésped (Motilidad, amortiguación del pH, adhesión y quimiotaxis, factores citotóxicos y evasión del sistema inmunológico) para manipular y evadir las defensas del huésped, para de esta manera, asegurar su supervivencia en el ambiente gástrico u otros nichos **(Kumar *et al.,* 2016).** En este sentido, en este capítulo hemos centrado nuestros esfuerzos en caracterizar y comparar los genomas centrales tanto entre cepas de *H. pylori,* como con otras especies de NHPH. Esta estrategia es interesante ya que, además del número mínimo de genes para mantener la vida celular, es necesario proporcionar los genes indispensables para la supervivencia y establecimiento del microorganismo, además del potencial de estos genes para constituirse como posibles blancos para aplicaciones de diagnóstico y terapias.

2.5.1 Características del genoma de *H. pylori*

2.5.1.1 Tamaño del genoma

Tras la comparación de 53 genomas de *H. pylori* **(Tabla 1)**, se ha observado que los genomas tuvieron tamaños comprendidos entre 1.709.911 (Gambia94/24) de máximo con 1.613 genes y mínimo de 1.494.183 (Aklavik86 de origen amerindio) como un contenido génico de 1.430 con una media de 1.621.671 ±43.785.

El número mínimo de genes los presentó la cepa Aklavik86 con 1.430 como podríamos esperar al ser la cepa que presenta el menor tamaño del genoma. Sin embargo la cepa Rif1 de origen europeo que presenta el mayor número de genes 1.640 no es la de mayor tamaño aunque si está próximo 1.667.883.

La variación en la longitud del tamaño del genoma y el contenido génico entre las cepas de *H. pylori* puede indicar que en la evolución de esta bacteria suceden fenómenos de pérdida y ganancia de elementos genómicos, un fenómeno común que ocurre tanto en bacterias como en arqueas **(Iranzo *et al.,* 2019)**.

Así, uno podría preguntarse sí, ¿la variabilidad entre genomas refleja la adquisición de elementos genómicos (secuencias repetidas cortas, genes, fagos, secuencias de inserción) por

algunas cepas o la perdida por otras, o es una combinación de ambos procesos? En este sentido, en *H. pylori* son diversos los mecanismos que pueden contribuir a esta variación **(Suerbaum & Josenhans, 2007; Draper *et al.,* 2016; Bubendorfer *et al.,* 2016)**.

En primer lugar, se ha podido comprobar que la ganancia de genes se puede producir por transferencia horizontal (HGT), mecanismo que permite que algunos de elementos genómicos estén presentes solo en unas pocas cepas, pero no en todas formando así, parte del genoma accesorio **(Tettelin *et al.,* 2005).** En este sentido, hemos podido podido comprobar que tanto los genes de la isla patogenicidad cag PAI: *cagA, cagB, cagD, cag23, cagE y cagB,* así como el transposon *IS606* transposase *(que solo* se limita a cepas asiáticas) están presentes en unas cepas y otras no. Por lo tanto, probablemente estos componentes genómicos fueron importados por HGT en algún momento de la evolución de esta especie a partir de otra diferente. Los genes adquiridos a través de HGT pueden mantenerse mediante presiones de selección del nicho local, ya que la retención de estos genes accesorios está generalmente relacionada con la adaptación a los ambientes extremos, como es el caso de *H. pylori* **(Moulana *et al.,* 2020).**

En segundo lugar, la perdida de genes puede deberse a mutaciones o eventos de recombinación que provocan deleciones **(Björkholm *et al.,* 2001; Aras *et al.,* 2003; Dong *et al.,* 2014)**. La reducción del número de genes puede ser más frecuente para genes no esenciales o redundantes **(Medonca *et al.,* 2011)**. *H. pylori*, presenta una gran cantidad de genes que codifican proteínas con función desconocida (hypothetical proteins), o elementos génicos como ADN repetido, los cuales pueden no poseer características funcionales esenciales y por lo tanto es posible que estos genes tiendan a ser eliminados **Gressmann *et al.,* 2005)**. Sin embargo, es inusual que una bacteria como es *H. pylori* que solo coloniza el ambiente gástrico humano pierda tantos genes. Puede ser probable que la variación del genotipo del huésped, la respuesta inmune e inflamatoria del estómago influyan en las propiedades fisicoquímicas y genéticas de *H. pylori*, lo que conllevaría a alteraciones genómicas **(Dong *et al.,* 2014)**. En este sentido es de destacar los tamaños pequeños que presentan los genomas de las cepas africanas, 2017 (1484 genes), 2018 (1494 genes9 y 908 (1479 genes) próximos al número de genes de la cepa amerindia aklavik86 con menor número de genes (1430), pudiendo haber asociación entre la región geográfica con la pérdida de genes. En este sentido ***Dong et al.* (2014)** señalan contrariamente a lo que decimos, la reducción más evidente de daba en las cepas asiáticas.

Dong *et al.* (2014) proponen que las cepas asiáticas en comparación con cepas occidentales presentaban un menor tamaño de genoma lo que está en contradicción con nuestros resultados. Esta discrepancia puede ser explicada por dos hechos. Por un lado, en el trabajo mencionado

incluyen en el mismo clado cepas de origen asiático y amerindio, mientras que en esta tesis los hemos considerado como dos clados distintos lo que permite interpretar los resultados de forma diferente.

La tercera explicación, para la reducción del número de genes consistiría en hecho de que un número de genes supondría un menor gasto energético para el mantenimiento de la estructura y replicación del ADN (**Renea *et al.,* 2005; Carril & Martin, 2010; Renea *et al.,* 2005**).

2.5.1.2 Sesgo de GC en el genoma

El contenido medio de GC para los genomas completos de *H. pylori* fue del 40%. No se encontraron variaciones apreciables, ya que estos valores oscilaron entre cepas entre el 39% al 40%. Este contenido es más bajo que en otras especies. Se ha sugerido que, miembros de taxones de vida libre como Actinobacteria, Acidobacteria, Betaproteobacteria y Deinococcus-Thermus, con genomas grandes y con capacidad de adquirir ADN transferido lateralmente poseen un contenido medio de GC mucho más alto (> 65) **(Mann & Chen., 2010)**. Por otro lado dentro de las Enterobacterias, grupo al que pertenece *H. pylori* también presentan un contenido de GC de alrededor del 50% **(Hershberg &Petrov, 2010)**, más alto que el porcentaje que presenta *H. pylori* (40%

Varias hipótesis se han planteado para dar explicación a los cambios en el contenido de GC en los genomas bacterianos. Una explicación que puede aducirse podría ser la tendencia observada en bacterias con ambiente especifico de presentar un tamaño de genoma menor de 3Mb y con un mayor contenido rico en AT **(Bentley & Parhill, 2004)**, **(Wixon, 2001)**.

Otra explicación alternativa, estaría ligada a procesos adaptativos tales como:, factores como la temperatura corporal **(Bernardi *et al.,* 1988)**, temperatura ambiental **(Kawaga *et al.,* 1984)**, condiciones halófilas **(Kennedy *et al.,* 2001)**, condiciones aeróbicas **(Naya *et al.,* 2002)**, ambientes de alta radiación (**Singer & Ames, 1970),** costes energéticos **(Rocha & Danchin, 2002)** y ambientes bajos en contenido de nitrógeno **(Dufresne *et al.,* 2005)**. Y por otro lado están las explicaciones neutralistas, que invocan cambios en el sesgo de mutación como causa de variaciones en el contenido de GC del genoma **(Freese, 1962; Sueoka, 1962).** En este sentido, se ha propuesto que la presencia o ausencia de genes de la vía de reparación del ADN, aumenta la tasa de mutación, lo que peude dar lugar a cambios en el contenido de GC. Se ha demostrado que la deleción de los genes de reparación *mutM* y *mutY*, que son componentes de

la vía de reparación de mutaciones por escisión de bases (BER), pueden influir en el contenido de GC del genoma, ya que ambos corrigen las mutaciones de GC → AT **(García-Gonzalez *et al.,* 2012).** Estos dos genes están ausente en *H. pylori* en el que solo se encuentra un encontraron un homólogo de *mutY* en *H. pylori* y la deleción de este influyo en la elevación de las mutaciones GC → AT. **Kulick *et al.* (2008)**

En los genomas bacterianos está bien establecido que la guanina es más abundante que la citosina a lo largo de la cadena principal de cada brazo cromosómico, lo que da como resultado un sesgo de GC promedio positivo (sesgo de GC= (G-C) / (G+C)) **(Lobry, 1996; Grigoriev, 1998; Fonseca *et al.,* 2008; Hendrickson & Lawrence, 2006). En este trabajo hemos calculado** el número genes codificados por la cadena principal 753 ± 31, los genes codificados en la cadena rezagada 798 ± 37, así como el sesgo de GC en cada una de las cadenas mediante la metodología de **Merrikh & Merrikh. (2018).**

2.5.1.3 Pangenoma *H. pylori*

Se ha sugerido que *H. pylori* posee un pan-genoma abierto **(Fischer *et al.,* 2010; Kawai *et al.,* 2011).** Es decir, tiene la capacidad y la maquinaria para adquirir ADN exógeno **(Medine *et al.,* 2005)**. En este sentido, la naturaleza abierta o cerrada de un pan-genoma está ligada al estilo de las especies, que en el caso de *H. pylori* es simpátrico facultativo **(Rouli *et al.,* 2015)**. Y aunque no tiene un genoma grande, esta especie se caracteriza por la panmixia, mosaicismo y una alta tasa de HGT **(Salaün *et al.,* 1998).** Así mismo, el tipo de pangenoma se suele relacionar al estilo de vida de la bacteria, en el caso del pangenoma abierto está relacionada con la gran capacidad para adquirir nuevos genes y muchos de ellos pueden aportar ventajas adaptativas.

Nuestro análisis ha revelado un promedio de 1.551 ± 42 por genoma completo y una media de 820 grupos de genes compartidos en el genoma central, representando el 53% del tamaño del genoma. Nuestros resultados para el genoma central difieren notoriamente de los estudios hasta ahora publicados, ya que el promedio de genes para el genoma central de estos estudios ronda los ~1.189 genes **(Salama *et al.,* 2000; Ali *et al.,* 2015; You *et al.,* 2015; Kumar *et al.,* 2015; Uchiyama *et al.,* (2016); Cao *et al.,* 2016); van Vliet, 2017).** Estas diferencias pueden deberse a dos motivos, por un lado el número de cepas utilizadas en cada estudio así como los parámetros de fiabilidad utilizados en cada uno ellos. Por ejemplo **Cao *et al.,* (2016)**, implemento un método muy similar al utlizado en este trabajo para el agrupamiento de ortólogos denominado “orthoMCL”, las comparaciones de BLASTP de todos contra todos para obtener el

conjunto de datos de proteínas, tuvo en cuenta una cobertura para el alineamiento de secuencias a partir del 50%. **Ali *et al.,* (2015)** estimo el genoma central mediante solo empleo de alineamiento con BLASTP y al igual que **Cao *et al.,* (2016)**, la cobertura de los alineamientos fue a partir del 50%. Similar estrategia, utilizó **van Vliet (2017)**, en su caso implementó un punto de corte del 90% de identidad como porcentaje mínimo para BLASTP, sin embargo, no tuvieron en cuenta la cobertura para el alineamiento de secuencias. Mientras que en este trabajo se impuso una cobertura a partir del 75%, lo que supone un valor más estricto para detectar verdaderos ortólogos al momento de detectar secuencias del genoma central y, por lo tanto, un menor número de grupos de genes que constituirán el genoma central se vería reflejado, como es el caso. Mis datos

2.5.1.4 Filogenia, estructura poblacional y flujo génico de *H. pylori*

La historia evolutiva de *H. pylori* es fascinante por su larga e íntima asociación con las migraciones humanas a través del trazado de su filogeografía y, hay suficiente evidencia que ha demostrado que los humano modernos emigraron de África a la península Arábiga hace aproximadamente 60.000-150.000 años, y posteriormente de forma independiente a Europa y Asia **(Megraud, Lahours & Vale., 2016; Waskito & Yamahoka., 2019)**.

En la última década se han desarrollado múltiples herramientas para determinar la estructura poblacional y reconstruir la historia evolutiva de esta especie **(Megraud, Lahours & Vale., 2016)**. En 1998, se propuso el método de tipificación de secuencias multilocus (MLST) para la caracterización de especies bacterianas monomórficas, cuyo método está basado en la utilización de fragmentos de genes de mantenimiento de aproximadamente 470 pb **(Achtman, 2008)**. Así **Achtman *et al.,* (1999)** fueron los primeros en aplicar MLST en *H. pylori* utilizando los siguientes genes: *atpA, efp, mutY, ppa, trpC, ureI, yphC, cagA y vacA*. Sus resultados, demostraron que se podía distinguir la descendencia clonal, a pesar de las frecuentes recombinaciones que reflejan las diferentes regiones geográficas. Estos resultados fueron confirmados por **Falush *et al.,* (2003a)** mediante la implementación del paquete STRUCTURE **(Falush *et al.,* 2003b)** pero, además establecieron que *H. pylori* es una bacteria polimórfica. Con el aumento en la disponibilidad de datos, ha sido necesario desarrollar nuevos enfoques que consideren todos los datos del genoma, y que permitan además diferenciar a una escala más fina entre subpoblaciones. En este sentido surgió el paquete fineSTRUCTURE **(Lawson, Myers**

& Falush, 2012), que que cumple con las características para llevar a cabo estas clasificaciones y que hemos implementado en este trabajo.

De nuestros analisis se pueden valorar diferentes aspectos respecto al estudio genomico *de H. pylori.* El primer aspecto es la alta tasa de variablidad que exhibe esta especie, lo que nos permite realizar asignaciónes en y entre regiones geograficas, diferenciando de forma muy exacta el origen y el asilamiento de una cepa ( ).

Con los tres métodos de analisis filogenticos utilizados se obtiene una topología idéntica en cuanto a la asignación de dos grandes clados, África/Europa y Asia/Amerindia coincidentes con lo reportado por **Falush *et al.,* (2003)** mediante análisis de genes MLST y por **Yahara *et al.,* (2013)** Las asignaciones realizadas para subpoblaciones en este trabajo, fueron mejor establecidas por fineSTRUCTURE y más coherentes con la filogenia. Así, las cepas se agruparon conjuntamente según su ascendencia genética, **Figuras 3** y **4.**

Sin embargo, cinco excepciones a estas asignaciones fueron evidenciadas por la existencia de cepas de origen hibrido. Dos de ellas, PeCan4 un hibrido entre cepas amerindios y cepas de genomas europeas y SMJ180 híbrido entre hpAfrica1 y los subgrupos europeos ya descritas previamente por **Yahara *et al.* (2013);** la cepa ELSE37, un asilamiento de origen de Centro América es un hibrido con ascendencia principalmente desde Europa, seguido de África y America, la cepa PeCan18 un asilamiento de Perú, pero con ascendencia principalmente de desde África y Europa. Y UM037 cepa de origen malayo, con *K=5* mostró una ascendencia principalmente desde Europa y África,

Nuestros resultados también son consistentes en relación a la rama larga del árbol de la población hpAfrica2 (cepas sudafricanas), lo cual indica que es la más divergente, **Falush *et al.,* (2003b)**. Por otro lado, las cepas BM012A, BM012S incluidas en este estudio (aislamientos de origen australiano), y no incluidas por **Yahara *et al.* (2013)**, **Kumar *et al.* (2014)** y **Montana *et al.* (2015)** presentan una ascendencia principalmente desde hpEurope (hpEurope_sg53: Lithuania75 y HPAG1), indicativo de una mezcla reciente **(Tabla 4).**

También revelamos que el subgrupo hspEAsia_sg3 (UM032, UM298, UM299, cepas de origen malayo), está claramente diferenciado en un clado en la matriz de coascendencia **(Figura 4)** y en su composición genética de los otros subgrupos asiáticos **(Figura 2 y 4).**

Entre los aislamientos amerindios, nuestros análisis nos permitieron distinguir hasta ocho subgrupos, dos más que los reportados por **Yahara *et al.* (2013) (Tabla 4)** en los siguientes

subgrupos: hspAmerind_sg1 (Puno135: Perú, Puno120: Perú) igual a hspAmerind_sg1, hspAmerind_sg2 (Puno470: Perú, Sat464: Perú) con hspAmerind_sg2, hspAmerind_gs3 (Cuz20), nosotros adicionamos a este grupo la cepa Shi112 (Perú), singleton hspAmerind_sg4 (v225d: Venezuela) igual a hspAmernid_sg4. Adicionalmente, en este trabajo se caracterizaron los subgrupos, hspAmerind_sg4 (Shi169: Perú, Shi417: Perú) y hspAmerind_sg5 (Aklavik86, Aklavik117, ambas de una comunidad de aborígenes del Ártico, Canadá). Interesantemente, la cepa Aklavik86 mostró tener una ascendencia desde cepas de Malasia (UM299, UM298, UM032) **(Figura 2 y 4)** y tanto el árbol filogenético como la filogenia asociada a la matriz de coascendencia, mostraron que estas forman un clado diferente **(Figura 4** y **5).**

En general la exportación de segmentos de una cepa A, a una cepa B es mayor que la importación de una cepa B, a una cepa A. En este sentido es interesante destacar el flujo asimétrico observado en ciertas poblaciones. Por ejemplo, en la **Figura 3**, observamos que la subpoblación malaya (hspEAsia_sg3) dona a todas las subpoblaciones asiáticas, sin embargo, importa poco de las mismas. También observamos esto para la subpoblación hpEurope_sg1, la cual dona a casi todas las poblaciones, sin embargo, recibe poco importa poco desde otras poblaciones. Este mismo fenómeno es apreciable desde la subpoblación hspWAfrica_sg2 la cual dona a diversas regiones geográficas, pero importa poco o casi nada. También hpEuropa_sg3 (BM012A y BM012S, aislamientos australianos), mostró este comportamiento donando a diversas regiones geográficas, sin embargo, la importación es muy baja. Estos hallazgos pueden ser interesantes, ya que puede ser posible que el mecanismo que causa el flujo genético asimétrico pueda ser menos eficiente en ciertas poblaciones o que un entorno selectivo diferente haya disminuido las importaciones **(Yahara *et al.,* 2013).**

2.6 H. pylori con siete especies de *Helicobacter* no pylori (NHPH)

2.6.1 Características del genoma de las cepas *Helicobacter* y las especies NHPH

El género *Helicobacter* comprende actualmente más de 40 especies, las cuales han establecido relaciones simbióticas en el tracto gastrointestinal de uno o más huéspedes **(Smet *et al.,* 2019)**. Muchas de estas bacterias son de importancia patógena tanto para los seres humanos como para animales **(Flahou, Haesebrouck, Smet, 2016)**. En el momento de la obtención de los genomas de las especies NHPH con los cuales hemos realizado este trabajo, solo estaban disponibles los genomas completos de los genomas analizados aquí.

El tamaño medio de los genomas para las especies NHPH fue 1.795.004. El genoma con menor tamaño lo presentó *H. acinonychis* (1.553.927 pb) y el máximo *H. cinaedi* (2.240.130 pb). La primera es la especie más cercana a la espcie *H. pylori* y el tamaño de su genoma no esta muy distante del tamaño medio de *H. pylori* (1.621.671 pb).Se considera que estos genomas son relativamente pequeños y compactos en comparación con otras bacterias, lo que puede indicar una adaptación específica para sus estilos de vida como patógenos obligados **(Cao *et al.,* 2016)**.

El promedio de genes para los 60 genomas (*H. pylori* y NHPH) fue de ~1.587 (parecido al de la especie *H. pylori* 1.551 ±42), con un mínimo de genes de 1.430 (Akavik86) y un máximo de 2.368 (*H. Cinaedi*). Mientras que, para las NHPH fue de 1.865 ± 209.241 genes, con mínimo de 1.467 (*H. mustelae*) y un máximo de 2.386 (*H. Cinaedi*). Interesantemente el número de genes presentes en los genomas de las especies *H. acinonychis* (1.655) y *H. cetorum* (1.733) está mucho más cercano a la especie *H. pylori,* lo cual podría estar vinculado a su relación cercana con esta especie, además de ser especies también gástricas. Mientras que casi todas las otras especies superaron el rango de los 1900 genes, excepto *H. mustelae* (1.467).

El contenido medio de GC para los 60 genomas fue del 40%. Mientras que para las NHPH fue del 41%. Estos valores no representan variaciones considerables a nivel de promedios de los genomas completos, al comparar las cepas de *Helicobacter* con las especies NHPH. Sin embargo, estos estos valores si oscilaron drásticamente cuando revisamos el punto de corte más bajo para GC, donde el mínimo hallado en las especies NHPH fue del 36% (*H. Cetorum*) y el máximo del 48% (*H. Heilmannii*). Lo cual podría tener alguna relación con el tipo de ambiente en que habitan.

En el caso de las NHPH, cinco de las siete especies mostraron que los nucleótidos de citosina superan a los nucleótidos de guanina. Solo las especies *H. bizzozeronii* y *H. mustelae* mostraron el patrón contrario, similar al que presentan las cepas de *H. pylori*. Curiosamente, una de ellas fue la que menos número de genes presentó en su genoma (*H. mustelae*, 1.467). En las cepas que los nucleótidos citosina superan a los nucleótidos de guanina, observamos que, para *H. bizzozeronii* hay un número mayor de genes que son codificados en la cadena rezagada (HO) con un sesgo de GC negativo, además esta cepa exhibió el segundo promedio de GC más alto, un 46%.

Mientras que, *H. mustelae* mostró un mayor número de genes codificados en la cadena principal (CD) con un sesgo de GC negativo. Curiosamente, este patrón es el que exhiben las cepas de *H. pylori* y es interesante, el hecho de que esta cepa presente un número de genes más cercano a

las cepas de *H. pylori* y un contenido medio de GC también muy similar a *H. pylori*, un 39%. Sin embargo, al no tener incluidas más cepas NHPH con resultados similares para corroborar este hallazgo, no pudimos establecer una correlación.

El pan-genoma para los 60 genomas de *Helicobacter* aumento en 3.637 genes, el genoma central disminuyo en 315 genes, en relación a *H. pylori*. El genoma variable aumento en 3.803 grupos de genes, lo que se traduce en 18 nuevos genes aportados por cada una de las especies NHPH. Los resultados del pan-genoma para las especies NHPH sugieren ser más abiertos y más diversos, que los de *H. pylori,* como también indica **Cao et al. (2016).** Sin embargo, recientemente **Smet *et al.* (2019)** calcularon el genoma central en 399 grupos de genes, para las cepas de *Helicobacter*. Esta diferencia en gran medida se debe a que estos incluyeron más cepas de NHPH. Aunque podemos resaltar que nuestros resultados son más precisos utilizando menos genomas, en comparación a los resultados obtenidos por **Cao *et al.* (2016)** utilizando 99 genomas, 39 más que nosotros. También es interesante destacar la diferencia que existe entre el tamaño del genoma de *H. pylori* cuando incluimos las especies NHPH, pasando de 820 a 505, respectivamente (un 39% menos con respecto a *H. pylori*). Esto puede indicar la presencia de familias de genes únicas compartidas por las *H. pylori* con funciones exclusivas o muy relevantes para adaptación y patogénesis.

2.6.1.2 Filogenia, estructura poblacional y flujo génico de los genomas de *Helicobacter*

La filogenia basada en el genoma central, mostró una clara división entre las cepas por nichos ecológicos. Se puede observar que la especie NHPH gástrica estrechamente relacionada con *H. pylori, H. Acinonychis* se agrupa con las cepas sudafricanas, seguida de la también gástrica *H. cetorum*. Las otras especies también gástricas *H. heilmannii* y *H. bizzozeronii* formaron su propio grupo, al igual que las enterohepáticas *H. hepaticus* y *H. cinaedi* y, la gastrointestinal *H. mustelae*. Nuestros resultados, aunque con menos cepas de NHPH son consistentes con los reportes previos hechos por **Cao *et al.* (2016)** y **Smet *et al.* (2019)**.

Como hemos mencionado en párrafos anteriores, las altas tasas de recombinación son una característica de las cepas de *Helicobcater* **(Fraser *et al.*, 2011).** Así, en general la asignación de poblaciones de las especies NHPH, estuvo bien correlacionada con los diferentes clados del árbol filogenético y la filogenia asociada a la matriz de co-ascendencia **(Figura #)**. Las demarcaciones ecológicas observadas en la filogenia y la filogenia asociada a la matriz de co-ascendia podrían subrayar los limites en los eventos de recombinación y desempeñar un papel

evolutivo en el proceso de especiación al definir la estructura de la población de las especies bacterianas **(Kennemann *et al.,* 2011; Smet *et al.,* 2019)**. De hecho, en este estudio, el intercambio genético entre las poblaciones gástricas y enterohepáticas de *Helicobacter*, no fue evidenciado **(Figura #, rectángulo 1)**, como también reportan **Smet *et al.* (2019).** Reforzando la idea de la presencia de una barrera ecológica que podría haberse producido cuando la primera parte del tracto digestivo se especializo (es decir, el estómago con producción de ácido), sin embargo, no se puede desestimar una barrera genética como una posible explicación alternativa **(Smet *et al.,* 2019)**.

Para el grupo gástrico (*H. acinonychis* y *H. cetorum*) más cercano a la especie *H. pylori*, se produjo un flujo de genes considerable entre estas especies dentro del clado. Aunque estas cepas no comparten el mismo hospedador, es muy probable que estos fenómenos representen señales restantes de ascendencia compartida.

Cuando se investigó el flujo de genes de este grupo gástrico, con el otro grupo gástrico (*H. bizzozeronii/H. Heilmannii),* no se encontró evidencia de flujo génico desde *H. acinonichys* hacia *H. heilmannii* y *H. bizzoronii,* a pesar de que ambas especies pueden infectar felinos, pero si desde *H. cetorum* hacia estas dos, lo que sería indicativo de señales restantes de ascendencia compartida. Así mismo, se ha demostrado que las especies gástricas que infectan animales domésticos (*H. bizzozeronii/H. Heilmannii)* han evolucionado de forma paralela, siguiendo un camino muy distinto **(Smet et al., 2019)**, como también lo refleja la falta de señales de mezcla entre las cepas de *H. pylori* y estas especies **(Figura #)**. Estas señales de intercambio génico entre especies diferentes también podrían representar características genéticas implicadas en el metabolismo, ya que muchas de estas especies tienes requisitos de crecimiento in vitro similares **(Flahou *et al.,* 2016; Smet *et al.,* 2013; Baele *et al.,* 2008)**.

Al igual que **Smet *et al.* (2019)**, el intercambio genético canino y felino fue evidente entre *H. heilmannii* y *H. bizzozeronii* **(Figura #)**, mostrando uno de los niveles más altos de recombinación entre especies.

Otro hallazgo interesante, fueron las señales de ascendencia dentro del clado de *H. pylori* hpAfrica2 (cepas sudafricanas). Donde se observó, la mezcla entre las cepas ancestrales de hpAfrica2 con *H. acinonychis* **(Figura #)**. Estos hallazgos están de acuerdo con hallazgos previos reportados por **Moodley *et al.* (2012)** y **Smet *et al.* (2019)**, donde se demostró que los ascendientes de esta antigua población de *H. pylori* son el origen de *H. acinonychis*. Así mismo se observó flujo génico con *H. cetorum* desde hpAfrica2, lo cual supondría un ancestro común de

ambas especies el cual tendría su origen en el continente africano **(Moodley *et al.,* 2012; Smet *et al.,* 2019)**. En este sentido ya se ha estimado que para *H. pylori* y *H. acinonychis*, la edad mínima de asociación es de aproximadamente ~100 kya, además de que *H. acinonychis* fue el resultado de un salto posterior de huéspedes humanos (hpAfrica2) a grandes felinos ca. 43-5 kya **(Moodley *et al.,* 2012)**. Y para *H. cetorum* se ha estimado en ca. 600 kya, en cuyo periodo de tiempo, los humanos modernos y los neandertales también habían divergido **(Green *et al.,* 2006; Smet *et al.,* 2019)**. Esto supondría que, la presencia de *H. pylori* en humanos es el resultado de un salto desde un huésped animal **(Moodley *et al.,* 2012; Smet *et al.,* 2019)**.

La matriz de co-ascendencia reveló también tales asimetrías en el flujo génico entre las 53 cepas de *H. pylori* y las otras especies. Por ejemplo, en la **Figura #**, observamos que la *H. acinonychis* la especie NHPH más cercana a *H. pylori,* ha donado moderadamente segmentos de ADN a la subpoblación hpAfrica2 (S.Africa 7, S.Africa 20), sin embargo, la importación de segmentos por parte de *H. acinonychis* es más alta a partir de hpAfrica2 **(Figura #, rectángulo 2)**. Otra especie cercana a *H. pylori*, como lo es *H. cetorum* dona poco o nada a la subpoblación hpAfrica2 sin embargo está, si importa segmentos de ADN desde hpAfrica2 **(Figura #, rectángulo 1)**

Por otro lado, las especies NHPH también gástricas, *H. heilmannii* y *H. bizzozeronii* donan segmentos a las especies *H. hepaticus* y *H. cinaedi* (enterohepáticas), siendo simétrico el flujo génico con *H. cinaedi* y asimétrica con *H. hepaticus*. También es asimétrico el flujo génico hacia *H. mustelae (gastrointestinal),* siendo mucho mayor la donación que, la importación de segmentos **(Figura 8, rectangulo 1)**.

