UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE QUÍMICA

“ANÁLISIS DE LA ASIMETRÍA TRANSCRIPCIONAL EN GENOMAS PROCARIONTES”

TESIS

QUE PARA OBTENER EL TÍTULO DE

QUÍMICO FARMACÉUTICO BIÓLOGO

PRESENTA

JOEL CORONA PACHECO

MÉXICO, D.F. **AÑO 2014**

JURADO ASIGNADO:

PRESIDENTE: Profesor: Marisol López López

VOCAL: Profesor: Ruth Edith Martín Fuentes

SECRETARIO: Profesor: Víctor Manuel Valdés López

1er. SUPLENTE: Profesor: María Benita Leonor Fernández Salgado

2° SUPLENTE: Profesor: Nancy Monroy Jaramillo

SITIO DONDE SE DESARROLLÓ EL TEMA:

Laboratorio de Biología Molecular y Genómica

Facultad de Ciencias UNAM

ASESOR DEL TEMA:

Dr. Víctor Manuel Valdés López

SUSTENTANTE:

Joel Corona Pacheco

ÍNDICE.

RESUMEN _____ **4**

*INTRODUCCIÓN*_____ **6**

*JUSTIFICACIÓN*_____ **31**

*OBJETIVOS*_____ **32**

*MATERIALES Y MÉTODOS*_____ **33**

*RESULTADOS Y DISCUSIÓN*_____ **38**

*CONCLUSIONES*_____ **51**

*BIBLIOGRAFÍA*_____ **52**

RESUMEN.

Los genomas procariontes convencionales consisten de una sola molécula circular de DNA y su tamaño es variado. Sobre esta molécula de DNA los genes se encuentran dispuestos de manera secuencial, con relativamente poco espacio intergénico. Dado que el genoma es circular, la ubicación de genes se puede dar de acuerdo a su posición en pares de bases o en referencia a sus coordenadas en el mapa genómico circular. Sin embargo, la dirección de la transcripción no es única. Algunos genes se transcriben en el sentido de las manecillas del reloj y otros en sentido contrario. Convencionalmente, las coordenadas genómicas parten del sitio de origen de la replicación y dado que ésta va a proceder de manera bidireccional, a una cadena se le denomina positiva (líder) y a la otra negativa (retrasada), en referencia al avance de las horquillas de replicación. Cuando el genoma replica, el proceso termina aproximadamente en el lugar opuesto al sitio de origen, lo cual le da al genoma una división simétrica respecto a este proceso. Desde hace tiempo, se notó que existe una asimetría en la dirección transcripcional, en el sentido de que un mayor número de genes se transcriben en la misma dirección global del proceso de replicación en cada una de las dos mitades del genoma. La hipótesis más simple es que si la replicación y la transcripción proceden al mismo tiempo, pero en direcciones opuestas, se darían “colisiones” entre la

DNA polimerasa y la RNA polimerasa; entonces la asimetría dependería simplemente del nivel de expresión, de tal manera que aquellos genes con mayor nivel de transcripción se ubicarían en la dirección global de la replicación. Una hipótesis alternativa propone que los genes que se transcriben en la misma dirección que la replicación corresponden a genes “esenciales” para la supervivencia del organismo, si bien no es sencillo definir qué significa que un gen sea esencial ya que esto depende de las condiciones de desarrollo y las características del propio organismo.

Las predicciones y resultados de ambas propuestas siguen debatiéndose. Nuestro enfoque para evaluar la controversia fue cuantificar la asimetría transcripcional en 41 genomas procariontes representativos de diferentes grupos, incluyendo arqueas, y llevar a cabo una comparación de los genes que se transcriben en una u otra cadena, para determinar si existe una correlación entre los diferentes genomas. Dicho de otro modo, se hicieron comparaciones para establecer si en diferentes genomas se encuentra una relación entre los genes y su asimetría transcripcional. Nuestros resultados confirman las asimetrías reportadas y aportan un nuevo enfoque sobre las causas de las mismas.

INTRODUCCIÓN.

Los seres vivos están integrados por células y éstas a su vez por moléculas. Cuando se examinan individualmente, estas moléculas se ajustan a todas las leyes físicas y químicas que rigen el comportamiento de la materia inerte. Los sistemas biológicos son altamente complejos y tienen un alto grado de organización, poseen estructuras internas que contienen muchas clases de moléculas complejas, de bajo peso molecular -tales como lípidos, carbohidratos etc.- y de alto peso molecular -como las proteínas y el DNA-. En especial, el DNA ha sido y sigue siendo de gran interés para la comunidad científica[1].

El descubrimiento del DNA se remonta al invierno de 1868-1869 cuando un joven doctor suizo llamado Friedrich Miescher, (el cual trabajaba en el laboratorio de Felix Hoppe-Seyler en la Universidad de Tübingen) realizaba experimentos sobre la composición química de leucocitos. En sus experimentos, Miescher notó un precipitado de una sustancia desconocida. Sus propiedades durante el procedimiento de su aislamiento y su resistencia a la digestión por proteasas indicaron que la nueva sustancia no era una proteína o un lípido. En trabajos subsecuentes con otros tejidos tales como hígado, testículos, riñón, eritrocitos nucleados y levaduras, Miescher mostró que la molécula era un componente característico de todos

los núcleos y a partir de sus observaciones y el aislamiento de ésta, la llamó “nucleína”.

Miescher estaba ansioso de caracterizar mejor el precipitado y escribió varios protocolos para obtener mayores cantidades de nucleína. Después de haber desarrollado el protocolo con el cual aisló nucleína con suficiente pureza y en grandes cantidades, Miescher se dio a la tarea de establecer su composición elemental. El procedimiento involucraba el calentamiento de la sustancia para ser analizada en presencia de varios agentes químicos que reaccionaban selectivamente con los diferentes elementos constitutivos. Los productos de las reacciones fueron pesados para cuantificar cada elemento. Miescher estuvo muy consciente que el análisis elemental era crucial para ayudar a descubrir la verdadera naturaleza de la nucleína y si ésta en realidad era distinta a otras moléculas orgánicas.

Encontró varios elementos típicos de las moléculas orgánicas como carbono, hidrógeno, oxígeno, nitrógeno y notablemente grandes cantidades de fósforo. Esto reveló que era diferente a las proteínas y que como Miescher confirmó más tarde, carecía de azufre, lo cual lo llevó a concluir que “... estamos trabajando con una entidad *sui generis* no comparable con cualquier grupo conocido hasta ahora” (Miescher, 1871).

Durante un cuarto de siglo, Miescher trabajó con la nucleína y desarrolló varias teorías sobre sus funciones. La idea inicial fue que la nucleína servía como almacenamiento de fósforo dentro de la célula o actuaba como un precursor para la generación de otras moléculas. Después de haber descubierto nucleína en células germinales, por ejemplo, su gran abundancia en espermatozoides, Miescher llegó a sospechar que la nucleína estaba involucrada en el mecanismo de fertilización (la fusión de dos células derivadas de un macho y una hembra).

El final del siglo XIX fue un tiempo de intensa investigación y especulación sobre los mecanismos que controlan la fecundación y la transmisión de rasgos hereditarios de una generación a otra. En este contexto, las publicaciones de Miescher de 1874 sobre la aparición de grandes cantidades de nucleína en esperma de diferentes especies de vertebrados causaron gran interés dentro de la comunidad científica. El mismo Miescher estuvo cerca de contestar correctamente a la pregunta. En su publicación Die Spermatozoen einiger Wirbeltiere escribió: "...si uno quiere asumir que una única sustancia es la causa específica de la fertilización, entonces se debe sin duda considerar primero a la nucleína" (Miescher, 1874). Sin embargo, Miescher descartó la idea de que la nucleína contenía la información hereditaria porque pensaba que era

poco probable que la misma sustancia pudiera dar como resultado la diversidad de diferentes especies animales.

Miescher presumía que la información hereditaria podría ser codificada en el estado estereoquímico del átomo de carbono. Al igual que un alfabeto de 24 a 30 letras es suficiente para representar todas las palabras y conceptos en diferentes lenguajes, estos estereoisómeros podrían ser usados para crear moléculas conteniendo diferente información. El enorme número de carbonos asimétricos a lo largo de las moléculas orgánicas, tales como el de las proteínas, podrían permitir un inmenso número de estereoisómeros. Una proteína que contenga 40 carbonos asimétricos podría tener 2^{40} estereoisómeros: un número suficientemente grande que pudiera codificar para la información hereditaria de todas las diferentes formas de vida.

Miescher contrajo tuberculosis a principios de los 1890s y cayó seriamente enfermo y tuvo que dejar de trabajar. Estuvo en un sanatorio en Davos donde no recuperó su salud y finalmente el 26 de agosto de 1895 Friedrich Miescher murió.

Después de la muerte de Miescher, otros investigadores continuaron con su trabajo de la naturaleza de la nucleína influenciados por Hoppe-Seyler, entre ellos, Albrecht Kossel y Richard Altmman, quienes desarrollaron protocolos que les permitieron purificar

nucleína libre de proteína y de esa manera pudieron examinar los componentes químicos. Finalmente, Altmman logró la separación del DNA de las proteínas en sus preparaciones y en 1889, basado en el hecho de que se comportaba como un ácido, él nombró a la sustancia “Nucleïnsäure” (ácido nucleico). Kossel, a su vez, llevó a cabo un trabajo pionero identificando la construcción fundamental de la nucleína y determinó que estaba compuesta por cinco bases diferentes -dos bases derivadas de la purina: adenina y guanina; y tres bases de la pirimidina: timina, citosina y uracilo- además de un azúcar y ácido fosfórico. Asimismo confirmó que esto era exclusivo del núcleo[2]. Más adelante el bioquímico ruso Phoebus Levene estudió la estructura y función de los ácidos nucleicos y encontró que existían diferentes tipos químicos (DNA y RNA). Sus modelos de estudio fueron levaduras y células del timo.

Dado que las levaduras tienen un citoplasma comparativamente más grande que su núcleo el ácido nucleico en mayor abundancia es RNA, a diferencia de los timocitos que tienen un núcleo más grande que su citoplasma siendo el ácido nucleico mayoritario el DNA.

Él demostró que la pentosa que aparecía en la nucleína de la levadura era ribosa. Esto fue porque sus desarrollos experimentales los realizó en medio alcalino para hidrolizar el ácido nucleico, y con dicho procedimiento, en realidad el estudio se enfocó en la

composición y estructura del RNA porque el DNA no se hidroliza en medio alcalino[3].

