UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN CIENCIAS BIOLÓGICAS
FACULTAD DE CIENCIAS
BIOLOGÍA EVOLUTIVA

DINÁMICA EVOLUTIVA DE LOS GENOMAS EN ORGANISMOS HIPERTERMOFÍLICOS

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTOR EN CIENCIAS

PRESENTA:

HÉCTOR GILBERTO VÁZQUEZ LÓPEZ

TUTOR PRINCIPAL DE TESIS:
DR. ARTURO CARLOS II BECERRA BRACHO FACULTAD DE CIENCIAS, UNAM

COMITÉ TUTOR:
DR. RAFAEL CAMACHO CARRANZA INSTITUTO DE INVESTIGACIONES BIOMEDICAS, UNAM
DR. PEDRO MIRAMONTES VIDAL FACULTAD DE CIENCIAS, UNAM

MÉXICO, D.F. MARZO, 2016

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN CIENCIAS BIOLÓGICAS
FACULTAD DE CIENCIAS
BIOLOGÍA EVOLUTIVA

DINÁMICA EVOLUTIVA DE LOS GENOMAS EN ORGANISMOS HIPERTERMOFÍLICOS

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTOR EN CIENCIAS

PRESENTA:

HÉCTOR GILBERTO VÁZQUEZ LÓPEZ

TUTOR PRINCIPAL DE TESIS:
DR. ARTURO CARLOS II BECERRA BRACHO FACULTAD DE CIENCIAS, UNAM

COMITÉ TUTOR:
DR. RAFAEL CAMACHO CARRANZA INSTITUTO DE INVESTIGACIONES BIOMEDICAS, UNAM
DR. PEDRO MIRAMONTES VIDAL FACULTAD DE CIENCIAS, UNAM

MÉXICO, D.F. MARZO, 2016

UNAM
POSGRADO
Ciencias Biológicas

COORDINACIÓN

POSGRADO EN CIENCIAS BIOLÓGICAS
FACULTAD DE CIENCIAS
DIVISIÓN DE ESTUDIOS DE POSGRADO

OFICIO FCIE/DEP/030/16

ASUNTO: Oficio de Jurado

Dr. Isidro Ávila Martínez
Director General de Administración Escolar, UNAM
P r e s e n t e

Me permito informar a usted que en la reunión ordinaria del Comité Académico del Posgrado en Ciencias Biológicas, celebrada el día **28 de septiembre de 2015**, se aprobó el siguiente jurado para el examen de grado de **DOCTOR EN CIENCIAS** del (la) alumno (a) **VÁZQUEZ LÓPEZ HÉCTOR GILBERTO** con número de cuenta **96504585** con la tesis titulada: **"Dinámica evolutiva de los genomas en organismos hipertermofílicos"**, realizada bajo la dirección del (la) **DR. ARTURO CARLOS II BECERRA BRACHO:**

Presidente:	DR. GERMINAL COCHO GIL
Vocal:	DR. LUIS FELIPE JIMÉNEZ GARCÍA
Secretario:	DR. RAFAEL CAMACHO CARRANZA
Suplente:	DRA MARÍA COLÍN GARCÍA
Suplente	DRA. ALICIA NEGRÓN MENDOZA

Sin otro particular, me es grato enviarle un cordial saludo.

A T E N T A M E N T E
"POR MI RAZA HABLARA EL ESPÍRITU"
Cd. Universitaria, D.F. a 14 de enero de 2016

DRA. MARÍA DEL CORO ARIZMENDI ARRIAGA
COORDINADORA DEL PROGRAMA

POSGRADO EN CIENCIAS BIOLOGICAS
COORDINACIÓN

MCAA/MJFM/ASR/ipp

Agradecimiento

Al Posgrado de Ciencias Biológicas, UNAM.

Al apoyo por parte de CONACYT por medio de la beca de doctorado.

Agradezco al Doctor Arturo Carlos II Becerra Bracho, por su tutoría.

Al Comité Tutoral conformado por: el Doctor Rafael Camacho Carranza y el Doctor Pedro Miramontes Vidal: gracias por su tutoría.

Agradecimiento a título personal

A Cuauhtemoc, por su amistad, fraternidad y enseñanza.

A sus hijos por enseñarme mas de lo que yo veía en mi.

A la familia Tola, que me han dado su apoyo, asilo y sobretodo, que me han enseñado el significado de una familia.

A mi madre, Virginia López, por apoyarme mas allá de su propia existencia, gracias mamá.

A mi padre y su esposa, por presentarme un ejemplo y una filosofía de vida.

A todos mis alumnos de preparatoria, por que me hicieron recordar mi cariño por lo que nunca pense recordar: mi gusto por las matemáticas.

A mis alumnos de licenciatura, gracias por retarme y hacerme ser mejor profesor cada día.

Al Sagrado Dominus Et Magister, que ha logrado apoyarme de corazón siempre.

A Arturo Becerra por su tutoria y amistad constante.

Al laboratorio de Origen, gracias por enseñarme y apoyarme siempre con una sonrisa y demostrarme que la amistad y la academia estan unidas.

Índice

Índice de figuras

Índice de tablas

Resumen

La estructura y composición de los genomas es moldeada por el ambiente y por interacciones biológicas. El caso de los organismos hipertermófilos no es la excepción. Para analizar el impacto de estos cambios, nos enfocamos a medir índices de heterogeneidad, regiones en tándem, secuencias de alta flexibilidad y pangenómica comparada. El resultado fue el reconocer que a) Los índices de Quintana reflejan una proporción de índices H,L y V similar tanto dentro de los cromosomas como en los plásmidos, que b) el arreglo que presenta en regiones de alta flexibilidad es diferencial a lo reportado presentàndose tanto entre cromosoma como plásmido; c) no se presenta ninguna secuencia común a este estilo de vida, aunque la estructura de su arreglo y composición tenga una huella compartida en el índice V y d) es posible que se presenten eventos compartidos de duplicación. La conclusión a este trabajo es que a pesar de que se presenta carentes evidencias de genes y metabolismos, la estructura de los cromosomas hipertermofílicos denota motivos con propiedades comunes a los genomas que evidencían una convergencia hacia la estructura y composición de los cromosoma y plásmidos presentes. De manera colateral, ha sido posible el reconocer que el estilo de vida halófilo presenta rasgos comunes a nivel estructural y pangenómico.

Abstract

The structure and composition of the genomes is shaped by the environment and biological interactions.The case of the hyperthermophilic organisms is no exception. To analyze the impact of these changes, we aim to measure levels of heterogeneity regions tandem sequences and comparative genome-wide high flexibility. The result was the recognition that a) Quintana rates reflect a proportion of indices H, L and V similarly both within the chromosomes and plasmids, which b) the arrangement presented in regions of high flexibility is differential to that reported presenting both between chromosome and plasmid; c) any sequence common to this lifestyle is not presented, although the structure of the arrangement and composition has a shared mark on the V and d) may share duplication events occur. The conclusion of this work is that despite lacking evidence of genes and metabolism occurs, the structure of chromosomes hyperthermophilic denotes motifs common to the genomes show a convergence towards the structure and composition of the chromosome and plasmids present properties. Collaterally, it has been possible to recognize that the halophilic lifestyle have common structural and genome-wide pangenomic features.

Dinámica evolutiva de los genomas en organismos hipertermofílicos

1 Capítulo I Introducción

1.1 La dinámica evolutiva en los seres vivos

La dinámica evolutiva se puede definir como la diversificación, adquisición, modificación y pérdida de caracteres moleculares o genómicos, los cuales están expuestos a diferentes factores que alteran su estructura y composición [Ochman y Jones, 2000]. Si bien la dinámica evolutiva se presenta como una herramienta para comparar las consecuencias que diferentes variables las cuales pueden alterar la configuración genómica de los seres vivos, en este estudio hemos decidido realizar una comparación a un nivel mas ínfimo, intentando, a partir de la cuantificación y medición de las propiedades del DNA usando diferentes índices, el reconocer ragos comunes a nivel de DNA genómico compartidos a un estilo de vida hipertermófilo, siendo estos el resultado de eventos evolutivos compartidos o el resultado de una adaptación frente a condiciones fisicoquímicas comunes.

1.2 Estudio de la estructura del DNA y los organismos extremófilos

La propuesta de Zuckerkandl y Pauling cambió el modo de como estudiamos a los seres vivos. Al usar a las moléculas bioinformacionales como documentos de la historia evolutiva [Zuckerkand y Pauling, 1965], se diversificó la manera en que analizamos la biología molecular. A partir de diferentes estudios, ha sido posible originar desde genomas completamente secuenciados (era genómica), como desarrollar métodos para identificar e interpretar la estructura, interacción y diversificación de diferentes marcadores moleculares (era post-genómica) [Tettelin *et al.*, 2008]. Los resultados de diferentes grupos de investigación han permitido concluir que: *a)* las moléculas bioinformacionales han divergido por diferentes mecanismos moleculares, incluso de manera reciente [Zhu *et al.*, 2013] y *b)* cada moléculas parece presentar una historia evolutiva particular [Doolittle y Brown, 1994].

Si bien es cierto que la biología molecular presenta una nueva oportunidad para estudiar a los seres vivos de manera indirecta, la importancia de la historia evolutiva obtenida por comparación de secuencias es limitada. Se

depende de secuencias o regiones específicas en las que se suponen cambios únicos y graduales a lo largo del tiempo [Martínez-Cano *et al.*, 2015] y que obedecen a una tasa de substitución constante, la cual se puede correlacionar con el tiempo geológico y la separación de diferentes grupos taxonómicos [Galtier y Duret, 2007].

Sin embargo, existen secuencias y regiones del genoma que pueden ser alteradas de manera directa por una presión de selección. Estos pueden ser: una región codificante afectada por un mecanismo de recombinación o un factor ambiental, el cual remodela de manera diferencial las secuencias. Tal es el caso de las algunas moléculas de tRNA en organismos hipertermofílicos, los cuales aumentan su concentración de guanina y citosina(contenido de GC o "GC amount"), favoreciendo con ello un aumento en su vida media [Marck y Grosjean, 2002]. Es por ello que muchas veces se ha optado por analizar el genoma de forma completa para así reconocer sesgos o rasgos moleculares compartidos por todas las regiones: codificantes y no codificantes.

En un intento por analizar la estructura del genoma y su historia evolutiva: *i)* se han desarrollado estrategias para analizar secuencias codificantes, apoyándose en la idea de que éstas, como marcadores moleculares, pueden reflejar la historia evolutiva del organismo que lo porta, vinculándolo con un estudio parcial de genómica comparada y marcadores moleculares obtenidos, analizándolos como genes o incluso como regiones codificantes o marcos de lectura (ORF: Open Reading Frame). Por otra parte, *ii)* se ha propuesto, basándose en estudios por cristalografía, interacciones parciales de decámeros y dodecámeros, índices y formas de medir y cuantificar interacciones entre los nucleótidos. Entre los índices más importantes que podemos mencionar son los índices de Quintana [Quintana *et al.*, 1992] y los de Miramontes [Miramontes *et al.*, 1995]. Ambos permiten reconocer aspectos estructurales del DNA, como los tipos de interacción que se realiza entre cada uno de los dinucleótidos, reconocer propiedades termodinámicas diferenciales y flexibilidad del DNA.

1.3 Índices de Miramontes o Índices de Heterogeneidad

Algunos análisis realizados sobre la estructura del DNA han demostrado que la conformación estructural es el promedio de un conjunto de cambios conformaciones locales [Miramontes *et al.*, 1995]. Una de las aproximaciones que se ha presentado dentro de diferentes grupos de estudio ha sido la teórica, basándose ésta en conocimientos de cristalografía e interacción bioquímica de decámeros frente a diferentes secuencias y condiciones ambientales [Miramontes *et al.*, 1995, Tereshko *et al.*, 1999, Dickerson y Ng, 2001].

Estos estudios han permitido aproximar el impacto de secuencias ricas en adenina y timina (ricas en AT o "AT rich") como aquellas ricas en guanina y citosina (ricas en GC o "GC rich") dando con esto desde reconocer propiedades diferenciales de flexibilidad torsión y respuesta a condiciones ricas en iones o incluso su interacción diferencial ante medios acuosos [Dickerson y Ng, 2001]. Complementando estos estudios, existen grupos de investigación que caracterizaron la presencia de ambos tipos de pares de nucleótidos para definir que tan heterogéneo es el DNA y por tanto, cuales son sus implicaciones en su estructura y su secuencia.