Capítulo 3. Evolución concertada y recurrente en genes de *H. pylori*

3.1 Introducción

El modelo de evolución concertada (EC) representa la co-evolución de secuencias de ADN dentro de los genes de una misma familia génica de forma que conservan altos niveles de similitud entre sí **(Pride & Blaser., 2002).** La EC se caracteriza generar homogeneidad y mayor similitud entre las secuencias parálogas (gen ancestral y gen duplicado), que entre ortólogos; Esto provoca que las secuencias parálogas se agrupen de forma monofilética **(Carson & Scherer., 2009; Wang & Chen., 2018)**.

Los parálogos en EC de la misma especie muestran una mayor similitud de secuencia entre sí que cualquiera de los ortólogos de otras especies y, a menudo se forman de forma monofílica en el árbol filogenético. Sin embargo, tal patrón también podría surgir de la duplicación de genes específicos del linaje. Para distinguir entre estos dos escenarios, es muy importante tener en cuenta la sintenia genética para resolver la ortología y la paralogía del gen **(Nei & Rooney, 2005; Mansai & Innan, 2010)**. Esto se debe a que es poco probable que los parálogos con sintenia compartida entre especies se deriven de la duplicación de genes independientes y, por lo tanto, deberían ser el resultado de una evolución concertada **(Scienski *et al.,* 2015)**.

La EC entre genes duplicados puede producirse por un intercambio genético llamado conversión de genes (CG) **(Nei & Rooney, 2005).** La CG es el intercambio no reciproco de material genético entre secuencias homólogas y este proceso puede desencadenar eventos tanto negativos como positivos **(Carson & Scherer., 2009)**. De manera beneficiosa, la conversión de genes puede disminuir la carga mutacional, eliminar mutaciones deletéreas y propagar alelos ventajosos, jugando de esta forma un rol en la evolución adaptativa **(Hansen *et al.,* 2000)**. De forma negativa, la conversión de genes puede producir fenotipos perjudiciales, por ejemplo, cuando las mutaciones disruptivas de un pseudogen se sustituyen en su duplicado funcional **(Boocock *et al.,* 2003; Tayebi *et al.,* 2003).**

Evidencia de EC tanto en procariotas como eucariotas hay mucha y se ha observado con mayor frecuencia en los ARNr **(Liao, 1999).** También se ha observado EC en bacterias relacionadas con *H. pylori*, como lo es *Campyobacter*. En este sentido **Meinersmann & Hiett. (2000)**, pusieron de manifiesto este tipo de eventos para los genes de flagelina, *flaA* y *flaB* (**Meinersmann & Hiett, 2000; sheppard & Maiden, 2015).**

En *H. pylori* son pocos los estudios que han puesto de manifiesto eventos de EC en genes. En este sentido **Pride & Blaser (2002)** pusieron de manifiesto eventos de EC, en los segmentos 3' de los genes *babA* y *babB*, los cuales forman parte de una amplia familia de proteínas de membrana, las cuales cumplen una importante función de adherencia **(Moore et al., 2011)**. Otro gen que muestra evolución concertada es *HomB*, otra proteína de membrana que posee un parálogos estrechamente relacionados, *HomA* **(Oleastro *et al.,* 2009; Castro & Ménard., 2013).** Los estudios previos en *H. pylori* no abordan el escenario de la sintenia, por lo tanto, es necesario buscar nuevos genes duplicados con EC en *H. pylori* y así mismo, y establecer su importancia en la relación huésped-patógeno.

3.2. Material y Métodos.

3.2.1 Selección de cepas

53 cepas de la especie *H. pylori* fueron seleccionadas en función de los datos genómicos disponibles en la base de datos del NCBI **(consultada por última vez en 2017)**. Estos genomas se analizaron en 5 grupos (África, Europa, Asia, amerindia e híbridos) ya que la alta variabilidad que exhibe la especie *H. pylori* no permitió realizar un análisis de los 53 genomas implementado Mauve desde iSeeCe en un solo lanzamiento para detectar patrones de evolución concertada. La división de los grupos se realizó en base a los resultados obtenidos a partir de la filogenia y el análisis con fineSTRUCTURE.

3.2.2 Identificación de evolución concertada mediante el programa IseeCe

El programa IseeCe identifica la evolución concertada en función de la filogenia y la sintenia génica con ayuda del programa Mauve **(Darling et al. 2008; Darling et al. 2010)**. Así, los parálogos de la misma cepa se agruparan de manera monofílica (evolución concertada) y los ortólogos hallados en otras cepas se posicionaran de forma sinténica en la filogenia y se identificaran como genes convertidos. Como la evolución concertada puede ser difícil de distinguir de la duplicación en tándem, el algoritmo IseeCe excluye las duplicaciones en tándem.

IseeCe agrupa los genes en familias utilizando **OrthoMCL v2.0.4 (Li *et al.,* 2003)**. Sin embargo, el resultado de OrthoMCL puede verse afectado por el índice de inflación de **Markov Clustering (mcl) (Li *et al.,* 2003)**. Para reducir este sesgo en la clasificación de la familia de genes, IseeCe ejecuta OrthoMCL (mcl) aplicando diferentes índices de inflación (1, 1.5, 2, 4 y 6), para

finalmente fusionarlos y dar como resultado familias de genes consenso. Posteriormente, cada familia de genes codificantes es alineada utilizando **MAFFT (Katoh *et al.,* 2013)** y un árbol filogenético es construido con **FastTree (Price *et al.,* 2010).** En este estudio, la selección inicial de todas las familias de genes, donde parálogos de la misma cepa forman una monofílica en al menos tres cepas basándose en la filogenia construido por **FastTree**. A continuación, para las cepas se identifican los ortólogos sinténicos apoyados por sintenia genética conservada en todas las cepas utilizando **Mauve (Darling *et al.* 2008; Darling *et al.,* 2010).** Se necesitaron al menos **dos** genes circundantes con ortólogos de distintas cepas para respaldar la sintenia, esto debido a la alta variabilidad que muestra *H. pylori*. Por último, para cada familia de genes que pasó la selección inicial, se construyó la filogenia usando **RAxML (Stamatakis, 2006)** con 500 pseudorreplicados de arranque e implementando el modelo evolutivo GTR + GAMMA como modelo de sustitución (-s entrada -n salida -m GAMMAGTR - # 500 -p 123 -x 123 -fa).

Los genes recurrentes en evolución concertada se definieron solo si se descubría que los parálogos experimentan una evolución concertada en al menos cinco cepas (--gc_count_min 5)

3.2.3 Diversidad genética y pruebas de selección natural para los genes duplicados con evolución concertada

Los estadísticos de diversidad poblacional para los genes duplicados con evolución concertada se calcularon mediante los programas MEGA X **(Kumar *et al.,* 2018),** SWAAP **(http://www.thepridelaboratory.org/software.html)** y DnasPv 6 **(Librado *et al.,* 2017)**. Estos estadísticos fueron: **1)** porcentaje de identidad entre las secuencias (%I), **2)** número de sitios segregantes (S), 3**)** número de haplotipos (H), 4**)** diversidad de haplotipos (Hd), **5)** sustituciones sinónimas (Ks), **6)** sustituciones no sinónimas (Ka), **7)** relación media Ks/Ka, **8)** tasa de transiciones (Ts), **9)** tasa de transversiones (Tv) y **10)** relación media de Ts/Tv.

3.2.4 Pruebas de recombinación

El programa RDP5 **(Martin *et al.,* 2021)** se utilizó para el análisis de recombinación. Este programa utiliza los estadísticos Maxchi **(Smith, 1992)**, Bootscan **(Padidam *et al.,* 1999)** Chimera **(Posada & Crandall, 2001),** Siscan **(Gibbs *et al.,* 2000)**, 3seq **(Bomi *et al.,* 2007)**, Lard **(Holmes *et al.,* 1999)** y Phylopro **(Weiller, 1998)**. También se implementó el programa DnasP **(Librado *et al.,* 2017)** para detectar eventos de recombinación mediante la metodología **Hudson & Kaplan (1985).** También el programa RDP5 **(Martin *et al.,* 2021)** mediante el algoritmo

GENECONV permite analizar eventos de conversión de genes **(Sawyer, 1989)**. Teniendo en cuenta que presentaron valores de $p < 0.05$.

3.2.5 Desviaciones del modelo de neutralidad de evolución molecular mediante la prueba de Tajima

Las desviaciones del modelo neutral de evolución molecular se probaron utilizando la prueba de Tajima's. Todas estas pruebas fueron estimadas con MEGA X **(Kumar *et al.,* 2018).**Con el programa MEGA X **(Kumar *et al.,* 2018)**

Para ello, se utilizan los siguientes parámetros de diversidad genética: (π) que cuantifica la diversidad de nucleótidos, que es el número de nucleótidos diferentes por sitio entre dos secuencias tomadas al azar y (θ), que se calcula de manera indirecta utilizando el número total de sitios segregantes en un grupo de secuencias (un sitio segregante, es un sitio donde las secuencias difieren) **(Castillo-Cobian, 2007).** Estos dos parámetros se utilizan para calcular la desviación el sesgo de sustituciones de nucleótidos con el parámetro *D*. Así, si *D* da como resultado un valor negativo quiere decir que θ posee un valor mayor que π, lo que indica presencia de mutaciones deletéreas. Si por el contrario, *D* resulta positiva quiere decir que π tiene un valor que θ, lo que indica que algunos alelos están bajo selección positiva. Por último, si D es igual a cero, indica que no existen diferencias entre π y θ, por lo tanto nos encontramos bajo selección neutral **(Tajima, 1989).**

3.2.6 Pruebas de selección.

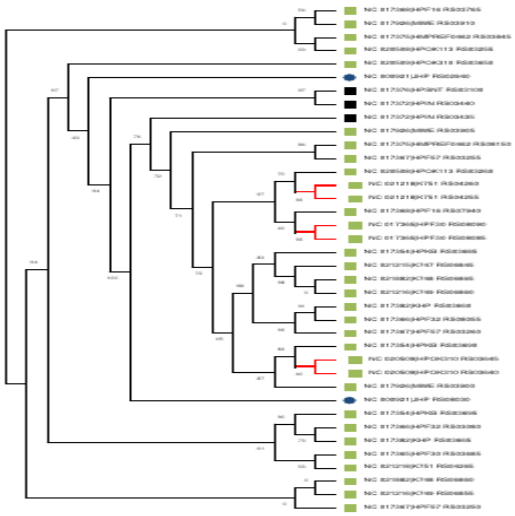
Se buscó evidencia de selección positiva usando el programa ETE3 **(Huerta-Cepas, Serra & Bork, 2016)** que implementa CODEML del paquete PAML **(Yang, 1997)** implementado el modelo M2 de tres categorías. Este modelo permite determinar tres clases de selección (negativa, neutra y positiva)

3.3 Resultados

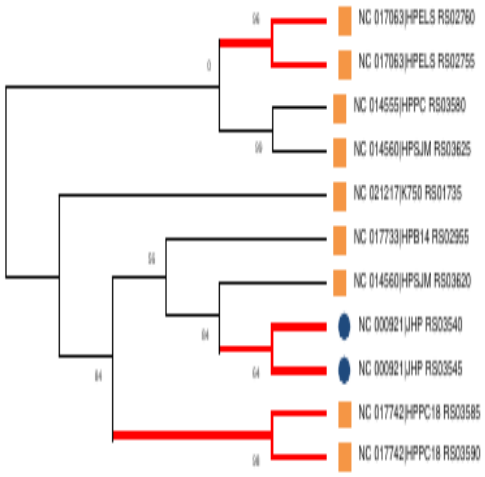
3.3.1 Identificación de genes duplicados con evolución concertada

Para detectar la EC se utilizó el programa iSeeCe **(Wang & Chen, 2018)** bajo los siguientes criterios: **1)** los genes duplicados en EC analizados deben de formar un grupo monofilético, **2)** los genes duplicados deben estar asociados a ortólogos sinténicos en el árbol filogenómico.

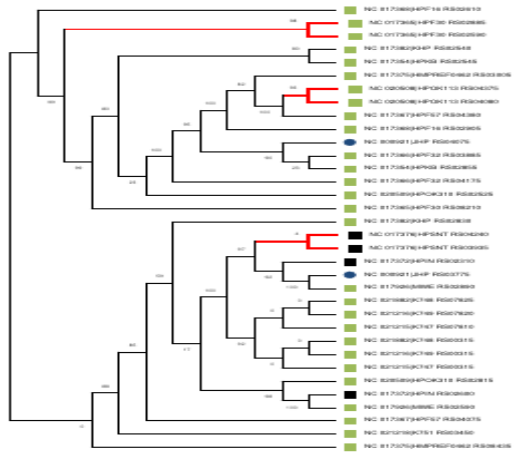
El análisis de los árboles filogenómicos obtenidos para los cinco grupos preestablecidos nos permitió detectar ocho familias de genes duplicados en EC **(Figura 1)**. Los grupos que presentaron cinco o más cepas con genes en EC fueron los siguientes para cada uno de los grupos: *oipA* (Asia= 11 sucesos de EC), alpha (1,3)-fucosyltransferase (Asia= 6 sucesos), hopJ/hopK (Asia= 12 sucesos), HopJ/HopK (Europa= 5 sucesos), hopJ/HopK (amerindia1= 10 sucesos), HopJ/HopK (amerindia2= 9 sucesos) y HopJ/HopK (híbridos= 6 sucesos).

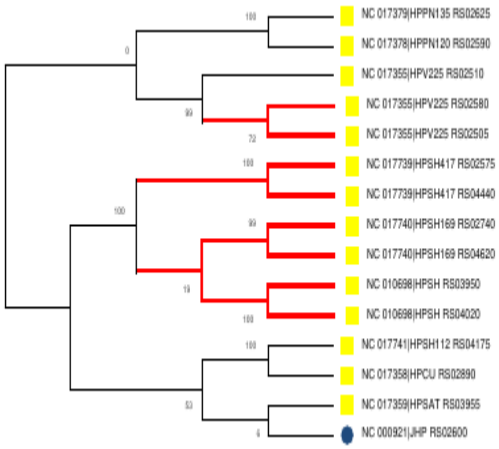


Glycosyltransferase (Asia)

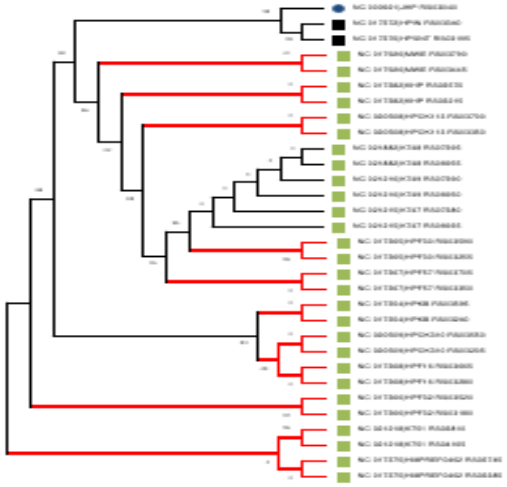


50S ribosome-binding GTPase (híbridos)

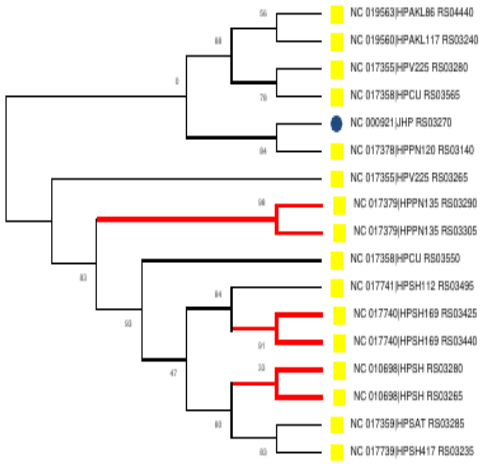




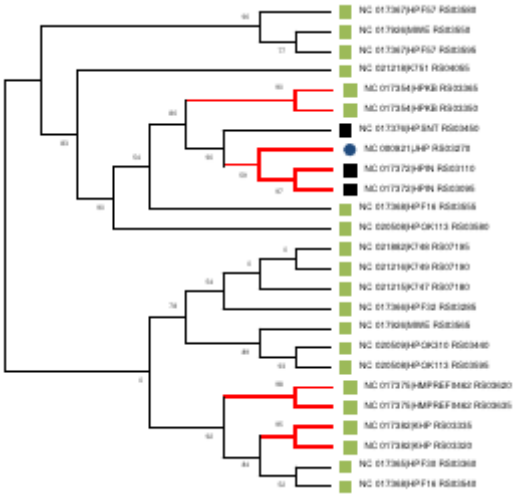
restriction endonuclease subunit S (Asia) (amerindia)



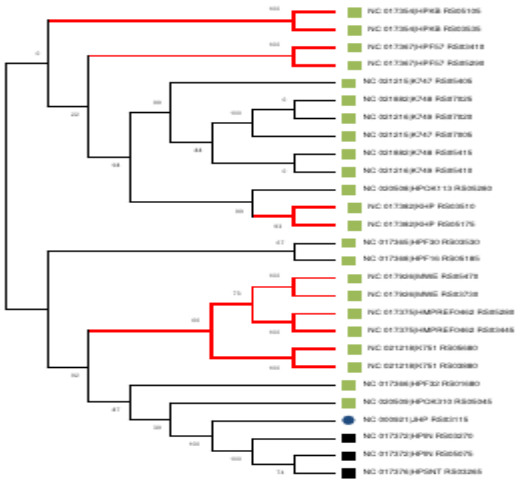
type IV secretion system oncogenic effector cagA



outer inflammatory protein OipA (Asia, EC recurrente)

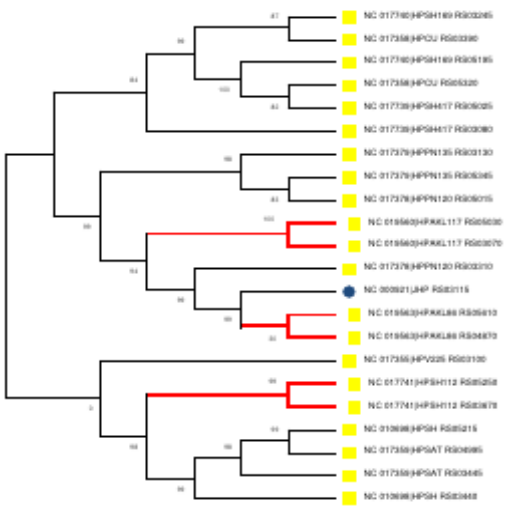


3'-5' exonuclease (Asia)

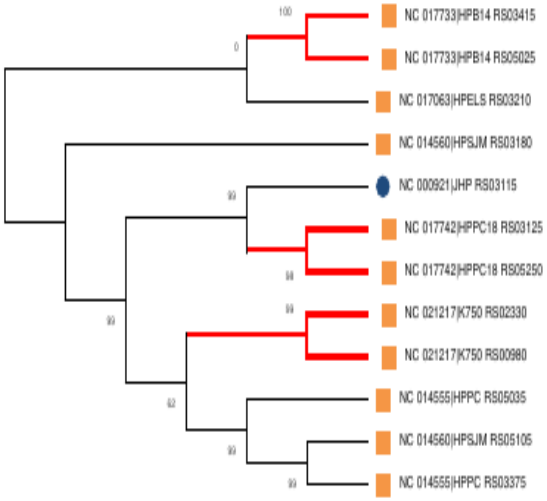


3'-5' exonuclease (amerindia)

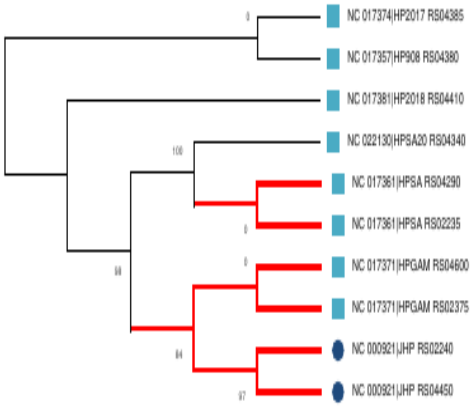
alpha-(1,3)-fucosyltransferase (Asia, EC recurrente)



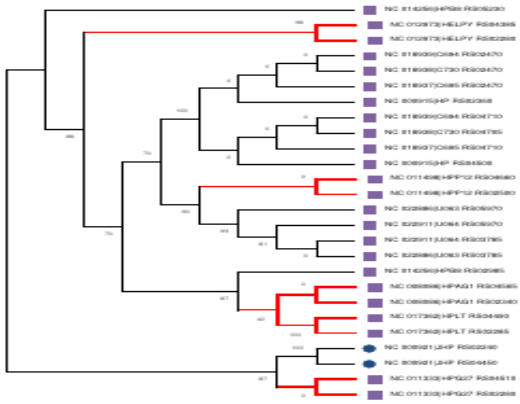
alpha-(1,3)-fucosyltransferase (amerindia)



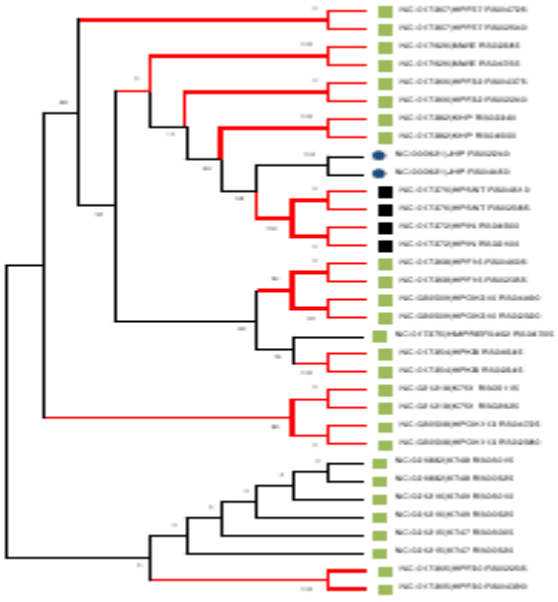
alpha-(1,3)-fucosyltransferase (híbridos)

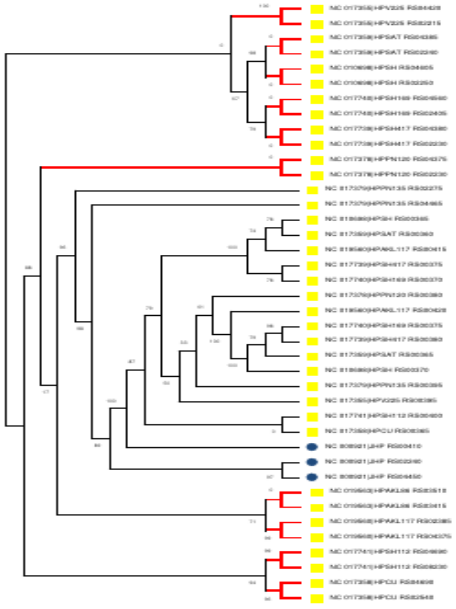


hopJ/HopK (África)

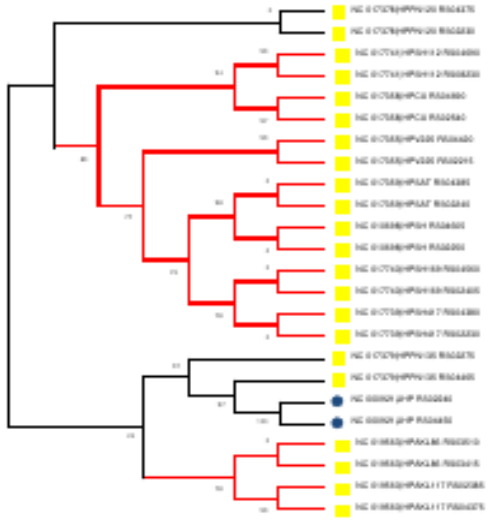


hopJ/HopK (Europa, EC recurrrente)

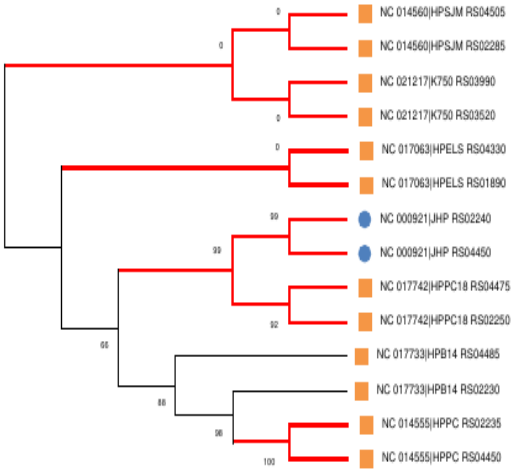




hopJ/HopK (Asia, EC recurrente)



hopJ/HopK (amerindia1, EC recurrente)



hopJ/HopK (amerindia2, EC recurrente)

hopJ/HopK (híbridos, EC recurrente)

Figura 1. Genes duplicados en evolución concertada. En la figura se recogen todas las familias de genes duplicados en evolución concertada. Los clados rojos representan los genes duplicados en EC agrupados de forma monofilética. Así mismo se muestran aquellas familias génicas con EC recurrente

En la **Tabla 1** esta se recogen los genes, el número de cepas que en cada grupo mostraron EC para cada uno de los genes y el porcentaje en las que ocurría. Se considera como EC recurrente aquel gen en el que suceso de conversión ocurre en cinco o más cepas.

Los genes en los que la EC fue detectada fueron los siguientes: **1)** Glycosyltransferase, **2)** 50S ribosome-binding GTPase, **3)** Restriction endonuclease subunit S, **4)** Type IV secretion system oncogenic effectos *cagA*. **5).** Outer membrane protien *OipA*, **6)** 3'-5' exonuclease, **7)** alpha-(1,3)-fucosyltransferase y **8)** *HopJ/HopK* **(Tabla 1).**

De estas ocho familias de genes, seis estuvieron presentes en el grupo de Asia, cuatro en Amerindia y tres en el grupo de los híbridos. Los grupos de cepas con origen en África y Europa, solo presentaron EC para el gen *hopJ/hopK*, siendo este el único que mostró evidencia de EC en los cinco grupos geográficos predefinidos.

Grupo	Familia génica	Número de cepas con genes duplicados en EC	% cepas con EC	Número de cepas analiza-das
Asia	glycosyltransferase	3	17%	17
Híbridos	50S ribosome-binding GTPase	3	38%	5
Asia	restriction endonuclease subunit S	3	17%	17
Amerindia	type IV secretion system oncogenic effector cagA	4	40%	12
Asia	outer inflammatory protein OipA	**11 EC recurrente**	65%	17
Asia	3'-5' exonuclease	3	25%	17
Amerindia	3'-5' exonuclease	4	23%	12
Asia	alpha-(1,3)-fucosyltransferase	**6 EC recurrente**	35%	17
Amerindia	alpha-(1,3)-fucosyltransferase	3	25%	12
Híbridos	alpha-(1,3)-fucosyltransferase	3	43%	5
África	hopJ/HopK	3	43%	7
Europa	hopJ/HopK	**5 EC recurrente**	39%	13
Asia	hopJ/HopK	**12 EC recurrente**	71%	17
Amerindia1	hopJ/Hopk	**10 EC recurrente**	74%	12
Amerindia2	hopJ/HopK	**9 EC recurrente**	75%	12
Híbridos	hopJ/HopK	**6 EC recurrente**	43%	5

Tabla 1. Genes duplicados con patrones de evolución concertada en los cinco grupos. Grupo donde se encontró evidencia de evolución concertada (EC), familia génica, número de cepas que con genes en EC y porcentaje de cepa en EC.

3.3.2 Diversidad y sustituciones dentro de los genes duplicados con evolución concertada

En la en la **Tabla 2** se recogen todos los parámetros de diversidad y sustitución nucleótidos analizados como: porcentaje de identidad (I) entre el gen ancestral y su parálogo, número de cepas analizadas (m), número de sitios segregantes ("S", sitios que muestran diferencias, polimorfismos), numero de haplotipos (H) o variaciones en el ADN, nivel de diversidad de los haplotipos (Hd), número de sustituciones sinónimas (Ks), número de sustituciones no sinónimas (Ka), relación media entre Ks/Ka, tasa de transiciones (Ts), tasa de transversiones (Tv) y relación media de Ts/Tv.

En general los análisis de identidad, mostraron valores altos de conservación, oscilando entre 84% y 98%. Excepto para tres genes que mostraron valores considerablemente mucho más bajos: el gen glycosyltransferase (56%), 50S ribosome-binding GTPase (52%) y restriction endonuclease subunit S (44%).