Al final de su carrera, Levene, asumiendo que los ácidos nucleicos eran macromoléculas, hizo un supuesto simplificador, en donde propuso que el DNA y el RNA eran “polímeros de tetranucleótidos”. Este principio se basó en su análisis de la composición molar de bases en el DNA y había llegado a la conclusión de que las proporciones de A, T, G, y C eran equimolares y conformaban una molécula básica denominada tetranucleótido, el cual constituía la unidad básica del DNA. Esta unidad se repetiría de manera invariante hasta formar la macromolécula. Esta interpretación enfatizó la idea de que los ácidos nucleicos no tenían un papel importante en la transmisión hereditaria (informacional) y que solo eran elementos estructurales. Esta idea la retomaron un grupo de científicos llamados “el grupo del fago” integrado por Max Delbrück y Salvador Luria, entre otros, donde consideraban que los ácidos nucleicos eran parte del andamiaje de los cromosomas por una razón intuitivamente aceptable: si las proteínas estaban constituidas por veinte aminoácidos (20 unidades) y los ácidos nucleicos por solo cuatro nucleótidos (4 unidades), lo más razonable era considerar que las instrucciones genéticas tendrían que ser codificadas por las proteínas y no por los ácidos nucleicos[4].

Mientras tanto, un bacteriólogo británico llamado Frederick Griffith, el cual era especialista en la epidemiología y patología de la neumonía bacteriana, en 1928 publicó un artículo demostrando la transformación bacteriana. En su trabajo inyectó ratones con dos tipos de neumococos: bacterias virulentas muertas por calor junto con bacterias no virulentas vivas. Muchos ratones murieron después de cierto tiempo y en el tejido cardiaco encontró gran cantidad de bacterias vivas virulentas, causantes de la muerte de los ratones. Lo que Griffith descubrió fue que durante la infección, las bacterias muertas por calor habían transmitido su capacidad patogénica a las bacterias no virulentas, es decir, las habían "transformado" y esta capacidad virulenta se mantenía a través de generaciones[5]. La hipótesis más probable fue que algún compuesto de las bacterias patógenas era el responsable de la transformación. Quedaba esclarecer entonces la naturaleza química de este principio transformante.

Quien se propuso dilucidar la naturaleza química del "principio transformante" fue un médico canadiense llamado Oswald Avery, y en 1943 demostró que la sustancia era DNA y no proteína. Esto lo logró mediante un estudio en el que hizo un análisis detallado del fenómeno de transformación de algunos tipos específicos de neumococos. De éste concluyó que la fracción aislada que contenía ácido desoxirribonucleico de neumococos tipo III capsulados

(patógenos), era capaz de transformar neumococos tipo II no capsulados (no patógenos) a capsulados. Concluyó que la información genética estaba contenida en genes: "…la sustancia que induce transformación al igual que el antígeno capsular, el cual se produce en respuesta a la misma ha sido considerada como un gen…"[5].

Posteriormente en 1952 Martha Chase y Alfred Hershey confirmaron los resultados de Avery. Ellos hicieron sus trabajos infectando bacterias de *Escherichia coli* utilizando el fago T2 (un virus que sólo infecta bacterias *E. coli* y se reproduce dentro de ellas). Dado que, las proteínas no contienen fósforo y el DNA no contiene azufre, marcaron el DNA del fago con fósforo radiactivo (P-32) y las proteínas con azufre radiactivo (S-35) para poder diferenciar cuál de los dos componentes entraba a la bacteria. Analizando los resultados se dieron cuenta de que las proteínas del fago se encontraban fuera de la bacteria, y su DNA se encontraba dentro de la misma[6]. Con base en estos resultados concluyeron que las proteínas del fago no tienen una función en el crecimiento intracelular del mismo y que el DNA es la parte del fago que infecta a las bacterias, por lo tanto, el material hereditario de los fagos era el DNA.

Después de que Oswald Avery identificó que el DNA era la molécula capaz de transmitir la información hereditaria, en 1944 el bioquímico austriaco Erwin Chargaff y colaboradores se interesaron en el estudio del DNA y en 1949 nuevamente analizaron la equivalencia de bases en dicha molécula. En su primer estudio reportaron la distribución de purinas y pirimidinas en el DNA derivado del timo y bazo de ternera[7]. Posteriormente realizaron estudios similares con *B. subtilis* y espermas de erizo de mar[8] y encontraron una relación en la proporción de cada una de las diferentes bases, en donde [A]=[T] y [G]=[C] lo cual los condujo a que $\frac{(Purinas)}{(Pirimidinas)} = 1$ y a esta proporción se le llama la relación de simetría de Chargaff. Dicha relación no es compatible con la hipótesis del tetranucleótido propuesta por Levene[7]. Por otro lado con los resultados obtenidos en sus trabajos también los condujeron a que $\frac{(A+T)}{(G+C)} \neq 1$, lo cual es una característica específica de cada genoma, llamando a ésta, la relación de asimetría[9]. Todas las investigaciones de Chargaff aportaron información sumamente importante para establecer la estructura física del DNA.

Mientras la evidencia genética y bioquímica iba acumulándose en torno al DNA como el material hereditario, los estudios sobre su estructura física se llevaron a cabo en Inglaterra utilizando la

difracción de rayos X, desarrollada por W. Henry y W. Lawrence Bragg, padre e hijo respectivamente, a principios del siglo XX.

Desde luego que el conjunto de técnicas implicadas en la cristalografía de rayos X requiere de un conocimiento profundo de física y matemáticas, razón por la cual, hacia principio de la década de 1950, existían muy pocos especialistas en esta área. Entre ellos se encontraban Rosalind Franklin, Maurice Wilkins y Francis Crick, entre otros.

Rosalind Franklin fue una química y cristalógrafa inglesa que en 1952 en el King's College en Londres, hizo estudios de difracción de rayos X de fibras de DNA. Ella describió dos formas de dichas fibras, a las que llamó forma A y B respectivamente[10]. Posteriormente, en 1953, escribió un artículo donde presentó la fotografía de difracción de rayos X de la forma B del DNA [11] que fue la clave necesaria para confirmar el modelo de la doble hélice propuesto por Watson & Crick en el mismo año.

Francis Crick fue un físico británico que había desarrollado diversos métodos matemáticos para la interpretación de los patrones de difracción de rayos X. Trabajando en el laboratorio Cavendish en la Universidad de Cambridge conoció a un biólogo norteamericano que se unió al laboratorio, llamado James Watson, y así comenzó una labor de cooperación entre ambos científicos.

Watson y Crick consideraban de gran interés biológico determinar la estructura física del DNA y en abril de 1953 publicaron un artículo en el cual proponen la estructura física y química del DNA, la cual describen de la siguiente manera: "...esta estructura tiene dos cadenas helicoidales cada una enrollada alrededor del mismo eje. Nosotros hemos hecho supuestos en la química, a saber, que cada cadena consiste de grupos fosfatos unidos por un enlace diéster, a su vez unidos a residuos de β-D-desoxiribofuranosa con enlaces 3',5'"[12]. También, basados en los resultados de Chargaff, ellos propusieron la complementariedad de las bases y las condiciones para la formación de los puentes de hidrógeno que mantienen unidas las cadenas de DNA. "... ellas se encuentran unidas en pares, una sola base de una cadena se enlaza a una sola base de la otra cadena mediante puentes de hidrógeno... uno de los pares debe ser una purina y la otra una pirimidina para que ocurra la unión... Los pares son: adenina (purina) con timina (pirimidina) y guanina (purina) con citocina (pirimidina)"[12].

Más tarde, en otra publicación del mismo año, propusieron que dicha estructura explicaba el mecanismo de copiado del material genético; este modelo tenía una gran relevancia en la transmisión de la información genética, el cual fue postulado de la siguiente manera: "...una cadena que conforma el DNA es complementaria a la otra. Por ese motivo dicha molécula podría duplicarse por sí misma,

porque, si las cadenas se separan rompiendo los puentes de hidrógeno, una cadena sería el molde para la formación de la otra cadena, entonces se tendrían dos cadenas donde antes sólo se tenía una"[13].

Después del modelo de la "doble hélice", todavía existían preguntas, una de ellas era: ¿Cómo el DNA dirige la síntesis de proteínas?. Ya que en ese tiempo no existía evidencia, sólo se sabía que el DNA era una molécula exclusiva del núcleo y las proteínas del citoplasma, se pensaba que debía de existir una molécula mensajera intermediaria entre estas dos. Los científicos Sydney Brenner, Francois Jacob y Matthew Meselson, explicaron e identificaron la naturaleza de la molécula intermediaria a la cual llamaron RNA mensajero -mRNA- y se dieron cuenta que dicha molécula era la que acarreaba la información genética del núcleo a los ribosomas en el citoplasma para la producción de proteínas[14].

Desde que los científicos determinaron que el mRNA servía como una copia de los genes en el DNA y con el conocimiento de la secuencia de residuos de aminoácidos de una proteína, inmediatamente surgieron más preguntas acerca de la formación de una proteína. Específicamente se sabía que las proteínas estaban compuestas por 20 diferentes aminoácidos y que sólo existían 4 nucleótidos en el mRNA -adenina (A), citocina (C), guanina (G) y

uracilo (U)-. Pero ¿cómo exactamente podrían 4 nucleótidos codificar para 20 aminoácidos?. Algunas de estas respuestas fueron dilucidadas por F. Crick y colaboradores. Los resultados de sus trabajos los llevaron a concluir que el mRNA debe ser descifrado en tripletes (3 nucleótidos) y a esto lo llamaron codón. También determinaron que existen 64 codones y que varios de ellos son redundantes -varios codones codifican para el mismo aminoácido- y a esto lo llamaron el código genético[15].

Aunque la estructura de la doble hélice ya se había propuesto, varias décadas pasarían antes de poder analizar una secuencia de nucleótidos del DNA en el laboratorio. El producto directo de la secuenciación es el ordenamiento de nucleótidos. El obtener información de estos nucleótidos se llama anotación, así se puede extraer información acerca de la composición de bases, su estabilidad termodinámica, posibles estructuras secundarias, periodicidades o repeticiones en la secuencia, entre otras[16], pero principalmente, saber cuál es la región codificante y reguladora. El primer gen que se secuenció por W. Fierst y colaboradores, codifica para una proteína de envoltura del bacteriófago MS2 (virus de RNA) por también fue el primer genoma completo de RNA secuenciado por el mismo investigador en 1976[17]. Con el paso del tiempo se fueron perfeccionando las técnicas de secuenciación hasta poder

secuenciar nucleótidos de DNA ya que se necesitaba otras técnicas diferentes para poderlo hacer.