La heterogeneidad y los patrones estructurales presentes en el DNA se han intentado explicar en diversas ocasiones. Según Miramontes [Miramontes *et al.*, 1995], los factores como el sesgo mutacional y el uso de codones, aunque afectan la forma de distribución de los dinucleótidos en las secuencias del DNA, no son los únicos factores que determinan este fenómeno. Dichos elementos, no son suficientes para explicar las diferencias globales entre los genomas de los diferentes organismos, las propiedades estructurales del DNA también juegan un papel esencial al respecto. Estas propiedades estructurales junto con lasus propiedades termodinámicas dependen en gran medida más que de la composición de nucleótidos, y de su distribución a lo largo de la secuencia. El dinucleótido es la unidad que se utiliza para medir la distribución de las bases que tiene un genoma tanto en los genomas procariontes omo eucariontes, al ser una medición teòrica que permite ver su transición como su flexibilidad de forma B a A [Marathe y Bansal, 2011].

Usando como fundamentos estos principios, es posible proponer que la energìa de giro y enrollamiento depende del número y tipo de interacciones que se presente dentro de de todos los ángulos intra nucleótidos, asi como las interacciones parciales dentro de los mismos. La sumatoria de interacciones parciales entre cada dinucleòtido al verlo como una gran columna espiral se define como energìa de apilamiento [Quintana *et al.*, 1992].

El hecho de que la energía de apilamiento de la doble hélice se distribuya de manera heterogénea en las secuencias, depende en primer lugar de la regla de apareamiento entre las bases fuertes GC (s), y las bases débiles AT (w). Así, se sugiere que los fragmentos de las secuencias de desoxirribonucleótidos con una mayor cantidad de pares de GC agregados, tenderían a ser estructuralmente más rígidos y termodinámicamente más estables [Miramontes *et al.*, 1995].

Otra característica estructural del DNA es la distribución de sus bases considerando su tamaño: las bases pequeñas, que son las (pirimidinas, Y), con respecto a las de mayor tamaño (purinas, R) pueden originar interacciones diferentes. De este modo si se encuentran regiones de purinas alternadas con pirimidinas geométricamente la estructura del DNA adquiriría una conformación más irregular.

Una tercera propiedad estructural de la doble hélice, aunque poco estudiada, es la distribución de las bases tipo M (A, C, que exponen un grupo amino, hacia la parte externa del zurco mayor), y las bases tipo K (T, G, que exponen un grupo cetona). A pesar de que no se sabe el efecto que tiene la agregación de dichas bases M y K, se especula que podrían facilitar las interacciones entre el DNA frente a metales y algunas proteínas. Aunque el significado biológico de los patrones que se encuentran a partir del análisis de la distribución de estos 3 tipos de dinucleótidos (WS, RY, MK), aun no se comprende del todo, se considera, que dicho orden podría funcionar como un factor importante en el reconocimiento regiones codificantes, interacciones no viables dentro de ciertas secuencias o patrones especie específicoso o frente a la estructura del genoma mismo[Miramontes *et al.*, 1995] frente al análisis particular en la estructura de genes y motivos de unión a proteínas en particular [Hong *et al.*, 2008].

A continuación se anexa una tabla en la que se hace un resumen de las caracterìsticas de los índices de heterogeneidad:

La forma binaria en que se pueden representar los elementos estructurales del DNA (W vs S; Y vs R; y M vs K), ha sugerido la formulación del índice IDH que expresa el nivel de heterogeneidad estructural que tienen las secuencias del DNA, al medir el nivel de agregación (1) o alternancia (-1) que tienen sus componentes [Miramontes *et al.*, 1995]. Con ello es posible establecer un sistema binario para medir estos índices y así obtener resultados promedio.

1.4 Índices de Quintana

Para describir la estructura del DNA, Quintana se basa en dos ideas principales:

a) La doble hélice no es una estructura rígida indeformable. Algunas regiones son más susceptibles a la deformación por la influencia del ambiente y tienden a modificarse más fácilmente que otras.

Tabla 1: Cuadro de relación de las variaciones de los Índices de Heterogeneidad basado en los trabajos de Miramontes y colaboradores [Miramontes *et al.*, 1995]

Propiedad	Equivalencia en estructura y secuencia	Valores
Energía dentro de pares de bases	La energìa de los pares de bases de GC tenderán a ser más rígidos y termodinámicamente mas estables	Bases Fuertes GC (s); Bases débiles AT (w)
Tamaño de los nucleótidos	Por ser mas grande una purina que una pirimidina, si estas están seguidas, producirán un arreglo irregular	Purinas (R); Pirimidinas (Y)
Orientaciòn de grupos de nucleótidos hacia el surco mayor	Dependiendo de el grupo ceto o amino es expuesto al surco	Base tipo M (amino) (A y C); Base tipo K (T y G)

b) La organización estructural tridimensional del DNA depende de las formas en las que los pares de bases puedan ordenarse espacialmente, y de la energía de apilamiento existente entre cada uno de ellos. [Quintana *et al.*, 1992]

De acuerdo con Quintana el dinucleótido es la principal unidad estructural del DNA, y estudia sus relaciones con las bases adyacentes. Según sus análisis, cada dinucleótido tiene diferentes características estructurales, como el ángulo de giro (Twist), su elevación (Rise), formar una copa (Cup) y modificar su abertura (Roll), esto simplemente por influencia de los peldaños vecinos (Figura 1) [Quintana *et al.*, 1992].

Gracias a esto, Quintana estableció las diferentes interacciones presentes, definiendo los tipos de interacción con letras diferentes y denotando relaciones entre ángulos y arreglos formados (Figura 1), definiendo asi que:

- Todas las interacciones purina - purina (R-R) se definen con la letra L excepto la interacción de guanina y adenina (un tipo de interacción H)
- Las interacciones purina - pirimidina se definen como I con excepción de la interacción guanina - citosina (tipo de interacción H)
- Las interacciones de pirimidina - purina se definen todas como V

Luego de proponer y subclasificar estas interacciones (Figura 2), las cuales son consecuencias de los niveles de energía de apilamiento que adoptan los dinucleótidos, fue posible concluir por Quintana, y por su grupo de trabajo, que cada uno de los pares de dinucleótidos está relacionado con los pares de dinucleótidos más cercanos, a patir de esta interacción se describe la hipótesis de la tétrada.

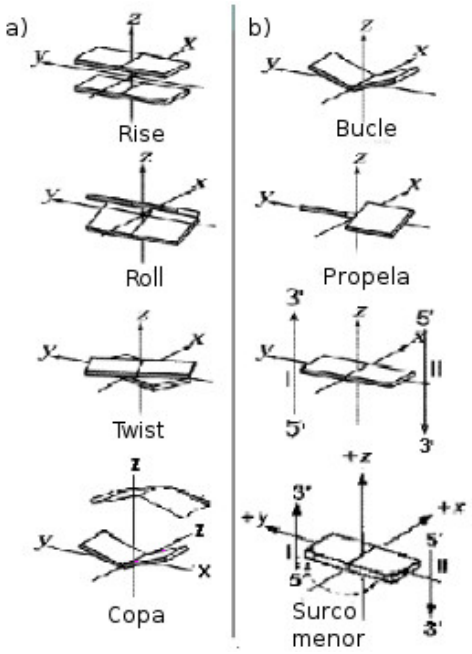


Figura 1: Fundamento por el que se establecen los estudios de Quintana. a) Variables establecidas dentro de dos dinucleótidos que definen las interacciones L, I, V y H respectivamente. b) Tipos de ángulos que puede haber entre un solo par de bases, es decir, entre una purina y una pirimidina. el dinucleótido, acorde a lo mostrado con esta figura, permite establecer la separación entre los dinucleótidos (Rise: levantamiento), la inclinación presente solamente en uno de los extremos del dímero (Roll: enrollamiento), el giro de uno de los pares de nucleótidos frente a otros (Twist: giro) y la separación de los mismos, incluso dando un cambio conformacional entre ellos (Cup: copa)

1.5 Hipótesis de la tétrada y flexibilidad del DNA

Ésta postula que si los escalones de los extremos de un tetranucleótido son de tipo H o I, el escalón central estaría forzado a ser de tipo L; mientras que los extremos no son H, sino L, el escalón del centro sería entonces de tipo H. Este postulado, aunque caracteriza de forma más contundente los escalones del DNA, no ha sido desarrollado lo suficiente [Quintana *et al.*, 1992].

Por último, el modelo de Quintana se puede explicar metafóricamente haciendo una analogía entre un pedazo largo del DNA y un brazo humano, de modo que el brazo humano, al igual que el DNA, no es una estructura estática y tiene tanto restricciones como flexibilidad estructural. Algunas partes del brazo, como la muñeca, pueden doblarse y girar. Otras, como los antebrazos, son más estáticas y no pueden realizar ninguna de las funciones anteriores.

Un aspecto que se considera para integrar las interacciones purina - purina y purina - pirimidina, es el factor de flexibilidad y de transición dentro de la molécula de DNA. Esto se relaciona directamente con los primeros trabajos que estaban enfocados en reconocer a partir de la cristalografía de decámeros y dodecámeros, diferentes propiedades relacionadas a este tópico [Dickerson y Ng, 2001]. Posterior a la propuesta relacionada por parte Quintana en 1992, el grupo de Dickerson continuó estudiando la cristalografía y propiedades diferenciales de decámeros, intentando vincular

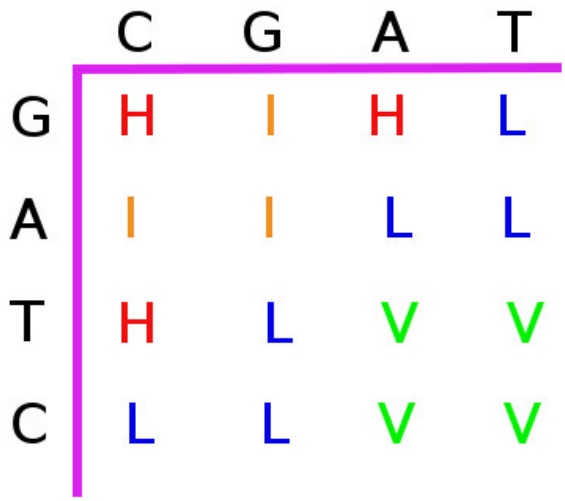


Figura 2: Interacciones entre dímeros de acuerdo con el modelo de Quintana y colaboradores [Quintana *et al.*, 1992]. En este esquema se mide la interacción entre dímeros, representada por los valores presentes en el marco horizontal y vertical. La combinatoria en este cuadro permite definir cuando se lleva a cabo cada interacción

en ellos las tendencias de las tres formas o estructuras básicas del DNA. Estas son descritas en la Tabla 3

La forma del DNA cuando se encuentra en sistemas vivos, en su mayoría esta representado por las formas B y A [Dickerson y Ng, 2001], del mismo modo se ha propuesto que el genoma de todos los seres vivos puede encontrarse fluctuante entre estas dos formas [Quintana *et al.*, 1992, Dickerson y Ng, 2001]. Esto se basa en lo previamente reportado por los trabajos de Goodsell y colaboradores [Goodsell *et al.*, 1993], en donde a partir de un contexto diferencial de purinas y pirimidinas, se encuentran variaciones en los valores de giro (twist) y enrollamiento (roll) [Goodsell *et al.*, 1993].

Para el año de 1999 [Tereshko *et al.*, 1999] se proponía que ciertas secuencias pudieran cambiar en sus valores de enrollamiento: es decir, cambiar de una forma B a una forma A. En trabajos de Ng y colaboradores [Ng y Dickerson, 2002] se propone que no hay una diferencia significativa entre ambas formas o al menos para ciertas secuencias repetitivas de guanina y citosina (G3C3 o GCGCGC), si se encuentran enmarcadas por regiones ricas en AT o con trímeros como CAG o CAC, éstas pueden cambiar su enrollamiento [Dickerson y Ng, 2001, Ng y Dickerson, 2002].