Así mismo, para aquellos genes cuya presencia fue detectada en al menos dos grupos, no hay correlación directa entre el número de cepas y el número de loci segregantes. Por ejemplo, el gen 3'-5' exonuclease (amerindia) en la que se analizaron 8 cepas mostraron casi el doble de

sitios segregantes (S= 33) que el mismo gen 68_3'-5' exonuclease en la grupo de Asia (S= 12), en la que el número de cepas analizados fue 6. Por el contrario, el gen HopK/HopK en África donde se analizan 6 cepas muestra un número de loci segregantes de S=154, mientras que la zona europea donde se analizan 10 cepas posee S= 147.

Los resultados también mostraron que para todos los genes analizados hay más de dos haplotipos, poniendo de manifiesto que el número de haplotipos se correlaciona con el número de cepas analizadas. Aunque, para determinados genes como *HopJ/HopK* y grupos geográficos como Asia y amerindia el número de haplotipos es prácticamente del 50% en relación al número de cepas, señalando por lo tanto la identidad entre algunas de las cepas de estas regiones. Cosa que confirma el estadístico Hd que mide la diversidad entre haplotipos.

También obtuvimos el número de sustituciones sinónimas (Ks) y no sinónimas (Ka), encontrando que en siete de los ocho genes, la tasa de Ka fue menor que la tasa de Ks **(Tabla 2).** Por lo tanto, la relación Ka/Ks siempre es positiva, variando desde valor 0,09 hasta 1,23.

En general las tasas de Ks y Ka presentan valores bajos excepto en los genes glycosyltransferase, 50S ribosome-binding GTPase y restriction endonuclease subunit S que presentaron valores mayores. Estos genes son los que presentan menor identidad entre cepas

Calculamos tanto la tasa de transiciones (Ts), transversiones (Tv) y la relación media de Ts/Tv. Encontramos que las transiciones de nucleótidos superaron las transversiones en siete de los ocho genes, excepto en el gen “restriction endonuclease subunit S” (Ts=0,11/Tv=0,13).

Familia génica	% I	*m*	S	H	Hd	ks	ka	ka/ks	Ts	Tv	Ts/Tv
glycosyltransferase (Asia)	56,3	6	260	6	1,00	0,54	0,31	0,58	0,28	0,24	1,19
50S ribosome-binding GTPase (híbridos)	51,6	6	204	6	1,00	0,31	0,24	0,77	0,27	0,15	1,82
restriction endonuclease subunit S (Asia)	44,5	6	57	5	0,93	0,16	0,16	0,98	0,11	0,13	0,85
type IV secretion system oncogenic effector cagA (Asia)	87,2	8	32	8	1,00	0,02	0,03	1,23	0,04	0,01	2,57
outer inflammatory protein OipA (Asia)	98,2	22	58	15	0,97	0,06	0,01	0,15	0,03	0,01	3,46
3'-5' exonuclease (Asia)	98,2	6	12	5	0,93	0,05	0,01	0,12	0,02	0,00	8,06
3'-5' exonuclease (amerindia)	96	8	33	7	0,96	0,14	0,01	0,09	0,04	0,01	5,67
alpha-(1,3)-fucosyltransferase (Asia)	88,9	12	162	12	1,00	0,09	0,06	0,71	0,09	0,04	2,48
alpha-(1,3)-fucosyltransferase (amerindia)	84,4	6	179	6	1,00	0,11	0,08	0,71	0,12	0,04	3,03
alpha-(1,3)-fucosyltransferase (híbridos)	84	6	162	6	1,00	0,25	0,05	0,19	0,11	0,04	2,85
hopJ/HopK (África)	91,2	6	154	4	0,86	0,26	0,06	0,23	0,12	0,06	1,84
hopJ/HopK (Europa)	93,6	10	147	7	0,93	0,19	0,04	0,19	0,09	0,04	2,44
hopJ/HopK (Asia)	93,4	24	194	16	0,97	0,17	0,04	0,27	0,09	0,04	2,42
hopJ/Hopk (amerindia)	96	20	121	13	0,96	0,10	0,02	0,25	0,05	0,02	2,12
hopJ/HopK (amerindia)	96,1	18	118	12	0,96	0,10	0,02	0,25	0,05	0,02	2,11
hopJ/HopK (híbridos)	91,4	6	175	5	0,93	0,26	0,05	0,21	0,13	0,06	2,20

Tabla 2. Estadísticos de diversidad para los genes duplicados con evolución concertada. **1) %I,** porcentaje de identidad entre las secuencias, **2)** m, número de cepas, **3)** S, número de sitios segregantes, **4)** H, número de haplotipos, **5)** Hd, diversidad de haplotipos, **6)** Ks, sustituciones sinónimas, **7)** Ka, sustituciones no sinónimas, **8)** relación media Ka/Ks, **9)** Ts, tasa de transiciones, **10)** Tv, tasa de transversiones y **11)** Ts/Tv relación media de transiciones/transversiones.

3.3.3 Localización de la variación intragénica dentro de los genes duplicados con evolución concertada.

A partir de los alineamientos obtenidos de las familias de genes mediante el programa Mafft **(Katoh *et al.*, 2013)** implementado a través del programa Geneious v6.18 **(Kearse *et al.*, 2012)** se obtuvieron las identidades de las regiones internas de las secuencias. En la **tabla 3** se muestra que la identidad entre regiones de los distintos genes es variable, existiendo genes con valores altos a lo largo de todo el gen mientras que otros existen diferencias de identidad por regiones. Se puede observar que los genes glycosyltransferase (Asia) y 50S ribosome-binding GTPase (híbridos) las identidades de sus regiones variaron entre el 41,5% al 63%. El gen restriction endonuclease subunit S (Asia) mostró una identidad más alta en su región 3' (95%). El gen type IV secretion system oncogenic effector *cagA* mostró en la región 5' el mayor grado de conservación. El gen *oipA* en general está altamente conservado a lo largo de la secuencia del gen (94%-95%).

En ambos casos, los dos genes 3'-5' exonuclease, mostraron su mayor grado de conservación a partir de la región media hasta la región 3' (a partir del 96%). Los tres genes de la familia alpha-(1,3)-fucosyltransferase coincidieron en tener la región media como las más conservada (a partir del 93%). En la familia de genes *HopJ/hopK* los grupos de África, Europa y Asia mostraron a partir de su región media hasta la región 3' el mayor porcentaje de identidad (a partir del 93%). Mientras que el grupo de amerindia presentó un alto grado de conservación a lo largo de la secuencia del gen (96%). Por último, los híbridos mostraron estar conservados a partir de la región media hasta la región 3' (92%).

Gen	Región 5'	Región media	Región 3'
glycosyltransferase (Asia)	44.5% (1-260)	60.4% (261-653)	57.8% (654-940)
50S ribosome-binding GTPase (híbridos)	41.5% (1-346)	63% (347-565)	54% (566-856)
restriction endonuclease subunit S (Asia)	1.1% (1-704)	50% (705-1.194)	95% (706-1.339)
type IV secretion system oncogenic effector cagA	97% (1-575)	85% (576-3.046)	92% (3.047-3.747)
outer inflammatory protein OipA (Asia)	94% (1-94)	Región media hasta 3' 99% (95-915)	
3'-5' exonuclease (Asia)	1-414 (Deleciones)	Región 98% (415-813)	
3'-5' exonuclease (amerindia)	1-414 (Deleciones)	96% (414-813)	
alpha-(1,3)-fucosyltransferase (Asia)	73% (1-161)	94% (162-1.116)	74% (1.117-1.412)
alpha-(1,3)-fucosyltransferase (amerindia)	47% (1-159)	94% (160-1.1668)	79% (1.169-1.385)
alpha-(1,3)-fucosyltransferase (híbridos)	45% (1-153)	93% (154-1.193)	83% (1.192-1.437)
hopJ/HopK (África)	68% (1-76)	Región media hasta 3' 93% (77-1.122)	
hopJ/HopK (Europa)	75% (1-50)	Región media hasta 3' 95% (51-1.113)	
hopJ/HopK (Asia)	73% (1-82)	Región media hasta 3' 95% (83-1.117)	
hopJ/Hopk (amerindia)	Toda la secuencia 96% (secuencia altamente conservada)		
hopJ/HopK (amerindia)	Toda la secuencia 96% (secuencia altamente conservada)		
hopJ/HopK (híbridos)	74% (1-50)	Región media hasta 3' 92% (51-1.116)	

Tabla 3. Resumen de las regiones variables como regiones conservadas detectadas en los duplicados con evolución concertada.

3.3.4 Recombinación de los genes duplicados en evolución concertada

Los análisis comparativos de secuencias a través de los programas RDP5 **(Martin *et al.*, 2021)** y DnaSP **(Rozas *et al.*, 2017)** proporcionaron evidencia de recombinación en todas las familias de genes. Estos programas localizaron los sitios de recombinación dentro de las secuencias genómica de cada gen. En la **figura** 2, se muestra mediante una escala de colores los porcentajes de identidad (grado de conservación) para cada región de cada gen, además se representa la recombinación (R) de las regiones del gen donde fue detectada la recombinación

las cuales coinciden con las regiones de mayor porcentaje de identidad de los genes. Así, para los genes glycosyltransferase y alpha-(1,3)-fucosyltransferase (Asia), los puntos calientes recombinación se localizaron en la región intermedia del alineamiento. Para los genes 50S ribosome-binding GTPase, outer inflammatory protein *OipA*, alpha-(1,3)-fucosyltransferase (amerindia), *hopJ/HopK* (Europa), *hopJ/HopK* (Asia), *hopJ/Hopk* (amerindia1), *hopJ/HopK* (amerindai2) y *hopJ/HopK* (híbridos) los sitios de recombinación se encontraron distribuidos a lo largo del alineamiento. Para los genes restriction endonuclease subunit S, 3'-5' exonuclease (Asia) y 3'-5' exonuclease (amerindia), estos puntos fueron localizados cercanos a la región 3'. En el gen type IV secretion system oncogenic effector *cagA*, mostró sitios de recombinación en la región 5'. En el gen alpha-(1,3)-fucosyltransferase (híbridos), los puntos de recombinación se localizaron tanto en 5' como en 3'. Y por último el gen *hopJ/HopK* (África), mostró recombinación en la región 5' y en la región intermedia **(Figura 2)**

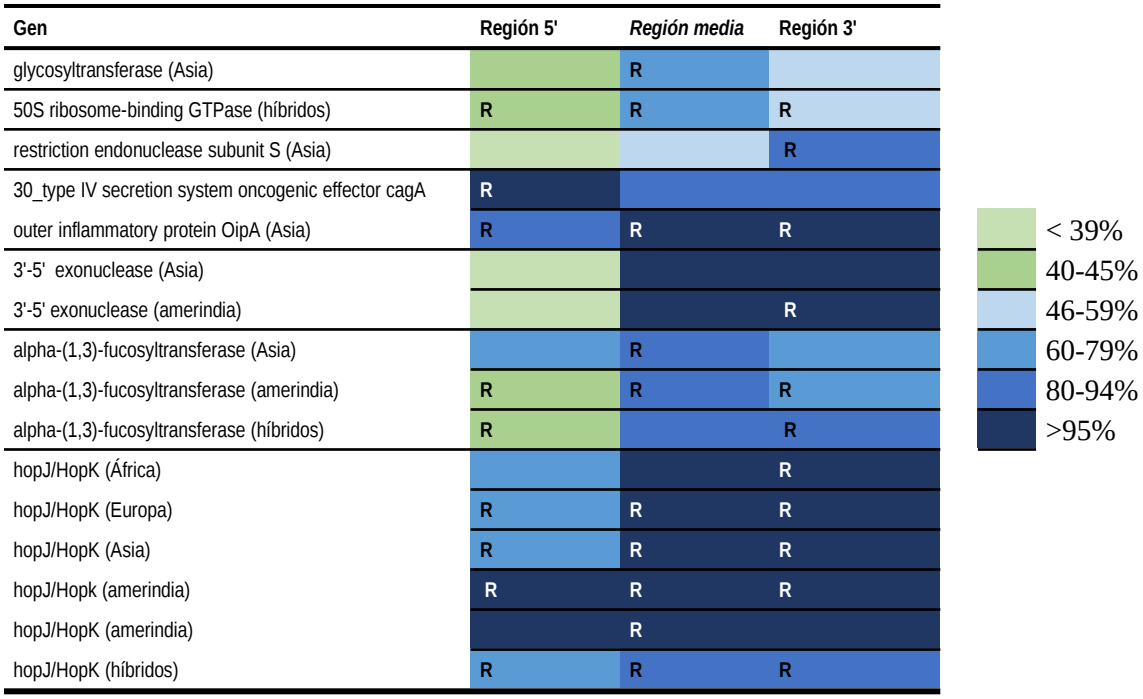


Figura 2. Porcentajes de identidad de las regiones internas para cada gen representado en escala de colores y regiones del gen donde se detectó recombinación. Los genes se analizaron en tres regiones: región 5', región media y región 3'. La escala de color a la derecha indica el porcentaje de identidad para las regiones. "R" representa la detección de recombinación en esa región del gen.

También, utilizando el programa RDP5 mediante la prueba estadística GENECONV **(Sawyer, 1989)**. Se detectó la conversión de genes. Este análisis identifico los mismos genes con eventos de EC que con los programas anteriormente mencionados. Este programa proporciona una probabilidad con significación estadística de < 0,05.

3.3.5 Desviaciones del modelo de neutralidad de evolución molecular

Los genes duplicados con EC en este estudio fueron sometidos a la prueba de neutralidad de Tajima's **(Véase Material y métodos)**. Los valores promedios para los parámetros *π y Θ* son variables entre genes y grupos. Los tres genes con porcentajes de identidad más bajo (glycosyltransferase, 50S ribosome-binding GTPase y restriction endonuclease subunit), los parametros de sustitución mayores que en el resto, seguidos por los genes alpha-(1,3)-fucosyltransferase y hopJ/HopK. El valor de *D* obtenido varía desde 1,936 en Glycosyltransferase hasta 0,138 en el gen outer inflammatory protein *OipA*, mostrando todos ellos valores positivos, lo que nos indicaría que se encuentran bajo la acción de la selección equilibrada o balanceada **(Tabla 4)**.

	Tajima's neutrality					
	S	**Ps**	**Θ**	**π**	**D**	**Interpretation**
glycosyltransferase (Asia)	260	0,500	0,218	0,284	1,936	Balancing selection
50S ribosome-binding GTPase	204	0,466	0,204	0,261	1,801	Balancing selection

(híbridos)						
restriction endonuclease subunit S (Asia)	57	0,219	0,096	0,117	1,466	Balancing selection
type IV secretion system oncogenic effector cagA (Asia)	32	0,055	0,021	0,022	0,360	Balancing selection
outer inflammatory protein OipA (Asia)	58	0,063	0,017	0,017	0,138	Balancing selection
3'-5' exonuclease (Asia)	12	0,030	0,013	0,015	1,170	Balancing selection
3'-5' exonuclease (amerindia)	33	0,082	0,031	0,040	1,397	Balancing selection
alpha-(1,3)-fucosyltransferase (Asia)	162	0,142	0,047	0,054	0,733	Balancing selection
alpha-(1,3)-fucosyltransferase (amerindia)	179	0,154	0,067	0,074	0,623	Balancing selection
alpha-(1,3)-fucosyltransferase (híbridos)	162	0,138	0,060	0,073	1,326	Balancing selection
hopJ/HopK (África)	154	0,140	0,061	0,075	1,482	Balancing selection i
hopJ/HopK (Europa)	147	0,134	0,047	0,056	0,991	Balancing selection
hopJ/HopK (Asia)	194	0,179	0,048	0,054	0,557	Balancing selection
hopJ/Hopk (amerindia)	121	0,112	0,031	0,035	0,430	Balancing selection
hopJ/HopK (amerindia)	118	0,109	0,031	0,034	0,384	Balancing selection
hopJ/HopK (híbridos)	175	0,159	0,069	0,080	1,012	Balancing selection

Tabla 4. Parámetros para el análisis de la acción de la selección natural en los genes con evolución concertada. S Número total de sitios; Ps, porcentaje de número de sitios segregantes por secuencia, Θ número total de sitios segregantes en un grupo de secuencias; π, diversidad de nucleótidos y *D* es el estadístico de Tajima.

3.3.6 Análisis de selección de los sitios variables en los genes con evolución concertada

El programa ETE3 **(Huerta-Cepas & Bork, 2016**) que implementa CODEML **(véase Material y métodos)** identifica el grado de selección y posiciones de aminoácidos donde esta ocurre. En la tabla # se recogen los resultados obtenidos. Según el análisis los genes donde el efecto de la selección es mayor son: type IV secretion system oncogenic effector cagA y 3'-5' exonuclease (Asia), 3'-5' exonuclease (amerindia).

3.4. Discusión

3.4.1 Evolución concertada y análisis de diversidad

Estudios previos ya han demostrado la existencia de EC en genes de *H. pylori* **(Pride & Blaser, 2002; Oleastro & Ménard, 2013)**. Específicamente se ha demostrado en la familia de proteínas de membrana (OMP) con función de adhesión, *BabA* y su parálogo *BabB* y en una familia más pequeña de proteínas de membrana denominada *Hom*, en cuyo caso se ha observado EC entre *homA* y su parálogo *homB*. En estudio no encontramos evidencia de EC en estos genes.

También **Kawai et al. (2011)**, propusieron que las similitudes encontradas dentro de una cepa para el gen *HopJ/HopK* podría deberse a EC, sin demostración alguna de esta.

Tras el análisis de 53 genomas de *H. pylori* mediante el programa iSeeCe **(Wang & Chen, 2018)** se han obtenido agrupaciones monofiléticas a nivel intragenómico y conservando la sintenia con sus ortólogos evidencias de EC en 8 genes **(Figura 2)**. Este fenómeno no es uniforme en todos los grupos ya que algunos genes muestran EC en una sola región mientras que otros pueden estar representados en varios grupos y uno *HopJ/HopK* en todas las zonas **(Tabla 1)** Esta observación confirmaría la sugerencia hecha por **Kawai *et al.* (2011)** de EC en este gen**.** Tres genes, dos únicamente presentes en Asia, alpha(1,3)-fucosyltransferase y *oipA* y el gen *HopJ/HopK*, muestran EC recurrente es decir que este proceso ocurre en cinco o más cepas dentro de la misma región. Esto nos indicaría que estos genes podrían estar bajo un proceso de evolución genética por EC podría no aleatorio, si no que estaría favorecido por la selección **(Kondrashov *et al.,* 2007, Lathe *et al.,* 2001).** Adicionalmente mediante la prueba estadística GENECONV del programa RDP5 confirmamos los eventos de EC para los mismo genes Este programa proporciona una probabilidad con significación estadística de < 0,05.

Del análisis de variabilidad de estos genes **(Tabla 2)** se observan dos grupos diferentes, Uno lo formarían los genes con una alta identidad entre genes e incluso entre grupos diferentes y por otro lado estarían los tres genes con una identidad mucho menor y en los que la EC solo ocurre en una sola región geográfica.

En cuanto al número de sitios segregantes podemos ver que también existen diferencias, pero esta es entre genes. Así los genes restriction endonuclease subunit S (Asia), type IV secretion system oncogenic effector cagA (Asia), outer inflammatory protein OipA (Asia), 3'-5' exonuclease (Asia) y 3'-5' exonuclease (amerindia) presentan un menor número de loci segregantes que van desde 12 hasta 58 mientras que el resto de genes presentan un número superior a 100 no siendo esta variación dependiente del número de cepas analizadas. En este sentido podemos destacar el dato para el gen *OipA* en el que el número de cepas analizadas es de 22 y sin embargo el número de loci segregantes es de 58.

En cuanto al número de haplotipos y diversidad de estos, podemos comprobar que existe una correlación directa entre el número de cepas y el de haplotipos con valores de Hd de 1 o próximos a 1 **(Tabla 1)**

Las relaciones K a / K s de todos los genes presentaron valores positivos variando desde 0,08 hasta 1,23. Asimismo las transiciones de nucleótidos excedieron a las transversiones, con relaciones que van desde 0,85 hasta 8,06. Estos dos datos nos pueden indicar que todos los genes podrían estar bajo un proceso de selección.

Cuando se divide cada uno de los genes por segmentos en cuanto a su secuencia **Tabla 2** podemos comprobar que los valores de identidad son diferentes entre ellas, existiendo zonas con valores superiores al 90% y otros con valores menores. Estos resultados sugieren que la homogenización puede producirse en segmentos concretos del gen y, por lo tanto, no resultar en una homogenización completa. En este sentido, el análisis de recombinación intragénica **(Figura 2)** entre parálogos nos permite comprobar que esta se realiza en las zonas de mayor identidad de los genes, lo que señalaría una relación directa entre recombinación y conversión génica favoreciendo la homogenización de estas zonas.

3.4.2 Análisis de Selección en los genes bajo evolución concertada

Dos tipos de análisis detectan selección balanceada en los genes que están bajo EC (**Tabla 4).** El primer análisis mediante la prueba de Tajima ‘s, detecta desviaciones del modelo de neutralidad en todo el gen. Mientras que el segundo utilizando el programa ETE3 permite detectar señales de selección en posiciones concretas del gen. Esta última observación supondría que dentro del mismo gen habría zonas que están evolucionando bajo EC mientras que otras estarían experimentando selección. Esto nos proporcionaría evidencias de que, los diferentes segmentos de la proteína pudieran codificar distintas dominios en cuanto a la función y a la estructura de la proteína la misma **(Oleastro *et al.,* 2009).**

Capitulo 4 Reorganizaciones genómicas en *H. pylori*

4.1 Introducción

Hoy en día se reconoce cada vez más que, además de las variaciones de un solo nucleótido, otros tipos de variaciones que incluyen grandes reordenamientos genómicos no son infrecuentes en los genomas bacterianos **(Darling *et al.,* 2008; Sun *et al.,* 2012; Gonzalez Torres at al 2019)**. Las variaciones estructurales abarcan una parte bastante grande de ADN y pueden surgir como resultado de diferentes mecanismos celulares, como la recombinación, la replicación y la reparación del ADN **(Hastings *et al.,* 2009b)**. Estos procesos de reordenamiento genómico introducen variabilidad en el número de copias de genes, la posición, la orientación e incluso las combinaciones de estos eventos **(Freeman *et al.,* 2006).**

A medida que evolucionan los genomas, las fuerzas mutacionales y las presiones selectivas introducen reordenamientos mediante procesos de inversión, transposición y duplicación / pérdida. La transferencia lateral puede introducir contenido de genes novedosos o, en el caso de la recombinación homóloga, introducir cambios más sutiles, como la sustitución alélica. En el caso de la comparación intraespecífica bacteriana, los genes que tienen patrones filogenéticos similares, es decir, que coexisten exclusivamente en un conjunto particular de cepas, a menudo se encuentran cerca unos de otros y forman una isla genómica.

En *H. pylori* la recombinación es frecuente durante la colonización entre múltiples cepas debido a la transformación del ADN, lo que da como resultados variantes dentro de los huéspedes individuales que difieren en el contenido de secuencia y la composición genómica. Como resultado de la recombinación frecuente, casi todos los aislamientos de *H. pylori* de hospedadores no relacionados poseen secuencias únicas, a diferencia de otras bacterias, donde se encuentran secuencias idénticas de genes centrales de mantenimiento en múltiples aislamientos. *H. pylori* difieren entre huéspedes individuales, pero se encuentran diferencias aún mayores cuando las secuencias se comparan con aislamientos de diferentes continentes, posiblemente reflejando la deriva genética durante el aislamiento geográfico, así como la adaptación a las diferencias genéticas entre diferentes grupos étnicos de humanos **(Suerbaum *et al.,* 1998, Suerbaum *et al.,* 2007)**

A través de la comparación del genoma, esperamos identificar primero las diferencias entre los organismos a nivel del genoma y luego inferir la importancia biológica de las diferencias y similitudes entre los organismos relacionados.

4.2 Material y métodos

Todas las secuencias del genoma se descargaron de la base de datos NCBI. Se realizaron alineaciones del genoma completo de las 53 cepas de prueba de *H. pylori* utilizando Mauve versión. Con los parámetros predeterminados de Mauve progressive **(Darlin *et al.,* 2004; Darlin *et al.,* 2010).** Este método utiliza alineaciones por pares o múltiples de secuencias conservadas para genomas completos. Se realizaron alineaciones locales para identificar múltiples coincidencias únicas máximas (multi-MUM), que posteriormente se utilizaron para calcular la construcción de un árbol guía. Luego se utilizaron subconjuntos de múltiples MUM como anclajes y se dividieron en bloques colineales locales. Cada bloque es una región de ADN homóloga de múltiples MUM, que carece de reordenamientos de secuencia y es compartida por dos o más genomas bajo análisis. El alineamiento de secuencias permite identificar el número de bloques colineales comunes utilizando la longitud de las regiones conservadas totales y la identidad de nucleótidos total entre las secuencias cromosómicas para cada par de cepas. Además, alineamos los nueveGenomas de *H. pylori* con el software MUMmer 3.0.

El alineador de Mauve filtra y clasifica las coincidencias identificadas internamente en bloques colineales localmente (LCB). Cada LCB representa una región de secuencia homóloga sin reordenamiento entre los genomas de entrada. Cada LCB debe separarse del siguiente mediante reordenamiento en al menos un genoma **(Darling *et al.,* 2004)**. Los límites de contig (bordes) representan bordes LCB potencialmente artificiales. Por lo tanto, encontrar el orden contig que minimiza el número de LCB causados por bordes contig es equivalente a encontrar un orden contig probable.

Usando los LCB de alineación Mauve, el proceso de reordenamiento ocurre en tres pasos: colocar contigs sin conflicto aparente en la información de pedido, ubicar contigs con información conflictiva en posiciones intermedias de anclaje y, finalmente, hacer coincidir los extremos LCB que se extienden a los límites de contig. Cada paso ocurre en un tiempo máximo de O (n 2), donde n es el

número de LCB, más el tiempo requerido para la alineación **(Darling *et al.,* 2004).** Mauve asume que los contigs están en el orden correcto al filtrar coincidencias, por lo que a medida que se optimiza el orden, los resultados de la alineación cambian. Por lo tanto, los resultados se refinan mediante la alineación iterativa hasta que no sea posible realizar más pedidos.

MCM genera una serie de alineaciones malva, cada una de las cuales representa una iteración del reordenamiento. Además de la salida estándar de Mauve, el proceso de reorden genera un archivo FastA que contiene el nuevo orden y orientación, así como una lista de contigs ordenados que incluyen el nombre y la ubicación de las coordenadas. La visualización estándar de Mauve se puede aplicar de formas novedosas para analizar el orden de los contig. Por ejemplo, lo hemos utilizado para identificar posibles ensamblajes incorrectos en contigs y para evaluar la presencia o ausencia de genes divididos por límites de contig o por reordenamientos. Si se crean FastAs que representan el orden producido por otros programas, Mauve también se puede utilizar para comparar resultados, como en Figura complementaria. Además, las anotaciones de la entrada en formato GenBank se pueden ver, incluso una vez que se reordenan.

4.3 Resultados

La estructura de los 53 genomas de *H. pylori* se ha analizado utilizando la cepa J99 de origen africano como cepa de referencia en todos los casos. Hemos llevado a cabo un análisis por separado para cada región geográfica para estudiar tanto el número de reordenaciones como su localización genómica y si existe algún patrón en la distribución geográfica.

Según las secuencias de ADN homólogas entre los genomas, cada genoma se ha dividido en bloques colineales. El rango de variación del número de bloques sinténicos es variable entre cepas del as misma región y entre regiones

El alineamiento conjunto de todos los genomas mostró que la sintenia del genoma ha sido interrumpida principalmente por inversiones, ademas de algunas deleciones y translocaciones.

En la **Figura 1** podemos observar que las cepas incluidas en la región africana presentan un número reducido de reordenaciones cromosómicas siendo las dos cepas sudafricanas (las dos últimas de la grafica) las que presentan un patrón algo diferente al resto en cuanto a l número de reordenaciones

genómicas (siete), aunque todas ellas de tamaño pequeño. También se pueden observar fenómenos de transposición de pequeños bloques en estas mismas cepas.