En este campo destaca un bioquímico británico llamado Frederick Sanger. Sus primeros trabajos se basaron en conocer la estructura de las proteínas, especialmente la estructura de la insulina. Él, junto con otro investigador llamado Alan Coulson, secuenciaron pequeños segmentos de nucleótidos de DNA de cadena simple, con una técnica a la que nombraron “Plus and minus”. En esta técnica sintetizaban pequeños iniciadores, marcaban los cuatro nucleótidos con fósforo radiactivo (P-32) y utilizaban una enzima llamada DNA polimerasa I de *E. coli,* para polimerizar el fragmento de DNA. Este método se utilizó para poder determinar dos secuencias del bacteriófago ΦX174[18].

Ellos sustituyeron la técnica “Plus and minus”, logrando desarrollar otro procedimiento más rápido y más exacto. Ésto gracias a la implementación de didesoxinucleótidos -nucleótidos que carecen del grupo 3'-OH en la desoxirribosa-; estas moléculas terminan la polimerización de la cadena de DNA debido a que no se forma el enlace fosfodiéster, y de esta manera podían controlar el didesoxinucleótido que se agrega al medio, encontrando diferentes tamaños de la misma cadena de DNA polimerizada al revelarlos en un gel de agarosa. Esta técnica secuenciaba fragmentos hasta de

600 bases, con ello pudieron secuenciar el genoma completo del bacteriófago ΦX174 (virus de DNA de cadena simple) que contiene 5,386 nucleótidos. A este método se le llamó “Sanger sequencing”[19].

En los años 1990s, con el comienzo de investigaciones que requerían secuenciación, se promovió la mejora de estos métodos y mediante la implementación de didesoxinucleótidos con fluorescencia, perfeccionamiento y automatización del proceso, se lograron secuenciar hasta 96 muestras de DNA en un tiempo corto, procesando entre 500 y 1,000 bases. Poco a poco, se han ido reportando más y más secuencias de genomas completos desde el genoma del virus ΦX174[20], hasta el genoma humano[21 22], pasando por genomas mitocondriales[23], de cloroplasto, plásmidos, virus, bacterias[24], arqueas, levaduras, moscas[25], gusanos[26] y plantas[27]. Si bien en cada genoma secuenciado se encuentran características peculiares, en una perspectiva global éstos se pueden separar en dos tipos biológicos: eucariontes y procariontes[28].

Las características del genoma procarionte (bacterias y arqueas) muchas veces sirve para contrastar y tomarlo como punto de referencia para abordar la descripción particular del genoma eucarionte, sin embargo, los genomas procariontes son muy diferentes a los eucariontes, en particular en su organización física dentro de la célula. El concepto general ha sido que el genoma

procarionte está contenido en una sola molécula circular de DNA, la cual se encuentra en el nucleoide.

Independientemente de cuál sea la función asignada a cada gen, los genomas procariontes comparten muchas particularidades tales como:

- Densidad genética alta. Entre genes contiguos existe una pequeña región de DNA intercistrónica, pero no existen grandes cantidades de DNA espaciador.

- El tamaño de los genes individuales es variable, unos cortos y unos largos, pero, en la mediana, el tamaño canónico de un gen procarionte es de 1,000 pares de bases.

- Los genes son unidades discretas en los genomas y por lo tanto, están definidos y delimitados. En este sentido, un gen ocupa una posición definida en el genoma. Sin embargo, un gen se tiene que transcribir y por lo tanto, independientemente de su ubicación, una de las dos cadenas del DNA va a servir como molde. Esto depende de la posición del promotor, el cual determina la dirección de la transcripción.

- Algunos genes están asociados, formando unidades de transcripción que llamamos operones

A pesar de estas características y otras, la organización estructural del genoma es altamente conservada (Figura 1). Por ejemplo, al hacer una comparación de los mapas genómicos de *E. coli* y *Bacillus subtilis* (bacterias), los cuales divergieron hace varios millones de años, éstos son más similares entre sí que dos genomas de levaduras (eucariontes), los cuales divergieron hace pocos millones de años. Por lo tanto, los cromosomas bacterianos tienden a una arquitectura en donde son complejos y plásticos al mismo tiempo [29].

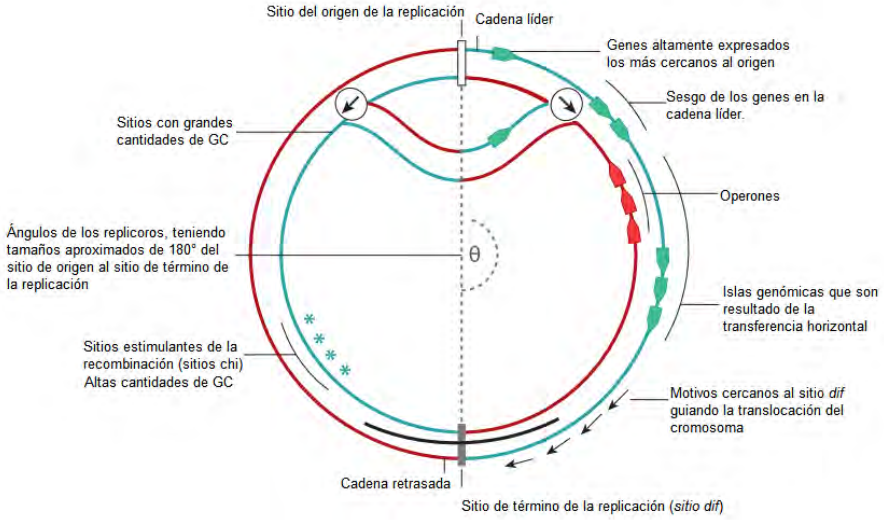


Figura 1| Arquitectura del genoma procarionte. Se muestran los elementos organizacionales que en general se han encontrado en la mayoría de los genomas bacterianos. Tomado y modificado [29].

La organización del genoma de las arqueas ha recibido menos atención, aunque éstas comparten muchas similitudes con las bacterias. Notablemente, las arqueas también tienen operones conservados y acoplamiento de transcripción-traducción. Sin embargo, a diferencia de las bacterias, muchas especies tienen sincronizados múltiples orígenes de replicación [30], aunque otras tienen orígenes facultativos que hasta ahora no se han encontrado entre las bacterias [31].

El origen de replicación en bacterias, sólo está en una región esencial *cis-acting* secuencias de DNA vecinas de un gen necesarias para su expresión[32].

En el sitio de origen se abren las cadenas de DNA y aparecen dos horquillas de replicación las cuales se desplazan siguiendo direcciones opuestas hasta llegar al sitio de término, el *sitio dif,* donde toma lugar la decatenación (separación de las cadenas) del cromosoma (Figura 1). Finalmente, el cromosoma es separado en dos mitades (replicoros) replicados a partir de las diferentes horquillas. El paso de las horquillas de replicación remodela la estructura del nucleoide y desplaza todas las moléculas que interaccionan físicamente con el cromosoma.

Existe la posibilidad de iniciar un nuevo evento de replicación antes de que una ronda previa finalice. Por ejemplo, en *E. coli* se tienen

rondas de replicación simultáneas, ya que la división celular toma 20 minutos y la replicación del cromosoma toma tres veces más tiempo. El número estimado de eventos de replicación simultáneos (R) es la razón entre el tiempo requerido para replicar todo el cromosoma y el tiempo entre dos divisiones celulares sucesivas. Si R es cercano a cero entonces la replicación del cromosoma es escasa. Cuando R>1, las células experimentan múltiples eventos de replicación. Cuando un gen está más cerca del origen, éste será en promedio 2^R más abundante en la célula que un gen cerca del término. Dicho de otro modo, los genes cerca del origen tienen un mayor número de copias cuando *R* es alto.

Ahora, si se compara el proceso de replicación en los genomas de las arqueas y los eucariontes, contra las bacterias, se observan diferencias, por ejemplo, la única DNA polimerasa reconocida en el genoma de *Methanococcus jannaschii* (arquea), es homóloga a la polimerasa de replicación nuclear de un eucarionte, pero estas enzimas no están relacionadas a su contraparte funcional en *E. coli* y otras bacterias -DNA polimerasa III- [33]. También otro ejemplo son los componentes de los aparatos de transcripción. Las arqueas y eucariontes son versiones similares entre sí, comparándolas con cualquiera de las bacterias. Las holoenzimas de las arqueas contienen un número adicional de subunidades que tienen contrapartes sólo en los eucariontes[34].

Desafortunadamente, en la actualidad se ignoran muchos mecanismos celulares básicos de las arqueas, complicando el análisis de sus efectos sobre la organización del genoma.

Se ha propuesto que la organización específica del genoma procarionte se puede deber a que algunos de los procesos celulares interaccionan directa o indirectamente con el DNA a diferentes escalas, local como la expresión de genes y global, tal como la replicación, afectando así la estructura del genoma. Las causas moleculares de estos procesos imponen restricciones y/o conducen a la selección de muchas configuraciones favorables para efectos genómicos. Naturalmente si dos procesos interaccionan en el cromosoma, entonces se requiere una fina y sincronizada organización. Como ejemplo de estas características en la organización, se incluye la sobreabundancia de genes en la cadena líder (Figura 2) que posiblemente es causada por la interacción antagónica entre la replicación y la transcripción. Se ha postulado que este sesgo en la distribución de genes posiblemente favorece la segregación del cromosoma, ya que, existirían menos encuentros físicos (colisiones) entre la RNA polimerasa y la DNA polimerasa, siendo así la segregación más eficiente [35] [36].