Otro factor que favorece este cambio y flexibilidad es la presencia de regiones ricas en AT. En el mismo trabajo de Ng del 2002 donde se identificó la presencia de regiones ricas en AT que interactúan con iones del medio, esto analizado a partir de cristalografía. De este estudio se concluye que los iones monovalentes cercanos a regiones ricas en AT, da cierta estabilidad a la estructura y permite dar origen a regiones en donde se anidan dipolos que dan estabilidad al DNA, denominándolos bolsillos o ”pockets” de AT, de estos, se ha propuesto que no son una causa, sino una consecuencia de las diferentes interacciones entre la secuencia misma y que el apilamiento de la secuencia puede ser

Tabla 2: Cuadro de relación de las variaciones de los Índices de Quintana, síntesis obtenida de la propuesta de Quintana ([Quintana *et al.*, 1992])

Valores	Pares de bases relacionados	Propiedades
Valor L	R - R o Y - Y	Baja torsión (*Low Twist*): los pares de nucleótidos consecutivos forman un menor ángulo entre ellos. Cup valor negativo
Valor I	R - Y (purina - pirimidina)	Torsión intermedia (*Intermediate Twist*): mayor valor de roll de todos los arreglos. Arreglo dependiente de la interaccion de los pares de base cercanos
Valor V	Y - R (pirimidina - purina)	Variable Twist: angulo varía en virtud de las condiciones ambientales
Valor H	G - C y G - A	Alta torsión (textitHigh Twist): ángulo de giro entre los pares de bases con mas diferencia entre un par a otro. Raise, intermedio y Cup con valor positivo

Tabla 3: Cuadro comparativo de las formas de DNA en sus características principales, basados en los trabajos de Diekmann [Diekmann, 1989] y Dickerson [Dickerson y Ng, 2001]

Propiedad	DNA B	DNA A	DNA Z
Sentido de torsión	Derecha	Derecha	Izquierda
Nucleótidos por vuelta	10	11	12
Aumento vertical por cada par de bases	3.4	2.56	3.7
Rotación por cada par de bases	34.7 grados	32.7 grados	-30 grados
Twist promedio de propela	16 grados	18 grados	0 grados
Inclinación de pares de bases al eje	-1.2 grados	+19 grados	-9 grados
Diámetro de la hélice (Armstrongs)	20	23	18
”Pocket” de azúcar	C3-endo	C2-endo	Citosina C2-endo; Guanina C3-endo

un factor para el cambio en la estructura de la columna de DNA frente a esos iones[Schuerman y van Meervelt, 2000, Ng y Dickerson, 2002].

Previamente se había reportado la importancia de estas secuencias en la interacción DNA-proteína dentro de la interacción con los telómeros [Schuerman y van Meervelt, 2000], así como interacciones frente a la regulación de expresión de genes [Rhodes y Klug, 1986, McCall *et al.*, 1986] de este modo, el identificar regiones en donde se presente una frecuencia mayor de regiones ricas en GC frente a un genoma completo nos permita evaluar el impacto de la fluctuación de valores en donde, los valores de deslizamiento o ”slide” pueden ir aumentado y gradualmente fluctuar los valores de enrollamiento o ”roll” (Figura 3).

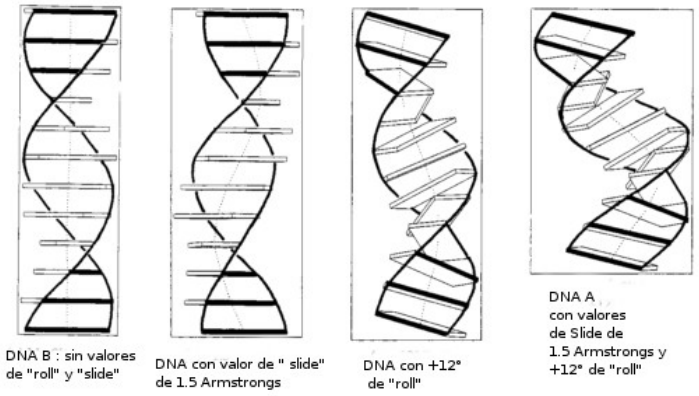


Figura 3: Comparativa entre los valores de roll y slide que presentan las formas B y A del DNA. Un DNA B con la flexibilidad de cambiar en estructura se comportará fluctuante entre la segunda y la tercera imagen, de izquierda a derecha, aumentando sus valores de ”roll” y ”slide”. Imagen modificada del trabajo de Dickerson y Ng [Dickerson y Ng, 2001].

1.6 Importancia del estudio de dímeros dentro de otras disciplinas. El caso eucarionte

Recientemente se ha aplicado el estudio de dímeros en diferentes campos de estudio. Tal es el caso de los estudios de dímeros de guanina y citosina. Estos, por una parte, han sido reportados como regiones que presentan mayor incidencia a la actividad de los complejos de reparación y recombinación [Sved y Bird, 1990, Schofield y Hsieh, 2003]. Por otra parte, se ha logrado reconocer procesos de epigenética implicados en estos dímeros. Ciertas enzimas logran metilar a las citosinas en estas secuencias, provocando un impedimento en el reconocimiento de algunas secuencias de inicio de transcripción y logrando con ello una regulación negativa en la expresión de ciertos genes [Karlin y Mrázek, 1997]. Adicionalmente en estos sistemas de regulación, se ha logrado describir que esta metilación logra incluso rearreglar los complejos de DNA e histonas presentes en los eucariontes [Murrell *et al.*, 2005].

Este mecanismo se ha reportado predominantemente en genomas eucariontes y en especial, como un factor importante en los genomas de primates como del ser humano [Yavartanoo y Choi, 2013], sin embargo, la metilación de citosina y guanina parece no ser un factor presente en la regulación y estructura de los genomas procariontes.

Wojciechowski y colaboradores [Wojciechowski *et al.*, 2013] evaluaron la incidencia de estos dímeros como la distribución diferencial en tres diferentes especies procariontes, con ello identificaron al menos dos papeles importantes.

El primero es que la incidencia de estas secuencias facilita, en el caso de organismos patógenos, el poder evadir la respuesta inmune específica. La segunda importancia es al parecer meramente estructural y esta vinculada a la carga de material genético presente en los grupos de proteobacterias con alto contenido de GC [Wojciechowski *et al.*, 2013]

Es por ello que despues de haber presentado la importancia desde las regiones ricas en GC como sus dímeros hemos plateado utilizar los Índices de Heterogeneidad de Miramontes y los índices propuestos por Quintana, la cual forma parte del grupo de trabajo de Dickerson, para así buscar una correlación entre ciertas regiones de GC o AT frente a diferentes estilos de vida.

2 Objetivos de tesis

2.1 Objetivos generales

Reconocer a nivel estructural rasgos compartidos dentro del genoma de organismos hipertermófilos a nivel de DNA o genes.

2.2 Objetivos específicos

- Calcular e identificar los índices de Quintana y Miramontes para cromosomas y plásmidos para así reconocer arreglos comunes al estilo de vida hipetermofílico.
- Calcular e identificar regiones de alta flexibilidad en hipertermófilos y mesófilos para así establecer una comparativa de su estructura y composición.
- Calcular e identificar regiones repetitivas en tándem en hipertermófilos y mesófilos para establecer una comparativa de su estructura y composición.
- Realizar una genómica comparada para identificar elementos comunes del estilo de vida hipertermófilo usando desde el programa BLAST hasta con *gethomologues* para así reconocer elementos compartidos y comprobar valores comunes al estilo de vida extremófilo.

3 Capítulo II. Articulo publicado.

3.1 Introducción. Estilos de vida, Mesofília y extremofilia

Un estilo de vida se define como aquellas condiciones que permiten el desarrollo óptimo de un organismo dentro de un medio. Los procariontes presentan la característica de que a pesar de que sean de linajes diferentes, pueden converger en características genómicas similares [Altermann, 2012]. Por otra parte, especies bacterianas filogenéticamente

relacionadas, incluso sepas o variedades de una misma especie, cuando se encuentran en diferente ámbito ecológico diferente, pueden presentar asimismo, una diversidad genómica substancial [Dutta y Paul, 2012], con ello se puede elucidar una historia que en la que el genoma, ha sido modificado a través de generaciones en virtud del nicho que ocupan ciertos organismos. Las consecuencias producidas por el estilo de vida de un microbio pueden constribuir de manera significativa en la estructura del genoma, estableciendo elementos diferenciales únicos, así como elementos que tienen impacto para las comunidades de organismos de las cuales forma parte. Uno de los estilos de vida que se puede definir de manera mas cercana y práctica es el estilo de vida mesófilo.

El estilo de vida mesófilo se ha definido a partir del punto de vista parcial del estilo de vida eucarionte, definiéndose en temperaturas que no superan ambientes mayores de 40^oC o menor a 10^oC [Stetter, 2006], en ambientes que no superen 1 atm de presión y condiciones de salinidad que no excedan la concentración 1M de NaCl [Pikuta *et al.*, 2007]. Como puede observarse, la mayoría de los eucariontes requerimos estas condiciones para poder desarrollarnos de forma adecuada, a diferencia de los grupos de organismos procariontes, los cuales presentan al menos dos grandes grupos en condiciones ambientales: los mesófilos e extremófilos.

Los extremófilos son todos aquellos organismos que viven en condiciones que sobrepasan las condiciones mesófilas. En un intento por reconocer variaciones o motivos comunes a un estilo de vida se han estudiado los estilos de vida hipertermófilo, termofílico y halófilo por diferentes líneas de investigación, nosotros nos enfocamos en este estudio a las características de los hipertermófilos y de manera colateral, a los organismos halófilos.

3.1.1 Parámetros estructurales del DNA en organismos hipertermófilos

Los organismos termófilos e hipertermófilos son organismos que prosperan en condiciones de 50^o a 80^o C y ma-yores a 80^o C respectivamente [Stetter, 2006, Boussau *et al.*, 2008]. Se han caracterizado estrategias adaptativas a nivel de genómica y proteómica. Una de estas estrategias es el intentar establecer la relación entre la temperatura óptima de crecimiento (OGT, por sus siglas en inglés: *Optimal Growth Temperature* frente a la estructura y composición de nucleótidos. Por una parte se ha logrado caracterizar que los genes de tRNA y rRNA presentan un mayor contenido de GC, al grado de establecer una correlación positiva con su temperatura óptima de crecimiento [Lightfield *et al.*, 2011, Dutta y Paul, 2012], lo cual se debe a que esta composición permite dar mas estabilidad en su estructura secundaria, provocando del mismo modo un aumento en su tiempo de vida media [Su *et al.*, 2013], sin embargo, no se presenta una correlación directa cuando se analiza su estructura genómica. Lineas de investigación han intentado establecer una correlación entre la OGT y los estilos de vida termofílicos, mesofílicos y psicrofílicos.

Uno de los trabajos que logra establecer de manera clara ambos factores es el trabajo de Dutta y Paul [Dutta y Paul, 2012], donde se logra reconocer una relación directa entre la composición de dímeros de purina y pirimidina en la siguiente manera:

RR + YY - RY - YR

así, es posible generar un coeficiente de correlación lineal de 0.66, lo cual permite correlacionar un cambio en la arquitectura del DNA frente a la extremofilia [Dutta y Paul, 2012].

Otra evidencia entre la OGT y estructura del genoma de extremófilos es el trabajo de Das y colaboradores [Das *et al.*, 2006], en donde a partir del estudio de la arqueobacteria *Nanoarchaeum equitans* se reconoce un sesgo en la proporción de purinas frente a pirimidinas en todos los marcos de lectura, en ambas cadenas del DNA. La incidencia compartida de este fenómeno en la mayoría de genomas hipertermofílicos arqueobacterianos ha permitido reconocer regiones codificantes en los proyectos de genoma completo de hipertermófilos. La última vertiente que estudia la estructura y su dinámica evolutiva se basa en la incidencia y frecuencia del uso de codones.