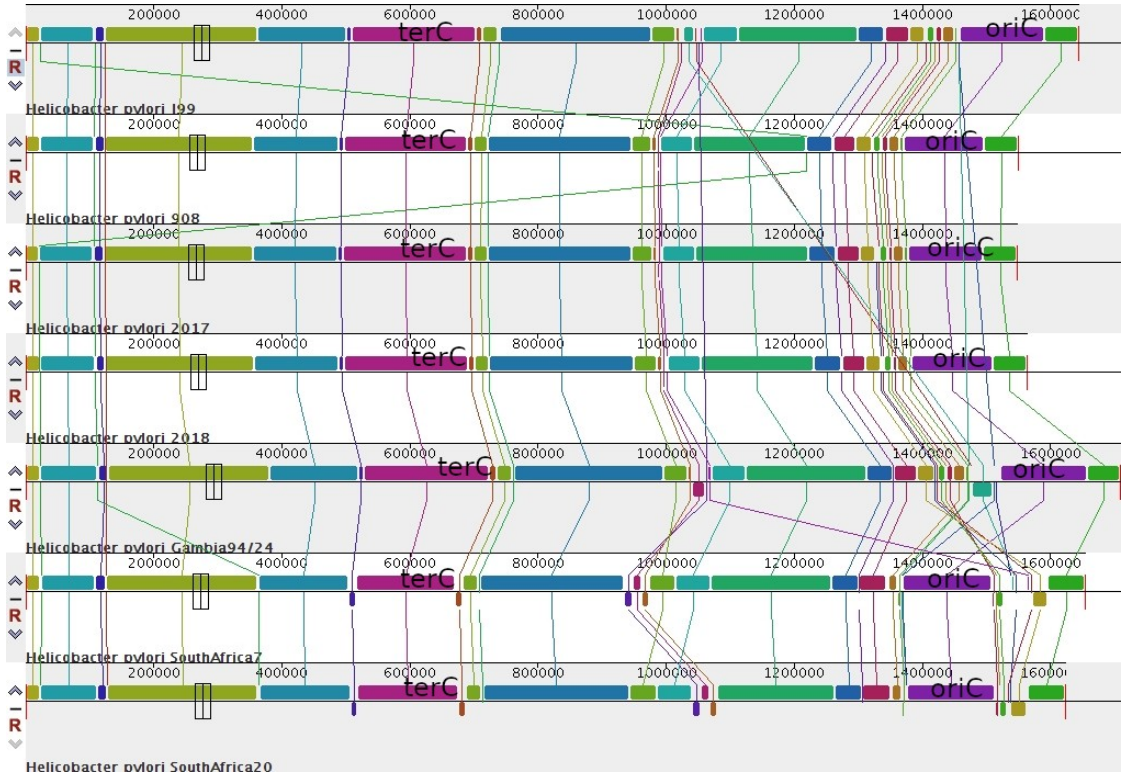


Figura 1. Alineamiento Mauve para la región geográfica de África. Se indican el inicio (oriC) y termino (terC) de la replicación.

En cuanto a las poblaciones europeas en la **Figura 2** podemos ver que existen varios subgrupos de cepas según su patrón de reorganizaciones genómicas. Así, un primer grupo lo constituirían tres cepas que presentan la mayor similitud con la cepa de referencia. Un segundo grupo estaría formado por cuatro cepas con un patrón común de inversiones relativamente grandes además de otras pequeñas adicionales. Un tercer grupo estaría formado por las dos cepas aisladas en Australia, pero de origen europeo y que presentan un patrón idéntico de inversiones. Por último se ha incluido en este grupo la cepa B8 que presenta grandes inversiones. También en el grupo europeo se observa r la presencia de fenómenos de transposición de pequeños bloques.

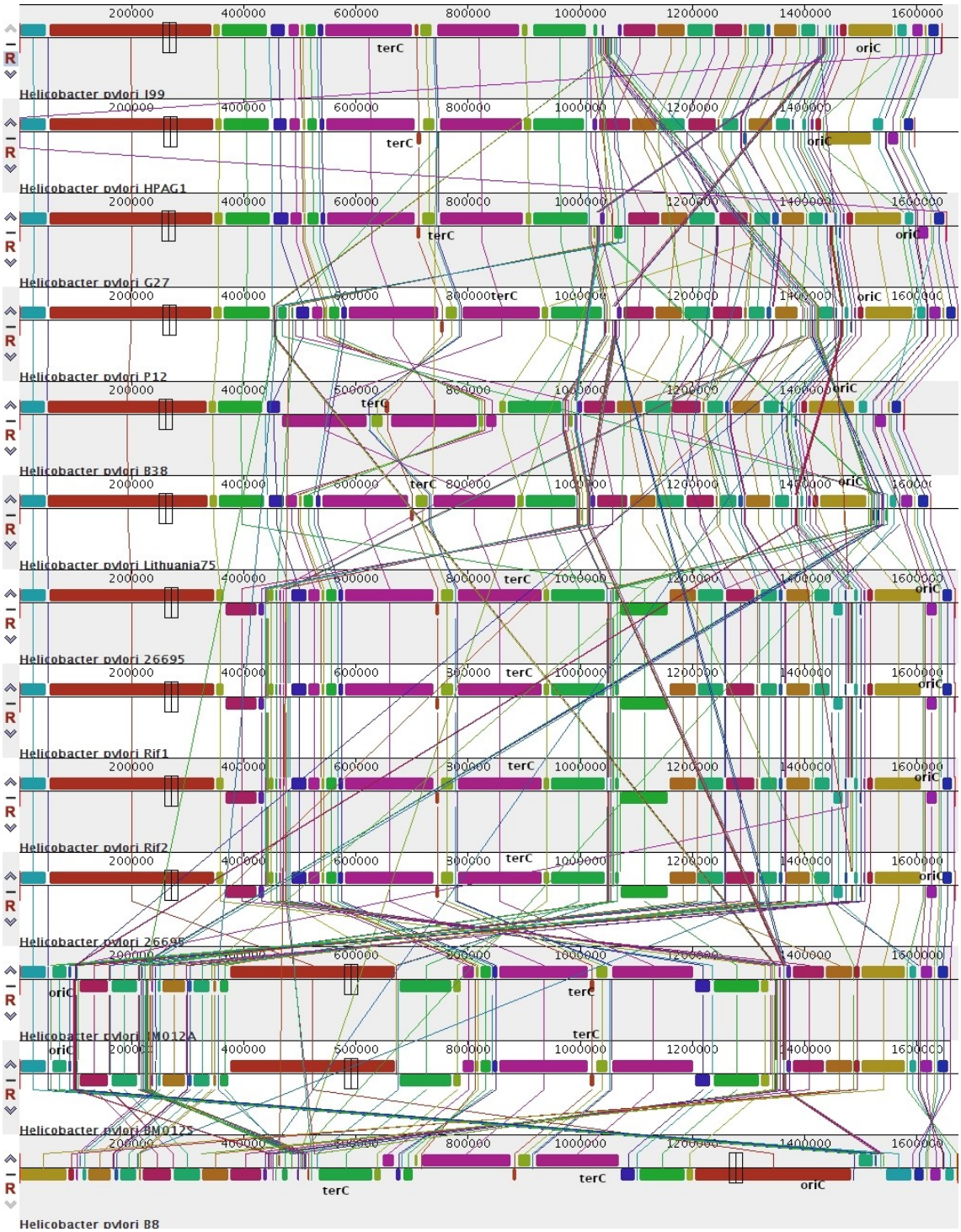


Figura 2. Alineamiento Mauve para la región geográfica de Europa. Se indican el inicio (oriC) y termino (terC) de la replicación

En la **Figura 3** podemos ver las cepas de origen asiático. En este conjunto de cepas también se pueden observar varias agrupaciones en función de la presencia o ausencia de mayoritariamente

inversiones cromosómicas. El primer agrupamiento lo formarían tres cepas que presentan un gran parecido a la cepa africana de referencia con una o dos pequeñas inversiones. El grupo más importante lo componen nueve cepas todas ellas con una gran inversión común que afecta a la región central del cromosoma y algunas de ellas parecen haber sufrido pequeñas inversiones posteriores a la primera por lo que estas presentan bloques en su posición original. En otro grupo formado por tres cepas afectado por una gran inversión que afecta a más del 50% del genoma y que incluye otras pequeñas reordenaciones posteriores a la de mayor tamaño.

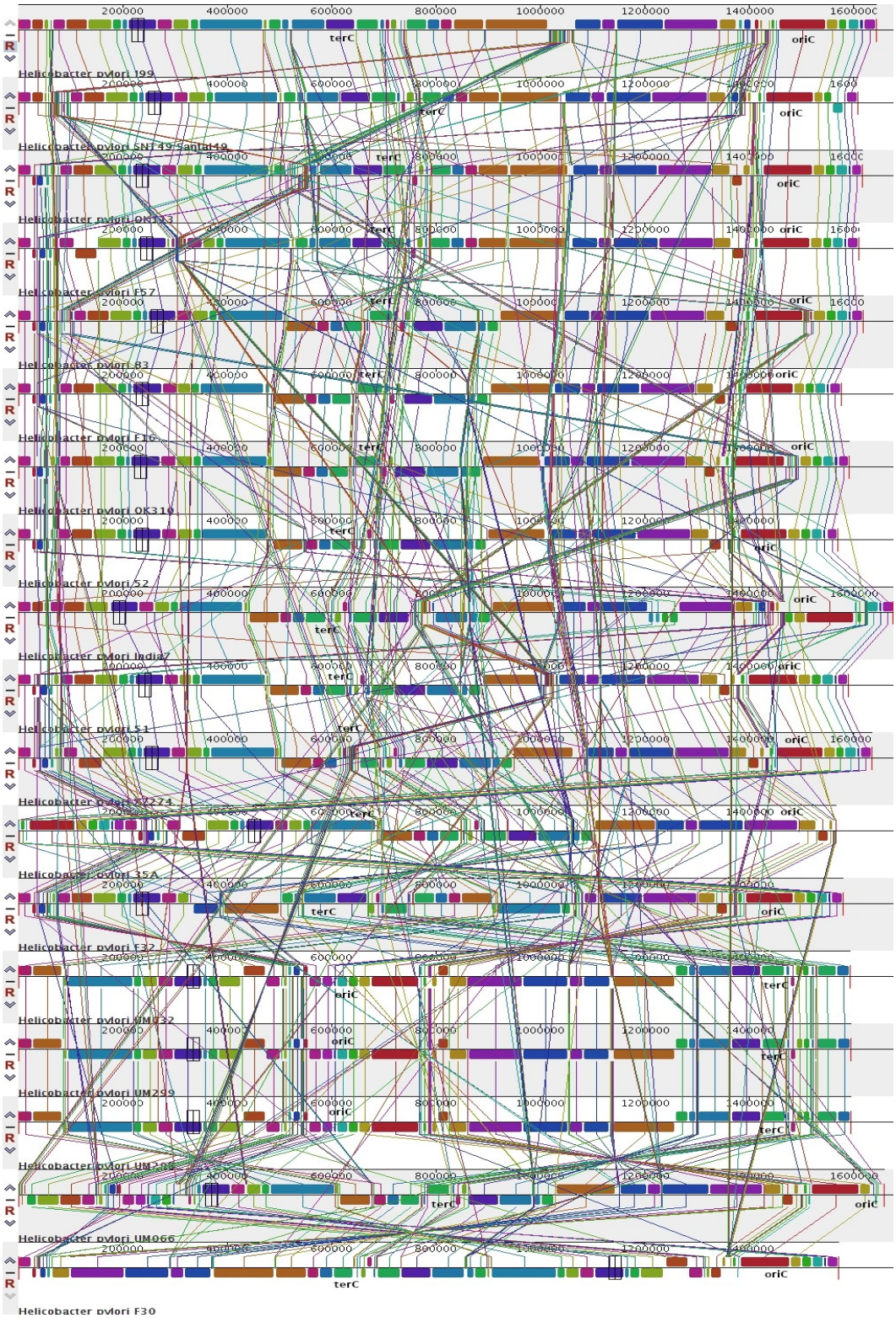


Figura 3. Alineamiento Mauve para la región geográfica de Asia. Se indican el inicio (oriC) y termino (terC) de la replicación

En la **Figura 4** se pueden observar las reordenaciones ocurridas en las cepas de origen amerindio. En general podemos observar que la mayoría de ellas son muy parecidas a la cepa de referencia excepto tres de ellas que presentan una inversión de tamaño intermedio común. La cepa Aklavik86 presenta numerosas reordenaciones, pero pequeñas en cuanto a su longitud.

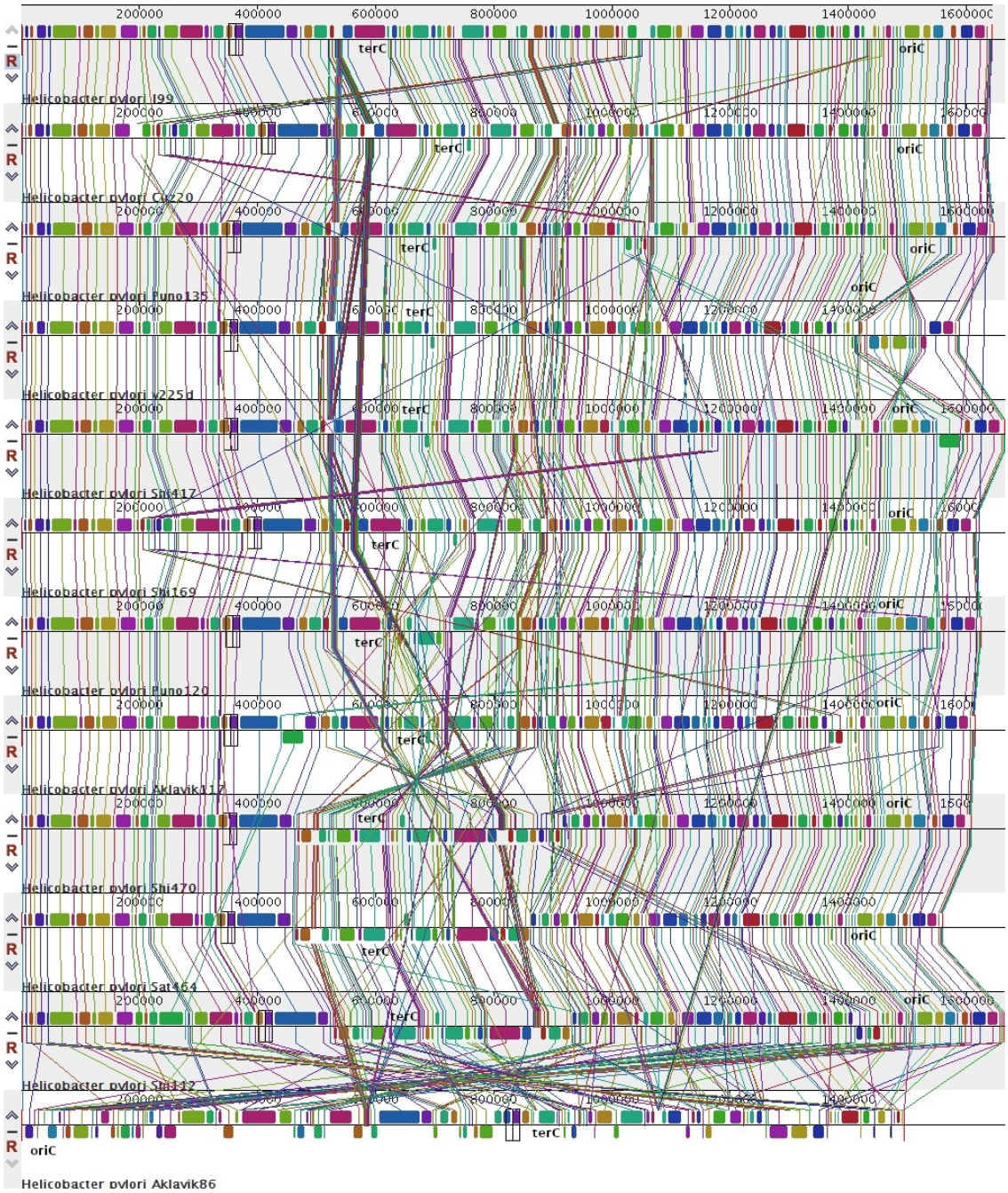


Figura 4. Alineamiento Mauve para la región geográfica de amerindia. Se indican el inicio (oriC) y termino (terC) de la replicación

Por último, para terminar con este análisis de sintenia hemos comparado las cepas cuyo origen se califica como híbrido **Figura 5**. En ella podemos ver que también dentro de este grupo se presenta cierta diversidad en cuanto a la presencia de reordenaciones cromosómicas. Las cepas SMJ180, PeCan18 y PeCan4 presentan una gran similitud con la cepa referencia. La cepa SMJ180 incluye algunas pequeñas inversiones una de ellas cercana al origen de replicación y una translocación que incluye varios grupos sinténicos pequeños. La cepa PeCan4 presenta una inversión de tamaño pequeño cercano al sitio de replicación y la cepa PeCan18 muestra una inversión de tamaño mediano tambien cercano al lugar de terminación de la replicación. HUPB14 presenta una inversión grande cercana al sitio de terminación y una pequeña translocación cercana al lugar de origen de la replicación. Las otras dos cepas ELS37 y UM037 presentan una relación curiosa. En la primera ha ocurrido una gran inversión que incluye ter aunque se puede comprobar posiblemente posteriormente han ocurrido nuevas reordenaciones internas que han cambiado la orientación de una serie de grupos sinténicos pequeños. En la cepa UM37 se puede ver la presencia de esta misma inversión, pero posteriormente ha sufrido tanto eventos de translocación como de nuevas inversiones dentro de la

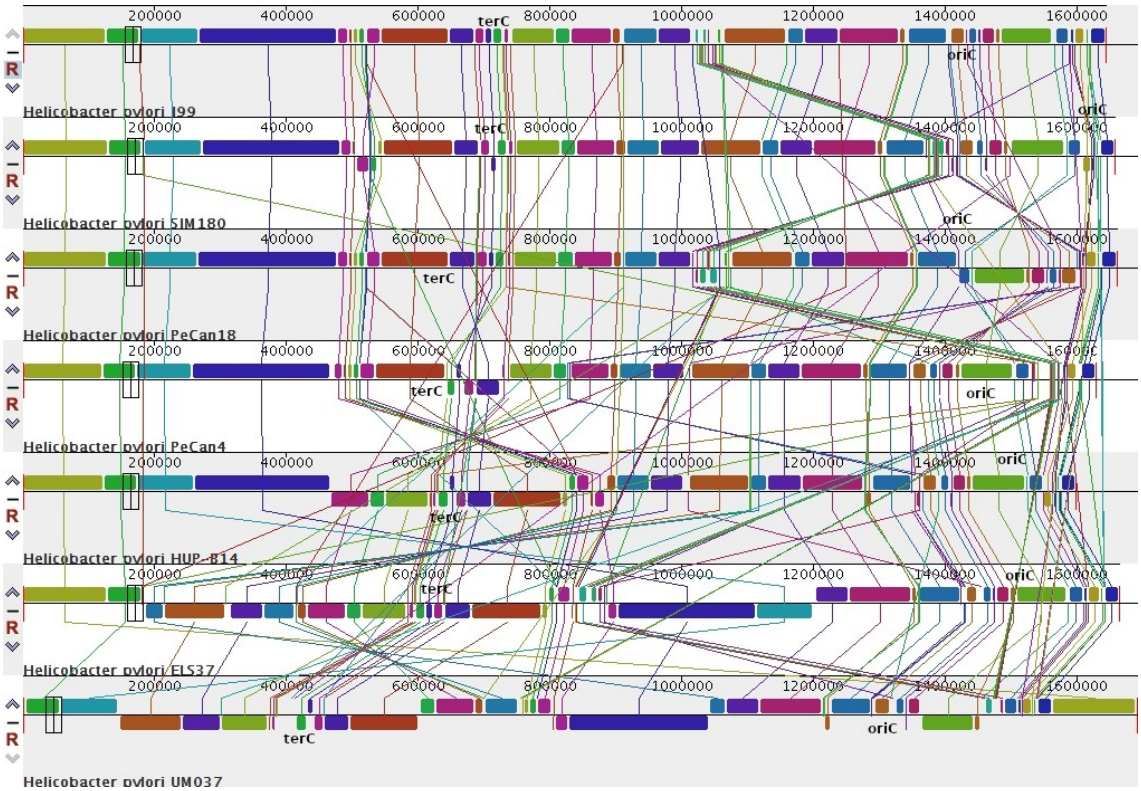


Figura 5. Alineamiento Mauve para el grupo de cepas híbridas. Se indican el inicio (oriC) y termino (terC) de la replicación

4.4 Discusión

En este trabajo presentamos el análisis de la presencia y localización geográfica de diferentes reordenaciones cromosómica en *H.pylori*. En esta especie ya habían sido señalada este fenómeno previamente en diferentes estudios tales como (Furuta et al., 2011; Rodriguez, Birth and death of genes linked to chromosomal inversión

De nuestro análisis se desprende que la ocurrencia de reordenaciones cromosómica es diferente según la zona geográfica que se considere. Así podemos ver que la región africana es la más constante en ordenación sinténica aunque dentro de ella las dos poblaciones sudafricanas se separan del resto. A continuación, le sigue en orden creciente de número de reordenaciones las cepas amerindias en donde la mayoría de ellas (8) presentan prácticamente la misma ordenación con pequeñas inversiones. También en este grupo se incluyen dos cepas con una inversión común de tamaño intermedio. Por otro lado, en las poblaciones europeas existen dos grupos de cepas en los que aparecen diversas reordenaciones cromosómicas. Uno formado por 4 cepas que presentan dos inversiones de tamaño moderado además de otro tamaño pequeño y el segundo incluye las dos poblaciones sudafricanas idénticas entre sí, pero claramente diferentes del resto en cuanto a reordenaciones cromosómicas. Este patrón de agrupamiento por regiones es compatible con fenómenos de deriva genética en la que es claro el efecto fundador como podemos observar en las dos cepas incluidas en el grupo europeo pero aislado en Australia y que muestran exactamente el mismo patrón de reordenaciones. Esta observación también vendría apoyada por el hecho de que las diferentes inversiones observadas son exclusivas de cada región no observándose ninguna de ellas compartida entre regiones.

La presencia de estas inversiones y su mantenimiento en las poblaciones es muy significativa ya que puede tener una gran importancia funcional y evolutiva.

En cuanto al papel funcional se ha señalado qu las inversiones podían jugar un papel importante en la prevención de problemas de funcionamiento en la fisiología de las bacterias ya que, al ser la replicación y la transcripción del ADN simultáneas en las bacterias, esto provoca las colisiones entre

la replisoma y las ARN polimerasas colapsando la horquilla de replicación, lo que provoca rotura de las hebras de ADN y aumentan la mutagénesis. Los encuentros codireccionales (CD) ocurren dentro de los genes codificados en la cadena principal cuando las bifurcaciones de replicación superan a los RNAP. Las consecuencias de estos encuentros son mucho menos graves que los conflictos frontales (HO), que ocurren en genes codificados en la cadena rezagada. Esto haría pensar que la presión selectiva para minimizar las colisiones de replicación-transcripción de HO impulsa la organización del genoma. Una reducción de los genes HO se debería producir a través de la deleción de estos genes o su inclusión en una inversión genética lo que cambiaría su orientación a CD. Entre estas dos posibilidades, solo los eventos de inversión son teóricamente capaces de prevenir conflictos de HO sin resultar en la pérdida de genes importantes. Como tal, los eventos de inversión genética son un medio óptimo por el cual la carga del conflicto celular podría reducirse a lo largo del tiempo evolutivo. En este sentido la presencia de inversiones en los genomas de *H. pylori* sería la respuesta de a la resolución de conflictos genómicos de replicación y transcripción.

Sin embargo, los datos presentados aquí contradicen directamente esta expectativa. Encontramos que *H. pylori* presenta un porcentaje igual o superior genes HO (en relación con los genes codireccionales) por lo que se podría pensar que este proceso se encuentran bajo selección positiva. Esto sugiere que la mayor tasa de mutación de los genes HO podría ser beneficiosa para la célula a través de una generación más rápido de nuevas mutaciones. Si las células que albergan el alelo HO obtienen primero una mutación beneficiosa (una posibilidad significativa debido a la mayor tasa de mutación), estas células deberían aumentar en abundancia, amplificando la prevalencia del alelo HO. En otras palabras, la retención de los alelos HO debería ser el resultado de una selección de segundo orden, también conocido como efecto acompañante. Este modelo también sirve para explicar cómo los genes bajo selección positiva podrían enriquecerse en la orientación HO; las células que albergan alelos HO que no están bajo selección positiva deberían disminuir dentro de las poblaciones debido a la presión de selección negativa conferida por los conflictos HO. Por el contrario, aquellos con alelos HO bajo una fuerte selección positiva deberían aumentar en prevalencia (esto supone que el beneficio de la mutación HO es suficiente para superar la selección negativa conferida por el aumento mediado por el conflicto en el estrés de replicación).

En cuanto al origen de estas reordenaciones cromosómicas se ha señalado en primer lugar que surgen como un subproducto de eventos de recombinación ilegítima (recombinación homóloga) así

como también por un mecanismo de reparación no homólogo impreciso durante la replicación aberrante del ADN para reparar horquillas de replicación rotas **(Hastings *et al.,* 2009)**

En *H. pylori* se ha detectado uno los mayores niveles de recombinación en bacterias patógenas Impacto de la recombinación homóloga en la evolución de los genomas centrales procariotas altos niveles de recombinación en patógenos pueden relacionarse con la presión de selección impuesta por el sistema inmunológico del huésped para aumentar la variabilidad antigénica, que se ha demostrado que activa los sistemas de recombinación y la inducción de los sistemas de competencia y reparación. También, señalan estos mismo autores que las especies que codifican más sistemas de RM tienden a adquirir más material genético por HR con lo cual pueden mayores fenómenos de reordenaciones cromosómicas cosa que ocurre precisamente en *H. pylori*. Esta observación es compatible con el papel de los sistemas de RM que generan extremos recombinogénicos de ADN que pueden promover la HR, aunque se ha observado que los intercambios intraespecíficos son limitados entre cepas que codifican diferentes sistemas de tipo RM **(Wroblewski *et al.,* 2016)**

En segundo lugar también se han propuesto variadas predisposiciones moleculares que permiten que ocurran las reordenaciones. Esto incluye una amplia variedad de contextos cromosómicos tales como motivos secuenciales y estructurales, elementos repetidos, elementos de secuencia de inserción (IS) y elementos de transposón (TE). En organismos con ADN repetitivo, segmentos repetitivos homólogos dentro de un cromosoma o en diferentes cromosomas pueden servir como sitios para cruces ilegítimos. El ADN bacteriano consta de una amplia gama de secuencias repetitivas, que subyacen significativamente a la inestabilidad genómica y contienen puntos calientes de recombinación **(Aras *et al.,* 2003; Treangen *et al.,* 2009).** Las repeticiones de ADN aumentan las posibilidades de reordenamiento a través de la recombinación, amplificación y eliminación de material genético, lo que conduce a la plasticidad del genoma **(Aras *et al.,* 2003; Bao *et al.,* 2014).**

Capítulo 5 Análisis del sistema Pan-inmune del género *Helicobacter*

5.1 Introducción

Las bacterias y arqueas se encuentran constantemente bajo el riesgo de muerte celular e invasión por diferentes elementos genéticos, que incluyen fagos (los virus más abundantes del planeta) y plásmidos **(Forsberg & Malik, 2018)**. Para protegerse de la abundancia y diversidad de fagos, tanto las bacterias como arqueas has desarrollado múltiples mecanismos de defensa que en su conjunto pueden denominarse "el sistema inmunológico procariota o pan-inmune" **(Berheim & Sorek, 2020).**

Las primeras investigaciones sobre los sistemas de defensa procarióticos se centraron principalmente en los sistemas de restricción modificación (RM) y de infección abortiva (Abi), mientras que en la última década el centro de atención se trasladó a los sistemas CRISPR-Cas **(Berheim & Sorek, 2020)**. En los últimos años la caracterización de nuevos sistemas de defensa procarióticos ha ido en aumento y se estima que los sistemas descubiertos hasta fecha, solo representan una fracción de la diversidad que realmente existe en la naturaleza **(Forsberg & Malik, 2018)**; además, sus mecanismos de acción son mucho más complejos de lo que se percibía con anterioridad, ya que se ha aportado evidencia de mecanismos de defensa químicos **(Kronheim *et al.,* 2018)** y de señalización intracelular que regulan el resultado final de la defensa **(Niewoehner *et al.,* 2017; Kazlauskiene *et al.,* 2017)**, así como de una gran cantidad de los cuales a un se desconoce su accionar **(Doron *et al.,* 2018)**

Los sistemas de defensa pueden dividirse en aquellos que se dirigen a los ácidos nucleicos (por ejemplo, RM y CRISPR-Cas, sistemas Abi que llevan al huésped a suicidarse y otros tipos de sistemas**.** De estos sistemas, los más abundantes y complejos son los que dirigen a los ácidos nucleicos, posiblemente esto se deba a que el ácido nucleico suele ser el primer componente viral que se introduce en la célula tras la infección **(Koonin, Makarova & Wolf, 2017; Oliveira, Touchon & Rocha, 2014; Hille *et al.,* 2018)**.