La naturaleza de los encuentros entre el complejo de replicación del DNA y una RNA polimerasa activa, depende de la orientación

relativa del gen en relación con la horquilla de replicación. Si un gen tiene su sitio de inicio de la transcripción en el mismo sentido que el movimiento de las horquillas de la replicación, ambas polimerasas se moverán a lo largo del templado en la misma dirección. Sin embargo, ya que las horquillas de replicación se mueven diez veces más rápido que los complejos transcripcionales, una DNA polimerasa podría alcanzar y chocar con una RNA polimerasa por la parte trasera ya que ésta se mueve más lentamente. Tal encuentro podría tener dos consecuencias, una podría ser que al chocar la DNA polimerasa con la RNA polimerasa, la moviera del templado de DNA y como consecuencia se abortaría prematuramente la transcripción. La otra sería que la DNA polimerasa disminuyera su velocidad de síntesis al alcanzar a la RNA polimerasa, así las dos sintetizarían a la misma velocidad[37] y no se abortaría ningún proceso.

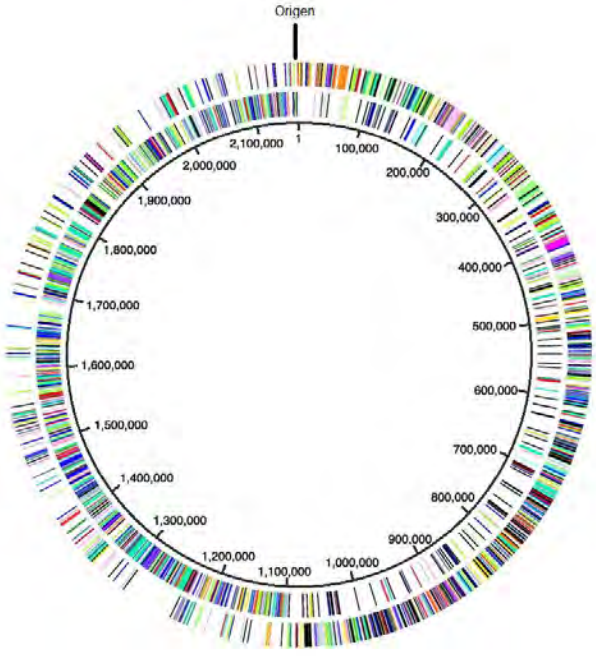


Figura 2| Ejemplo esquemático de la asimetría transcripcional en genomas bacterianos. Las barras de los dos círculos externos representan los genes, iniciando en el origen de la replicación (arriba). El círculo más externo muestra a los genes que se transcriben en el sentido de las manecillas del reloj y el siguiente círculo, los genes que se transcriben en sentido contrario. En ambas orientaciones, la gran mayoría de genes están situados en la cadena líder.

Del mismo modo, cuando un gen transcripcionalmente activo, se transcribe en sentido opuesto a la dirección de las horquillas de

replicación, el complejo de la DNA polimerasa se encontrará de frente con la RNA polimerasa (colisión de frente). Esto se debe a que el complejo de proteínas que forman la horquilla de replicación incluye una o más proteínas que están localizadas a lo largo de la cadena retrasada de DNA con una polaridad opuesta a la dirección de la polimerización. Estas proteínas son esenciales para el avance de la horquilla de replicación, como la helicasa que desenrolla el DNA de doble cadena y la primasa que provee de iniciadores para la polimerización de la cadena retrasada del DNA. Debido a que se translocan a lo largo de la cadena retrasada, dichas proteínas podrían encontrarse de frente con cualquier RNA polimerasa que se esté moviendo en dirección opuesta a la horquilla de replicación (Figura 3). Sin un mecanismo de regulación esto podría impedir que continúe el movimiento de ambas [38]. Por esta razón se ha planteado que los genes altamente expresados son posicionados preferentemente en la cadena líder para permitir una rápida replicación y minimizar la interrupción de la transcripción [39]. Entonces, la alta expresividad tendría como resultado un rápido crecimiento de la célula. Por lo tanto, células de rápido crecimiento tendrían un alto sesgo de sus genes en la cadena líder. A pesar de esto, existe evidencia que recientemente ha cambiado este énfasis. Primero en estudios en *E. coli* y *B. subtilis,* se encontró que en promedio el 78% de los genes que codifican en la cadena líder, no

son altamente expresados. Segundo, el mayor sesgo de genes en la cadena líder que se ha encontrado, ha sido en bacterias con tasas bajas de crecimiento, por ejemplo *Borrelia burgdorferi*.

Sin embargo, existe una hipótesis que sugiere que el posicionamiento preferencial de los genes en la cadena líder, es debido a la esencialidad de los mismos, no a su expresividad[40].

Si bien el concepto de gen "esencial" es ambiguo, ya que en principio todos los genes son necesarios para el suministro básico de metabolitos que son indispensables para la célula bajo circunstancias generales o especiales. Lo que se ha propuesto como un gen esencial, es un grupo de genes que están involucrados particularmente en la transferencia de la información hereditaria y la compartamentalización del genoma[40].

Una manera de agrupar a los genes se logra al determinar su función, de esta manera se les asigna un código que los identifica -Cluster Ortology Groups (COG)- independientemente de la especie de que se trate. Así se ha llegado ubicar que, en la cadena líder, preferencialmente se posicionan genes que codifican para la estructura del ribosoma, traducción, biogénesis, control del ciclo celular y división celular[41].

También se ha reportado que los genes en la cadena retrasada son en promedio 48% más cortos que los ubicados en la cadena líder y que los genes que codifican para proteínas para un tamaño mayor a 200 aminoácidos son muy escasos en la cadena retrasada. Se piensa que probablemente el tamaño está correlacionado al nivel de conflicto entre la interacción de la replicación y la transcripción[42].

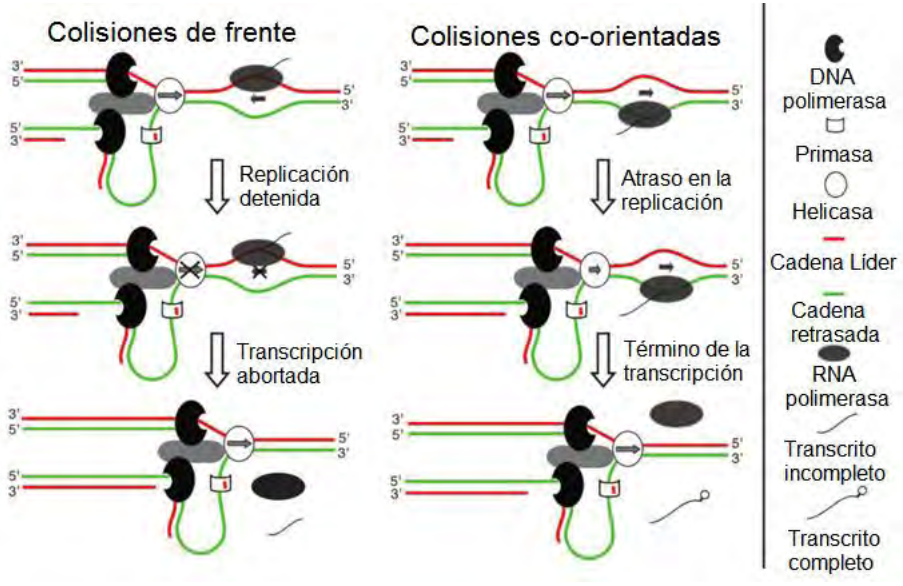


Figura 3| Tipos de colisiones entre la DNA polimerasa y RNA polimerasa. En esta figura se muestran los diferentes resultados que se esperarían cuando chocan la DNA polimerasa y la RNA polimerasa. Las colisiones de frente se dan cuando la RNA polimerasa está operando en la cadena retrasada y las colisiones co-orientadas se dan cuando la RNA polimerasa opera en la cadena líder. Cabe mencionar que hay elementos de las polimerasas que no se representan para no complicar el esquema. Tomado y modificado [35].

JUSTIFICACIÓN

Evolutivamente los genomas procariontes tienen un nivel de organización altamente conservado. Entre muchas características, una que resalta es el posicionamiento preferencial de genes en la cadena líder con respecto a la cadena retrasada.

Se han propuesto varias hipótesis para poder atribuir dicho sesgo en la cadena líder, sin embargo, sigue sin establecerse a qué se debe esta tendencia.

Por tal razón, los análisis que realizan comparaciones gen por gen de genomas completos, son una estrategia más fina que aún no se ha explorado exhaustivamente y ayudaría a fortalecer o refutar las teorías antes planteadas y poder así proponer otras.

OBJETIVOS.

- Corroborar que existe una asimetría transcripcional en los genomas procariontes.

- Identificar los genes anotados en las bases de datos y analizar su ubicación en la cadena líder y/o retrasada.

- Inferir el nivel de expresión de algunos genes seleccionados en ambas cadenas utilizando el Codon Adaptation Index (CAI), para genes que son considerados esenciales y no esenciales.

MATERIALES Y MÉTODOS.

Se seleccionaron 30 genomas de bacterias que no fueran parásitos intracelulares (que no dependa de otro organismo para su supervivencia) y 11 genomas de arqueas; éstos tenía que estar anotados y reportados en las bases de datos de National Center for Biotechnology Information (NCBI) y Kyoto Encyclopedia of Genes and Genomes (KEGG); (Tabla 1) con el fin de hacer más robusto el análisis.