3.1.2 Uso de codones, incidencia de aminoácidos e hipertermofília

En los organismos hipertermofílicos se ha postulado que los sesgos dentro de el uso de codones obedece a dos variaciones en particular: por un lado, se ha postulado que el factor que altera de manera directa la selección y uso de codones, así como variaciones en las proporciones de aminoácidos, se debe predominantemente al plegamiento y termoestabilidad frente a condiciones internas de los hipertermófilos. Con esto, se logra identificar en proteínas ortólogas, que hay un aumento en la frecuencias de aminoácidos cargados como el ácido glutámico, glicina y arginina frente a la frecuencia de aminoácidos polares, los cuales se ven disminuidos [Jaenicke y Böhm, 1998, Dutta y Paul, 2012]. Por otro lado, se ha reconocido la presencia de cisteína, aminoácido que promueve la formación de puentes disulfuro [Mallick *et al.*, 2002], permitiendo con ello mas tolerancia a la degradación y óptimo plegamiento de enzimas, frente al aumento de temperatura. Esto permite suponer que únicamente el plegamiento de las proteínas es el factor que remodela la estructura y la composición de los genomas, sin embargo esto no es del todo cierto.

En una propuesta integrativa, se ha logrado vincular la estructura del DNA, de codones prefe-renciales que dan más estabilidad a las moléculas de RNA por un lado, por otro lado se presenta una relación gradual frente a los codones y el caracter de cada uno de los aminoácidos Ejemplos que postulan que estos factores se presentan como una constante son los trabajos de Dutta y Paul [Dutta y Paul, 2012] y el trabajo reportado por Li y colaboradores [Li *et al.*, 2007]. Adicionalmente se ha propuesto que la presencia de estos sesgos se encuentran delimitados por la concentración de GC presente en la estructura y contenido de GC que se encuentra en cada uno de los linajes de organismos analizados [Dutta y Paul, 2012].

Aunado a este escenario, se han podido caracterizar otras estrategias y posibles adaptaciones las cuales complementan las características antes descritas. Estas responden a la presencia de un estrés termico u osmótico, estas son:

- La presencia de proteínas tipo histonas las cuales protegen el genoma y regulan el sistema de transcripción.

Tabla 4: Proteínas relacionadas a la termotolerancia reportadas en literatura reciente

Proteína	**Característica que da termoestabilidad**	**Organismo del que se obtuvo**	**Referencia**
Acetil CoA sintetasa	Uso de ion Mg++	*Pyrococcus furiosus*	[Vieille y Zeikus, 2001]
Glutamato deshidrogenasa	Puentes salinos	*P. furiosus*	[Kurz, 2008, Zeldovich *et al.*, 2007]
Sac7d	Lisina monometilada	*Sulfolobus acidocaldarius*	[She *et al.*, 2001]
Adenilato cinasa pirofosfatasa superóxido dismutasa	Modificación en al estructura y secuencia. Incierta	*S. acidocaldarius*	[Stetter, 1996]
Proteína tipo Histona rHMfB	Extremos amino de prolina, estructura predominantemente hidrofóbica	*Methanothermus fervidus*	[Pereira y Reeve, 1998]
Superóxido dismutasa	Puentes de sal, estructura hidrofóbica, tetrámero polimérico	*Aquifex pyrophilus*	[Deckert *et al.*, 1998]
Fosforibosil antranilato isomerasa	Homodímero modificado	*Thermotoga maritima*	[Greaves y Warwicker, 2007]

[Ronimus y Musgrave, 1996].

- La incidencia de sistemas del reparación de DNA con componentes únicos para el dominio Archaea.
- La adición de radicales dentro de la estructura de los ácidos grasos complejos que conforman la membrana plasmática, esto permite conformar una membrana más rígida y estable ante la temperatura y el intercambio iónico [Konings *et al.*, 2002] aislante al medio y
- la presencia de solutos compatibles. Es este último rasgo el que se desea abordar más a fondo ya que permite hilar a este grupo de microrganismos con aquellos de los que se han obtenido resultados positivos.

Otras características descritas dentro de la literatura se anexan en la Tabla 4.

Ademas de la identificación de rasgos moleculares dentro de diferentes proteínas (Tabla 4) se ha podido reconocer que todas las secuencias de aminoácidos provenientes de hipertermofílicos presentan una sustitución y preferencia de aminoácidos polares frente a los aminoácidos cargados en colecciones de genes ortólogos, esto es evidente al compararla con secuencias homólogas de origen mesófilo [Kumar y Nussinov, 2001].

Del mismo modo se ha logrado identificar un sesgo estructural, detevctando un aumento en aminoácidos hidrofóbicos en sus proteínas [Cambillau y Claverie, 2000], esto se ha justificado por que esta composición parece dar mas estabilidad y favorece la interacción frente a las moléculas de agua, las cuales presentan un contacto y afinidad limitados en estos ambientes [Tekaia y Yeramian, 2006]. Los conocimientos que permiten comprender la termoestabilidad en la estructura terciaria han hecho posible el desarrollo de enzimas cuya aplicación en el campo de la biotecnología, ha sido invaluable. Sin embargo la importancia a la que nos enfocaremos es dirigida hacia su proceso evolutivo.

3.1.3 Los organismos hipertermófilos y su impacto en la comprensión de la dinámica evolutiva

Para el caso de organismos hipertermofílicos, las condiciones y principales factores que pueden fungir como presiones de selección son las altas temperaturas ambientales ya que su temperatura óptima de crecimiento supera los 50^oC, los ambientes en donde se presentan en mayoría son anaeróbicos o cuentan con una baja concentración de oxígeno y en algunos casos, la presencia de un pH ácido es una limitante adicional. Aunado a esto, las necesidades y las limitaciones inherentes al estilo de vida se presentan vinculadas a elementos diferenciales frente a compuestos relacionados al azufre, nitrógeno, compuestos orgánicos e incluso óxidos metálicos [Stetter, 1996].

Sobre los organismos hipertermofílicos, se ha logrado reconocer que hay más de una estrategia para resolver el problema de termoestabilidad (Stetter, 1999) Del mismo modo, no hay un factor predominante o principal que se vincule con todas las especies conocidas dentro de este estilo de vida. Un ejemplo es que a pesar de encontrar solutos compatibles con una distribución universal, no es en todas las especies en que se encuentran ni todos los pasos de la ruta metabólica y estos en ninguno de los casos se presentan como única estrategia.

Otro caso antes vinculado como un caracter definitorio del estilo de vida extremófilo, es la presencia y expresión de la enzima reverso girasa, una enzima que induce el superenrollamiento positivo del DNA, dando con ello mayor estabilidad al material genético y con ello evitando su desnaturalización. Un estudio del 2004, pudo refutar su caracter vital a la extremofília debido a que una especie fue alterada, limitando su expresión y no con esto se evitó la supervivencia ante altas temperaturas. [Atomi *et al.*, 2004]. Esto nos permite reconocer que aun hay propiedades y mecanismos desconocidos de termoregulación y estabilidad presentes en diferentes especies.

Varias estrategias se han usado para reconocer o proponer nuevos rasgos moleculares comunes que va desde proteínas de choque térmico (heat shock proteins) [Trent, 2000], hasta el reconocer rutas metabólicas compartidas [Allers y Mevarech, 2005], estas no permiten esclarecer las relaciones filogenéticas y estilo de vida de ambos dominios procariontes.

De los elementos y rasgos indirectos los cuales pueden citarse por diferentes grupos de investigación se encuentran rasgos a nivel de incidencia y presencia de nucleótidos y aminoácidos comunes a más de un grupo filogenético. Dentro de los trabajos de Groussin y Gouy [Groussin y Gouy, 2011], se pueden reconocer dos elementos moleculares comunes a todos:

a) La tasa de cambio y mutación está sesgada de manera directa dentro de todos los genes que codifican para proteínas frente a todos aquellos genes que codifican RNA. Todos los genes de RNA presentes en organismos hipertermófilos presentan un aumento significativo de contenido de GC frente a los genes codificantes de proteínas [Klein *et al.*, 2002].
b) Es posible correlacionar la temperatura óptima de crecimiento con sesgos presentes dentro de secuencias codificantes y no codificantes, con ello, se ha identificado un aumento en contenido de GC en moléculas de tRNA [Groussin y Gouy, 2011]. Asimismo, este sesgo se encuentra presente en las secuencias codificantes de las amino acil

tRNA sintasas, las cuales presentan un uso de codones sesgados hacia un mayor contenido de GC y relacionando este aumento con la temperatura óptima de crecimiento de los organismos [Klipcan *et al.*, 2006].

c) Por último, gracias a un estudio comparativo de Agarwal y Grover del 2008 [Agarwal y Grover, 2008], en el que se analiza el uso e incidencia de purina dentro de los genomas hipertermófilos fue posible identificar cambios en la estructura del genoma al uso de codones y la presencia de ciertos aminoácidos mayoritariamente cargados o acídicos. Sin embargo esto no presenta una constante que nos permita extrapolar resultados hacia un conjunto de genes que codifiquen a proteínas que evidencíen más elementos comunes hacia un ancestro común al estilo de vida hipertermófilo.

La dificultad principal de estudiar este estilo de vida se presenta principalmente por el hecho de encontrarse dentro de los dos dominios procariontes. Los elementos limitantes de estas condiciones ambientales suponen el transporte horizontal selectivo. Dos casos se pueden citar a este respecto son el estudio comparativo de Zhaxybayeva en 2009 [Zhaxybayeva *et al.*, 2009], en donde se logró reconocer varios caracteres comunes relacionados a la termotolerancia en Thermotogales, los cuales se comprueba que son compartidos en todo el linaje, esto hace posible suponer que todos estos elementos provienen de un origen bacteriano. El otro trabajo es el estudio de Martins y colaboradores en 1996, en donde logró reconocer que la síntesis de ciertos compuestos resuelven la tolerancia ante el aumento en la temperatura y en la salinidad [Martins *et al.*, 1996].

Lo cual permite esclarecer que:

- Existen caracteres que se originaron dentro de un solo grupo filogenético que resuelven la tolerancia a altas temperaturas. Esto permite que estrategias para este estilo de vida debieron de haberse originado en mas de un grupo filogenético en mas de una ocasión.
- Algunos compuestos y estrategias que ayudan a aumentar la tolerancia a altas temperaturas, del mismo modo, pueden resolver otras condiciones ambientales extremas a las que se enfrentan.

Para evaluar los elementos bioinformacionales presentes en estas bacterias y reconocer su dinámica evolutiva hemos retomado argumentos establecidos dentro de los trabajos de Conant del 2008 [Conant y Wolfe, 2008]

La plasticidad que reconoce este investigador en diferentes especies y linajes procariontes hace mas compleja la manera en que se visualiza como cambia el material genético y como remodelan su estructura genómica a nivel de rutas metabólicas e interacciones [Conant y Wolfe, 2008].

La subclasificación de genes en genoma núcleo (core) y genoma flexible ha permitido reconocer procesos diferenciales dentro del genoma: por una parte el genoma núcleo agrupa todos aquellos genes housekeeping, que hacen posible el metabolismo basal y estructural de los seres vivos, estos genes pueden estar compartidos dentro de todos los representantes de una población [Conant y Wolfe, 2008].

Para el caso del genoma flexible, se identificaron todos los genes implicados dentro de la estructura y función desarrollada en la dinámica de bienes públicos (public goods) que es la presencia de aquellos genes incidentes dentro

del proceso de interacción metabólica con las demás especies presentes dentro del genoma, en donde se permite tanto un intercambio como una interacción local e interespecífica en las mismas poblaciones de un genoma. Este genoma flexible se encuentra completamente expuesto a mayor eventos de transporte horizontal, recombinación o relacionado a regiones ”hotspots” [Cordero y Polz, 2014].

En este estudio, del mismo modo se logró integrar la incidencia de elementos del genoma flexible predominantemente en el material extracromosómico que dentro del genoma mismo, es más, se estableció que la presencia de islas génicas, las cuales se trasponen de manera modular dentro de las moléculas bioinformacionales, se encuentra presente en mayor escala en los plásmidos que en el cromosoma mismo. Es en los plásmidos y dentro del genoma fle-xible en donde se puede reconocer genes de baja frecuencia poblacional y al mismo tiempo, al correlacionarlos con el ambiente, son capaces de evidenciar las condiciones locales a las que han sido expuestos [Coleman *et al.*, 2006, Cordero y Polz, 2014].