Las especies bacterianas individuales pueden llegar a albergar simultáneamente múltiples sistemas de defensa diferentes y a menudo estos son codificados en loci genómicos, denominados "islas de defensa" **(Makarova, Wolf & Koonin, 2013; Berheim & Sorek, 2020).** Estas islas, son sitios predilectos para la HGT y están sometidos a recombinación frecuente, lo que ayuda a la rápida diversificación del repertorio de genes de defensa.

El género *Helicobacter* que incluye a la especie patógena humana *H. pylori*, comprende al menos 40 especies diferentes aunque su taxonomía aún sigue siendo confusa **(Amorim *et al.,* 2015)**, En la especie *H. pylori*, los sistemas RM son inusualmente abundantes, y representan el 2% del número total de genes en el genoma y como se ha mencionado anteriormente los sistemas RM se consideran clásicamente como un mecanismo de defensa **(Tomb *et al.,* 1997; Alm *et al.,* 1999; Vasu & Nagaraja, 2013).** Entre los diversos tipos de sistemas RM, los tipo II son los más abundantes en *H. pylori*, y las razones para esto continúan siendo desconocidas. Otra característica que exhiben estos sistemas en *H. pylori*, es el alto grado de diversidad entre cepas, y cada cepa puede llegar a tener un conjunto único de genes MTases **(Tomb *et al.,* 1997)**. Así, por ejemplo *H. pylori* F30 codifica tres sistemas RM de tipo I, once sistemas RM de tipo II, un sistema RM tipo III, un sistema RM tipo III y un sistema RM de tipo IV **(Oliveira et al., 2014).**

5.2 Material y métodos.

Para el estudio de sistemas de defensa en el género *Helicobacter* fueron obtenidas todas secuencias disponibles de la base de datos NCBI. Se obtuvieron 531 genomas completos (noviembre 2021), para un total de 15 especies diferentes del género *Helicobacter.*

El análisis de genomas se llevó a cabo con el programa PADLOC **(Leighton *et al.,* 2021)**. Este programa identifica sistema de defensa antivirales contrastando las secuencias de los genomas contra una base datos de perfiles ocultos de Markov (HMM) **(Eddy, 1998)**. Estos perfiles contienen estados de coincidencia, inserción o eliminación que sirven para modelar una familia de secuencias y así, clasificarlos y anotarlos basado en la homología de la secuencia y la arquitectura genética de esta. Los sistemas incluidos en PADLOC hasta la última actualización son 28 (28 de Septiembre 2021) **(Tabla suplementaria 1)**. Sin embargo, hay otros sistemas de defensa que aún no han sido incluidos en esta base datos tales como: BREX **(Goldfarb *et al.,* 2015)**, pVips **(Bernheim *et al.,* 2021),** retrón **(Rodríguez-Mestre *et al.,* 2020)** y el sistema de restricción modificación RM. Para los tres primeros las secuencias de referencia se obtuvieron a partir de las publicaciones, excepto para los sistemas RM, cuyas secuencias fueron obtenidas desde **https://github.com/kblin/ncbi-genome-download (Zhan *et al.,* 2019).**

A partir de las secuencias de proteínas obtenidas se construyeron los perfiles HMM para los respectivos sistemas y se siguió el protocolo establecido por **Payne *et al.* (2021)** en **https://github.com/padlocbio/padloc-db (PADLOC-DB)**, para incluir estos en una base de local.

5.3 Resultados

En la **Tabla 1**, se muestran los resultados obtenidos para la presencia de los diferentes sistemas de defensa en las especies analizadas. En total 17 de los 31 sistemas de defensa fueron detectados. La especie con mayor número de sistemas diferentes es *H. cetorum* que presenta ocho sistemas diferentes, los más representados son *RM, pVips y AbiL,* siendo de destacar la presencia de sistemas tipo CRISPR cas en esta especie. Asimismo otras especies que también presentan sistemas tipo cas son: *H. apodemus, H. pullorum, H. thyplonius y H. mustelae y H. cianedi,* que presenta tres sistemas cas diferentes.

H. pylori presenta siete sistemas diferentes. El sistema más representado en esta especie es el sistema RM, con un total de 7.515 en un total de 473 genomas con un promedio de 16 gnes por cepa. También están bien representados en esta especie los sistemas AbiL (1.224), pVips (872) y Abi (429). El resto de sistemas en esta especie son: DRT type IV, kiwa y Cas other.

°	RMS	AbiL	AbiE	AbiU	AbiO	AbiD	DRT type IV	pVips	Kiwa	cas type II-C	cas type III-D	cas type III-A	Cas other	zorya other	Hachiman type I	cbass type I	GAO19
H. apodemus	30	4						6		4				8	6		
H. bilis	18							2						18			
H. cholecystus	25					1		5									4
H. cinaedi	337							470		18	36		6	60			
H. hepaticus	18							18						4	2		
H. pullorum					2					4	6			10			
H. thyplonius	6				2					4				6			
H. winghamen-sis	42	18						18						4			
H. acinonychis	6																
H. cetorum	60	24		8				60	14			12	4			14	
H. felis								18									
H. suis	60		4					30									
H. pylori	7515	1224	420				196	872	30				8				
H. himalayensis								18						2			
H. mustelae								11		2							
Total	8117	1270	424	8	4	1	196	1528	44	32	42	12	18	112	8	14	4

Tabla 1. Resumen de los sistemas de defensa detectados para el género *Helicobacter.*

5.4 Discusión

Del análisis para la búsqueda de diferentes sistemas inmunes en los 531 genomas del conjunto de 15 especies de género *Helicobacter* podemos destacar varios aspectos. En primer lugar, hemos encontrado que todas ellas excepto *H. felis* y *H. acinonychis* que tienen un solo sistema y *H. himalayensis que presenta 2, contienen* al menos tres sistemas de defensa diferentes. Y en segundo lugar si se contemplan todas las especies en conjunto están representados 17 de los 31 sistemas inmunes analizadas que representan el 55% de todos los sistemas.

El primer aspecto, la presencia de varios sistemas de defensa simultánea en un mismo genoma esto ha sido ya señalado por ejemplo en *E. coli* y *Pseudomonas aureginosa* por lo que una sola cepa puede codificar diversas estrategias de defensa. Los posibles beneficios para que una bacteria codifique múltiples sistemas de defensa, incluso si estos sistemas se superponen en la gama de virus a los que se dirigen son evidentes. En primer lugar, una bacteria puede ser invadida simultáneamente por varios virus diferentes y estos además pueden codificar mecanismos de contradefensa, En este sentido se ha comprobado que los fagos pueden evolucionar para eliminar secuencias específicas, como los motivos dirigidos por las enzimas de restricción o las secuencias PAM que son esenciales para la defensa CRISPR-Cas. También, los fagos a menudo codifican proteínas que se expresan al inicio de la infección e inhiben los sistemas de defensa. También se ha señalado que, las proteínas anti-restricción inhiben las enzimas de restricción: por ejemplo, la T4 IPI (proteína interna I) inhibe los sistemas de RM tipo IV , mientras que las proteínas DarA y DarB del fago P1 se unen a los sitios de restricción en el genoma del fago y los enmascaran de la escisión por el sistema de RM tipo I de *E. coli.* Por estos motivs las bacterias no pueden depender de un único sistema de defensa y, por lo tanto, deben presentar varias líneas de defensa como estrategia de supervivencia.

En segundo lugar, como podemos observar existe una gran variabilidad en cuanto al tipo y número de sistemas de defensa entre las distintas especies como podemos comprobar no necesariamente el aumento en el número de genomas significa un aumento en el número de sistemas de defensa. De hecho, la especie H. cetorum con ocho sistemas de defensa solo se han analizado 69 genomas analizados. Dados los costes de aptitud que suponen los sistemas antivirales, es probable que ninguna cepa bacteriana o de arqueas pueda codificar, a largo plazo, todos los sistemas de defensa posibles sin sufrir serias desventajas competitivas. Sin embargo, si estas especies se mezclan como parte de una megapoblación, el pangenoma de

esta población codificaría un "potencial inmunológico" que incluiría todos los sistemas representados.

Se ha señalado que los sistemas de defensa y respuesta al estrés, en particular, los sistemas de RM y TA pueden considerarse partes especiales del mobiloma. El mobiloma una propuesta de **(Doron *et al.,* 2018)** incluiría bacteriófagos, plásmidos, elementos transponibles y genes que a menudo se asocian con ellos y se convierten regularmente en pasajeros, como los sistemas de restricción-modificación (RM) y toxina-antitoxina (TA). El análisis comparativo de estos sistemas muestra evidencia de evolución rápida y HGT frecuente, y se encuentran con frecuencia en genomas de plásmidos y bacteriófago. Parece natural que, en la medida en que los virus y plásmidos son móviles por definición, también lo sean los sistemas de defensa. El mobiloma está indisolublemente conectado con los cromosomas procarióticos "principales". Los virus (bacteriófagos) y muchos plásmidos se integran sistemáticamente en los cromosomas, ya sea de forma reversible.

Como estos sistemas pueden estar fácilmente disponibles mediante HGT, dada la alta tasa de HGT de los sistemas de defensa, la población alberga en efecto un depósito accesible de sistemas inmunitarios que pueden ser adquiridos por los miembros de la población **(van Houte *et al.,* 2016)**. Cuando la población se somete a la infección, esta diversidad garantiza que al menos algunos miembros de la población serían codificar el sistema de defensa apropiada, y estos miembros podrían sobrevivir y formar la base para la perpetuación de la población. Por tanto, planteamos la hipótesis de que parte de la selección de los sistemas de defensa se produce a nivel de grupo.

Por último, es claro que en el caso de *H. pylori* existe un sesgo hacia la presencia de RM. Y no presenta sistemas CRISPR Cas. Estos dos hechos pueden ser fundamentales para favorecer la adaptación de esta especie ya que los sistemas RM, como hemos señalado anteriormente pueden ser importantes en los procesos de recombinación homóloga y por otro lado la ausencia de sistemas CRISPR favorece la integración de genomas foráneos en el genoma con lo cual pueden aumentar los fenómenos de transferencia horizontal y la adquisición de nuevos genes que pueden ser favorables para su adaptación.

Capítulo 6 Detection and variability analyses of CRISPR-like loci in the *H. pylori* genome

6. 1 Introduction

The genus *Helicobacter* comprises 40 formally validated species. Within this group *H. pylori* is particularly important for being a Gram-negative human pathogenic bacterium present in about half of global population (**Gangwer *et al*., 2010)**. This bacterium can induce superficial gastritis and constitutes a risk factor for the development of peptic ulcer disease, gastric adenocarcinome, and gastric mucosa-associated lymphoide tissue lymphoma **(Gangwer *et al*., 2010).** Due to its characteristics can be considered as a model organism for the study of genetics and evolution. *H. pylori* has a great genomic plasticity, presenting high rates of mutation and recombination that allows for the generation of new alleles, allowing it to adapt to relatively specific and well-defined habitats such as the stomach and the duodenum **(Backert & Yamaoka, 2015)**. It is well established that *H. pylori* is a highly competent bacterium, and different strains can be found living together in the gastric environment, bringing the populations of *H. pylori* closer to panmixia **(Kang & Blase, 2006; Suerbaum *et al*., 1998)**. Genome comparative analyses from diverse origins have shown that this bacterium shows a high degree of genetic diversity, ranging from nucleotide polymorphisms to genetic mosaicism **(Zawilak-Pawlik, & Zakrzewska-Czerwińska, 2017)**.

The CRISPR-Cas system is a defense mechanism against foreign genetic elements derived from bacteriophages, plasmids or extracellular chromosomal DNA **(Mojica et al., 2005; Sampson & Weiss, 2013)**. The CRISPR-Cas loci are variable in number between bacteria and strains **(Grissa *et al.,* 2008)**, and its typical structure is characterized by a CRISPR matrix, a nearby Cas-gene locus, and an AT-rich leader region **(Zhang *et al*., 2017)**. This system is also characterized by its rapid evolution and variability which makes its classification a highly complex task, due to the frequent modular recombination of the CRISPR **(Delaney *et a*l., 2012)** matrix, which may mean that not all CRISPR systems carry the same components **(Grissa *et al*., 2008; Delaney *et al*., 2012)** or fulfill the same functions **(Sampson & Weiss, 2013)**.

The CRISPR-Cas systems have been identified in approximately 40% of the bacteria and 90% of the archaea. However, **Burstein et al. (2016)** recently proposed that CRISPR-Cas systems are present in only 10% of the archaea and bacteria. This difference in the presence of the CRISPR-Cas system in prokaryotes could due to the fact that the system may not exist in the main non-cultivable bacterial lineages and in those whose lifestyle was symbiotic **(Burstein *et al*., 2016; Burstein *et al*., 2017)**.

In the genus *Helicobacter*, the CRISPR-Cas system has only been detected in *H. cinaedi* and *H. mustelae* **(Kersulyte *et al*., 2013; Tomida *et al*., 2017)**, both pathogenic species, but not in *H. pylori*. However, **Bangpanwimon et al. (2017)** have more recently described CRISPR-like sequences in the genome of *H. pylori*, more precisely located in the vacA-like paralogue gene (*VlpC*, HP0922), that could be related to the ability to colonize the stomach **(Foegeding *et al.,* 2016)**, suggesting that they could have a regulatory role **(Albert *et al*., 2005)**. In fact, in recent years, hypotheses involving CRISPR loci in the regulation of genes, a function analogous to the functions of RNAi in eukaryotes **(Bondy-Denomy & Davidson, 2014)**, have appeared where the CRISPR spacers coincided with genes from the genome itself, with important cellular functions (housekeeping) **(Bondy-Denomy & Davidson, 2014; Stern *et al*., 2010)**. Another relationship established between CRISPR and pathogenicity has been discussed in strains of *E. coli* and other species, where the interference of CRISPR prevented the acquisition of virulence genes **(García-Gutiérrez *et al*., 2015)**. On the other hand, a reduced content of CRISPR repeats has also been correlated with a greater likelihood that a strain exerts pathogenicity (potential ability to cause disease) **(García-Gutiérrez *et al*., 2015)**. All of these data exemplify the versatility of CRISPR-Cas systems and suggest roles beyond canonical interference against strange genetic elements **(Hatoum-Aslan & Marraffini, 2014)**. The presence of CRISPR orphans of non-vestigial subtype I-F and E in *E. coli* (CRISPR without cas-genes) have been attributed to a possible habitat change, where their presence would be counterproductive **(Almendros *et al*., 2016)**, granting them a regulatory role, whose spacers could prevent the acquisition of cas (anti-cas) genes, thus facilitating the acquisition of genetic material and increasing biological aptitude **(Almendros *et al*., 2016)**.

In this work, we analyze the presence and variability of CRISPR-like sequences in *H. pylori* by studying 53 strains, finding that there are several CRISPR-like sequences in their genomes, which are relatively conserved among strains and can be grouped by geographic area. We discuss their possible role in the generation of variability as well as in the regulation of the genes into which they are inserted.

6.2 Materials and methods

For an analysis of CRISPR-like loci in *Helicobacter pylori*, the sequences of 53 complete genomes **(Table 1)** (GenBank and fasta formats) of different *H. pylori* strains were downloaded from the genomic resource database of the National Center Biotechnology Information (NCBI)

(Benson *et al.,* 2011) (ftp: //ftp.ncbi.nlm.nih.gov/genomes/). To characterize the CRISPR region in the *H. pylori* genomes we used the CRISPRFinder program with default parameters **(Grissa *et al*., 2007) (http://crispr.i2bc.paris-saclay.fr/Server/)** (Last update on May 9, 2017). In addition, to characterize the CRISPR-like regions in all the genomes analyzed in this work, multiple alignments were created with the Muscle program **(Edgar, 2004)**.

We used CRISPRsBlast (E-value: 0.01) to determine the similarity between the direct repeats sequences (DRs) and spacers of the CRISPR loci detected in *H. pylori* and the sequences of DRs and confirmed spacers deposited in the BLAST CRISPR database **(http://crispr.i2bc.paris-saclay.fr/crispr/BLAST/CRISPRsBlast.php) (Grissa *et al*., 2008)**.

The spacers were also blasted with default parameters against the CRISPRTarget server, which predicts the most likely targets of the CRISPR RNAs **(http://bioanalysis.otago.ac.nz/CRISPRTarget/crispr_analysis.html) (Biswas *et al*., 2013)**. The databases used were: mobile genetic elements and phages, viruses.

The alignments of CRISPR-like sequences were carried out by Muscle **(Grissa *et al*., 2007)** and Geneious v 6.1.8 **(Kearse *et al.,* 2012)** softwares. The secondary RNA structure and minimum free energy of the DR sequences were predicted using RNAfold WebServer **(http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi)** with default parameters **(Zuker & Stiegler, 1981)**.

For phylogenetic analyses we used the Mega7 program **(Kumar *et al.,* 2016)** with the following parameters: Neighbor-joining method with bootstrap of 1000 replications and Jukes Cantor model.

6.2.1 Identification of operons linked to CRISPR-like and Cas domains

The research on operons linked to the CRISPR-like structure was carried out using the OperonDB database **(http://operondb.cbcb.umd.edu/cgi-bin/operondb/operons.cgi) (Haft *et al*., 2005)**.

For the identification of cas domains, the HMMs profiles (Markov Hidden Models Profile) of the Cas families were downloaded from TIGRFAM **(ftp://ftp.jcvi.org/pub/data/TIGRFAMs/)** as well as the Cas proteins described by **Haft et al., 2005**. The search of cas proteins was carried out with HMMER software v3.1b2 **(Eddy, 1998)**, implementing the option 'hmmscan' (search in proteins against collections of proteins of the 53 genomes), with an E-value 10e-5.

6.2.2 Identification of vacA-like gene (*VlpC*)

To identify and determine the presence of vacA-like gene (*VlpC*) in the 53 genomes of *H. pylori* used in this study, the reference sequence of strain J99 **(Chanto *et al.*, 2002; Lara-Ramirez *et al.*, 2011)** was downloaded from NCBI: WP_000874591.1 (*VlpC*). VlpC gene is located in strain J99 at position 945.691 to 952.890. This sequence was blasted against the 53 *H. pylori* genomes with the following parameters: E- value: 10e-5, query coverage >75%.

Also, to determine if the corresponding mRNA of the *VlpC* gene of the different strains of *H. pylori* was expressed, the cDNA sequences of the 53 genomes were downloaded via FTP **(http://bacteria.ensembl.org/info/website/ftp/ index.html)** and used as a target to be blasted with the CRISPR-like sequences detected in the *VlpC* gene, using an E-value of 10e-5

.In addition, for genes that showed a CRISPR-like sequence outside of *VlpC*, their presence in all genomes was verified using blastn with Geneious v 6.1.8 **(Pertea et al., 2009)** using an E-value of 10e-5.

To determine genomic rearrangements and possible break-point involved in recombination events, the Mauve software was used for complete alignment of genomes **(Haft *et al*., 2005)**.

6.3 Results

6.3.1 CRISPR-like loci identification

A total of 53 *H. pylori* assembled and annotated genomes from different geographical regions were analyzed with CRISPRFinder software. Twenty-two CRISPR-like loci were found in 20 strains, with 19 of them exhibiting one CRISPR-like locus and only one strain, SJM180 **(Table 1)** showing three CRISPR-like loci. Of all loci, 16 were located within a vacA-like gene (VlpC gene), with four DRs and three spacer sequences. This gene was integrated in an operon with the genes *OMP*, *4-oxalocrotonate tautomerase*, *recR, truD, htpX, folE, IspA* and, *surE*, in this order [24]. The remaining six CRISPR-like loci were present in other locations of the genome. More specifically, they were located in: a) the BM012A (Australian origin) and Shi470 (Peru origin) strains in a Poly E-rich gene rich protein; b) the Shi417 and Shi112 (both Peru origin) strains within a hypothetical protein (with GO term COG119), and; c) the SJM180 (Peru origin) strain, with two additional loci, with these located in two different hypothetical protein genes **(Table 1)**. For these 22 loci, which were detected with CRISPRFinder, 95 direct repeat sequences (DRs) were identified, being present in 4 to 7 sequences per CRISPR-like locus and

ranging from 23 to 36bp in length **(Table S1)**. No similarities were found when these sequences were blasted against the CRISPRsBlast database. A total of 73 spacers were detected ranging in number from 3 to 6 sequences per locus, with lengths ranging between 16 to 69bp. Using CRISPRTarget software, 5 spacers showed similarities to phage, plasmids or viruses sequences **(Table S2)**.

Accesion number	Strain	Origin/isolation	Diagnosis	Gene with CRISPR locus	cas 2	cas 3	cas 4	Csa 3
NC_000921	J99	Africa/USA	Duodenal ulcer	VlpC				
NC_017374	2017	Africa/France	Duodenal ulcer	VlpC				
NC_017381	2018	Africa/France	Duodenal ulcer	VlpC				
NC_017357	908	Africa/France	Duodenal ulcer	VlpC				
NC_017371	Gambia94/24	Africa/Gambia	unknown					
NC_017742	PeCan18	Africa/Peru	gastric cancer					
NC_017361	S. africa7	South Africa	unknown					
NC_022130	S. africa20	South Africa	unknown					
NC_017063	ELS37	America/El Salvador	Gastric cancer					
NC_017733	HUP-B14	Europe/Spain	unknown					
NC_014560	SJM180	America/Peru	Gastritis	Hypothetical protein -VlpC- Hypothetical protein				
NC_011498	P12	Europe/German	Duodenal ulcer					
NC_012973	B38	Europe/France	MALT lymphoma	VlpC				
NC_011333	G27	Europe/Italy	unknown					
NC_021217	UM037	Asia/Malasya	unknown					
NC_014256	B8	unknown	Gastric ulcer	VlpC				
NC_017362	Lithuania75	Europe/Lithuania	unknown					
NC_000915	26695	Europe/UK	Gastritis	VlpC				
NC_018939	26695	unknown	unknown	VlpC				
NC_018937	Rif1	Europe/German	unknown	VlpC				
NC_018938	Rif2	Europe/German	unknown	VlpC				
NC_008086	HPAG1	Europe/Sweden	Atrophic gastritis	VlpC				
NC_022886	BM012A	Oceania/Australia	Asymptomatic -reinfection	Poly E-rich protein				
NC_022911	BM012S	Oceania/Australia	Asymptomatic -reinfection					
NC_017372	India7	Asia/India	Peptic ulcer	VlpC				
NC_017376	SNT49	Asia/India	Asymptomatic	VlpC				
NC_017926	XZ274	Asia/China	Gastric cancer	VlpC				
NC_020509	OK310	Asia/Japan	unknown					
NC_017367	F57	Asia/Japan	Duodenal ulcer	VlpC				
NC_017360	35A	Asia/Japan	unknown					
NC_017368	F16	Asia/Japan	Gastritis					
NC_021218	UM066	Asia/Malasya	unknown					
NC_021215	UM032	Asia/Malasya	peptic ulcer					
NC_021216	UM299	Asia/Malasya	unknown					
NC_021886	UM298	Asia/Malasya	unknown					

2				
NC_01736	F30	Asia/Japan	Duodenal	
5			ulcer	
NC_02050	OK113	Asia/Japan	unknown	
8				
NC_01737	83	unknown	unknown	
5				
NC_01738	51	Asia/Korea	Duodenal	
2			ulcer	
NC_01736	F32	Asia/Japan	Gastric cancer	
6				
NC_01735	52	Asia/Korea	unknown	
4				
NC_01455	PeCan4	America/Peru	gastric cancer	
5				
NC_01737	Puno135	America/Peru	Gastritis	
9				
NC_01737	Puno120	America/Peru	Gastritis	
8				
NC_01735	Sat464	America/Peru	unknown	
9				
NC_01069	Shi470	America/Peru	Gastritis	Poly E-rich protein
8				
NC_01774	Shi169	America/Peru	unknown	
0				
NC_01773	Shi417	America/Peru	unknown	Hypothetical protein
9				
NC_01774	Shi112	America/Peru	unknown	Hypothetical protein
1				
NC_01735	Cuz20	America/Peru	unknown	
8				
NC_01735	v225d	America/	Gastritis	
5		Venezuela		
NC_01956	Aklavik86	America/Canada	Gastritis	
3				
NC_01956	Aklavik117	America/Canada	Gastritis	
0				

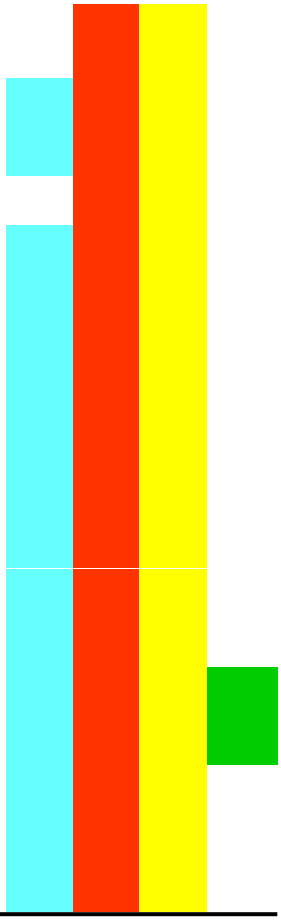


Table 1. Characteristics of the CRISPR-like loci detected with CRISPRFinder in the 53 strains of *H. pylori*. The different colors represent the presence of cas domains in the analyzed genomes: blue: Cas2, red: Cas3, yellow: Cas4 and green: Csa3. Cas1 domains were not detected

Consensus DRs for each locus, were generated by CRISPRFinder and sequences for each spacer were used to carry out a phylogenetic study. For DRs, two main groups were observed in the phylogenetic tree: one, including the DRs of the six CRISPR-like loci located out of *VlpC* gene, and the other, with the strains that had the loci within the *VplC* gene **(Figure 1)**.

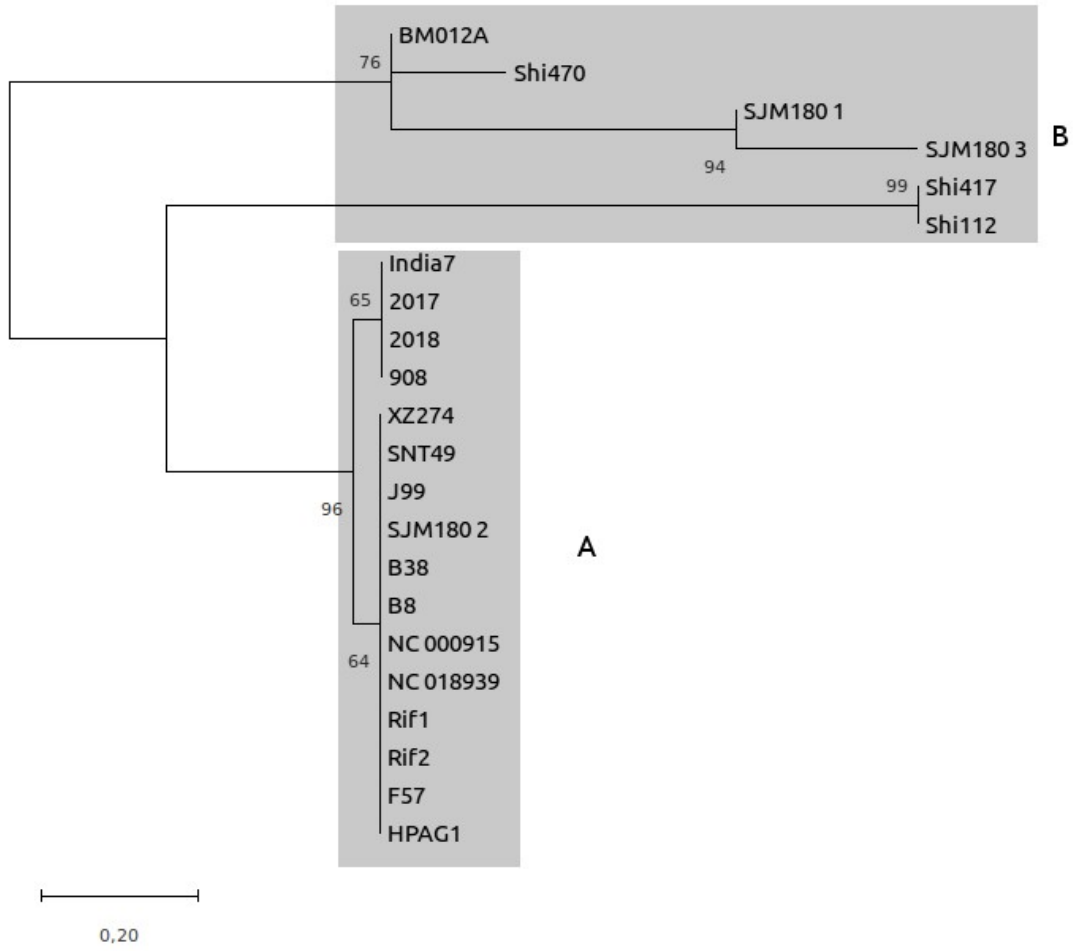


Figure 1. Classification of the repeated consensus sequences obtained from CRISPRFinder. Phylogenetic tree of there peated consensus sequences obtained from the CRISPR loci confirmed to establish evolutionary relationships and classify these sequences. The MEGA7 software was implemented for this analysis. The evolutionary distance scale is 0.2 Jukes-Cantor model. **(A)** CRISPRlocatedwithintheVlpCgene.**(B)** CRISPR located within genes other than the *VlpC* gene.