Tabla 1| Genomas utilizados[a]

	Organismo	Clave KEGG	Clave NCBI	Tamaño del genoma (pb)	Número de Genes
Bacteria	*Francisella philomiragia*	fph	NC_010336	2045775	1911
	Runella slithyformis	rsi	NC_015703	6568739	5458
	Nostoc punctiforme	npu	NC_010628	8234322	6086
	Desulfurispirillum indicum	din	NC_014836	2928377	2571
	Nitrosococcus halophilus	nhl	NC_013960	4079427	3749
	Escherichia coli K-12	eco	NC_000913	4639675	4146
	Pusillimonas T7 7	put	NC_015458	3883605	3696
	Acidiphilium cryptum	acr	NC_009484	3389227	3063
	Candidatus Koribacter versatilis	aba	NC_008009	5650368	4777
	Geobacter lovleyi	glo	NC_010814	3917761	3606
	Rhodoferax ferrireducens	rfr	NC_007908	4712337	4169
	Escherichia coli (EHEC)	ece	NC_002655	5528445	5298
	Methylococcus capsulatus	mca	NC_002977	3304561	2956
	Actinobacillus succinogenes	asu	NC_009655	2319663	2079
	Nitrobacter winogradskyi	nwi	NC_007406	3402093	3122
	Sinorhizobium meliloti	sme	NC_003047	3654135	3359
	Bradyrhizobium sp.	bra	NC_009445	7456587	6717
	Yersinia enterocolitica palearctica	yep	NC_015224	4552107	3936
	Pseudogulbenkiania NH8B	pse	NC_016002	4332995	4012
	Corynebacterium efficiens	cef	NC_004369	3147090	2838
	Erwinia tasmaniensis	eta	NC_010694	3883467	3427
	Pelodictyon phaeoclathratiforme	pph	NC_011060	3018238	2707
	Sulfurimonas denitrificans	tdn	NC_007575	2201561	2096
	Nautilia profundicola	nam	NC_012115	1676444	1730
	Dictyoglomus thermophilum	dth	NC_011297	1959987	1912
	Exiguobacterium sibiricum	esi	NC_010556	3034136	3007
	Bacillus subtilis	bsu	NC_000964	4215606	4176
	Clostridium acetobutylicum	cac	NC_003030	3940880	3671
	Clostridium botulinum	cbo	NC_009495	3903260	3592
	Leuconostoc citreum	lci	NC_010471	1796284	1702
Arquea	*Nitrosopumilus maritimus*	nmr	NC_010085	1645259	1796
	Methanocaldococcus jannaschii	mja	NC_000909	1664970	1714
	Caldivirga maquilingensis	cma	NC_009954	2077567	1963
	Sulfolobus tokodaii	sto	NC_003106	2694756	2826
	Thermoproteus neutrophilus	tne	NC_010525	1769823	1966
	Hyperthermus butylicus	hbu	NC_008818	1667163	1603
	Methanosarcina acetivorans	mac	NC_003552	5751492	4540
	Pyrobaculum arsenaticum	pas	NC_009376	2121076	2299
	Methanococcus aeolicus	mae	NC_009635	1569500	1490
	Methanothermobacter thermautotrophicus	mth	NC_000916	1751377	1873
	Thermococcus kodakaraensis	tko	NC_006624	2088737	2306

[a]En la columnas se presenta el nombre del organismo (organismo), la clave con la cual se identifica en la base de datos del KEGG (clave KEGG), la clave con la cual se identifica en la base de datos del NCBI (clave NCBI), el tamaño del genoma en pares de bases (tamaño del genoma en pb) y el número total de genes que tiene cada organismo en su genoma (número de genes). Para bacterias y arqueas

Se obtuvieron las listas de genes de cada genoma completo en el servidor FTP del NCBI y se realizó la revisión de las características de cada gen, tales como ubicación en el genoma, cadena en la que está codificado, tamaño, código COG (Cluster of Orthologous Groups) y producto. Asimismo se corroboró su ubicación en el mapa genómico del NCBI correspondiente (Figura 4).

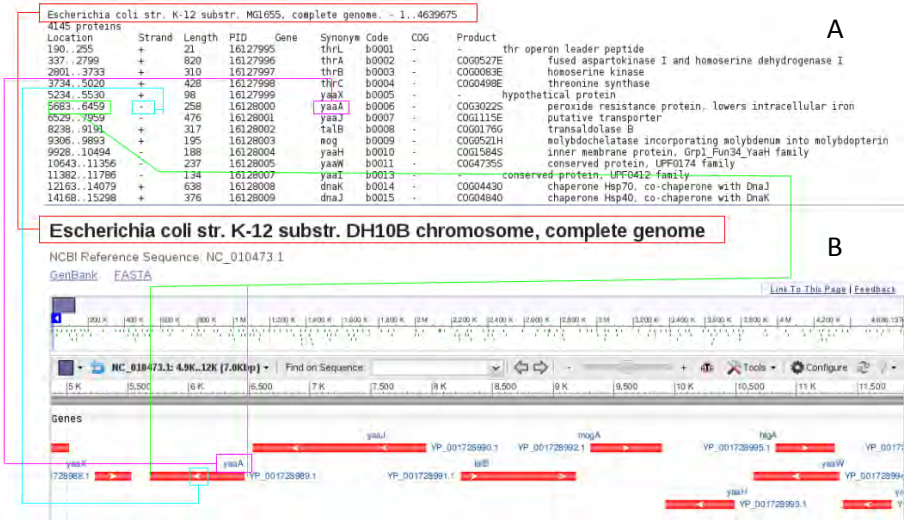
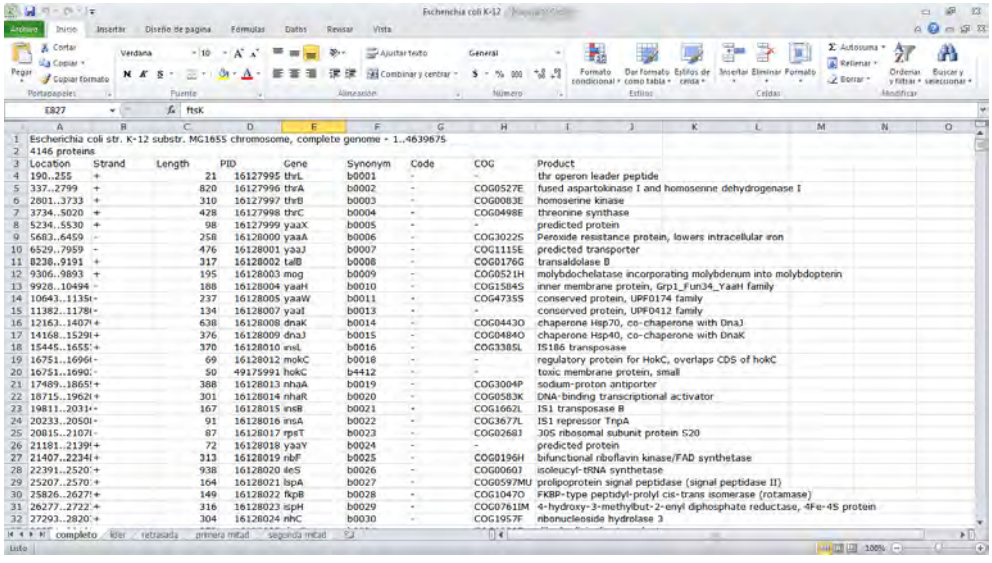


Figura 4| Corroboración de los genomas anotados. En la imagen A se presenta el genoma anotado de *E. coli* K-12 en el servidor FTP del NCBI y en la figura B se muestra el mapa genómico correspondiente, esto con el fin de corroborar que no existieran errores en la anotación.

Una vez corroborado que los datos coincidían, se cargaron los archivos del servidor FTP del NCBI en hojas de cálculo independientemente (Figura5).



Escherichia coli str. K-12 substr. MG1655 chromosome, complete genome - 1..4639675
4146 proteins

Location	Strand	Length	PID	Gene	Synonym	Code	COG	Product
190..255	+	21	16127995	thrL	b0001	-	-	thr operon leader peptide
337..2799	+	820	16127996	thrA	b0002	-	COG0527E	fused aspartokinase I and homoserine dehydrogenase I
2801..3733	+	310	16127997	thrB	b0003	-	COG0083E	homoserine kinase
3734..5020	+	428	16127998	thrC	b0004	-	COG0498E	threonine synthase
5234..5530	+	98	16127999	yaaX	b0005	-	-	predicted protein
5683..6459	-	258	16128000	yaaA	b0006	-	COG3022S	Peroxide resistance protein, lowers intracellular iron
6529..7959	-	476	16128001	yaaJ	b0007	-	COG1115E	predicted transporter
8238..9191	+	317	16128002	talB	b0008	-	COG0176G	transaldolase B
9306..9893	+	195	16128003	mog	b0009	-	COG0521H	molybdochelatase incorporating molybdenum into molybdopterin
9928..10494	-	188	16128004	yaaH	b0010	-	COG1584S	inner membrane protein, Grp1_Fun34_YaaH family
10643..11358	+	237	16128005	yaaW	b0011	+	COG4735S	conserved protein, UPF0174 family
11382..11781	-	134	16128007	yaaI	b0013	-	-	conserved protein, UPF0412 family
12163..14079	+	638	16128008	dnaK	b0014	-	COG0443O	chaperone Hsp70, co-chaperone with DnaJ
14168..15294	+	376	16128009	dnaJ	b0015	-	COG0484O	chaperone Hsp40, co-chaperone with DnaK
15445..16557	+	370	16128010	insL	b0016	-	COG3385L	IS186 transposase
16751..16964	-	69	16128012	mokC	b0018	-	-	regulatory protein for HokC, overlaps CDS of hokC
16751..16902	-	50	49175991	hokC	b4412	-	-	toxic membrane protein, small
17489..18657	+	388	16128013	nhaA	b0019	-	COG3004P	sodium-proton antiporter
18715..19624	+	301	16128014	nhaR	b0020	-	COG0583K	DNA-binding transcriptional activator
19811..20317	+	167	16128015	insB	b0021	+	COG1662L	IS1 transposase B
20233..20505	-	91	16128016	insA	b0022	-	COG3677L	IS1 repressor TnpA
20815..21076	-	87	16128017	rpsT	b0023	-	COG0268J	30S ribosomal subunit protein S20
21181..21399	+	72	16128018	yaaY	b0024	-	-	predicted protein
21407..22341	+	313	16128019	nifF	b0025	-	COG0196H	bifunctional riboflavin kinase/FAD synthetase
22391..25203	+	938	16128020	ileS	b0026	-	COG0060J	isoleucyl-tRNA synthetase
25207..25704	+	164	16128021	lspA	b0027	-	COG0597MU	prolipoprotein signal peptidase (signal peptidase II)
25826..26271	+	149	16128022	fkpB	b0028	-	COG1047O	FKBP-type peptidyl-prolyl cis-trans isomerase (rotamase)
26277..27224	+	316	16128023	sspH	b0029	-	COG0761M	4-hydroxy-3-methylbut-2-enyl diphosphate reductase, 4Fe-4S protein
27293..28204	+	304	16128024	nhC	b0030	-	COG1957F	ribonucleoside hydrolase 3

Figura 5| Transferencia de los genomas anotados en FTP de NCBI a hojas de cálculo. Se ejemplifica la transferencia de los datos anotados en el servidor FTP del NCBI para su posterior tratamiento Bioinformático. Todos los genomas anotados en el del NCBI que se seleccionaron (Tabla 1) fueron transferidos a hojas de cálculo.