Esto complementa el conocimiento previo que se tiene de algunos organismos hipertermofílicos, los cuales, en algunos casos pueden estar expuestos a eventos de recombinación y de transporte horizontal [Wiezer y Merkl, 2005, Krupovic *et al.*, 2013] como de conversión y duplicación génica [Archibald y Roger, 2002], haciendo posible preguntar como es el arreglo que se presenta del genoma flexible dentro de los hipertermófilos siendo este diferencial o con la misma dinámica evolutiva que sus contrapartes mesófilas.

Por ello, se desarrolló de manera teórica y bioinformacional una metodología comparativa en la que es posible reconocer de una forma dentro de las bases de datos del NCBI a todos los genomas de hipertermofílicos, los cuales presentan tanto cromosomas como plásmidos. Al compararlos con regiones de alta flexibilidad como de su uso de codones y la incidencia de regiones y secuencias repetitivas, fue posible reconocer la presencia de patrones diferenciales acordes dentro de los genomas y plásmidos, así como la incidencia de procesos locales cambio dentro del cromosoma y el plásmido.

3.2 Material y métodos

Para identificar diferencias entre el cromosoma y el plásmido, asi como la flexibilidad y estructura en organismos hipertermófilos, se usaron todos los genomas de hipertermofílicos presentes en el sitio de internet del NCBI (ftp://ncbi.nlm.nih.gov) que contasen con cromosomas y plásmidos para asíc compararlos entre sí.

La información compilada se presenta dentro de la siguiente tabla:

Siete organismos hipertermófilos fueron seleccionados de la base de datos de NCBI, del mismo modo se seleccionó uno de los metanógenos, para tenerlo como control y un punto de comparación frente a los valores de Quintana, uso de codones y estructura de material extracromosómico.

Tabla 5: Organismos hipertermofílicos analizados dentro del estudio. Dentro de la tabla se señala con una c los valores de cromosoma y con p los valores de plásmidos. Los valores de temperatura óptima de crecimiento y los valores estructurales se tomaron de NCBI (www.ncbi.nlm.nih.gov/genome)

Subdivisión Dominio Archaea o Bacteria	Especies	Tamaño del Cromosoma y Plásmido (Mb)	Contenido de GC	Temperatura óptima de crecimiento (OGT) (°C)
Aquificales - (B)	*Aquifex aeolicus* VF5	1.55(c)+0.039(p)	43.5(c), 36.4(p)	93(°C)
Crenarchaeota - (A)	*Sulfolobus islandicus* L.D.8.5	2.72(c)+0.026(p)	35.3(c), 36.1(p)	85(°C)
	Sulfolobus islandicus Y.N.15.51	2.81(c)+0.04(p)	35.3(c), 36.1(p)	85(°C)
	Thermofilum pendens Hrk5	1.78(c)+0.031(p)	57.7(c), 56.5(p)	95(°C)
Euryarchaeota - (A)	*Archaeoglobus profundus* DSM 5631	1.56Mb(c)+0.0028(p)	42(c), 39.8(p)	102(°C)
	Methanococcus maripaludis C5	1.78Mb(c)+0.008(p)	33.0(c), 27.2(p)	45(°C)
	Pyrococcus abyssi	1.77Mb(c)+0.003(p)	44.7(c), 43.4(p)	102
	Thermococcus barophilus MP	2.01Mb(c)+0.054(p)	41.8(c), 38.3(p)	85(°C)

Los valores de Quintana H, V y L fueron cuantificados usando un script en perl de hechura del laboratorio del que formó parte, (codon.pl), intentando reconocer una correlación entre los valores de temperatura óptima de crecimiento y los índices de Quintana, se realizó la prueba de coeficiente de correlación de Pearson, usando el software de Sigmaplot.

Cabe mencionar que los valores I no fueron tomados en consideración ya ue en esta muestra en particular se logró establecer que los valores de V los valores de I se comportan con la misma tendencia y relación, frente a los demás índices (información mostrada en Anexo 3).

Adicionalmente, para cuantificar regiones de alta flexibilidad y secuencias repetitivas (en tándem) se utilizó el software de UGENE para Windows [Okonechnikov *et al.*, 2012]. Se calculó la incidencia de estas regiones dentro del cromosoma y dentro del plásmido usando los valores establecidos de forma basal, evaluando regiones mayores de 20 nucleótidos. Este tamaño está basado en la actividad de trasposasas y proteínas CRISP, las cuales estan presentes en genomas arqueobacterianos [van der Oost *et al.*, 2014].

Por otra parte, se tomaron como referencia los aminoácidos propuestos por Zeldovich y colaboradores (2007) como referencia y siendo los aminoácidos: isoleucina, valina, tirosina, triptófano, arginina, glutamato y leucina [Zeldovich *et al.*, 2007], relacionándose estos al estilo de vida hipertermófilo usando una correlación lineal normalizada.

3.3 Resultados

3.3.1 Contenido de GC y propiedades de la muestra

Al comparar los valores de contenido de GC en la muestra analizada ha sido posible el reconocer que las especies de Crenarchaeota presentan un valor estable tanto en los cromosomas como en los plásmidos. Esto coincide con lo que Dutta y Paul [Dutta y Paul, 2012] proponen, es decir, es posible una comparación adecuada de los valores y las proporciones del comportamiento de DNA.

Debido a que el número de genomas bacterianos con cromosoma y plásmido en organismos hipertermófilos es limitado, no es posible extrapolar los resultados obtenidos de forma global. Sin embargo, para los genomas pertenecientes a Euryarchaeota, existe un número suficiente de genomas y plásmidos para identificar un patrón de comportamiento extracromosómico.

3.3.2 Valores comparativos en los Índices de Quintana

Para los valores V, L y H, ha sido posible establecer una localización (Figura 4) en un mismo cuadrante para todos los hipertermofílicos, con excepción del genoma de *M. maripaludis* y del hipertermófilo *T. pendens*.

Los valores obtenidos de V nos permiten diferenciar entre la estructura de cromosomas y plásmidos, sobretodo dentro de los genomas de *A. aeolicus* y *A. profundus*. El valor V implica un giro variable dependiente de condiciones ambientales internas.

3.3.3 Regiones de alta flexibilidad y repeticiones en tándem

Los resultados obtenidos se muestran en la siguientes tabla:

Tabla 6: Secuencias de alta flexibilidad reconocidas en cromosomas y plásmidos. a partir del programa UGENE. Todas las secuencias reportadas se presentan localizadas a lo largo del material genético y se representan con una secuencia tipo

Especies	Regiones de alta flexibilidad	
	Número y longitud de las secuencias	Ejemplo del patrón
Methanococcus maripaludis cromosoma	103-106(5); 141(1) 120(1) and 153(1) nucls.	-5’ GTAATATTAATTTTAAT 3’-
Methanococcus maripaludis plásmido	130(1) nucls	-5’ GATATTTTTTTATATAT 3’-
Sulfolobus islandicus Y.N.15.51 cromosoma	101-110(7), 118(1) 128-133(2), 174(1) nucls.	-5’ GATATATTTGGTGGTTA 3’-
Sulfolobus islandicus L.D.8.5 cromosoma	32(1), 60(2), 101-107(3) 116(1) 125(2) 174(1) nucls.	-5’ GTA AATATATGCATATA 3’-

La incidencia de motivos de alta flexibilidad solamente se encuentra limitado dentro del cromosoma y el plasmido

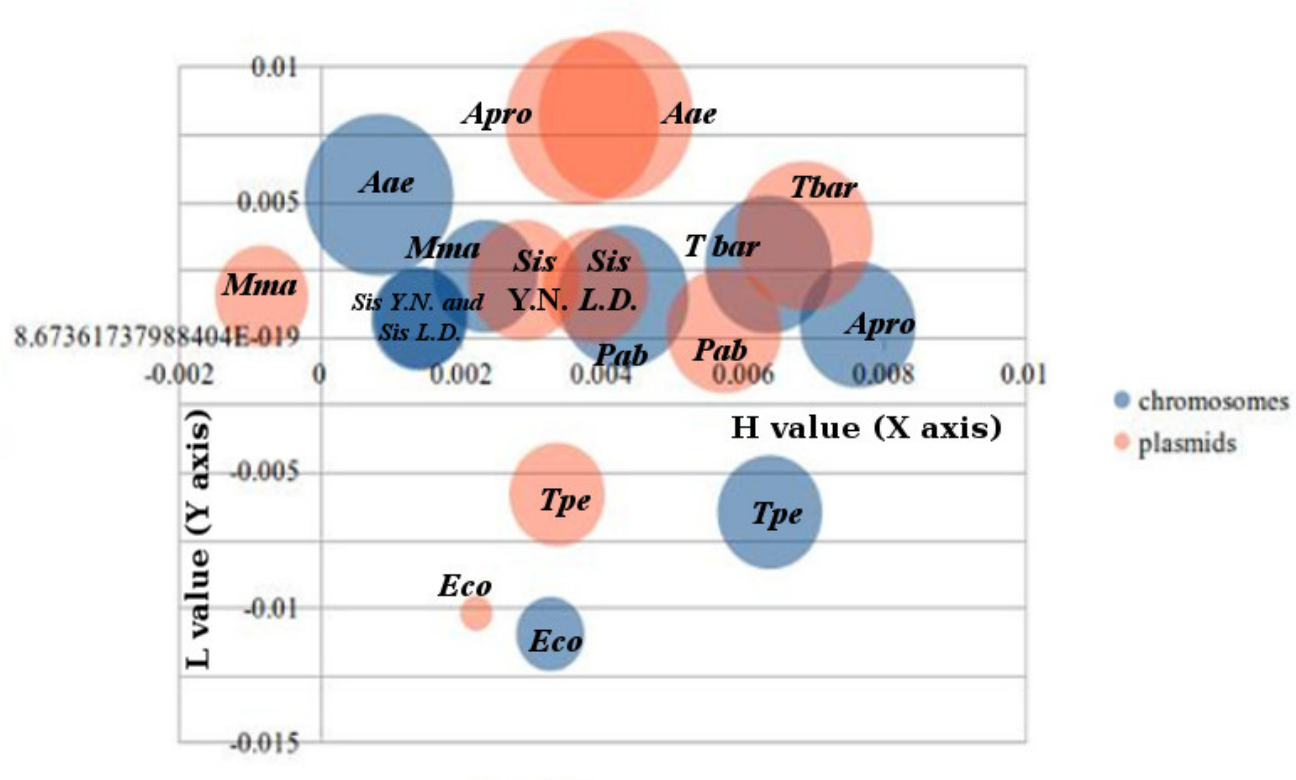


Figura 4: Grafica en donde se integran los tres Índices de Quintana obtenidos para cromosomas y plásmidos. Los valores H y L fueron usados para colocalizar los parámetros en los ejes X y Y. Los valores de V fueron representados en el diámetro de la burbuja. los nombres de las especies se abrevian de la siguient manera: Aae: *Aquifex aeolicus* VF5; Sis: *Sulfolobus islandicus* (L.D. = L.D.8.5.; Y.N.= Y.N.15.51); Tpe: *Thermofilum pendens* Hrk5; Apro: *Archaeoglobus profundus* DSM 5631; Mma: *Methanococcus maripaludis* C5; Pab: *Pyrococcus abyssi* y Tbar: *Thermococcus barophilus* MP. Los valores de los cromosomas son representados en gris, los valores plásmidos son representados en azul.

de la especie metanógena analizada y en ambos cromosomas de las dos especies de *Sulfolobus* analizadas, denotando un comportamiento diferente al esperado.

Por otra parte, la presencia de tanto secuencias repetidas en tándem , solamente se limita dentro del cromosoma y no es reconocible dentro de los plásmidos. La presencia de secuencias presentes en los cromosomas del género *Sulfolobus*, como en *M. maripaludis* y *A. aeolicus* la incidencia en el tamaño y secuencia de estos motivos cambia considerablemente en cada uno de los casos

3.3.4 Uso de codones y contenido de aminoácidos

En este proyecto se usó como referencia el trabajo de correlación lineal de Dutta y colaboradores [Dutta y Paul, 2012], en donde se estableció una correlación entre la temperatura óptima de crecimiento y los caracteres moleculares a partir de valores iguales o mayores a 0.6. Tomando ésto como referencia, se estableció que los codones usados en los cromosomas que presentan valores cercanos al 0.6 son los codones de arginina (AGG), glutamina (GAG) y triptofano (UGG), los cuales se relacionan de manera directa con la temperatura óptima de crecimiento; mientras que otro codón de arginina(CGA), e isoleucina (AUU) correlacionan de manera inversa con el aumento del valor de OGT. Las gráficas y los valores de correlación se presentan en Figura 5.