The spacers in the phylogenetic tree could be divided into four main groups: three of them corresponded to the group of spacers present in the first, second and third position within the CRISPR-like loci located within the *VlpC* gene. The fourth group corresponded to the spacers of the CRISPR-like loci found in other genes within the SJM180 (CRISPR1-like and CRISPR3-like), BM012A, Shi417 and Shi112 strains **(Figure 2)**

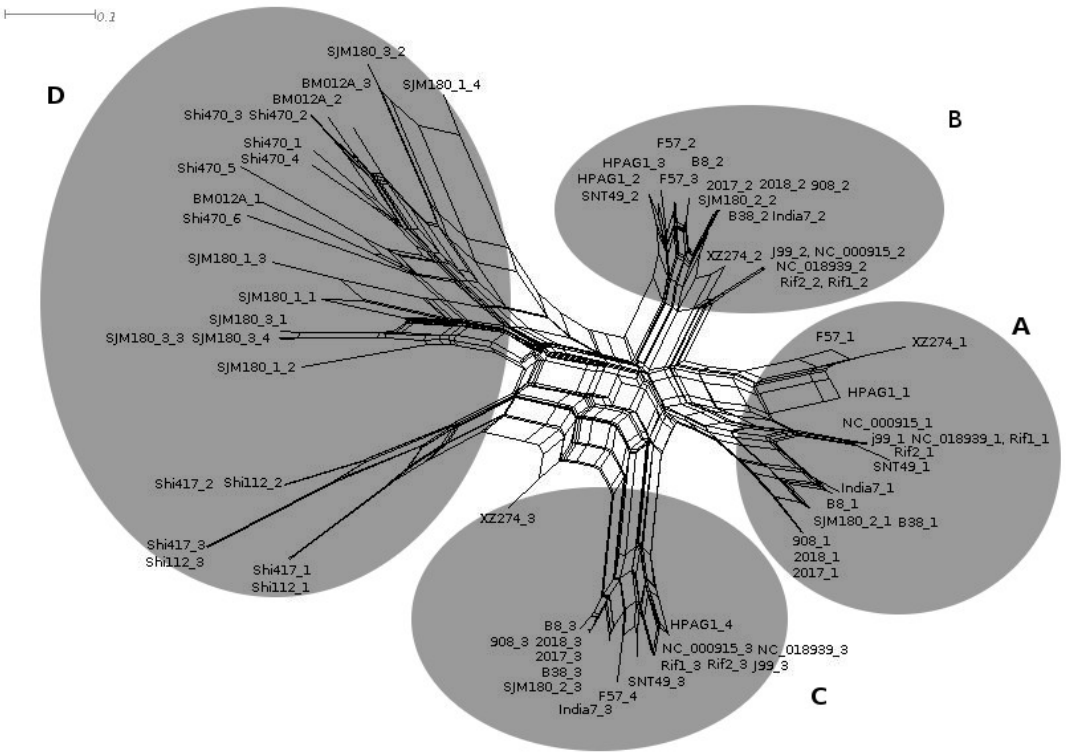


Figure 2. **Classification of the spacers sequences obtained from CRISPRFinder**. Phylogenetic tree for the classification of the spacers sequences obtained from the confirmed CRISPR, based on evolutionary relationships implementing the MEGA7 and software. The evolutionary distance scale is 0.1 Jukes-Cantor model. **(A, B, and C)** represent the spacers located within the VlpC gene. **(D)** Represents the spacers located within genes other than *VlpC*.

6.3.2 Analysis of CRISPR-like sequences located within *VlpC*

The *VlpC* gene was present in all genomes, except for strain Aklavik86. A manual construction of multiple alignments allowed us to determine the presence of a CRISPR-like structure within the *VlpC* gene for all genomes. Only the South Africa20 strain showed the *VlpC* gene but not the CRISPR-like locus, as the gene is truncated in the 5' region where this structure would be found. The CRISPR-like locus possessed different degrees of variability between strains. The alignment allowed for an in-depth study of DRs and spacers for this locus. It was observed that the variation of the CRISPR-like structure in the *VplC* gene was mainly due to the complete duplication and/or deletion of spacers and DRs **(Figure S1A to S1F)**. The sequences from the 51 CRISPR-like loci detected in *VlpC* were used to carry out a phylogenetic analysis. Three clusters were observed, created by grouping the sequences according to their geographical origins **(Figure 3)**. The first group included the Africa and Europe strains (group A), the second included the Asia (group B) strains and with the last being the Amerindian strains (group C).

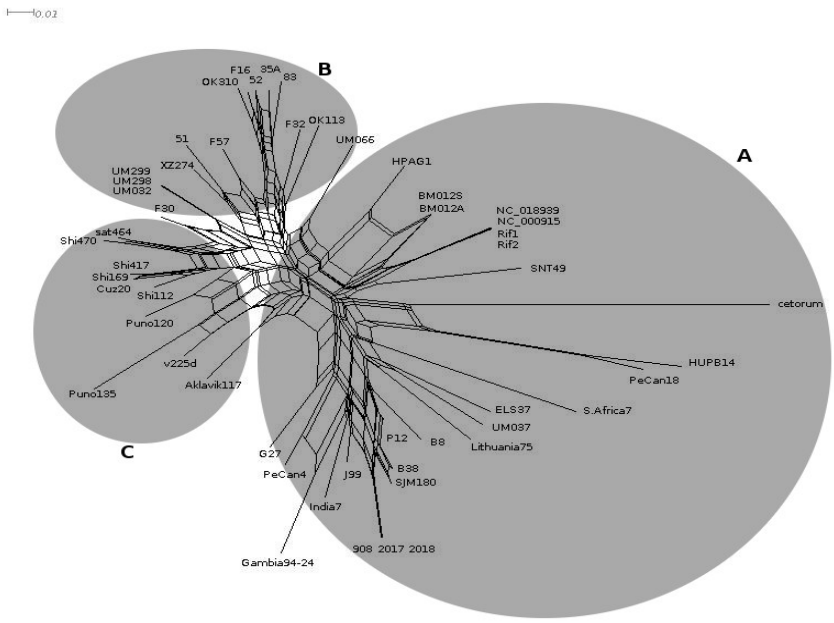


Figure 3. Classification of CRISPR-like in VlpC gene. Phylogenetic tree constructed with the 51 CRISPR-like sequences present and located inside the *VlpC* gene, which evidences a phylogeographic differentiation of the CRISPR-like loci. Analysis executed with MEGA software. The evolutionary distance scales is 0.01 Jukes-Cantor model. (A) Group of African and European geographical origin.(B) Geographical group of Asian origin and (C) Amerind geographic group.

Despite the great variability detected, when the transcriptomes of the *H. pylori* strains were analyzed, it was found that the gene corresponding to *VlpC* mRNA was expressed in 50 of the 52 genomes that possessed this gene, including the CRISPR-like sequence **(Table S3)**.

When a blastn (E-value: 10e-5, query coverage> 75%) was performed using the *VlpC* gene sequence from *H. pylori* against the genomes of other *Helicobacter* species, only *H. cetorum* showed the presence of this gene. This gene had the CRISPR-like structure, similar to *H. pylori* and an identity above 80% in DRs, indicating that it was the same locus.

6.3.3 Analysis of CRISPR-like sequences located outside the *VlpC* gene

In addition to the 16 CRISPR-like loci detected in the *VlpC* gene by CRISPRFinder, we detected two additional loci in the Shi417 and Shi112 (WP_000536430 and Shi112 WP_000536429 hypothetical protein, respectively) strains, which had identical sequences in their DRs and spacers. These were located in the 5' region of a gene from a hypothetical protein, between the positions 55,000 to 56,000 of the genome. The ontology analysis showed that this protein had domains related to cell division and cycle control. The CRISPR-like locus

of this gene had a length of 150bp, with four 23bp DRs and three 19bp spacers. No similarities were found with other types of genetic element.

When blastn (E-value: 10e-5, query coverage > 75%) was performed using the sequence of this gene against the remaining 51 genomes, it was found in 12 more strains. The 5´ regions were low conserved, and even three strains (aklavik86, aklavik117 and P12) had this region truncated. Whereas the 3' region was highly conserved (85%) for the twelve strains **(Figure. S2)**. All the genes had a CRISPR-like locus in their sequence but, as the CRISPR-like locus is located in the 5' region of the genes, they were degenerate (56% of identity). The origins of these 14 strains with CRISPR-like locus were Amerindian (6) European (6) and African (2), and the phylogenetic tree constructed, using the CRISPR-like sequences, clearly separated these three groups **(Figure 4a)**.

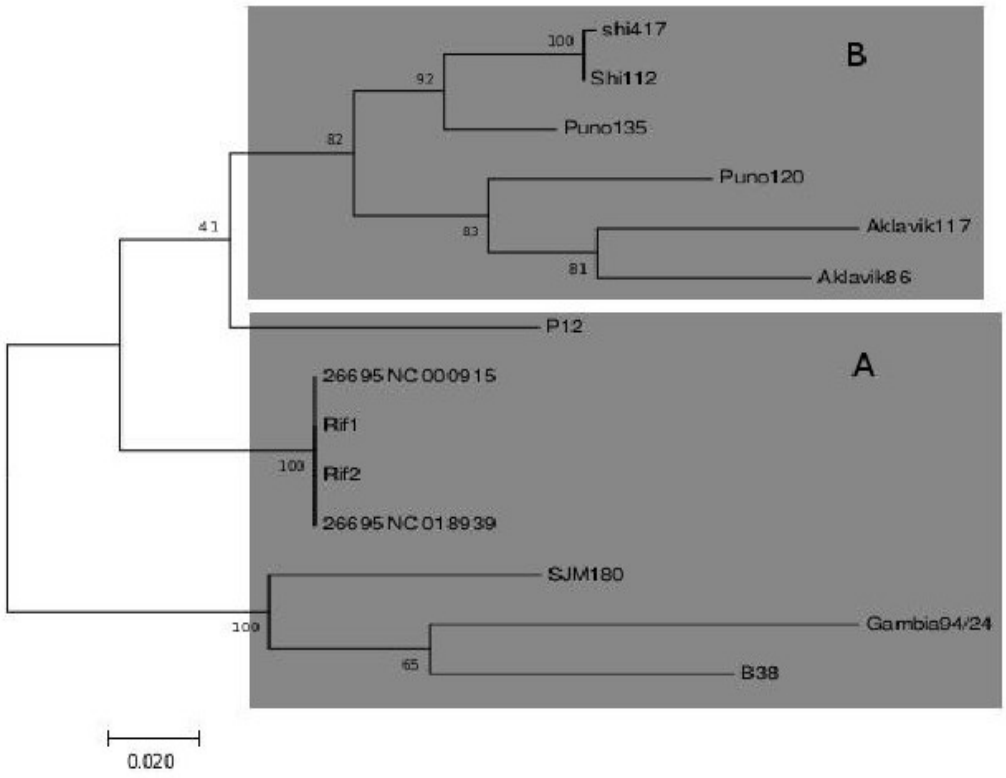


Figure 4a. Phylogenetic tree constructed with the 14 CRISPR-like loci detected in the gene coding for a hypothetical protein. A phylogeographic differentiation of CRISPR-like loci is observed. Analysis performed using MEGA7 software. Evolutionary distance scale: 0.02 model of Jukes-Cantor. (A) Group of African and European geographical origin. (B) Amerind geographic group.

The Shi470 and BM012A strains showed a CRISPR-like locus within a Poly-E rich protein gene (WP_00078209, WP_023591955 respectively). In Shi470, this gene was located between the positions 320.726 and 322.187. In the case of BM012A, it was between the positions 659.636 and 661.240. In a genomic structural analysis of these two strains with Mauve **software (Darling *et al*., 2010)**, it was verified that it was the same gene present in a syntenic region but affected by a genomic rearrangement. This gene was included in an inverted segment and near a breakpoint **(Figure 4b).** The alignment of this gene from both strains revealed a middle location of the CRISPR-like locus, with a 70% similarity. The divergences could be explained for the different number of DRs and spacers detected among them **(Figure 4b)**.The CRISPR-like locus of Shi470 had a length of 660bp with seven DRs and six spacers while that for BM012A was 174bp in length with four DRs and three spacers. It was interesting to note that the spacers of the Shi470 strain showed similarity with mobile elements and phages **(Table S2)**, while no similarities were found for those from the BM012A strain.

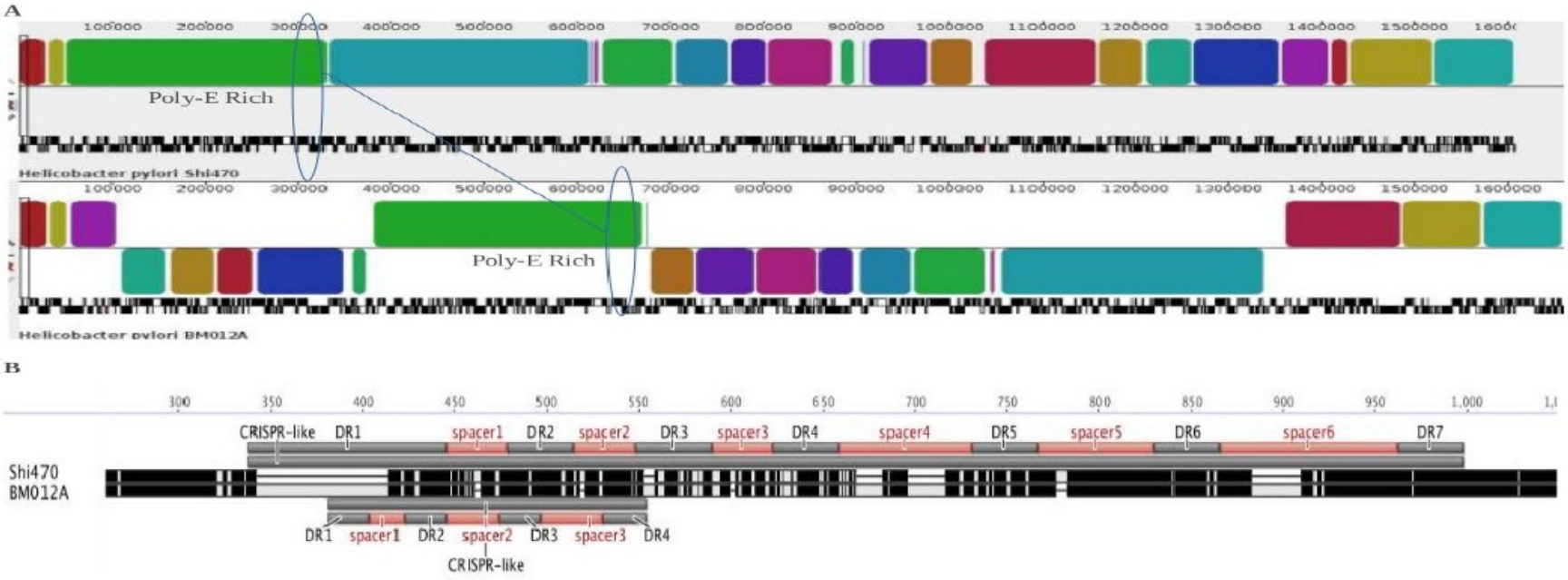


Figure 4b. Alignment of the gene for the Poly-E rich protein. (A) Alignment of the Shi470 and BM012A genomes using Mauve software. Figures show the genome region where Poly-E rich protein gene is included. The alignment revealed that this gene was in a region close to the breaking point of an inversion that affects these strains (blue oval). **(B)** Alignment of the Color indicates the degree of variation in both the gene and its CRISPR-like loci. Dark (high values of pairwise % identity), light (low pairwise % identity). Alignment was performed with Muscle software. Repeated direct sequence (DR). Solid line indicates the presence of gaps. The alignment suggests that the differences observed can be explained by the number of DR sequences and spacers in which they differ.

A blastn (previous parameters) search with the rest of the genomes (51) allowed us to identify this gene in 33 more strains, all of them with CRISPR-like features. The genes showed a high identity (close to 80%) in their sequence except in the CRISPR-like region (60% identity) **(Figure S3)**. The phylogenetic tree, constructed with the CRISPR-like sequences, clearly separates the four geographic regions **(Figure 4a).**

In relation with the two additional CRISPR loci detected by CRISPRFinder in SJM180 strain, these were called CRISPR1-like and CRISPR3-like and were found in two different hypothetical proteins (WP_000446591-CRISPR1-like; WP_013356447-CRISPR3-like). Their percentage of identity was not significant for considering that they were the same gene. The CRISPR1-like loci was inserted in the middle of the gene and was located in position 128.894 to 128.614 of this strain's genome, with a length of 314bp, five DRs (with an average length of 25bp), and four spacers (with a length of 34bp). Spacers 1_1 and 3_1 showed similarity with plasmids and viruses, respectively **(Table S2)**. A blastn search for this gene revealed that this protein is present in 39 strains. Also, it was observed that the sequence from region 5´ to the beginning of CRISPR1-like (approximately 380bp) was highly conserved (91%), while the region corresponding to CRISPR-like was degenerate (63%), with the 3' region (approximately 600bp) being highly conserved (90%) as well **(Figure S7).** The phylogenetic tree using the 39 CRISPR1-like sequences showed, in this case, a mixture of the strains in relation to their geographical origin. **(Figure S8).**

The CRISPR3-like region, with a length of 266bp, was also inserted in the middle of the gene (positions 1.201.946 to 1.201.720 of the genome) and showed five DRs (average length of 23bp) and four spacers (ranging between 19 and 31bp), with spacers 1 and 3 showing similarity with plasmids **(Table S2)**. The blastn analysis revealed this protein to be in 27 strains. This hypothetical protein was highly conserved (96%) from the 5` region to the beginning of the CRISPR-like region (approximately 380bp) while the CRISPR3-like region was degenerate (61%) and the 3` region (approximately 520bp) was conserved (81%) **(Figure S9).** The phylogenetic tree created with these 27 sequences showed, as in the previous case, a mixture of the strains of different geographical origins **(Figure S10).**

Additionally, the DRs consensus from all CRISPR-like loci found in *H. pylori* were analysed using RNAfold Server **(Zuker & Stiegler, 1981)**, and the secondary RNA structure predicted. Also, the minimum free energy was found to range from -0.74 to -7,73 kcal/mol **(Figure S9)**.

6.4.4 Cas Domain detection

Cas3 and Cas4 domains were identified in 100% of the analyzed strains, whereas Cas2 domains were found in 32 strains (60.4%), and the Csa3 domain only in 2 strains (4%) **(Table 1, Table S4)**. These domains were found in various locations in the different strains.

6.5 Discussion

In the genome of the human pathogenic bacterium Gram negative, *H. pylori*, the CRISPR-Cas system is not functional and does not exist by forming an operon structure as it is known for other organisms. The lack of this system in some prokaryotes has been related to the increase in the capacity to integrate exogenous DNA in the genome of these bacteria and, resulting in the acquisition of new functions, which can confer an adaptive advantage to these strains, particularly during their transition to pathogenesis **(Sampson & Weiss, 2013; Sampson & Weiss, 2014)**. But recently, **Bangpanwimon *et al*. (2017)** reported the presence of CRISPR-like sequences inserted into the *VlpC* gene of *H pylori*. In that study, the detection was performed by PCR in partial regions of the genome of Thailand isolates **(Bangpanwimon *et al*., 2017)**. Each isolated strain showed a CRISPR-like locus with similar DRs sequences. However, results from other strains from different geographical regions, the variability of this locus, or the possibility of the presence of other CRISPR-like loci in the genome of *H. pylori* were not analyzed.

In this work, we show the analysis of 53 strains of *H. pylori* which comprise all the continents. The phylogenetic analyses carried out using the sequences of the CRISPR-like locus found by CRISPRFinder revealed the existence of additional loci to the CRISPR-like locus inserted into the VlpC gene described by **Bangpanwimon et al. (2017) (Figures S2, S5, S7, and S9)**.

Of the 53 genomes analyzed, 51 of them showed a locus similar to the CRISPR-like locus found in *VlpC* gene, with DRs and spacer sequences similar to those detected in **Bangpanwimon *et al*. (2017)**. In the phylogenetic tree, using the CRISPR-like sequences present in this gene, we observed that the strains that corresponded to an African and European origin formed a differentiated cluster with respect to the Asian and Amerindian strains **(Figure 3)**. This fact would indicate that the strains furthest from the African origin, such as those of Asian and Amerindian origin, have undergone a process of greater differentiation. **Duncan *et al*. (2013)** proposed that the different strains of *H. pylori* were subject to different selective pressures depending on their

environmental conditions and according to their phylogeographic origin, and this can lead to the diversification of certain genomic regions, as seems to be the case here **(Duncan *et al*., 2013)**.

The presence of CRISPR-like loci caused changes in the sequence of the genes where they are inserted into, truncating it or varying its sequence close to the insertion point **(Figures S2, S5, S7, and S9)**. In addition, the CRISPR loci themselves showed great variability between strains because DRs and spacers were variable in number, even with reverse positions in several genomes **(Figures S2, S3, S5, and S7)**. These variations indicate recombination phenomena that involve the CRISPR-like locus. In this sense, the CRISPR-like loci could be considered as repetitive sequences involved in intra- and inter- genomic recombination, contributing to the diversity of *H. pylori*. In fact, the variability found between strains, with duplications and deletions within DRs and spacers, could be the result of both types of recombination. In addition, in this work, we showed the presence of a CRISPR-like locus in a region near the breaking point of a large inversion that affects several strains (Shi470 and BM012A), and may therefore be involved in this process (**Figure 4b).** In *Helicobacter* there have been reports about the implications of repeated sequences in this type of rearrangement events **(Kang & Blaser, 2016; Aras *et al*., 2003; Suerbaum & Josenhans, 2007)**. The implication of CRISPR-like loci in the recombination process could also be supported by the presence of a RecR gene, which is implicated in recombination and repair processes **(Mandin *et al*., 2008)**, in the same operon as the *VlpC* gene.

All of these processes would be part of the mechanisms that infer the extreme genome plasticity of *H. pylori* through mutation and recombination intra e inter genomic, exhibiting genetic mosaicism **(Suerbaum *et al*., 1998)**.

Currently, it is hypothesized that degenerated CRISPR-Cas systems, or their individual components, as in this case, could derive into diverse roles in a wide range of processes **(Mojica *et al*., 2005)**. Thus, if a novel function of a CRISPR system, or one of its components, confers a competitive advantage in the environment in which the organism evolved (that is, it is adaptive) its maintenance and propagation in populations could be a direct result of natural selection. The analysis of secondary structure of direct repeat RNA showed that all sites can form stable RNA secondary structure, exhibiting stem-loop structures and low minimum free energies. The presence of these structures would suggest a possible role in recognition-mediated contact between gap targeted RNA, DNA or protein **(Zhao, Yu & Xu, 2018)**.

In fact, it has been shown that orphan CRISPRs loci may be involved in gene regulation. In *Listeria monocytogenes*, orphan CRISPR affected virulence through the FeoAB iron transport

system **(Mandin *et al*., 2007)**. In this sense, the constant presence of this repeated and mutable structure in these genes of *H. pylori*, and more specifically in the *VlpC* gene, which is part of the central genome, could be related to the regulation of its expression, as they are located in the promoter region. The integration of the CRISPR-like structure into the *VlpC* gene would allow the bacteria to be less sensitive to the host defense mechanisms as indicated by **Bangpanwimon *et al*. (2017)**, and would confer the ability to adapt to different stomach areas, facilitating the capacity to adhere to the gastric epithelium **(Harvey *et al*., 2014)**. Similar situations have been described in *Staphylococcus aureus*, in which case the absence of the CRISPR-Cas system conferred the ability to acquire new genes and be more virulent, or as *Enterococus fecalis*, where the modification of their CRISPR-Cas systems made their strains more resistant to antibiotics **(Zuker & Stiegler, 2013).**

Although clustered Cas genes were not detected, and therefore a functional CRISPR-Cas system was also not found, in this work the presence of cas domains in the genome of *H. pylori* was found **(Table S4)**. This presence could signify that the presence of this system is ancestral. This theory could be strengthened by the fact that in other *Helicobacter* species, CRISPR-Cas systems are present and active **(Kersulyte *et al*., 2013; Tomida *et al*., 2017)**. The cas domains in the *H. pylori* genome could be performing other functions. In fact, it has been reported, in *H. pylori*, that the *VapD* protein, associated with a ribonuclease function, is phylogenetically related to Cas2 proteins. Specifically, the HP0315 protein, a member of the VapD family, has a structural similarity to Cas2 and appears to be an evolutionary intermediate between Cas2 and a gene from the Toxin-Antitoxin system **(Kwon *et al*., 2012)**.

The loss of the functional system is also supported by the fact that in the evolutionary process the number of repetitions present in a CRISPR locus depends on the level of decay of the associated genes **(Touchon & Rocha, 2010)**, as is the case of *H. pylori*, in which the number of DRs observed is low and only cas domains are found, which may be remnants of the original system.

From the analysis carried out, the presence of a CRISPR-like locus within several genes of *H. pylori* was demonstrated. The origin and evolution of these types of sequences is still uncertain. However, for the case of the structure found in the *VlpC* gene, data is available that has helped with inferring its evolutionary history. In this sense, when comparing the genomes of different *Helicobacter* species, it was found that the *VlpC* gene was only found in *H. pylori* and *H. cetorum* and with a high degree of similarity. This could indicate that this gene was acquired after the separation of the common ancestor of *H. pylori* and *H. cetorum* from the rest of the species, by duplication from the *vacA* gene **(Foegeding *et al*., 2016)**. After this event, the acquisition of the

CRISPR-like sequences could have taken place in the *VlpC* gene. These structures, as pointed out, are in a state of constant flow **(Marraffini, 2013)**, and therefore they can appear and disappear depending on the selective forces of the environment. During the speciation process of *H. pylori* and *H. cetorum*, the differentiation of CRISPR-like loci occurred between both species. In this sense, it could be said that although the DRs of both species have a high degree of similarity, indicating the common origin, the spacer sequences are variable. It has also been suggested that CRISPR loci can evolve rapidly in some environments, in accordance with the new role played in their antagonistic coevolution **(Westra *et al*., 2016)**.

The CRISPR-like loci in *H. pylori* have evolved independently of those of *H. cetorum* (sympatrically), supporting this type of antagonistic coevolution.

In addition, and due to the high degree of change found in these sequences (**Figures S1A to S1F**), CRISPR-like loci can be used to determine a strain's origin. Different genomic regions have been used for phylogenetic analyses of *H. pylori* as Multi Locus Sequence Typing (MLST), *housekeeping genes and* genes of the central genome **(Falush Stephens & Pritchard, 2003; Falush *et al*., 2003; Yahara *et al*., 2013)**. In our case, the phylogeny, using DRs and spacers of CRISPR-like locus within the *VlpC* gene, groups the strains by geographic origin **(Figure 3)**, relating the African ones with the European ones, separating them from the Asian and Amerindian ones (of more recent origin). This same situation was observed for the CRISPR-like loci of the Poly-E rich poly genes and for one of the hypothetical proteins (with cell division function), with a grouping by geographical origin (**Figure S3, and S4).** For this latter protein, the absence of this gene in all strains of the Asian clade was highlighted. Lastly, the sequences of CRISPR1-like and CRISPR3-like loci did not have a geographical grouping (**Figure S6, and S8)**, showing a process of variation that was independent of the geographical origin.

Conclusiones

1. Mediante el análisis comparativo de 53 genomas completos de cepas de *H. pylori* con distinto origen geográfico hemos determinado que la longitud media del genoma de esta especie es de 1.621.671 ±43.745 nucleótidos y una media de 1.551 ±42 genes estando constiutido el pangenoma de esta especie por 5.134 genes

2. Entre los distintos genomas analizados existe una gran variabilidad tanto en la longitud de nucleótidos desde 1.709.911 a 1.494.183, así como al contenido de genes que va desde 1.640 el de mayor a 1.430 el de menor contenido.

3. La proporción en G/C es del 40% en el conjunto del genoma. La cadena principal presenta un menor número de genes 753 ± 31 que la cadena retrasada con 738 ± 37. Son mayoritarios los genes con GC positivo en ambas cadenas.

4. El conjunto de genes que forman el genoma central de la especie está constituido por 802 genes. Este número se puede considerar normal dentro de las bacterias patógenas obligadas como *H. pylori* adaptadas a un ambiente extremo como es el estómago humano. El genoma variable esta constituido por 3.972 genes.

5. El análisis de variabilidad realizado en estos 53 genomas muestra una tasa alta ya que en los 802 genes del genoma central existen un total de 149.123 SNPs con una media de 186 por gen.