Para determinar qué genes se ubican en la cadena líder y cuáles en la cadena retrasada, se dividió el tamaño del genoma a la mitad a partir del origen de la replicación, de este modo se tuvieron dos mitades. En la base de datos FTP del NCBI, de toda la información que arroja la anotación de los genomas, un dato de gran relevancia es en qué cadena se ubica cada gen. Cuando un gen se localiza en la cadena sentido, se indica con un signo “+” y cuando se encuentra en la cadena antisentido se indica con un signo “-”. Entonces, al tener el genoma dividido en dos mitades, los genes que se encuentran en la primera mitad y están anotados con el signo positivo, corresponden a genes ubicados en la cadena líder, y los que tienen el signo negativo, corresponden a los genes ubicados en la cadena retrasada. De lo contrario, en la segunda mitad, los genes anotados con signo positivo corresponden a genes ubicados en la cadena retrasada y los genes anotados con el signo negativo, son genes que se ubican en la cadena líder.

Posteriormente se contaron de los genes que codifican en la cadena líder y en la cadena retrasada, en cada genoma.

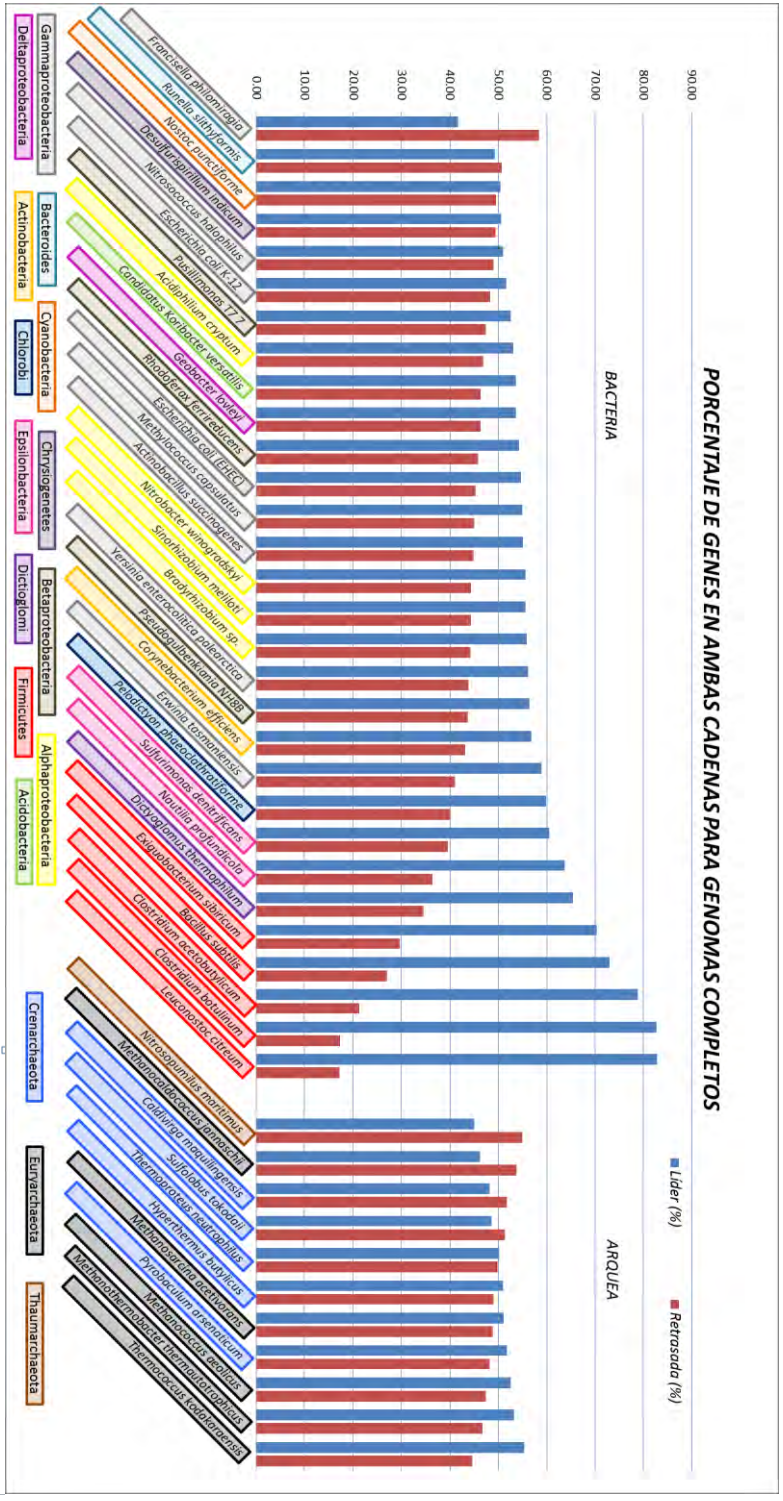
Finalmente, para todos los genomas analizados, se realizó una macro en donde se le dio las instrucciones necesarias para que ubicara cada gen, en que genomas se encontraba y la cadena en la que codificaba.

RESULTADOS Y DISCUSIÓN.

A partir de los archivos tenían el nombre del gen y la cadena en que se localiza, se llevó a cabo la separación de los genes que se ubican en la cadena líder y en la cadena retrasada y se cargaron independientemente en nuevas hojas de cálculo donde se contaron el número de genes ubicados en cada cadena y se hizo una relación porcentual con respecto al número total de genes en todo el genoma. Esto se repitió para todos los genomas y se obtuvieron dos listas independientes para cada genoma una en la que se tienen los genes codificantes en la cadena líder y otra que contiene los genes codificantes en la cadena retrasada.

Asimetría transcripcional.

Con el fin de observar si existe una asimetría en la transcripción en los genomas analizados, se realizó un histograma de los genes que codifican en la cadena líder y en la cadena retrasada, el cual muestra una tendencia a contener más genes en la cadena líder (Gráfica 1).



Gráfica 1| Se muestra una clara tendencia a contener mayor cantidad de genes en la cadena líder (arriba de 50%) para la mayoría de los genomas de las bacterias; no así para los genomas de las arqueas.

Como se puede apreciar en la Gráfica 1, efectivamente se confirma un sesgo transcripcional hacia la cadena líder en la mayoría de los genomas bacterianos; no así en los genomas de arqueas. Por esta razón sólo se continuó el trabajo con bacterias. Cabe mencionar que una característica de los genomas de las arqueas, es que, a diferencia de las bacterias, pueden tener varios orígenes de replicación. Otro genoma analizado parcialmente fue el de *Saccharomyces cerevisiae* y de igual manera no se percibe un sesgo transcripcional (datos no mostrados).

Tomando en cuenta el sesgo, se quiso analizar si existe alguna relación filogenética. Para esto, se observó que las Proteobacterias tienen genomas más simétricos que los Firmicutes, en donde existen tendencias muy marcadas a contener más genes en la cadena líder (Figura 6).

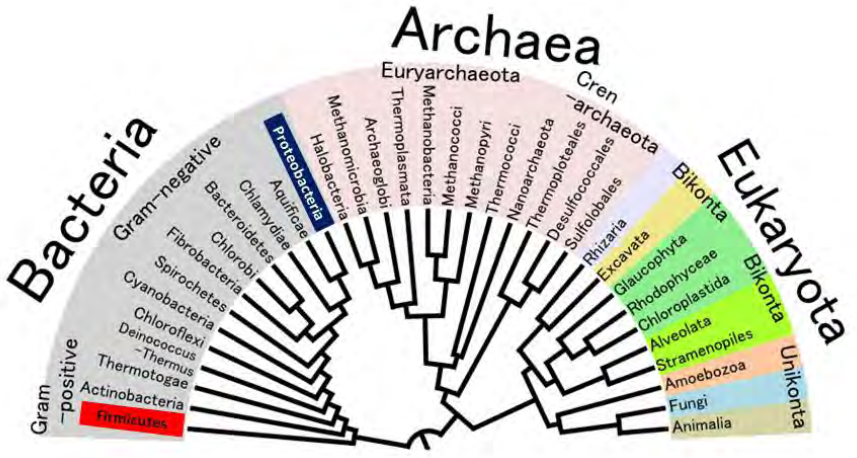


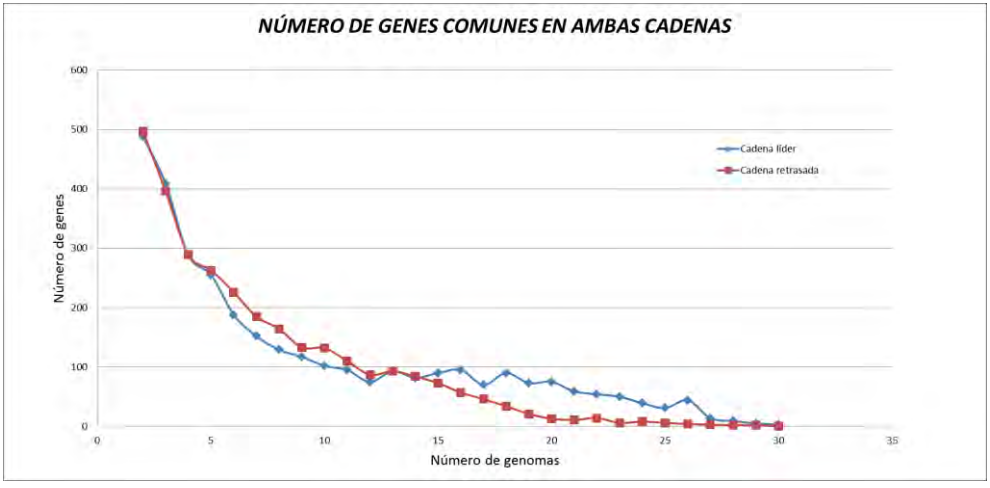
Figura 6| En rojo se muestran los Firmicutes y en azul las Proteobacterias, observando que se encuentran alejados filogenéticamente, ubicándose en los extremos del árbol para el Dominio Bacteria. Tomado [47].

Validación del sesgo en las cadenas.