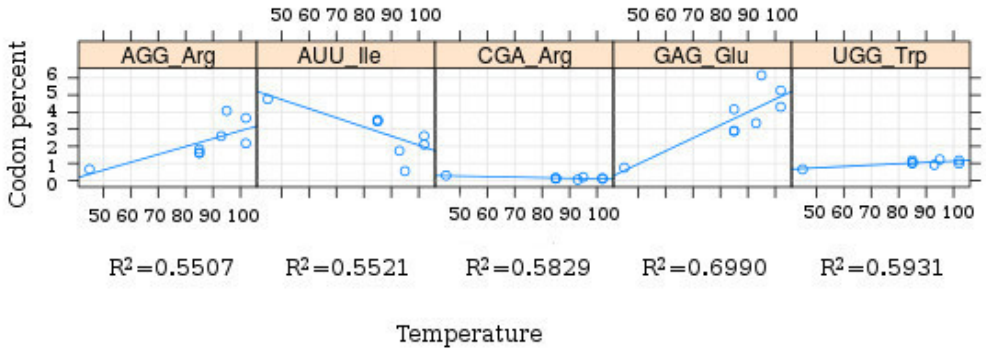


Figura 5: Incidencia de codones de valores mayores a 0.5 de correlación lineal. En estos análisis se detecta que aunque se preente un indice de corelación lineal solo en unos cuantos presentan una variación significativa entre la incidencia del codón y la temperatura óptima de crecimiento. La figura maneja siglas de aminoácidos: Arg: arginina, Glu: ácido glutámico, Ile: isoleucina, Trp: triptofano.

En los plásmidos, los codones que resultaron directamente proporcionales al aumento de temperatura óptima de crecimiento los codones de leucina (CUC) y triptofano (UGG) y una relación inversa frente a los codones de Leucina (UAA) e Isoleucina (AUA). Las gráficas y los valores de correlación se presentan en Figura 6.

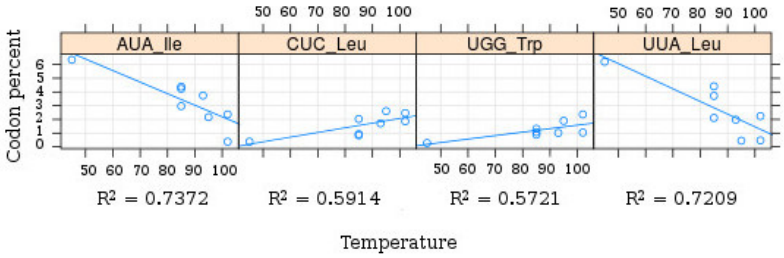


Figura 6: Incidencia de codones de valores mayores a 0.5 de correlación lineal dentro de las secuencias plasmídicas. En estos análisis se detectan índices de corelación que denotan una diferencia significativa desde lo que es termófilo a hipertermófilo. La figura maneja siglas de aminoácidos: Ile: isoleucina, Leu: leucina, Trp: triptofano.

3.4 Discusión

3.4.1 Valores H, L y V y su correlación con el estilo de vida hipertermófilo

A pesar de que la muestra no es significativamente grande, es posible correlacionar la variación dentro de genoma y plásmido o de los genomas de hipertermofílicos, estos resultdos son mostrados en la Figura 4. En esta es posible reconocer al menos elementos comunes a todos los genomas, los cuales son constantes en las estructuras bioinformacionales: por una parte se presenta la presencia de valores cercanos en V y L en un 75% de la muestra

Los valores diferenciales encontrados incluso al analizar los casos excepcionales de *Aquifex*, *M.maripaludis* y *T. pendens*, el valor en el que se presentan con mas variación es la presencia e incidencia de los dímeros GC y GA (valor H), dando con esto la relación argumentada previamente por Dutta, en donde los motivos y proporciones de las regiones de purinas y pirimidinas están correlacionados con los contenidos de GC.

El comportamiento de *T. pendens* parece seguir este patrón no solo al tener un valor diferente de GC, sino que esta proporción puede ser el resultado de eventos de perdida y recombinación de genes [Anderson *et al.*, 2009], así como eventos de transferencia horizontal implicando mayoritariamente genes bioinformacionales [Chan *et al.*, 2011].

Los procesos de recombinación presentes en *M. maripaludis* presentan una estructura diferencial tanto por ser un metanógeno el cual puede estar expuesto a eventos de conversión génica [Hildenbrand *et al.*, 2011] como por ser el único de la muestra que es un organismo termotolerante y no uno hipertermofílico.

La cercanía identificada dentro de los valores de H, V y L para los plásmidos de *Archaeoglobus profundus* y *Aquifex aeolicus*, hacen posible el sospechar, eventos de transporte horizontal o un evento de convergencia ante condiciones ambientales. Sin embargo será necesario realizar estudios de pangenómica comparada y reconocer si es posible identificar genes comunes a estos procesos.

3.4.2 Secuencias de alta flexibilidad y secuencias repetidas en tándem

El reconocer la incidencia de ambos tipos de secuencias nos ha permitido identificar que en algunas especies de hipertermofílicos, se pueden encontrar regiones puntuales resultado de la recombinación y transposición de elementos, así como posiblemente la identificación de ”hotspots” o regiones con una alta incidencia de eventos de mutación.

La incidencia de un cromosoma y un plásmido con secuencias repetidas en tándem así como de secuencias de alta flexibilidad coincide con lo reportado en la literatura, en donde los metanógenos presentan un mayor índice de reparación y conversión [Johnson *et al.*, 2013, Cai *et al.*, 2009] que los genomas hipertermófilos mismos. Además, estos resultados permiten identificar que los plásmidos y cromosomas podrían ser regulados de formas diferentes, al menos para los hipertermófilos de cada dominio.

3.4.3 Uso de codones e incidencia de aminoácidos

El análisis realizado dentro de este trabajo ha permitido reconocer que los patrones de correlación entre la temperatura óptima de crecimiento y cada uno de los codones de un aminoácido correlacionan de forma diferencial. Esto es una pauta que puede realizarse para encontar diferencias en el proceso de adaptación a ambientes hipertermófilos.

Un ejemplo de esto es la correlación negativa del codón de isoleucina, AUU, la cual es común tanto a la estructura de los plásmidos como a la de sus cromosomas. Existen codones que no importando el dominio o la especie, se presentan en esta aproximación temprana como codones no seleccionados para este estilo de vida.

Por otra parte, ha sido posible reconocer que se presentan codones específicos sin sesgo de GC que son favorecidos frente al estilo de vida común de manera diferente en el cromosoma que en el plásmido, tal es el caso del codón CUU, que aumenta en función de la temperatura de crecimiento en los cromosomas, y el codón CUC cuya incidencia aumenta en los plásmidos. Ambos codones codifican para leucina.

Del mismo modo ha sido posible identificar tendencias diferenciales que se presentan como únicas en los plásmidos, tal es el caso del Triptofano el cual previamente se reportaba como disminuido en los genomas de hipertermófilos [de Champdoré *et al.*, 2007].

Estos resultados hacen posible reconocer señales positivas y negativas diferenciales en la estructura y estilo de vida hipertermófilo, permitiendo con esto, reevaluar al menos de manera parcial la propuesta de Zeldovich,
[Zeldovich *et al.*, 2007], reconociendo con esto que es posible identificar una incidencia diferencial entre el plásmido y el cromosoma. Cabe recordar que la propuesta de Zeldovich permite establecer un aumento en los aminoácidos antes mencionados (Ile, Val, Tyr, Trp, Arg, Glu, Leu) mientras que la temperatura óptima de crecimiento aumenta, si bien esto se mantiene relativamente constante cabrá el reestablecer el uso de estos aminoácidos dentro de las secuencias de plásmidos y otros elementos móviles.

3.5 Conclusión

El estudio de la dinámica evolutiva de los oganismos hipertermófilos permite recrear desde valores diferenciales de giro y enrollamiento en la estructura de los cromosomas y plásmidos, evidenciando tendencias diferenciales en los procesos de recombinación, reparación y transporte horizontal.

Las tendencias diferenciales presentes en el uso de codones dentro de cromosomas y plásmidos, permiten dar un caracter complementario a los estudios de genómica comparada, ésto nos da información adicional acerca de como pueden reaccionar las entidades bioinformacionales (es decir cromosomas, plásmidos y genomas en general), como lo son los plásmidos y los cromosomas frente a una misma presión de selección.

Por último, el punto importante presentado dentro de este estudio comparativo es el reconocer que los orga-nismos hipertermofílicos denotan una dinámica evolutiva independiente entre dominios e incluso dentro de diferentes especies. Un ejemplo que recrea esta propuestas es el que se presenten especies que dentro del mismo género (*Sulfolobus*) puedan encontrarse organismos que presentan un arreglo y dinámica evolutiva diferencial incluso entre especies (*Sulfolobus solfataricus* frente a *Sulfolobus acidocaldarius*) [Hua-Van *et al.*, 2011].

Este trabajo sugiere que el estilo de vida hipertermofílico es convergente y que varios grupos filogenéticos han originado a uno o mas rasgos que les han dado un caracter de termófilos o hipertermofílicos, o que les ha permitido ser pionero de algún estilo de vida extremófilo, colonizando asi diferentes condiciones de vida y nichos diversos. Esta propuesta es evidente frente a los estilos de vida multi-extremofílicos que son alcalifilicos y halófilos o acidófilos e hipertermofílicos [Gribaldo y Brochier-Armanet, 2006, Ollivier *et al.*, 1994].

4 Anexo I

Manuscrito aceptado: **DNA structure and architecture in the chromosome and plasmid of hyperthermophilic organisms: a theoretical approach** , a publicarse en el Boletín de la Sociedad Geológica de México, revista indizada en Web Of Science y con factor de impacto reconocido por Thomson Reuters.

DNA structure and architecture in the chromosome and plasmid of hyperthermophilic organisms: a theoretical approach

Héctor Gilberto Vázquez-López[1], and Arturo Becerra[1]

[1]Laboratorio de Origen de la Vida

Facultad de Ciencias, UNAM, Ciudad Universitaria

Apartado postal 70-407, 04510 México D.F., México

Corresponding author: abb@ciencias.unam.mx

Resumen

Los organismos hipertermofílicos han sido reconocidos tanto como un elemento importante en el origen y la evolución temprana de la vida en la tierra, como un modelo en estudios exobiología. Analizar sus dinámicas moleculares nos pueden ayudar a entender este importante estilo de vida. Se comparó la composición del DNA en los cromosomas y plásmidos, y analizamos su estructura, su sesgo en el uso de aminoácidos y su flexibilidad. Algunos cromosomas, y en una medida inferior, sus plásmidos, presentaron características diferenciales de flexibilidad, y sesgos en su tasa de mutación, proponiendo que sólo algunos elementos moleculares muestran altos niveles de variabilidad.

Palabras claves: cromosoma, plásmidos, valores de Quintana, giro de DNA, uso de codones, hipertermófilos

Abstract

Hyperthermophilic organisms have been recognized as an important element in origin and early evolution of life on earth, but also as a model on exobiology studies. Analyze their molecular dynamics can help to understand this important lifestyle. We compared their DNA composition in chromosomes and plasmids, and analyzed their structure, amino acid bias and flexibility. Some chromosomes, and in a lower measure, a number of plasmids, shown features of flexibility, and patterns of mutation, proposing that only some molecular elements show high values of variability.

Keywords: chromosome, plasmids, Quintana values, DNA twisting, codon usage, hyperthermophiles

1 Workfield area

Our area of work focuses primarily on origin and early evolution of life using comparative genomics. We are especially interested in the evolution of the extremophile genomes.