6. Mediante los análisis llevados a cabo para estudiar de la estructura poblacional con los SNPs detectados en el genoma central encontramos que existe una agrupación casi exacta de las cepas por su origen geográfico. Sin embargo es posible detectar que durante el proceso evolutivo de la especie se han producido fenómenos de deriva genética.

7. Del análisis sinténico de los diferentes genomas podemos concluir que las inversiones son los fenómenos más frecuentes que provocan reordenaciones genómicas en *H. pylori*. Esta alta ocurrencia puede venir provocada por varios factores como: secuencias cortas repetidas y elementos mobiles.

8. Hemos comprobado que en determinados segmentos de algunos genes duplicados se producen eventos de evolución concertada acompañado de la acción de la selección. Esto indicaría que dentro de un mismo gen las diferentes regiones que lo conforman 3', intermedia y 5' pueden sufrir diferentes procesos evolutivos.

9. En el conjunto de especies analizadas podemos señalar que el género *Helicobacter* presenta 19 de los 31 sistemas inmunes detectados en bacterias hasta ahora. Existe una gran variabilidad en la presencia o ausencia de sistemas, aunque este dato puede estar sesgado por el número de genomas analizada por especie. No obstante, podemos señalar que todas las especies presentan más de un sistema siendo *H. cetorum* la que más presentaba con un total de ocho.

10. En la especie *H. pylori* el sistema más representado es el RM seguido por AbiL, pVips y AbiE, no presentado sistemas CRISPR-cas.

11. Hemos detectado en los genomas de la especie *H. pylori* la presencia de los denominados CRISPR-like que son secuencias genómicas que presentan la estructura de estos sistemas es decir secuencias palíndrómica cortas intercalada con segmentos espaciadoras pero no hemos detectado la presencia de genes tipo cas que podrían hacer pensar en la funcionalidad del sistema.

12. Estas estructuras se localizan en el gen vlpC que se he relacionado con la resistencia a los antibioticos y de la comparación de sus secuencias se deduce que entre ellas se producen recombinación provocando duplicaciones y deleciones.

Bibliografía

Aihara, E., Closson, C., Matthis, A.L., Schumacher, M. A., Engevik, A.C., Zavros, Y., Ottemann, K.M., & Montrose, M.H. 2014. Motility and chemotaxis mediate the preferential colonization of gastric injury sites by *Helicobacter pylori*. PLoS pathogens, 10(7), e1004275. https://doi.org/10.1371/journal.ppat.1004275

Åberg, A., Gideonsson, P., Vallström, A., Olofsson, A., Öhman, C., Rakhimova, L., Borén, T., Engstrand, L., Brännström, K., & Arnqvist, A. 2014. A repetitive DNA element regulates expression of the *Helicobacter pylori* sialic acid binding adhesin by a rheostat-like mechanism. PLoS pathogens, 10(7), e1004234. https://doi.org/10.1371/journal.ppat.1004234

Albert, T.J., Dailidiene, D., Dailide, G., Norton, J.E., Kalia, A., Richmond, T.A., Molla, M., Singh, J., Green, RD.,_& Berg, DE. 2005. Mutation discovery in bacterial genomes: Metronidazole resistance in *Helicobacter pylori*. Nature Methods, 2(12), 951–953. DOI: 10.1038/nmeth805.

Ahmed, N., Tenguria, S., & Nandanwar, N. 2009. *Helicobacter pylori*--a seasoned pathogen by any other name. Gut pathogens, 1, 24. https://doi.org/10.1186/1757-4749-1-24

Algood, H.M., & Cover, T.L. 2006. *Helicobacter pylori* persistence: an overview of interactions between *H. pylori* and host immune defenses. Clinical microbiology reviews, 19(4), 597–613. https://doi.org/10.1128/CMR.00006-06

Alm, R.A., Ling, L.S., Moir, D.T., et al. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature. 1999 Jan 14;397(6715):176-80. doi: 10.1038/16495.

Almendros, C., Guzmán, N.M., García-Martinez, J., & Mojica, FJM. 2016. Anti-cas spacers in orphan CRISPR4 arrays prevent uptake of active CRISPR-Cas I-F systems. Nature Microbiology, 1(8), 1–8. D

Amorim, I., Smet, A., Alves, O., Teixeira, S., Saraiva, A. L., Taulescu, M., Reis, C., Haesebrouck, F., & Gärtner, F. (2015). Presence and significance of *Helicobacter* spp. in the gastric mucosa of Portuguese dogs. Gut pathogens, 7, 12. https://doi.org/10.1186/s13099-015-0057-1 OI: 10.1038/nmicrobiol.2016.81.

Anba, J., Bidnenko, E., Hillier, A., Ehrlich, D. and Chopin, M.C. 1995. Characterization of the lactococcal abiD1 gene coding for phage abortive infection. Journal of Bacteriology, **177**, 3818–3823. doi: 10/gndc3w

Andersson, A.F., Lindberg, M., Jakobsson, H., Bäckhed, F., Nyrén, P., & Engstrand, L. 2008. Comparative analysis of human gut microbiota by barcoded pyrosequencing. PloS one, 3(7), e2836. https://doi.org/10.1371/journal.pone.0002836

Akopyants, N.S., Clifton, S.W., Kersulyte, D. et al. 1998. Analyses of the cag pathogenicity island of *Helicobacter pylori*. Mol Microbiol. Apr; 28(1):37-53. doi: 10.1046/j.1365-2958.1998.00770.x. PMID: 9593295.

Aras, RA., Fischer, W., Perez-Perez, G.I., Crosatti, M., Ando, T., Haas, R., & Blaser, MJ. 2003. Plasticity of Repetitive DNA Sequences within a Bacterial (Type IV) Secretion System Component. The Journal of Experimental Medicine, 198(9), 1349–1360. DOI: 10.1084/jem.20030381.Backert, S., & Yamaoka, Y. 2016. *Helicobacter* pylori research: From bench to bedside. *Helicobacter pylori* Research: From Bench to Bedside, 1–613.

Aspholm-Hurtig, M., Dailide, G., Lahmann, M., et al. 2004. Functional adaptation of BabA, *the H. pylori* ABO blood group antigen binding adhesin. Science. 2004 Jul 23; 305(5683):519-22. doi: 10.1126/science.1098801. PMID: 15273394.

Bangpanwimon, K., Sottisuporn, J., Mittraparp-arthorn, P., Ueaphatthanaphanich, W., Rattanasupar, A., Pourcel, C., & Vuddhakul, V. 2017. CRISPR-like sequences in *Helicobacter pylori* and application in genotyping. Gut Pathogens, 9, 65. DOI: 10.1186/s13099-017-0215-8.

Barabas, O., Ronning, D.R., Guynet, C., Hickman, A.B., Ton-Hoang, B., Chandler, M., & Dyda, F. (2008). Mechanism of IS200/IS605 family DNA transposases: activation and transposon-directed target site selection. Cell, 132(2), 208–220. https://doi.org/10.1016/j.cell.2007.12.029

Benson, DA., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., & Sayers, EW. 2011. GenBank. Nucleic Acids Research, 39(Database issue), D32–D37. DOI: 10.1093/nar/gkq1079.

Bergsland, K.J., Kao, C., Yu, Y.T.N., Gulati, R. and Snyder, L. 1990. A site in the T4 bacteriophage major head protein gene that can promote the inhibition of all translation in *Escherichia coli*. Journal of Molecular Biology, **213**, 477–494. doi: 10/fv36tb

Bernheim, A., Sorek, R. 2020. The pan-immune system of bacteria: antiviral defence as a community resource. Nat Rev Microbiol. Feb;18(2):113-119. doi: 10.1038/s41579-019-0278-2. Bernheim, A., Millman, A., Ofir, G., Meitav, G., Avraham, C., Shomar, H., Rosenberg, M.M., Tal, N., Melamed, S., Amitai, G., & Sorek, R. 2021. Prokaryotic viperins produce diverse antiviral molecules. Nature, 589(7840), 120–124. https://doi.org/10.1038/s41586-020-2762-2

Biswas, A., Gagnon, JN., Brouns, S.J.J., Fineran, P.C., & Brown, CM. 2013. CRISPRTarget: Bioinformatic prediction and analysis of crRNA targets. RNA Biology, 10(5), 817–827. DOI: 10.4161/rna.24046

Blaser, M. J., & Berg, D. E. 2001. *Helicobacter pylori* genetic diversity and risk of human disease. The Journal of clinical investigation, 107(7), 767–773. https://doi.org/10.1172/JCI12672

Bondy-Denomy, J., & Davidson, A.R. 2014**.** To acquire or resist: the complex biological effects of CRISPR-Cas systems. Trends Microbiol. 22 (2014), pp. 218-225. DOI: 10.1016/j.tim.2014.01.007

Boni, M.F., Posada, D., Feldman, M.W. 2007 An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics 176, 1035 – 1047. (doi:10.1534/genetics.106.068874).

Boocock G.R., Morrison, J.A, Popovic, M., Richards, N, Ellis, L., Durie, P.R, Rommens, J.M. 2003. Mutations in SBDS are associated with Shwachman-Diamond syndrome. Nat Genet.33:97–101. doi: 10.1038/ng1062.

Bouchard, J.D., Dion, E., Bissonnette, F. and Moineau, S. 2002. Characterization of the Two-Component Abortive Phage Infection Mechanism AbiT from *Lactococcus lactis*. Journal of Bacteriology, 184, 6325–6332. doi: 10/btq97w

Borges, A. L., Davidson, A. R., & Bondy-Denomy, J. 2017. The Discovery, Mechanisms, and Evolutionary Impact of Anti-CRISPRs. Annual review of virology, 4(1), 37–59. https://doi.org/10.1146/annurev-virology-101416-041616.

Bower, E., Cooper, L. P., Roberts, G. A., White, J. H., Luyten, Y., Morgan, R. D., & Dryden, D. 2018. A model for the evolution of prokaryotic DNA restriction-modification systems based upon the structural malleability of Type I restriction-modification enzymes. Nucleic acids research, 46(17), 9067–9080. https://doi.org/10.1093/nar/gky760.

Bubendorfer, S., Krebes, J., Yang, I., Hage, E., Schulz, T.F, Bahlawane, C., Didelot, X., Suerbaum, S. 2016. Genome-wide analysis of chromosomal import patterns after natural transformation of *Helicobacter pylori*. Nat Commun. un 22;7:11995. doi: 10.1038/ncomms11995. PMID: 27329939;

Butterer, A., Pernstich, C., Smith, R. M., Sobott, F., Szczelkun, M. D., & Tóth, J. 2014. Type III restriction endonucleases are heterotrimeric: comprising one helicase-nuclease subunit and a dimeric methyltransferase that binds only one specific DNA. Nucleic acids research, 42(8), 5139–5150. https://doi.org/10.1093/nar/gku122

Burstein, D., Sun, CL., Brown, CT., Sharon, I., Anantharaman, K., Probst, A.J., Thomas, BC., & Banfield, JF. 2016. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nature Communications, 7, 1–8. DOI: 10.1038/ncomms10613

Burstein, D., Harrington, LB., Strutt, SC., Probst, AJ., Anantharaman, K., Thomas, BC., Doudna, JA., & Banfield, JF. 2017. New CRISPR-Cas systems from uncultivated microbes. Nature, 542(7640), 237–241. DOI: 10.1038/nature21059

Carson, A. R., & Scherer, S. W. 2009. Identifying concerted evolution and gene conversion in mammalian gene pairs lasting over 100 million years. BMC evolutionary biology, 9, 156. https://doi.org/10.1186/1471-2148-9-156

Castillo-Cobián. 2007. Ecologia molecular. Capítulo 1 La selección natural a nivel molecular. Primera edición. Instituto Nacional de Ecología, Semarnat.

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P. C., Vergnaud, G., Gautheret, D. and Pourcel, C. 2018. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res, **46**, W246–W251. doi: 10/ggdjdf

Contreras-Moreira, B., & Vinuesa, P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Applied and environmental microbiology, 79(24), 7696–7701. https://doi.org/10.1128/AEM.02411-13

Cullen, T.W., Giles, D.K., Wolf, L. N., Ecobichon, C., Boneca, I.G., & Trent, M. S. 2011. *Helicobacter pylori* versus the host: remodeling of the bacterial outer membrane is required for survival in the gastric mucosa. PLoS pathogens, 7(12), e1002454. https://doi.org/10.1371/journal.ppat.1002454

Cram, D., Ray, A. and Skurray, R. 1984. Molecular analysis of F plasmid pif region specifying abortive infection of T7 phage. Mol Gen Genet, **197**, 137–142. doi: 10/fg7s8g

Cluzel, P.J., Chopin, A., Ehrlich, S.D. and Chopin, M.C. 1991. Phage abortive infection mechanism from *Lactococcus lactis* subsp. lactis, expression of which is mediated by an Iso-ISS1 element. Applied and Environmental Microbiology, 57, 3547–3551. doi: 10/gndc3t

Chanto, G., Occhialini, A., Gras N., Alm, RA., Mégraud, F., Marais, A. 2002. Identification of strain-specific located oustside the plasticity zone in nine clinical isolates of *Helicobacter pylori*. Microbiology, 148, 3671-3680. DOI: 10.1099/00221287-148-11-367Czapinska, H., Kowalska, M.,

Zagorskaite, E., Manakova, E., Slyvka, A., Xu, S. Y., Siksnys, V., Sasnauskas, G., & Bochtler, M. 2018. Activity and structure of EcoKMcrA. Nucleic acids research, 46(18), 9829–9841. https://doi.org/10.1093/nar/gky731

Dai, G., Su, P., Allison, G. E., Geller, B. L., Zhu, P., Kim, W.S. and Dunn, N.W. 2001. Molecular Characterization of a New Abortive Infection System (AbiU) from *Lactococcus lactis* LL51-1. Applied and Environmental Microbiology, 67, 5225–5232. doi: 10/cwbxdg

Darling, A. C., Mau, B., Blattner, F. R., & Perna, N. T. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome research, 14(7), 1394–1403. https://doi.org/10.1101/gr.2289704

Darling, A. E., Miklós, I., & Ragan, M. A. 2008. Dynamics of genome rearrangement in bacterial populations. PLoS genetics, 4(7), e1000128. https://doi.org/10.1371/journal.pgen.1000128

Darling, A. E., Mau, B., & Perna, N. T. (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PloS one, 5(6), e11147. https://doi.org/10.1371/journal.pone.0011147.

Draper, J.L., Hansen, L. M., Bernick, D.L., Abedrabbo, S., Underwood, J.G., Kong, N., Huang, B. C., Weis, A.M., Weimer, B.C., van Vliet, A.H., Pourmand, N., Solnick, J.V., Karplus, K., & Ottemann, K. M. 2017. Fallacy of the Unique Genome: Sequence Diversity within Single *Helicobacter pylori* Strains. mBio, 8(1), e02321-16. https://doi.org/10.1128/mBio.02321-16

Delaney, N.F., Balenger, S., Bonneaud, C., Marx, C.J., Hill, G.E., Ferguson-Noel, N., Tsai, P., Rodrigo, A., & Edwards, SV. 2012. Ultrafast Evolution and Loss of CRISPRs Following a Host Shift in a Novel Wildlife Pathogen, *Mycoplasma gallisepticum*. PLoS Genetics, 8(2), e1002511. DOI: 10.1371/annotation/b5608bc6-aa54-40a7-b246-51fa7bc4a9db.

Deng, Y. M., Liu, C.Q. and W. Dunn, N. 1999. Genetic organization and functional analysis of a novel phage abortive infection system, AbiL, from *Lactococcus lactis*. Journal of Biotechnology, 67, 135–149. doi: 10/bwc9k8

Deng, Y.M., Harvey, M.L., Liu, C Q. and Dunn, N.W. 1997. A novel plasmid-encoded phage abortive infection system from *Lactococcus lactis biovar*. diacetylactis. FEMS Microbiology Letters, 146, 149–154. doi: 10/d24tcq

Deveau, H., Barrangou, R., Garneau, J. E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D. A., Horvath, P., & Moineau, S. 2008. Phage response to CRISPR-encoded resistance in

Streptococcus thermophilus. Journal of bacteriology, 190(4), 1390–1400. https://doi.org/10.1128/JB.01412-07

Dinsmore, P.K. and Klaenhammer, T. R. 1994. Phenotypic Consequences of Altering the Copy Number of abiA, a Gene Responsible for Aborting Bacteriophage Infections in *Lactococcus lactis.* Applied and Environmental Microbiology, **60**, 1129–1136. doi: 10/gndc3s

Domingues, S., Chopin, A., Ehrlich, S.D. and Chopin, M.C. 2004. The Lactococcal Abortive Phage Infection System AbiP Prevents both Phage DNA Replication and Temporal Transcription Switch. Journal of Bacteriology, **186**, 713–721. doi: 10/bjjwc6

Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G. and Sorek, R. 2018. Systematic discovery of antiphage defense systems in the microbial pangenome. Science, 359, eaar4120. doi: 10/ggqhzm

Du, M.Z., Zhang, C., Wang, H., Liu, S., Wei, W., & Guo, F. B. 2018. The GC Content as a Main Factor Shaping the Amino Acid Usage During Bacterial Evolution Process. Frontiers in microbiology, 9, 2948. https://doi.org/10.3389/fmicb.2018.02948

Duncan, SS., Valk, PL., McClain, MS., Shaffer, CL., Metcalf, JA., Bordenstein, SR., & Cover, TL. 2013. Comparative Genomic Analysis of East Asian and Non-Asian *Helicobacter pylori* Strains Identifies Rapidly Evolving Genes. PLoS ONE, 8(1), e55120. DOI: 10.1371/journal.pone.0055120.

Durmaz, E. and Klaenhammer, T.R. 2007. Abortive Phage Resistance Mechanism AbiZ Speeds the Lysis Clock To Cause Premature Lysis of Phage-Infected *Lactococcus lactis*. Journal of Bacteriology, 189, 1417–1425. doi: 10/dnndhw

Dy, R.L., Przybilski, R., Semeijn, K., Salmond, G.P.C. and Fineran, P.C. 2014. A widespread bacteriophage abortive infection system functions through a Type IV toxin–antitoxin mechanism. Nucleic Acids Research, 42, 4590–4605. doi: 10/f5zkmw

Eaton, K.A., & Krakowka, S. 1994. Effect of gastric pH on urease-dependent colonization of gnotobiotic piglets by *Helicobacter pylori*. Infection and immunity, 62(9), 3604–3607. https://doi.org/10.1128/iai.62.9.3604-3607.1994

Edgar, RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5), 1792–1797. DOI: 10.1093/nar/gkh340.

Emond, E., Holler, B.J., Boucher, I., Vandenbergh, P.A., Vedamuthu, E.R., Kondo, J.K. and Moineau, S. 1997. Phenotypic and genetic characterization of the bacteriophage abortive infection mechanism AbiK from *Lactococcus lactis*. Applied and Environmental Microbiology, **63**, 1274–1283. doi: 10/gndc4n

Emond, E., Dion, E., Walker, S.A., Vedamuthu, E.R., Kondo, J.K. and Moineau, S. 1998 AbiQ, an Abortive Infection Mechanism from *Lactococcus lactis*. Applied and Environmental Microbiology, 64, 4748–4756. doi: 10/gndc4p

Eddy, S. 1998. Profile hidden Markov models. Bioinformatics, 14(9), 755–763. DOI: 10.1093/bioinformatics/14.9.755

Falush, D., Stephens, M., Pritchard, JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 164:1567-87.

Falush, D., Wirth, T., Linz, B., Pritchard, JK., Stephens, M., Kidd, M., Blaser, MJ., Graham, DY., Vacher, S., Perez-Perez, GI., Yamaoka, Y., Mégraud, F., Otto, K., Reichard, U., Katzowitsch, E., Wang, X., Achtman, M., & Suerbaum, S. 2003. Traces of human migrations in *Helicobacter pylori* populations. Science; 299:1582-5. DOI: 10.1126/science.1080857

Farzi, N., Yadegar, A., Sadeghi, A., Asadzadeh Aghdaei, H., Marian Smith, S., Raymond, J., Suzuki, H., & Zali, M. R. 2019. High Prevalence of Antibiotic Resistance in Iranian *Helicobacter pylori* Isolates: Importance of Functional and Mutational Analysis of Resistance Genes and Virulence Genotyping. Journal of clinical medicine, 8(11), 2004. https://doi.org/10.3390/jcm8112004

Fehlings, M., Drobbe, L., Moos, V., Renner Viveros, P., Hagen, J., Beigier-Bompadre, M., Pang, E., Belogolova, E., Churin, Y., Schneider, T., Meyer, T. F., Aebischer, T., & Ignatius, R. 2012. Comparative analysis of the interaction of *Helicobacter pylori* with human dendritic cells, macrophages, and monocytes. Infection and immunity, 80(8), 2724–2734. https://doi.org/10.1128/IAI.00381-12

Foegeding, N.J., Caston, R.R., McClain, MS., Ohi, MD., & Cover, T.L. 2016. An Overview of *Helicobacter pylori* vacA Toxin Biology. Toxins, 8(6), 173. DOI: 10.3390/toxins8060173

Forsberg K.J., Malik, H.S. Microbial Genomics: 2018.The Expanding Universe of Bacterial Defense Systems. Curr Biol. Apr 23;_28(8):R361-R364. doi: 10.1016/j.cub.2018.02.053. PMID: 29689213.

Fouts, D.E., Brinkac, L., Beck, E., Inman, J., & Sutton, G. 2012. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic acids research, 40(22), e172. https://doi.org/10.1093/nar/gks757

Gao, L., Altae-Tran, H., Böhning, F., Makarova, K.S., Segel, M., Schmid-Burgk, J.L., Koob, J., Wolf, Y. I., Koonin, E.V. and Zhang, F. 2020. Diverse enzymatic activities mediate antiviral immunity in prokaryotes. Science, **369**, 1077–1084. doi: 10/gpsx

Gangwer, K.A., Shaffer, CL., Suerbaum, S., Lacy, DB., Cover, TL., & Bordenstein, SR. 2010. Molecular evolution of the *Helicobacter pylori* vacuolating toxin gene vacA. Journal of bacteriology, 192(23), 6126-35. DOI: 10.1128/JB.01081-10.

Garvey, P., Fitzgerald, G.F. and Hill, C. 1995. Cloning and DNA sequence analysis of two abortive infection phage resistance determinants from the lactococcal plasmid pNP40. Applied and Environmental Microbiology, 61, 4321–4328. doi: 10/gndc3x.

García-Gutiérrez, E., Almendros, C., Mojica, F.J.M., Guzmán, N.M., & García-Martínez, J. 2015. CRISPR Content Correlates with the Pathogenic Potential of *Escherichia coli*. PLoS ONE, 10(7), e0131935. DOI: 10.1371/journal.pone.0131935.

Gangwer, K. A., Mushrush, D. J., Stauff, D. L., Spiller, B., McClain, M. S., Cover, T. L., & Lacy, D. B. 2007. Crystal structure of the *Helicobacter pylori* vacuolating toxin p55 domain. Proceedings of the National Academy of Sciences of the United States of America, 104(41), 16293–16298. https://doi.org/10.1073/pnas.0707447104

Gibbs, M.J., Armstrong, J.S., Gibbs, A.J. 2000 Sisterscanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics 16, 573–582. (doi:10.1093/bioinformatics/16.7.573)

Grissa, I., Vergnaud, G., & Pourcel, C. 2008. CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats. Nucleic Acids Research, 36(Web Server issue), W145–W148. DOI: 10.1093/nar/gkn228.

Goldfarb, T., Sberro, H., Weinstock, E., Cohen, O., Doron, S., Charpak-Amikam, Y., Afik, S., Ofir, G., & Sorek, R. 2015. BREX is a novel phage resistance system widespread in microbial genomes. The EMBO journal, 34(2), 169–183. https://doi.org/10.15252/embj.201489455

Goh, K. L., Chan, W. K., Shiota, S., & Yamaoka, Y. 2011. Epidemiology of *Helicobacter pylori* infection and public health implications. *Helicobacter*, 16 Suppl 1(0 1), 1–9. https://doi.org/10.1111/j.1523-5378.2011.00874.x

Grissa, I., Vergnaud, G., & Pourcel, C. 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Research, 35(Web Server issue), W52–W57. DOI: 10.1093/nar/gkm360.

Haaber, J., Moineau, S., Fortier, L.C. and Hammer, K. 2008. AbiV, a Novel Antiphage Abortive Infection Mechanism on the Chromosome of *Lactococcus lactis* subsp. cremoris MG1363. Applied and Environmental Microbiology, 74, 6528–6537. doi: 10/cnwcq7

Haley, K.P., & Gaddy, J.A. 2015. Metalloregulation of *Helicobacter pylori* physiology and pathogenesis. Frontiers in microbiology, 6, 911. https://doi.org/10.3389/fmicb.2015.00911

Hansen, T.F, Carter AJ, Chiu CH. 2000. Gene conversion may aid adaptive peak shifts. J Theor Biol. 207:495–511. doi: 10.1006/jtbi.2000.2189.

Haft, D.H., Selengut, J., Mongodin, E.F., & Nelson, KE. 2005. A guild of 45 CRISPR-associated (Cas) protein families .and multiple CRISPR/cas subtypes exist in prokaryotic genomes. PloS Computational Biology, 1(6), 0474–0483. DOI: 10.1371/journal.pcbi.0010060.

Hansen, T.F., Carter, J.R., Chiu, C.H. 2000. Gene conversión maya id adaptative peak shifts. Journal of theoretical biology. Volume 207, Issue 4, 21, Pages 495-511.

Harvey, V.C., Acio, CR., Bredehoft, A.K., Zhu, L., Hallinger, DR., Quinlivan-Repasi, V., Quinlivan-Repasi, V., Harvey, S.E., & Forsyth, MH. 2014. Repetitive Sequence Variations in the Promoter Region of the Adhesin-Encoding Gene sabA of *Helicobacter pylori* Affect Transcription. Journal of Bacteriology, 196(19), 3421–3429. DOI: 10.1128/JB.01956-14.

Hatakeyama M. 2017. Structure and function of *Helicobacter pylori* CagA, the first-identified bacterial protein involved in human cancer. Proceedings of the Japan Academy. Series B, Physical and biological sciences, 93(4), 196–219. https://doi.org/10.2183/pjab.93.013

Hatoum-Aslan, A., & Marraffini, LA. 2014. Impact of CRISPR immunity on the emergence and virulence of bacterial pathogens. Current Opinion in Microbiology, 0, 82–90. DOI: 10.1016/j.mib.2013.12.001.

Hille, F., Richter, H., Wong, S.P, Bratovič, M., Ressel, S., Charpentier, E. The Biology of CRISPR-Cas: Backward and Forward. Cell. 2018 Mar 8; 172(6):1239-1259. doi: 10.1016/j.cell.2017.11.032.

He, S., Guynet, C., Siguier, P., Hickman, A.B., Dyda, F., Chandler, M., & Ton-Hoang, B. 2013. IS200/IS605 family single-strand transposition: mechanism of IS608 strand transfer. Nucleic acids research, 41(5), 3302–3313. https://doi.org/10.1093/nar/gkt014

Holmes, E.C., Worobey, M., Rambaut, A. 1999 Phylogenetic evidence for recombination in dengue virus. Mol. Biol. Evol. 16, 405 – 409. doi:10.1093/ oxfordjournals.molbev.a026121.

Hudson, R. R., & Kaplan, N. L. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics, 111(1), 147–164. https://doi.org/10.1093/genetics/111.1.147.

Huerta-Cepas, J., Serra, F., & Bork, P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Molecular biology and evolution, 33(6), 1635–1638. https://doi.org/10.1093/molbev/msw046

Jabbar, M.A. and Snyder, L. 1984. Genetic and physiological studies of an *Escherichia coli* locus that restricts polynucleotide kinase- and RNA ligase-deficient mutants of bacteriophage T4. Journal of Virology, **51**, 522–529. doi: 10/gndc4t

Junaid, M., Al-Gubare, Muhammad Yousef, S.A., Ubol, M.N., et al. 2014. "Sequence and Apoptotic Activity of VacA Cytotoxin Cloned from a *Helicobacter pylori* Thai Clinical Isolate", BioMed Research International, vol. 2014, Article ID 398350, 8 pages, 2014. https://doi.org/10.1155/2014/398350

Kao, C.Y., Sheu, B.S., & Wu, J. J. 2016. *Helicobacter pylori* infection: An overview of bacterial virulence factors and pathogenesis. Biomedical journal, 39(1), 14–23. https://doi.org/10.1016/j.bj.2015.06.002

Kao, J.Y., Zhang, M., Miller, M.J., Mills, J.C., Wang, B., Liu, M., Eaton, K.A., Zou, W., Berndt, B.E., Cole, T.S., Takeuchi, T., Owyang, S.Y., & Luther, J. 2010. *Helicobacter pylori* immune escape is mediated by dendritic cell-induced Treg skewing and Th17 suppression in mice. Gastroenterology, 138(3), 1046–1054. https://doi.org/10.1053/j.gastro.2009.11.043

Kabir, S. 2009. Effect of *Helicobacter* pylori eradication on incidence of gastric cancer in human and animal models: underlying biochemical and molecular events. *Helicobacter*. Jun;14(3):159-71. doi: 10.1111/j.1523-5378.2009.00677.x

Kawai, M., Furuta, Y., Yahara, K., Tsuru, T., Oshima, K., Handa, N., Takahashi, N., Yoshida, M., Azuma, T., Hattori, M., Uchiyama, I., & Kobayashi, I. (2011). Evolution in an oncogenic bacterial

species with extreme genome plasticity: *Helicobacter pylori* East Asian genomes. BMC microbiology, 11, 104. https://doi.org/10.1186/1471-2180-11-104.