Con el fin de establecer si es representativo el sesgo que se muestra en la Gráfica 1 y validar la tendencia existente en los genomas de Bacterias, se realizó una prueba Ji cuadrado. Dicha prueba arrojó como resultado un valor de Ji cuadrado de 101.17, usando una frecuencia esperada de 50. El resultado se comparó con los valores críticos de la distribución de Ji cuadrado para un intervalo de confianza de 99.9% y como parámetro 29 grados de libertad, siendo 58.30. El resultado obtenido es mayor al reportado en las tablas de Ji cuadrado, por lo tanto, es rechazada la hipótesis de que la

densidad génica es igual en ambas cadenas para los genomas de bacterias**.**

Adicionalmente, se llevó a cabo una comparación de los genes que se transcriben en una u otra cadena, utilizando el COG (Cluster Orthologous Groups) que es un código específico para cada gen, con el propósito de saber su función y poderlo asociar con las categorías reportadas en la literatura. Dicho de otro modo, se hicieron comparaciones para establecer si en diferentes genomas se encuentra una relación entre los genes y su asimetría transcripcional, ya que una hipótesis alternativa propone que los genes que se transcriben en la misma dirección que la replicación corresponden a genes “esenciales” para la supervivencia del organismo. Para esto se realizó una macro en Excel, en la cual se buscaron asociaciones del mismo COG en las treinta listas que corresponden a cada genoma bacteriano por cadena. Se realizó un histograma con los resultados obtenidos (Gráfica 2).



Gráfica 2| Los resultados arrojaron tres genes que siempre se ubican en la cadena líder en los 30 genomas de las bacterias y 2 genes que se ubican en 29 genomas en la cadena retrasada.

Tabla 2| Nombre de los genes comunes encontrados.

Cadena Líder			Cadena Retrasada		
COG	**Gen**	**Número de genomas en los que se encuentra**	**COG**	**Gen**	**Número de genomas en los que se encuentra**
COG1028IQR	3-oxoacyl-ACP reductase (related to short-chain alcohol dehydrogenases)	30	COG0642T	Signal transduction histidine kinase	29
COG2226H	Methylase involved in ubiquinone/menaquinone biosynthesis	30	COG0745TK	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	29
COG0642T	Signal transduction histidine kinase	30			

Dado el sesgo transcripcional reportado en los genomas bacterianos, si la mayor cantidad de los genes ubicados en la mayoría de los genomas bacterianos tienden a estar en la cadena líder, se podría esperar que la conservación de genes comunes fuese mayor en la cadena líder con respecto a la cadena retrasada en cada evento de comparación. Como se muestra en la Gráfica 2, el porcentaje de genes comunes localizados en la cadena líder y en la cadena retrasada, disminuye de manera similar cuando se comparan mayor número de genomas.

Así, se puede predecir que si aumentamos el número de genomas analizados, habrá un momento en el cual ya no existan genes que sean comunes en ninguna de las cadenas.

Este hecho indica la no conservación en la localización de genes individuales en la cadena líder o retrasada en los 30 genomas analizados.

Esto sugiere que la posición de los genes (cadena líder o cadena retrasada) en genomas bacterianos es un proceso de ubicación aleatoria.

Un parámetro no analizado fue la ubicación física de los genes en el mapa genómico, con respecto a otros genomas.

Estos resultados contradicen la hipótesis de la esencialidad, planteada por algunos investigadores, ya que en el análisis realizado, los genes que siempre se encontraron ubicados en la cadena líder, no codifican para productos relacionados con la estructura del ribosoma, traducción, biogénesis, control del ciclo celular o división celular[41].

Por el contrario, nuestro análisis muestra que son pocos los genes que se presentan con alta frecuencia en la cadena líder y retrasada de los genomas analizados. En el caso de la histidincinasa, que está localizada con una alta frecuencia en ambas cadenas, es probable que represente una familia multigénica derivada por eventos de duplicación.

Los dos genes encontrados en la cadena retrasada, se encuentran en el operón llamado Vic. Este operón contiene 3 genes. A los dos presentados en la Tabla 2, se les ha encontrado función, mientras que el tercero aún no se ha descrito y aparece anotado como proteína hipotética [43]. Sirven para adaptarse a cambios en el ambiente, teniendo distintas vías para una amplia variedad de estímulos, incluyendo nutrientes, estados celulares redox, cambios en la osmolaridad, quorum sensing, antibióticos y más [44].

Asimismo, se hizo un análisis comparativo para poder determinar si los genes comunes encontrados en ambas cadenas están presentes

con alta frecuencia en todos los genomas utilizados, independientemente de si están ubicados en la cadena líder o retrasada (Tabla 3).

Tabla 3| Número de copias por genoma[b].

Bacteria	COG			
	1028IQR	2226H	0642T	0645KT
Acidiphilium cryptum JF-5	32	9	7	11
Actinobacillus succinogenes 130Z	6	3	3	3
Bacillus subtilis subsp. subtilis str	25	8	15	14
Bradyrhizobium sp. ORS 278	51	19	36	17
Candidatus Koribacter versatilis	14	18	21	26
Clostridium acetobutylicum ATCC 824	8	9	21	22
Clostridium botulinum A ATCC 3502	5	5	25	27
Corynebacterium efficiens YS-314	8	7	5	6
Desulfurispirillum indicum S5	4	7	23	8
Dictyoglomus thermophilum	7	2	1	3
Erwinia tasmaniensis Et1/99	9	4	14	10
Escherichia coli K-12 MG1655	15	6	17	14
Escherichia coli O157:H7 EDL933 (EHEC)	17	8	18	15
Exiguobacterium sibiricum 255 15	19	3	11	16
Francisella philomiragia	9	2	2	3
Geobacter lovleyi	3	15	51	16
Leuconostoc citreum	8	2	2	6
Methylococcus capsulatus str	5	13	10	7
Nautilia profundicola	4	4	4	11
Nitrobacter winogradskyi Nb-255	8	10	14	9
Nitrosococcus halophilus	6	15	8	7
Nostoc punctiforme	29	29	68	38
Pelodictyon phaeoclathratiforme	8	7	5	2
Pseudogulbenkiania NH8B	13	7	17	14
Pusillimonas T7 7	32	5	13	15
Rhodoferax ferrireducens	24	11	26	19
Runella slithyformis DSM	31	16	26	15
Sinorhizobium meliloti 1021	26	5	11	8
Sulfurimonas denitrificans	3	6	18	22
Yersinia enterocolitica palearctica	11	7	14	12

[b]Se muestra el número de copias que están presentes por genoma para los genes con alta frecuencia de la Gráfica 2. No se tomó en cuenta la cadena en la cual se ubican, por lo tanto, se utilizaron los genomas completos.

Con el fin de determinar si los genes comunes encontrados en ambas cadenas tienen la característica de ser altamente expresados, se obtuvieron los valores de CAI (Codon Adaptation Index), los cuales se asocian y pueden predecir el nivel de expresión de un gen. Comúnmente, los valores de CAI son usados como referencia de genes altamente expresados de una especie, determinando el uso de cada codón sinónimo [45]. Así, se obtuvieron los valores de CAI de *Nautilia profundicola* ya que se encuentran reportados [46].

Los resultados no fueron estadísticamente significativos (datos no mostrados); por lo anterior, decidimos hacer un análisis más enfocado.

Se eligieron algunos genes que se consideraron esenciales, los cuales codifican para la estructura del ribosoma, la estructura de la DNA polimerasa, algunas chaperonas y los genes comunes encontrados en el análisis anterior, para la cadena líder, hasta sumar un total de diez genes. Se obtuvieron las secuencias, se hizo un concatenado de las diez secuencias y se utilizó el programa llamado Codon Usage Database para traducir la secuencia concatenada y visualizar el uso de codones preferenciales, y así compararlos con los valores de CAI reportados. Lo mismo se hizo para la cadena retrasada (Tabla 4). Los resultados se presentan en la Tabla 5.

Tabla 4| Genes utilizados para hacer la secuencia concatenada[c].

Cadena líder	*Cadena retrasada*
50S ribosomal protein L25/general stress protein Ctc	*Chaperonin GroS*
DNA polymerase III delta prime subunit HolB	*Chaperonin GroEL*
Chromosomal replication initiator protein DnaA	*DNA polymerase III subunit alpha*
DNA-directed RNA polymerase subunit beta	*UDP-N-acetylglucosamine acyltransferase*
Replicative DNA helicase	*Sensory box/ggdef domain protein*
PAS/PAC sensor signal transduction histidine kinase	*Amidophosphoribosyltransferase*
Arabinose efflux permease	*DNA primase dnaG*
30S ribosomal protein S2	*Putative cupin domain protein*
Anthranilate/para-aminobenzoate synthases component I	*Dihydroxy-acid dehydratase*
Chaperone protein DnaJ	*Ribosomal large subunit pseudouridine synthase C*

[c]Indica los genes que se consideraron para hacer dos secuencias independientes concatenadas, una para la cadena líder (azul) y otra para la cadena retrasada (rojo).

Tabla 5/ Uso preferencial de codones.