2 Introduction

The studies focused in microorganisms have brought us an incredible contribution to the fields of biochemistry (Xu & Glansdorff, 2002; Conners et al., 2006), biotechnology (Vieille & Zeikus, 2001; Guiral et al., 2012) and early evolution, bringing the concept of a dynamic and diverse biosphere (Allers & Mevarech, 2005; Nisbet & Sleep, 2001) to enhance the limits of life. One of the group of organisms who bring us additional guidelines and parameters to research areas in exobiology includes the extremophiles.

The extremophiles are organisms that surpass the mesophilic intervals and limits of parameters such as temperature, pH, salinity, and cause us to recognize biochemical dependences of different elements and compounds like sulfur, nitrogen, organic and inorganic compounds and methane or even ferrous oxides respectively (Stetter et al., 1990). To make evident the differential respiration reactions of these organisms increase our vision about what can we find outside our planet (Trent, 2000). It is necessary to integrate a clear vision of the study of these extremophilic organisms to recognize new trends and molecular signatures in different environments. One of the extremophilic lifestyles that have an incredible effect in the number of studies, by their impact in early evolution (Islas et al., 2003) and their biochemical diversification, it is the group of hyperthermophilic prokaryotes from the Archaea and Bacteria domains (Stetter, 2006).

The high frequencies of studies involving hyperthermophilic prokaryotes are based on their multiple strategies for surveillance and their stability in high-temperature environments (Stetter, 1999), works have proven that there is no single molecule or metabolic pathway unique for this lifestyle (Atomi et al., 2004), and it has been complicated to find a pattern or common properties in all those described species (Trent, 2000; Allers & Mevarech, 2005).

However, after various researches, it is possible to associate some molecular traits at nucleotides and amino acids level, beyond their phylogenetic group. Such as Groussin & Gouy work's (2011), that recognizes two main processes in the hyperthermophilic lifestyle: *a)* the molecular evolutionary rate that is skewed by the optimal growth temperature among protein-coding and RNA-coding genes in a differential way and, *b)* that is possible to correlate the optimal growth temperature with multiple components of coding and non-coding sequences (Groussin & Gouy, 2011). Also, Klipcan and collaborators (2006) correlated the optimal growth temperature and the proportion and type of aminoacyl-tRNA synthases (aatRNAs). And finally, Agarwal & Grover (2008) recognizes in certain hyperthermophilic genomes a purine bias that modifies the amino acid frequency and codon usage.

Hence, it is impossible to recognize common genes for all the hyperthermophilic genomes, but is feasible to discovery biases and proportions that define a differential composition in this lifestyle. As it has been proposed (2014), that the hyperthermophilic genome can be studied in two main sections; *1)* the core genome that is composed by all the housekeeping genes involved in basic metabolism, and *2)* the flexible genome, which is shaped by genes implicated in habitat-specific properties, as well as interaction between viruses and predators. These genes have a variable presence

in the entire genome, and are involved in horizontal gene transfer events, gene loss and high rates of gene turnovers.

With this, it is possible to infer that the bacterial chromosomes are composed by genes from the core and flexible genome, but the extrachromosomal materials, like the plasmids are composed mainly of genes involved in the flexible genome.

Studying the plasmid genome, we can recognize the arrangement of the structure of the DNA and variance of codon usage in the same organism. Cordero and Polz (2014) and Berg (2002) proposed that to study the elements of the flexible genome, brings us the opportunity to recognize global or individual patches of genes that are the reflection of the prokaryotic community that shares the environmental conditions. It is even possible, according to the work of Cooper and collaborators (2010), that secondary bioinformational elements (like the secondary chromosomes, megaplasmids or plasmids) show a decrease in codon usage diversity.

There are different ways to measure the structure and topology of the genome, as the codon usage skews. One approach, which is familiar to us, it is to recognize dinucleotide interaction in the entire DNA sequence and correlate it with twist profile. As it is noted in Quintana and collaborators (1992) the DNA crystallography profiles bring us a correlation with the twist, the space and configuration among dimers. The **High twist profile** (H value) makes reference to the incidence of sequence of dinucleotides GC or GA and elevated values of twist and space among nucleotide dimers. In contrast, the **Low twist profile** (L value) is present where the sequence has a high frequency of CC, CT, TT, AA, AG and GG dimers. These configurations present the values of twist and separation between the dimers in low values. The **Variable twist profile** (V value) it is a combination of both configuration and is correlated with the incidence of pyrimidine and purine dimers that is a conformation strongly susceptible to the influence of the environment.

For the codon usage analysis, we have reviewed the works of Jaenicke and collaborators (1991), which implied that decreased in non-charged amino acids, that could bring us a feature in the composition of extremophiles and mesophile proteomes. And also, the work made by Zeldovich and collaborators (2007), where it is possible to recognize a direct correlation between seven amino acids (IVYWREL) and the increase in the optimal growth temperature.

Using both approaches: analysis in DNA twist profile and codon usage, we tried to identify a particular architecture of the plasmid and the chromosome from hyperthermophilic organisms to deduce some shared traits and prove that the plasmids and the genome show a differential structure, conformation and composition, which be common for the lifestyle.

3 Methods

In order to classify and recognize the optimal interval of temperature for the proposed archaeal and bacterial species, we use the National Center for Biotechnology and Information database (www.ncbi.nlm.nih.gov). These data are compiled in Table 1. A total of eight hyperthermophilic species with complete genome and plasmid sequences, were obtained from the ftp site of NCBI. Of which, three Crenarchaeota (*Sulfolobus islandicus* L.D.8.5, *Sulfolobus islandicus* Y.N.15.51, and *Thermofilum pendens* Hrk5), four are Euryarchaeota (*Archaeoglobus profundus* DSM 5631, *Methanococcus maripaludis* C5, *Pyrococcus abyssi* and *Thermococcus barophilus* MP), and one species from Bacteria (*Aquifex aeolicus* VF5). The H, V and L mean values were evaluated on genomes and plasmids, by perl script (codon.pl). To recognize a possible relation with these three values and the optimal growth temperature, a Pearson Coefficient probe was used from Sigmaplot software.

Additionally, using the UGENE software (Okonechnikov et al., 2012), we have calculated high flexibility areas and tandem repeat sequences in both chromosomes and plasmids. The module that would recognize the high flexible sequences was applied by using the default values, and the tandem repeat module was modified to recognize sequences greater than 20 nucleotides. The size of tandem repeat is based on previous reports (Van der Oost et al., 2014).

4 Study materials

In this study, we selected only those complete sequenced genomes that had a thermophilic or hyperthermophilic lifestyle reported in references (Stetter et al., 1990; Horneck & Baumstark-Khan, 2002), separated into chromosome and plasmid. Furthermore, we include the genome of *Methanococcus maripaludis* because of their thermophilic tolerance and because we thought that it would be important to have a comparative vision from a Methanobacterial genome.

5 Results

5.1 Genome sampling in the NCBI database

The information in Table 1 shows the available diversity of archaeal and bacteria hyperthermophilic species, with complete chromosome and its plasmid sequences. We recognized eight species: one from the bacterial domain, three from Crenarchaeaota subdomain, and four from the Euryarchaeota subdomain.

When we compare the GC amount in the three obtained groups, although is a short sample, we recognized that the value between the Crenarcheota is uniform and stable than other analyzed groups. The GC amount in the plasmids and chromosomes is more stable in the Crenarcheota than in other analyzed groups; and in this group, we can find that there are contained the two species with an increased genome size. With this, we can recognize that Crenarchaeota

presents a stable genomic structure, unlike the Euryarchaeota, which presents differences among the GC quantity and the correlation with the plasmid and genome size.

5.2 V, H and L comparative values

In the Figure 1, we have integrated the three main profiles of DNA twisting. According with the graphic, these results can be grouped into a same quadrant, where the H value and increases in diverse chromosomes and plasmids, and L value is positive and is similar for the 75%. The only exception of this grouping is the *Thermofilum pendens* genome and *Methanococcus maripaludis* plasmid.

Additionally, we have identified an important difference in the V values for the chromosome and plasmid of *Aquifex aeolicus* and also the plasmid of *Archaeoglobus profundus*. This does not correspond with a phylogenetic signal, or a similar optimal growth temperature or GC amount. The result was similar for *Aquifex* and *Archaeoglobus* plasmids who present overlapping and similar values of twisting.

5.3 Tandem repeat sequences and High flexibility regions

To recognize if all the plasmids present the same flexibility in their genome or if they share a common feature in this property, we analyzed the incidence of particular regions and coupled that with the search for tandem repeat (Allers & Mevarech, 2005; Norais et al., 2013). The result of this evidence is presented in Table 2.
The main result of this analysis is to recognize that the proposed flexibility of the plasmids; it was not found in them. Only in the plasmid of *M. maripaludis* is presented and shares a signal with the chromosome.

The structure of the chromosome allows high flexibility regions or tandem repeat regions. The chromosome of both *Sulfolobus* species shows both elements. This signal is not shared by any of the analyzed hyperthermophilic genomes.

5.4 Codon usage and amino acid values

For the correlating values, we can note that the major values are represented by the plasmids (Figure 2) with an $R^2 >$ 0.7 for leucine and isoleucine and for the chromosome (Figure 3) the highest value is presented by glutamic acid ($R^2 \approx$ 0.7).

6 Discussion

6.1 Genome sampling and further samples

Although the sample used is not representative, some important assumptions were drawn out from this work. Of the chromosomes and plasmids; gives us an approach about how the hyperthermophilic Bacteria and Archaea can be arranged for DNA twisting.

One case that needs further studies is the arrangement of Sulfolobales, where both chromosomes have an identical DNA twist value and a differential value for their plasmids, it still needs comparison with other strains to probe de their stability and to confirm it as a phylogenetic trend that could complement the previously reported vision of high flexibility genomes (Zillig et al., 1996; Farkas et al., 2011).

Increase the number of hyperthermophilic genomes, especially bacterial, will help to have a better perspective of the phenomenon.

6.2 H, L and V and its correlation with the hypermophilic lifestyle

Although not all values are correlated directly with the lifestyle, we can recognize similar values for the V and L value this implies for 75 % of the sample. This proportion is in both, chromosomes and plasmids, and may imply that the amount related to Low twist profile (L) arrangement and Variable twist profile (V) in the archaea hyperthermophilic genome is a main trend.

The eccentric position of the *T. pendens* genome, could be explained by the high incidence of events of gene loss (Anderson et al., 2008), and the effect of a recent split transfer genes involving bioinformational genes (Chan et al., 2011). *M. maripaludis* that shows a negative H value, we can suggest that their twist and DNA structure is associated to its thermotolerant and not hyperthermophilic lifestyle. Furthermore, it has been reported that the Methanobacteriales present a high incidence of gene conversion events (Hildenbrand et al., 2011) that could modify their base composition and gene arrangements.

The nearness position of the V and H value of the plasmids of *Archaeoglobus profundus* and *Aquifex aeolicus*, could be an evidence of the high horizontal gene transfer events between them (van Wolferen et al., 2013). However, it is necessary to develop further studies of pangenomic analysis and comparative studies of these sequences.

6.3 Tandem repeats and High flexibility sequences

Recognize the highest incidence of tandem repeats and high flexibility sequences in chromosomes, allowed us to identify, punctual regions involved in recombination and increased flexibility and possibly ”hotspots” of mutation and recombination.

The finding of only one sequence with repeats in the plasmid of *M. maripaludis*, suggest that plasmids from hyperthermophilic species need additional analyses. This result contrast with the general idea where tandem repeats and high flexibility sequences are correlated with recombination events (Johnson et al., 2013) and with the occurrence of DNA repair mechanisms (Cai et al., 2009), that shifts the structure of the plasmid DNA. Moreover, this suggests that plasmids in hyperthermophilic organisms might be regulated by different manners than in their genomes.

6.4 Codon usage and amino acid proportion

To compare the codon usage for the seven proposed amino acids that correlate with thermotolerance, have brought us a different pattern abou that was previously published. Although, it is impossible to associate the increase of certain codon amino acids to the role of thermal-stabilizers, we consider that this signal provides us additional information about the response of the genome to the thermostability and lifestyle.