Kang, J., & Blaser, MJ. 2006. Bacterial populations as perfect gases: Genomic integrity and diversification tensions in *Helicobacter pylori*. Nature Reviews Microbiology, 4(11), 826–836. DOI: 10.1038/nrmicro1528

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution, 30(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton, B., Meintjes, P., & Drummond, A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics, 28(12), 1647–1649. DOI: 10.1093/bioinformatics/bts199

Kennemann, L., Brenneke, B., Andres, S., Engstrand, L., Meyer, T.F., Aebischer, T., Josenhans, C., & Suerbaum, S. 2012. In vivo sequence variation in HopZ, a phase-variable outer membrane protein of *Helicobacter pylori*. Infection and immunity, 80(12), 4364–4373. https://doi.org/10.1128/IAI.00977-12

Kersulyte, D., Rossi, M., & Berg, DE. 2013. Sequence Divergence and Conservation in Genomes of *Helicobacter cetorum* Strains from a Dolphin and a Whale. PLoS ONE, 8(12), e83177. DOI: 10.1371/journal.pone.0083177.

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparitive studies of nucleotide sequences. J. Mol. Evol., 16 (1980), pp. 111-120

Koonin, E.V., Makarova, K.S., & Zhang, F. 2017. Diversity, classification and evolution of CRISPR-Cas systems. Current Opinion in Microbiology, 37, 67–78. DOI: 10.1016/j.mib.2017.05.008.

Koonin, E. V., Makarova, K. S., & Wolf, Y. I. 2017. Evolutionary Genomics of Defense Systems in Archaea and Bacteria. Annual review of microbiology, 71, 233–261. https://doi.org/10.1146/annurev-micro-090816-093830

Kazlauskiene, M., Kostiuk, G., Venclovas, Č., Tamulaitis, G., Siksnys, V. A. 2017. cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. Science. 2Aug 11;357 (6351):605-609. doi: 10.1126/science.aao0100.

Kersulyte, D., Lee, W., Subramaniam, D., Anant, S., Herrera, P., Cabrera, L., Balqui, J., Barabas, O., Kalia, A., Gilman, R. H., & Berg, D. E. 200). *Helicobacter Pylori's* plasticity zones are novel transposable elements. PloS one, 4(9), e6859. https://doi.org/10.1371/journal.pone.0006859

Kim, J., Kim, N., Park, J.H. et al. 2106. Analysis of Gastric Microbiota by Pyrosequencing: Minor Role of Bacteria Other Than *Helicobacter pylori* in the Gastric Carcinogenesis. *Helicobacter*. 2016 Oct; 21(5):364-74. doi: 10.1111/hel.12293.

Kwon, AR., Kim, JH., Park, .J., Lee, KY., Min, YH., Im, H., Lee, I., Lee, KY., & Lee, BJ. 2012. Structural and biochemical characterization of HP0315 from *Helicobacter pylori* as a VapD protein with an endoribonuclease activity. Nucleic Acids Research, 40(9), 4216–4228. DOI: 10.1093/nar/gkr1305

Kondrashov, F. A., Gurbich., T. A. & Vlasov, P. K.2007. Selection for functional uniformity of tuf duplicates in gamma-proteobacteria. Trends Genet. 23, 215–218.

Kristensen, D. M., Kannan, L., Coleman, M. K., Wolf, Y. I., Sorokin, A., Koonin, E. V., & Mushegian, A. 2010. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. Bioinformatics (Oxford, England), 26(12), 1481–1487. https://doi.org/10.1093/bioinformatics/btq229

Kronheim, S., Daniel-Ivad, M., Duan, Z., Hwang, S., Wong, A.I., Mantel, I., Nodwell, J.R., Maxwell, K.L. 2018. A chemical defence against phage infection. Nature. Dec; 564 (7735):283-286. doi: 10.1038/s41586-018-0767-x. Epub 2018 Dec 5. PMID: 30518855.

Kumar, S., Stecher G., & Tamura, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets, Molecular Biology and Evolution, Volume 33, Issue 7, 1 July 2016, Pages 1870–1874. DOI: 10.1093/molbev/msw054.

Labrie, S., Samson, J. & Moineau, S. 2010. Bacteriophage resistance mechanisms. Nat Rev Microbiol 8, 317–327. https://doi.org/10.1038/nrmicro2315

Lang, K.S., & Merrikh, H. 2018. The Clash of Macromolecular Titans: Replication-Transcription Conflicts in Bacteria. Annual review of microbiology, 72, 71–88. https://doi.org/10.1146/annurev-micro-090817-062514

Lara-Ramírez, EE., Segura-Cabrera, A., Guo, X., Yu, G., García-Pérez, CA., & Rodríguez-Pérez, M. A. 2011. New implications on genomic adaptation derived from the *Helicobacter pylori* genome comparison. PloS one, 6(2). DOI: 10.1371/journal.pone.0017300.

Lathe, W. C. & Bork, P. 2001. Evolution of tuf genes: ancient duplication, differential loss and gene conversion. FEBS Lett. 502, 113–11

Lassalle, F., Périan, S., Bataillon, T., Nesme, X., Duret, L., & Daubin, V. 2015. GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS genetics, 11(2), e1004941. https://doi.org/10.1371/journal.pgen.1004941

Li, L., Stoeckert, C.J., Jr, & Roos, D.S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research, 13(9), 2178–2189. https://doi.org/10.1101/gr.1224503

Leighton, J., Payne, Thomas C Todeschini, Yi Wu, Benjamin J Perry, Clive W Ronson, Peter C Fineran, Franklin L Nobrega, Simon A Jackson, Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types, Nucleic Acids Research, Volume 49, Issue 19, 8 November 2021, Pages 10868–10878, https://doi.org/10.1093/nar/gkab883

Li, L., Stoeckert, C. J., Jr, & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research, 13(9), 2178–2189. https://doi.org/10.1101/gr.1224503.

Liao, D. Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. 2000. J Mol Evol. Oct;51(4):305-17. doi: 10.1007/s002390010093. PMID: 11040282

Iida S, Streiff MB, Bickle TA, Arber W. 1987. Two DNA antirestriction systems of bacteriophage P1, darA, and darB: characterization of darA- phages. Virology. Mar; 157(1):156-66. doi: 10.1016/0042-6822(87)90324-2.

Lindahl, G., Sironi, G., Bialy, H. and Calendar, R. 1970. Bacteriophage Lambda; Abortive Infection of Bacteria Lysogenic for Phage P2. PNAS, 66, 587–594. doi: 10/fkgq7x

Makarova, K.S., Wolf, Y.I, Koonin, E.V. 2013. Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res. 2013 Apr; 41(8):4360-77. doi: 10.1093/nar/gkt157.

Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P., et al. 2020. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol, 18, 67–83. doi: 10/ggkfgj

Makarova, K.S., Timinskas, A., Wolf, Y.I., Gussow, A.B., Siksnys, V., Venclovas, Č. and Koonin, E.V. 2020. Evolutionary and functional classification of the CARF domain superfamily, key sensors in prokaryotic antivirus defense. Nucleic Acids Res, 48, 8828–8847. doi: 10/gg7qx6

McLandsborough, L.A., Kolaetis, K. M., Requena, T. and McKay, L. L. 1995. Cloning and characterization of the abortive infection genetic determinant abiD isolated from pBF61 of *Lactococcus lactis subsp. lactis* KR5. Applied and Environmental Microbiology, 61, 2023–2026. doi: 10/gndc4j

Merrikh, C.N., & Merrikh, H. 2018. Gene inversion potentiates bacterial evolvability and virulence. Nature communications, 9(1), 4662. https://doi.org/10.1038/s41467-018-07110-3

Moyat, M., & Velin, D. 2014. Immune responses to *Helicobacter pylori* infection. World journal of gastroenterology, 20(19), 5583–5593. https://doi.org/10.3748/wjg.v20.i19.5583

Millman, A., Melamed, S., Amitai, G. and Sorek, R. 2020. Diversity and classification of cyclic-oligonucleotide-based anti-phage signalling systems. Nature Microbiology, 5, 1608–1615. doi: 10/gg84nk

Mobley, H.L., Island, M.D., & Hausinger, R.P. 1995. Molecular biology of microbial ureases. Microbiological reviews, 59(3), 451–480. https://doi.org/10.1128/mr.59.3.451-480.1995

Mojica, F.J., Díez-Villaseñor, C., García-Martínez, J. 2005. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. J Mol Evol, 60: 174–82. DOI: 10.1007/s00239-004-0046-3.

Mandin, P., Repoila, F., Vergassola, M., Geissmann, T., & Cossart, P. 2007. Identification of new noncoding RNAs in *Listeria monocytogenes* &prediction of mRNA targets. Nucleic Acids Research, 35(3), 962–974. DOI: 10.1093/nar/gkl1096

Marsin, S., Mathieu, A., Kortulewski, T., Guérois, R., & Radicella, JP. 2008. Unveiling Novel RecO Distant Orthologues Involved in Homologous Recombination. PLoS Genetics, 4(8), e1000146. DOI: 10.1371/journal.pgen.1000146.

Martin, D.P., Posada, D., Crandal,l K.A., Williamson, C. 2005 A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res. Hum. Retroviruses 21, 98–102. (doi:10.1089/aid.2005.21.98)

Marraffini, LA. 2013. CRISPR-Cas Immunity against Phages: Its Effects on the Evolution and Survival of Bacterial Pathogens. PLoS Pathog 9(12): e1003765. DOI: 10.1371/journal.ppat.1003765.

.Mansai, S. P., & Innan, H. 2010. The power of the methods for detecting interlocus gene conversion. Genetics, 184(2), 517–527. https://doi.org/10.1534/genetics.109.111161.

Marshall, B.J, Warren, J.R. 1984. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. Lancet. Jun 16;1(8390):1311-5. doi: 10.1016/s0140-6736(84)91816-6.

Marshall, B.J., Armstrong, J.A., McGechie, D.B. Glancy, R.J. 1985. Attempt to fulfill Koch´s postulate for pyloric *Campylobacter* Med J Aust 142:436-439.

Martin, D.P, Arvind Varsani. Roumagnac, P., Botha, G., Maslamoney, S., Schwab, T., Kelz, Ze.na, Kumar V., Murrell., B. RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets, Virus Evolution, Volume 7, Issue 1, January 2021, veaa087, https://doi.org/10.1093/ve/veaa087.

Meinersmann, J., Hiett, K.L. 2000. Concerted evolution of duplicate fla genes in *Campylobacter*. Microbiology (Reading). Sep; 146 (Pt 9):2283-2290. doi: 10.1099/00221287-146-9-2283. PMID: 10974116.

Merrikh, H., Zhang, Y., Grossman, A. D., & Wang, J. D. 2012. Replication-transcription conflicts in bacteria. Nature reviews. Microbiology, 10(7), 449–458. https://doi.org/10.1038/nrmicro2800

Moodley, Y., Linz, B., Bond, R.P., Nieuwoudt, M., Soodyall, H., Schlebusch, C.M., Bernhöft, S., Hale, J., Suerbaum, S., Mugisha, L., van der Merwe, S.W., Achtman, M., 2012. Age of the association between *Helicobacter pylori* and man. PLoS Pathogens 8(5), e1002693. doi:10.1371/journal.ppat.1002693

Murray N.E. 2000. Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). Microbiology and molecular biology reviews : MMBR, 64(2), 412–434. https://doi.org/10.1128/MMBR.64.2.412-434.2000

Mruk. I., Kobayashi, I. To be or not to be: regulation of restriction–modification systems and other toxin–antitoxin systems, Nucleic Acids Research, Volume 42, Issue 1, 1 January 2014, Pages 70–86, https://doi.org/10.1093/nar/gkt711

Niewoehner, O., Garcia-Doval, C., Rostøl, J.T, Berk, C., Schwede, F., Bigler, L., Hall, J., Marraffini. L.A., Jinek, M. 2017. Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. Nature. Aug 31;548 (7669):543-548. doi: 10.1038/nature23467.

Niehus, E., Gressmann, H., Ye., Schlapbach, R., Dehio, M., Dehio, C., Stack, A., Meyer, T.F, Suerbaum, S., Josenhans, C. 2004. Genome-wide analysis of transcriptional hierarchy and feedback regulation in the flagellar system of *Helicobacter pylori*. Mol Microbiol. May;52(4):947-61. doi: 10.1111/j.1365-2958.2004.04006.x.

Nelson, M., Raschke, E., & McClelland, M. 1993. Effect of site-specific methylation on restriction endonucleases and DNA modification methyltransferases. Nucleic acids research, 21(13), 3139–3154. https://doi.org/10.1093/nar/21.13.3139

Noureen, M., Tada, I., Kawashima, T., & Arita, M. 2019. Rearrangement analysis of multiple bacterial genomes. BMC bioinformatics, 20(Suppl 23), 631. https://doi.org/10.1186/s12859-019-3293-4

Nei, M and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions Mol. Biol. Evol., 3 (1986), pp. 418-426.

Nei, M., & Rooney, A.P. 2005. Concerted and birth-and-death evolution of multigene families. Annual review of genetics, 39, 121–152. https://doi.org/10.1146/annurev.genet.39.073003.112240.

Nelson, M., Raschke, E., & McClelland, M. 1993. Effect of site-specific methylation on restriction endonucleases and DNA modification methyltransferases. Nucleic acids research, 21(13), 3139–3154. https://doi.org/10.1093/nar/21.13.3139

O'Connor, L., Coffey, A., Daly, C. and Fitzgerald, G.F. 1996. AbiG, a genotypically novel abortive infection mechanism encoded by plasmid pCI750 of *Lactococcus lactis subsp. cremoris* UC653. Applied and Environmental Microbiology, 62, 3075–3082. doi: 10/gndc4m.

Oliveira, P. H., Touchon, M., & Rocha, E. P. 2014. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. Nucleic acids research, 42(16), 10618–10631. https://doi.org/10.1093/nar/gku734

Oleastro., M., Cordeir., R., Ménard. 2009. Allelic diversity and phylogeny of homB, a novel co-virulence marker of *Helicobacter pylori*. BMC Microbiol 9, 248 https://doi.org/10.1186/147.

Parma, D.H., Snyder, M., Sobolevski, S., Nawroz, M., Brody, E. and Gold, L. 1992. The Rex system of bacteriophage lambda: tolerance and altruistic cell death. Genes Dev., 6, 497–510. doi: 10/b9xpsb

Parreira, R., Ehrlich, S.D. and Chopin, M.C. 1996. Dramatic decay of phage transcripts in lactococcal cells carrying the abortive infection determinant AbiB. Molecular Microbiology, 19, 221–230. doi: 10/b835bf

Pertea, M., Ayanbule, K., Smedinghoff, M., & Salzberg, SL. 2009. OperonDB: a comprehensive database of predicted operons in microbial genomes. Nucleic Acids Research, 37(Database issue), D479–D482. DOI: 10.1093/nar/gkn784

Owen, S.V., Wenner, N., Dulberger, C.L., Rodwell, E.V., Bowers-Barnard, A., Quinones-Olvera, N., Rigden, D.J., Rubin, E.J., Garner, E.C., Baym, M., et al. 2021. Prophages encode phage-defense systems with cognate self-immunity. Cell Host & Microbe, 29, 1620-1633. doi: 10/g29b

Padidam, M., Sawyer, S., Fauquet, C.M. 1999 Possible emergence of new geminiviruses by frequent recombination. Virology 265, 218– 225.(doi:10. 1006/viro.1999.0056.

Posada, D., Crandall, K.A. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl Acad. Sci. USA 98, 13 757– 13 762. (doi:10.1073/pnas.241370698).

Pflock, M., Finsterer, N., Joseph, B., Mollenkopf, H., Meyer, T. F., & Beier, D. 2006. Characterization of the ArsRS regulon of *Helicobacter pylori*, involved in acid adaptation. Journal of bacteriology, 188(10), 3449–3462. https://doi.org/10.1128/JB.188.10.3449-3462.2006

Prévots, F., Daloyau, M., Bonin, O., Dumont, X. and Tolou, S. 1996. Cloning and sequencing of the novel abortive infection gene abiH of *Lactococcus lactis ssp. lactis biovar. diacetylactis* S94. FEMS Microbiology Letters, 142, 295–299. doi: 10/dvjcb5.

Prevots, F. and Ritzenthaler, P. 1998. Complete Sequence of the New Lactococcal Abortive Phage Resistance Gene abiO. Journal of Dairy Science, 81, 1483–1485. doi: 10/cg69rg

Prévots, F., Tolou, S., Delpech, B., Kaghad, M. and Daloyau, M. 1998. Nucleotide sequence and analysis of the new chromosomal abortive infection gene abiN of *Lactococcus lactis subsp. cremoris* S114. FEMS Microbiology Letters, 159, 331–336. doi: 10/c94sd6

Price, M. N., Dehal, P.S., & Arkin, A. P. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PloS one, 5(3), e9490. https://doi.org/10.1371/journal.pone.0009490.

Pride, D.T., Blaser, M.J. 2002. Concerted evolution between duplicated genetic elements in *Helicobacter pylori*. J Mol Biol. Feb 22; 316(3):629-42. doi: 10.1006/jmbi.2001.5311. PMID: 11866522.

Posselt, G., Backert, S., & Wessler, S. 2013. The functional interplay of *Helicobacter pylori* factors with gastric epithelial cells induces a multi-step process in pathogenesis. Cell communication and signaling : CCS, 11, 77. https://doi.org/10.1186/1478-811X-11-77

Rodríguez Mestre, M.R., González-Delgado, A., Gutiérrez-Rus, L.I., Martínez-Abarca, F., Toro, N. 2020. Systematic prediction of genes functionally associated with bacterial retrons and classification of the encoded tripartite systems, Nucleic Acids Research, Volume 48, Issue 22, 16 December 2020, Pages 12632–12647, https://doi.org/10.1093/nar/gkaa1149

Roberts, R.J., Belfort, M., Bestor, T., Bhagwat, A.S. 2003. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nucleic acids research, 31(7), 1805–1812. https://doi.org/10.1093/nar/gkg274

Roesler, B.M., Rabelo-Gonçalves, E.M., & Zeitune, J.M. 2014. Virulence Factors of *Helicobacter pylori:* A Review. Clinical medicine insights. Gastroenterology, 7, 9–17. https://doi.org/10.4137/CGast.S13760

Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S. A. and Sørensen, S.J. 2020. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. The CRISPR Journal, 3**,** 462–469. doi: 10/gshm

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J.C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S.E., Sánchez-Gracia, A. 2017. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Datasets. Mol. Biol. Evol. 34: 3299-3302. DOI: 10.1093/molbev/msx248.

Salama, N. R., Hartung, M. L., & Müller, A. 2013. Life in the human stomach: persistence strategies of the bacterial pathogen *Helicobacter* pylori. Nature reviews. Microbiology, 11(6), 385–399. https://doi.org/10.1038/nrmicro3016

Salaün, L., Linz, B., Suerbaum, S., Saunders, N.J. 2004.The diversity within an expanded and redefined repertoire of phase-variable genes in *Helicobacter pylori*. Microbiology (Reading). Apr; 150(Pt 4):817-830. doi: 10.1099/mic.0.26993-0. PMID: 15073292.

Samson, J., Magadán, A., Sabri, M. et al. 2013. Revenge of the phages: defeating bacterial defences. Nat Rev Microbiol 11, 675–687. https://doi.org/10.1038/nrmicro3096

Sampson, TR., & Weiss, DS. 2013. Alternative Roles for CRISPR/Cas Systems in Bacterial Pathogenesis. PLoS Pathogens, 9(10), e1003621. DOI: 10.1371/journal.ppat.1003621

Sampson, TR., & Weiss, DS. 2014. CRISPR-Cas systems: new players in gene regulation and bacterial physiology. Frontiers in Cellular and Infection Microbiology, 4, 37. DOI: 10.3389/fcimb.2014.00037.

Scienski, K., Fay, J.C., & Conant, G.C. 2015. Patterns of Gene Conversion in Duplicated Yeast Histones Suggest Strong Selection on a Coadapted Macromolecular Complex. Genome Biology and Evolution, 7(12), 3249–3258. https://doi.org/10.1093/gbe/evv216

Shah, S.A., Alkhnbashi, O.S., Behler, J., Han, W., She, Q., Hess, W.R., Garrett, R.A. and Backofen, R. 2019. Comprehensive search for accessory proteins encoded with archaeal and bacterial type III CRISPR-cas gene cassettes reveals 39 new cas gene families. RNA Biology, **16**, 530–542. doi: 10/ggqv9p

Shmakov, S A., Makarova, K.S., Wolf, Y.I., Severinov, K.V. and Koonin, E.V. 2018. Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. PNAS, 115, E5307–E5316. doi: 10/gdpqwq

Smet, A., Yahara, K., Rossi, M., Tay, A., Backert, S., Armin, E., Fox, J. G., Flahou, B., Ducatelle, R., Haesebrouck, F., & Corander, J. 2018. Macroevolution of gastric *Helicobacter* species unveils interspecies admixture and time of divergence. *The ISME journal*, *12*(10), 2518–2531. https://doi.org/10.1038/s41396-018-0199-5

Sheppard, S.K., & Maiden, M.C. 2015. The evolution of *Campylobacter jejuni* and *Campylobacter coli*. Cold Spring Harbor perspectives in biology, 7(8), a018119. https://doi.org/10.1101/cshperspect.a018119

Smith, J.M., 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. 34, 126– 129. doi:10.1007/ BF00182389.

Smith, H.S., Pizer, L.I., Pylkas, L. and Lederberg, S. 1969. Abortive Infection of *Shigella dysenteriae* P2 by T2 Bacteriophage. Journal of Virology, 4, 162–168. doi: 10/gndc4r

Srikhanta, Y. N., Gorrell, R. J., Steen, J. A., Gawthorne, J. A., Kwok, T., Grimmond, S. M., Robins-Browne, R. M., & Jennings, M. P. 2011. Phasevarion mediated epigenetic gene regulation in *Helicobacter pylori*. PloS one, 6(12), e27569. https://doi.org/10.1371/journal.pone.0027569

Solnick, J.V., & Schauer, D.B. 2001. Emergence of diverse *Helicobacter* species in the pathogenesis of gastric and enterohepatic diseases. *Clinical microbiology reviews*, *14*(1), 59–97. https://doi.org/10.1128/CMR.14.1.59-97.2001

Su, P., Harvey, M., Im, H.J. and Dunn, N.W. 1997. Isolation, cloning and characterisation of the abiI gene from *Lactococcus lactis subsp. lactis* M138 encoding abortive phage infection. Journal of Biotechnology, 54, 95–104. doi: 10/ckwmpp

Suerbaum, S., Smith, JM., Bapumia, K., Morelli, G., Smith, NH., Kunstmann. 1998. Free recombination within *Helicobacter pylori*. Proceedings of the National Academy of Sciences of the United States of America, 95(21), 12619–12624. DOI: 10.1073/pnas.95.21.12619

Suerbaum, S., & Josenhans, C. 2007. *Helicobacter pylori* evolution and phenotypic diversification in a changing host. Nature Reviews Microbiology, 5(6), 441–452. DOI: 10.1038/nrmicro1658

Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006 Nov 1; 22(21):2688-90. doi: 10.1093/bioinformatics/btl446.

Stern, A., Keren, L., Wurtzel, O., Amitai, G., & Sorek, R. 2010. Self-targeting by CRISPR: gene regulation or autoimmunity? Trends in Genetics: TIG, 26(8), 335–340. DOI: 10.1016/j.tig.2010.05.008

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. Nov; 123(3):585-95. doi: 10.1093/genetics/123.3.585. PMID: 2513255; PMCID: PMC1203831.

Tayebi, N., Stubblefield, B.K., Park, J.K., Orvisky, E., Walke,r J.M., LaMarca, M.E., Sidransky, E. 2003. Reciprocal and nonreciprocal recombination at the glucocerebrosidase gene region: implications for complexity in Gaucher disease. Am J Hum Genet. 72:519–534. doi: 10.1086/367850.

Tettelin, H., Masignani, V., Cieslewicz, M. J. et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences of the United States of America, 102(39), 13950–13955. https://doi.org/10.1073/pnas.0506758102

Tomb, J.F, White, O., Kerlavage, A.R, Clayton, R.A, *et al.* 1997.The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature. Aug 7; 388 (6642):539-47. doi: 10.1038/41483.

Tomida, J., Morita, Y., Shibayama, K., Kikuchi, K., Sawa, T., Akaike, T., & Kawamura, Y. 2017. Diversity and microevolution of CRISPR loci in *Helicobacter cinaedi.* PLoS ONE, 12(10), e0186241. DOI: 10.1371/journal.pone.0186241.

Touchon, M., & Rocha, EP. C. 2010. The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. PLoS ONE, 5(6). DOI: 10.1371/journal.pone.0011126.

Vasu, K., & Nagaraja, V. 2013. Diverse functions of restriction-modification systems in addition to cellular defense. Microbiology and molecular biology reviews: MMBR, 77(1), 53–72. https://doi.org/10.1128/MMBR.00044-12

van Houte, S., Buckling, A., & Westra, E. R. 2016. Evolutionary Ecology of Prokaryotic Immune Mechanisms. Microbiology and molecular biology reviews: MMBR, 80(3), 745–763. https://doi.org/10.1128/MMBR.00011-16

Vernikos, G., Medini, D., Riley, D.R, Tettelin, H., 2015. Ten years of pan-genome analyses. Curr Opin Microbiol. 2015 Feb;23:148-54. doi: 10.1016/j.mib.2014.11.016. Epub 2014 Dec 5. PMID: 25483351.

Wang, S., Chen, Y. Phylogenomic analysis demonstrates a pattern of rare and long-lasting concerted evolution in prokaryotes. Commun Biol 1, 12 (2018). https://doi.org/10.1038/s42003-018-0014-x

Weiller, G.F. 1998 Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. Mol. Biol. Evol. 15, 326– 335. (doi:10.1093/oxfordjournals.molbev.a025929)

Westra, E.R., Dowling, A.J., Broniewski, JM., & van Houte, S. 2016. Evolution & Ecology of CRISPR. Annual Review of Ecology, Evolution, and Systematics, 47(1), 307–331. DOI: 10.1146/annurev-ecolsys-121415-032428.

Wroblewski, L.E, Peek, RM JR. 2016. *Helicobacter* pylori, Cancer, and the Gastric Microbiota. Adv Exp Med Biol. 2016; 908:393-408. doi: 10.1007/978-3-319-41388-4_19.

Yahara, K., Furuta, Y., Oshima, K., Yoshida, M., Azuma, T., Hattori, M., Kobayashi, I. 2013. Chromosome Painting In Silico in a Bacterial Species Reveals Fine Population Structure. Molecular Biology and Evolution, 30(6), 1454–1464. DOI: 10.1093/molbev/mst055.

Yadong, Zhang., Zhewen, Zhang., Hao, Zhang., Yongbing, Zhao., Zaichao, Zhang., Jingfa, Xiao., PADS Arsenal: a database of prokaryotic defense systems related genes, Nucleic Acids

Research, Volume 48, Issue D1, 08 January 2020, Pages D590–D598,https://doi.org/10.1093/nar/gkz916

Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 1997 Oct; 13(5):555-6. doi: 10.1093/bioinformatics/13.5.555. PMID: 9367129.

Zhao, X., Yu, Z., & Xu, Z. 2018. Study the Features of 57 Confirmed CRISPR Loci in 38 Strains of *Staphylococcus aureus*. Frontiers in microbiology, 9, 1591.

Zhang, Q., & Ye, Y. 2017. Not all predicted CRISPR–Cas systems are equal: isolated cas genes and classes of CRISPR like elements. BMC Bioinformatics, 18, 92. DOI: 10.1186/s12859-017-1512-4.

Zuker, M., & Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res.9, 133–148. DOI: 10.1093/nar/9.1.133

Zawilak-Pawlik, A., Zakrzewska-Czerwińska, J. 2017. Recent Advances in *Helicobacter pylori* Replication: Possible Implications in Adaptation to a Pathogenic Lifestyle and Perspectives for Drug Design. In: Tegtmeyer N., Backert S. (eds) Molecular Pathogenesis and Signal Transduction by *Helicobacter* pylori. Current Topics in Microbiology and Immunology, vol 400. Springer, Cham.