Cadena Líder																
AA	Codón	/1000	Número	AA	Codón	/1000	Número	AA	Codón	/1000	Número	AA	Codón	/1000	Número
Phe	U U U	37.2	162	Ser	U C U	9.4	41	Tyr	U A U	33	144	Cys	U G U	1.4	6
	U U C	10.6	46		U C C	2.1	9		U A C	8.3	36		U G C	1.4	6
Leu	U U A	28.2	123		U C A	10.8	47	Alto	U A A	2.1	9	Alto	U G A	0.2	1
	U U G	9.2	40		U C G	3	13		U A G	0	0	Trp	U G G	3.2	14
	C U U	41.3	180	Pro	C C U	14.9	65	His	C A U	10.6	46	Arg	C G U	3	13
	C U C	4.6	20		C C C	0.9	4		C A C	5.7	25		C G C	0.7	3
	C U A	7.8	34		C C A	6.9	30	Gln	C A A	14.5	63		C G A	0.2	1
	C U G	8.5	37		C C G	9.2	40		C A G	11	48		C G G	0	0
Ile	A U U	50	218	Thr	A C U	12.4	54	Asn	A A U	30.3	132	Ser	A G U	15.8	69
	A U C	18.1	79		A C C	5.5	24		A A C	22	96		A G C	6.9	30
	A U A	30.1	131		A C A	17.7	77	Lys	A A A	101	440	Arg	A G A	33.7	147
Met	A U G	24.3	106		A C G	6.2	27		A A G	8.7	38		A G G	5.7	25
Val	G U U	29.4	128	Ala	G C U	9.9	43	Asp	G A U	39.5	172	Gly	G G U	25.2	110
	G U C	2.5	11		G C C	8.7	38		G A C	18.4	80		G G C	3.2	14
	G U A	23.4	102		G C A	20.9	91	Glu	G A A	68.4	298		G G A	19.7	86
	G U G	8.3	36		G C G	8.3	36		G A G	19.3	84		G G G	6.9	30
Cadena Retrasada																
AA	Codón	/1000	Número	AA	Codón	/1000	Número	AA	Codón	/1000	Número	AA	Codón	/1000	Número
Phe	U U U	27.8	133	Ser	U C U	7.93	38	Tyr	U A U	21.08	101	Cys	U G U	5.43	26
	U U C	17.7	85		U C C	4.59	22		U A C	17.12	82		U G C	2.71	13
Leu	U U A	24.4	117		U C A	10.23	49	Alto	U A A	1.25	6	Alto	U G A	0.42	2
	U U G	4.59	22		U C G	2.5	12		U A G	0.42	2	Trp	U G G	1.18	9
	C U U	33	158	Pro	C C U	13.78	66	His	C A U	9.81	47	Arg	C G U	3.13	15
	C U C	9.18	44		C C C	2.92	14		C A C	9.39	45		C G C	1.25	6
	C U A	8.56	41		C C A	4.38	21	Gln	C A A	10.44	50		C G A	0.21	1
	C U G	9.18	44		C C G	14.61	70		C A G	12.73	61		C G G	0.42	2
Ile	A U U	36.7	176	Thr	A C U	11.06	53	Asn	A A U	24	115	Ser	A G U	5.84	28
	A U C	27.3	131		A C C	4.8	23		A A C	29.64	142		A G C	15.65	75
	A U A	32.8	157		A C A	18.37	88	Lys	A A A	91	436	Arg	A G A	25.88	124
Met	A U G	24	115		A C G	9.81	47		A A G	5.01	24		A G G	5.22	25
Val	G U U	20.3	97	Ala	G C U	16.7	80	Asp	G A U	32.77	157	Gly	G G U	20.87	100
	G U C	4.17	20		G C C	13.78	66		G A C	25.67	123		G G C	13.15	63
	G U A	25.5	122		G C A	26.3	126	Glu	G A A	76.81	368		G G A	25.26	121
	G U G	7.72	37		G C G	12.73	61		G A G	13.15	63		G G G	5.01	24

El análisis del uso de codones para diez genes en cada dirección transcripcional se muestra en la Tabla 5 para el mismo genoma, el cual es de *Nautilia profundicola*. En verde se muestran las coincidencias, las no coincidencias en rojo; y en amarillo se muestra cuando dos codones tienen una frecuencia aproximada. Como se

puede observar esta comparación muestra una preferencia similar en el uso de codones sinónimos, siendo una indicación de que no existe sesgo en los niveles expresión en ninguna de las cadenas y que ésta no es una fuerza mayoritaria para la ubicación de los genes.

CONCLUSIONES.

- Nuestros resultados confirman que existe un sesgo transcripcional en la distribución de genes en las cadenas de genomas de las 30 bacterias analizadas.

- En los genomas de las 11 arqueas estudiadas no se encontró un sesgo transcripcional.

- No se encuentra una dirección transcripcional específica para genes “esenciales”.

- Los genes que siempre se encuentran situados en la cadena líder son pocos, no encontrando sesgo por genes específicos, por lo tanto, no existe evidencia de que su distribución sea debido a su esencialidad.

- Se evaluó si el sesgo es debido a la expresividad, y los resultados indican que el uso de codones es parecido en ambas cadenas, por lo tanto, la expresividad no es un factor que influya en el posicionamiento de los genes en las cadenas.

- La asimetría transcripcional en una característica genómica evolutiva que no depende de los genes encontrados en el genoma.

BIBLIOGRAFÍA.

1. Nelson, D. L. & Cox, M. M. *Lehninger Principles of Biochemistry.* **4th edition,** (2005).

2. Dahm, R. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Hum. Genet.* **122,** 565–81 (2008).

3. Van Slyke, D. D., and Jacobs, W. Biographical Memoir of Phoebus Aaron Theodor Levene. *Natl. Acad. Sci.* **23,** 75–86 (1945).

4. Vilchis, A. & Valdés, V. Los 60 años del modelo de la doble hélice de Watson y Crick. *iBio* **1,** 2–6 (2013).

5. Griffith, F. The significance of pneumococcal types. *J. Hyg* **27,** 8–35 (1928).

6. Chase, M. & Hershey, A. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* **36,** 39–56 (1952).

7. Chargaff, E., Vischer, E., Doniger, R., Green, C. & Misani, F. The composition of the desoxypentose nucleic acids of thymus and spleen. *J. Biol. Chem.* **177,** 405–416 (1949).

8. Chargaff, E., Lipshitz, R. & Green, C. Composition of the desoxypentose nucleic acids of four genera of sea-urchin. *J. Biol. Chem.* **195,** 155–160 (1952).

9. Rudner, R., Karkas, J. D. & Chargaff, E. Separation of *B. subtilis* DNA into complementary strands, III. Direct analysis. *Proc. Natl. Acad. Sci. U. S. A.* **60,** 921–922 (1968).

10. Klug, A. Rosalind Franklin and the Discovery of the Structure of DNA. *Nature* **219,** 808–844 (1968).

11. Franklin, R. E. & Gosling, R. G. Molecular Configuration in Sodium Thymonucleate. *Nature* **171,** 740–741 (1953).

12. Watson, J. & Crick, F. Molecular structure of nucleic acids. *Nature* **171,** 737–738 (1953).

13. Watson, J. & Crick, F. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171,** 964–967 (1953).

14. Brenner, S., Jacob, F. & Meselson, M. An unestable carrying information from genes to ribosomes for protein synthesis. *Nature* **190,** 576–581 (1961).

15. Crick, F., Barnett, L., Brenner, S. & Watts-Tobin, J. General nature of the genetic code for proteins. *Nature* **192,** 1227–1232 (1961).

16. Vavak, F., Fogarty, T. C. & Jukes, K. A. A genetic algorithm with variable range of local search for tracking changing environments, parallel problem solving from nature IV. *Lect. Notes Comput. Sci.* **1141,** 376–385 (1996).

17. Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260,** 500–507 (1976).

18. Sanger, F. & Coulson, a R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94,** 441–8 (1975).

19. Sanger, F. & Nicklen, S. DNA sequencing with chain-terminating. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467 (1977).

20. Sanger, F. *et al.* Nucleotide sequence of bacteriophage φX174 DNA. *Nature* **265,** 687–695 (1977).

21. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

22. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–51 (2001).

23. Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408,** 708–13 (2000).

24. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512 (1995).

25. Adams, M. D. The Genome Sequence of *Drosophila melanogaster*. *Science (80-.).* **287,** 2185–2195 (2000).

26. Hodgkin, J., Plasterk, R. H. A. & Waterston, R. H. The nematode *Caenorhabditis elegans* and its genome. *Science.* **270,** 410–414 (1995).

27. Bevan, M. *et al.* Sequence and analysis of the *Arabidopsis* genome. *Curr. Opin. Plant Biol.* **4,** 105–10 (2001).

28. Valdés, V. & Vilchis, A. Genes y genomas. *Mensaje Bioquímico* **25,** 17–33 (2001).

29. Rocha, E. P. C. The organization of the bacterial genome. *Annu. Rev. Genet.* **42,** 211–33 (2008).

30. Lundgren, M., Andersson, A., Chen, L., Nilsson, P. & Bernander, R. Three replication origins in *Sulfolobus* species: synchronous initiation of chromosome replication and asynchronous termination. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 7046–51 (2004).

31. Norais, C. *et al.* Genetic and physical mapping of DNA replication origins in *Haloferax volcanii*. *PLoS Genet.* **3,** e77 (2007).

32. Kato, J. & Hashimoto, M. Construction of consecutive deletions of the *Escherichia coli* chromosome. *Mol. Syst. Biol.* **3,** 132 (2007).

33. Olsen, G. J. & Woese, C. R. Archaeal Genomics : An Overview Minireview. *Cell* **89,** 991–994 (1997).

34. Langer, D., Hain, J., Thuriaux, P. & Zillig, W. Transcription in archaea: similarity to that in eucarya. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 5768–72 (1995).

35. Rocha, E. P. C. The replication-related organization of bacterial genomes. *Microbiology* **150,** 1609–27 (2004).

36. Yuzhakov, a, Turner, J. & O'Donnell, M. Replisome assembly reveals the basis for asymmetric function in leading and lagging strand replication. *Cell* **86,** 877–86 (1996).

37. Brewer, B. J. When polymerases collide: replication and the transcriptional organization of the E. coli chromosome. *Cell* **53,** 679–86 (1988).

38. Liu, B. & Alberts, B. Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex. *Science.* **267,** 1131–1137 (1995).

39. Rocha, E. Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol.* **10,** 393–5 (2002).

40. Rocha, E. P. C. & Danchin, A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat. Genet.* **34,** 377–8 (2003).

41. Lin, Y., Gao, F. & Zhang, C.-T. Functionality of essential genes drives gene strand-bias in bacterial genomes. *Biochem. Biophys. Res. Commun.* **396,** 472–6 (2010).

42. Paul, S., Million-Weaver, S., Chattopadhyay, S., Sokurenko, E. & Merrikh, H. Accelerated gene evolution through replication-transcription conflicts. *Nature.* **495,** 512–5 (2013).

43. Wagner, C. & Saizieu, A. De. Genetic analysis and functional characterization of the *Streptococcus pneumoniae* vic operon. *Infect.* **70,** 6121–6128 (2002).

44. Wolanin, P. M., Thomason, P. A. & Stock, J. B. Protein family review Histidine protein kinases : key signal transducers outside the animal kingdom. *Genome Biol.* **3,** 1–8 (2002).

45. Sharpl, P. M. & Li, W. The codon adaptation index a mesure of directional synonymus codon usage bias, and its potential applications. *Nucleic Acids Research*. **15,** 1281–1295 (1987).

46. Codon Used Bias Database. at <http://cub-db.cs.umt.edu/whichalg.php?org_name=NautiliaprofundicolaAmH&fileRoot=large_scale/59345/59345>

47. Gribaldo S & Brochier-Armanet C. The origin and evolution of Archaea: a state of the art. *Philos Trans R Soc Lond B Biol Sci.* **361,** 1007-1022 (2006).