The analysis proposed here, where each codon is evaluated and correlates separately, have brought us to recognize different patterns in the same amino acid. An example of this differential pattern is the decreasing isoleucine, which is negatively correlated for chromosome and plasmid, although with different codon. On the other hand, leucine correlated positive (with codon CUC) and negative (with codon UUA) in plasmid. This might suggest that the proposal of Zeldovich, needs to be revised and consequently analyze another extremophilic group.

Furthermore, the increase in glutamic acid is significant in chromosome, while in the plasmid the UGG codon (tryptophan) increase in important way. It has been proposed that the glutamic acid is relevant in hyperthermophilic organisms, because of their charge, causing a difference in sidechain entropy, helping in the protein folding into extreme conditions (Greaves & Warwicker, 2007). However, the tryptophan, that it has been referenced as an amino acid which decrease the incidence into thermostable proteins (de Champdoré et al., 2007), shows a small increase in the plasmid.

With this feature, we could infer that proteins coded in the chromosome, shows a different performance than the coded proteins in the plasmid. This bias in the plasmid could be explained by their accessory role in the metabolism of the hyperthermophilic organisms, therefore, different selective pressure. However, it only would be clarified with the recognition of particular cases such as family genes with this trend or certain shared regions of the plasmids.

7 Conclusions

The study of the dynamic DNA structure of archaeal and bacterial hyperthermophiles, allowed us to learn more about the biology and relationships of this lifestyle. We can denote that the values of DNA twisting are similar; and possibly their processes of regulation, recombination and repair are different in chromosomes and plasmids.

The changing trend of codon usage in chromosomes and plasmids, bring us a complementary feature about the skews in the coding sequence. Although, this features it might be composed by few components, they could resemble us as indicators resulting of the interaction with horizontal gene events and environmental selective pressures.

Since the extreme environmental conditions that are found in planets similar to earth, it is important to have a better knowledge about hypertermophilic organisms, not only for possible microbiological contamination in future missions, but also to other astrobiological questions.

8 Acknowledgements

This paper constitutes a partial fulfillment of the Programa de Posgrado en Ciencias Biológicas of the Universidad Nacional Autónoma de México (UNAM). H.V. acknowledges the scholarship and financial support provided by the National Council of Science and Technology (CONACyT), and UNAM.

Also, we thanks for the support and the ideas of Dr. Pedro Miramontes and Dr. Germinal Cocho for the support in the knowledge and understanding in DNA twist and flexibility concepts, and Luis Delaye for the use and support of their script and their programming abilities.

References

Agarwal, S. & Grover, A. (2008). Nucleotide composition and amino acid usage in AT-rich hyperthermophilic species. *Open Bioinform J*, *2*, 11–19.

Allers, T. & Mevarech, M. (2005). Archaeal genetics - the third way. *Nature reviews. Genetics*, *6*(1), 58–73.

Anderson, I., Rodriguez, J., Susanti, D., Porat, I., Reich, C., Ulrich, L., Elkins, J., Mavromatis, K., Lykidis, A., Kim, E., Thompson, L., Nolan, M., Land, M., Copeland, A., Lapidus, A., Lucas, S., Detter, C., Zhulin, I., Olsen, G., Whitman, W., Mukhopadhyay, B., Bristow, J., & Kyrpides, N. (2008). Genome sequence of *Thermofilum pendens* reveals an exceptional loss of biosynthetic pathways without genome reduction. *Journal of Bacteriology*, *190*(8), 2957–65.

Atomi, H., Matsumi, R., & Imanaka, T. (2004). Reverse gyrase is not a prerequisite for hyperthermophilic life. *Journal of Bacteriology*, *186*(14), 4829–33.

Berg, O. & Kurland, C. (2002). Evolution of microbial genomes: sequence acquisition and loss. *Molecular biology and evolution*, *19*(12), 2265–76.

Cai, Y., Patel, D. J., Geacintov, N. E., & Broyde, S. (2009). Differential nucleotide excision repair susceptibility of bulky DNA adducts in different sequence contexts: hierarchies of recognition signals. *Journal of molecular biology*, *385*(1), 30–44.

Chan, P. P., Cozen, A. E., & Lowe, T. M. (2011). Discovery of permuted and recently split transfer RNAs in Archaea. *Genome biology*, *12*(4), R38.

Conners, S., Mongodin, E., Johnson, M., Montero, C., Nelson, K., & Kelly, R. (2006). Microbial biochemistry, physiology, and biotechnology of hyperthermophilic *Thermotoga* species. *FEMS microbiology reviews*, *30*(6), 872–905.

Cooper, V., Vohr, S., Wrocklage, S., & Hatcher, P. (2010). Why genes evolve faster on secondary chromosomes in bacteria. *PLoS computational biology*, *6*(4), e1000732.

Cordero, O. & Polz, M. (2014). Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Publishing Group*, *12*(4), 263–273.

de Champdoré, M., Staiano, M., Rossi, M., & D'Auria, S. (2007). Proteins from extremophiles as stable tools for advanced biotechnological applications of high social interest. *Journal of the Royal Society, Interface / the Royal Society*, *4*(13), 183–91.

Farkas, J., Chung, D., DeBarry, M., Adams, M., & Westpheling, J. (2011). Defining components of the chromosomal origin of replication of the hyperthermophilic archaeon *Pyrococcus furiosus* needed for construction of a stable replicating shuttle vector. *Applied and environmental microbiology*, *77*(18), 6343–9.

Greaves, R. B. & Warwicker, J. (2007). Mechanisms for stabilisation and the maintenance of solubility in proteins from thermophiles. *BMC structural biology*, *7*, 18.

Groussin, M. & Gouy, M. (2011). Adaptation to Environmental Temperature is a Major Determinant of Molecular Evolutionary Rates in Archaea. *Molecular Biology and Evolution*, *28*(9), 1–42.

Guiral, M., Prunetti, L., Aussignargues, C., Ciaccafava, A., Infossi, P., Ilbert, M., Lojou, E., & Giudici-Orticoni, M. (2012). The hyperthermophilic bacterium Aquifex aeolicus: from respiratory pathways to extremely resistant enzymes and biotechnological applications. *Advances in microbial physiology*, *61*, 125–94.

Hildenbrand, C., Stock, T., Lange, C., Rother, M., & Soppa, J. (2011). Genome copy numbers and gene conversion in methanogenic archaea. *Journal of Bacteriology*, *193*(3), 734–43.

Horneck, G. & Baumstark-Khan, C. (Eds.). (2002). *Astrobiology*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Islas, S., Velasco, A., Becerra, A., Delaye, L., & Lazcano, A. (2003). Hyperthermophily and the origin and earliest evolution of life. *International Microbiology : The Official Journal of the Spanish Society for Microbiology*, *6*(2), 87–94.

Jaenicke, R. (1991). Protein stability and molecular adaptation to extreme conditions. *European journal of biochemistry / FEBS*, *202*(3), 715–28.

Johnson, S., Chen, Y.-J., & Phillips, R. (2013). Poly(dA:dT)-rich DNAs are highly flexible in the context of DNA looping. *PloS one*, *8*(10), e75799.

Klipcan, L., Safro, I., Temkin, B., & Safro, M. (2006). Optimal growth temperature of prokaryotes correlates with class II amino acid composition. *FEBS letters*, *580*(6), 1672–6.

Nisbet, E. & Sleep, N. (2001). The habitat and nature of early life. *Nature*, *409*(6823), 1083–91.

Norais, C., Moisan, A., Gaspin, C., & Clouet-d'Orval, B. (2013). Diversity of CRISPR systems in the euryarchaeal Pyrococcales. *RNA biology*, *10*(5), 659–70.

Okonechnikov, K., Golosova, O., & Fursov, M. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics (Oxford, England)*, *28*(8), 1166–7.

Quintana, J., Grzeskowiak, K., Yanagi, K., & Dickerson, R. (1992). Structure of a B-DNA decamer with a central T-A step: C-G-A-T-T-A-A-T-C-G. *Journal of Molecular Evolution*, *225*, 379–395.

Stetter, K. (2006). Hyperthermophiles in the history of life. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *361*, 1474.

Stetter, K., Fiala, G., Huber, G., Huber, R., & Segerer, A. (1990). Hyperthermophilic microorganisms. *FEMS Microbiology Reviews*, *75*, 117–124.

Trent, J. (2000). Extremophiles in astrobiology: per Ardua ad Astra. *Gravitational and space biology bulletin : publication of the American Society for Gravitational and Space Biology*, *13*(2), 5–11.

Van der Oost, J., Westra, E., Jackson, R., & Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nature reviews. Microbiology*, *12*(7), 479–92.

van Wolferen, M., Ajon, M., Driessen, A. J. M., & Albers, S.-V. (2013). How hyperthermophiles adapt to change their lives: DNA exchange in extreme conditions. *Extremophiles : life under extreme conditions*, *17*(4), 545–63.

Vieille, C. & Zeikus, G. (2001). Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiology and molecular biology reviews : MMBR*, *65*(1), 1–43.

Xu, Y. & Glansdorff, N. (2002). Was our ancestor a hyperthermophilic procaryote? *Comparative biochemistry and physiology. Part A, Molecular & integrative physiology*, *133*(3), 677–88.

Zeldovich, K., Berezovsky, I., & Shakhnovich, E. (2007). Protein and DNA sequence determinants of thermophilic adaptation. *PLoS computational biology*, *3*(1), e5.

Zillig, W., Prangishvilli, D., Schleper, C., Elferink, M., Holz, I., Albers, S., Janekovic, D., & Götz, D. (1996). Viruses, plasmids and other genetic elements of thermophilic and hyperthermophilic Archaea. *FEMS microbiology reviews*, *18*(2-3), 225–36.

9 Tables and figures

9.1 Tables

Table 1. Thermophilic and hyperthermophilic genomes from the domain Archaea and Bacteria.

Archaeal or Bacterial subdivision	Species	Chromosome and plasmid amount (Mb)	GC amount	Optimal Growth Temperature (OGT) (°C)
Aquificales (Bacteria)	*Aquifex aeolicus* VF5	1.55(c)+0.039(p)	43.5(c), 36.4(p)	93
Crenarchaeota	*Sulfolobus islandicus* L.D.8.5	2.72(c)+0.026(p)	35.3(c), 36.1(p)	85
	Sulfolobus islandicus Y.N.15.51	2.81(c)+0.04(p)	35.3(c), 36.1(p)	85
	Thermofilum pendens Hrk5	1.78(c)+0.031(p)	57.7(c), 56.5(p)	95
Euryarchaeota	*Archaeoglobus profundus* DSM 5631	1.56Mb(c)+0.0028(p)	42(c), 39.8(p)	102
	Methanococcus maripaludis C5	1.78Mb(c)+0.008(p)	33.0(c), 27.2(p)	45
	Pyrococcus abyssi	1.77Mb(c)+0.003(p)	44.7(c), 43.4(p)	102
	Thermococcus barophilus MP	2.01Mb(c)+0.054(p)	41.8(c), 38.3(p)	85

Table 1. *Hyperthermophilic microorganisms.*

In this table are integrated the data from the NCBI database from the chromosomes (c) and plasmids (p), characteristics and the optimal growth temperature according with (Horneck & Baumstark-Khan, 2002)

Table 2. Tandem repeat sequences identified in genomic hyperthermophiles

Species	Tandem repeat sequences	
	Sequence size (number of sequences)	Pattern
Aquifex aeolicus (chromosome)	67(2) nucls.	-5' GATTGGGAAATTTTTTT 3'-
Methanococcus maripaludis chromosome	41(7); 315(1), 381(1), and 644(1) nucls.	-5' GCTGTCCTGTTAAGCAT 3'-
Sulfolobus islandicus Y.N.15.51 chromosome	54(2) and 99-105(4) nucls.	-5' GTCAAACGCCATTGCAT 3'-
Sulfolobus islandicus L.D.8.5 chromosome	60 nucls.(2)	- 5' GAAGTTTTAGTTTCTTT 3' -

Table 2. *Tandem repeats sequence identified in the chromosomes of hyperthermophiles.*
In this table are only recognized all those sequence and regions with the default values of UGENE. All these sequences with similar sizes are recognized in contiguous distribution among the chromosome and plasmids.

Table 3. High flexibility regions identified in genomic hyperthermophiles

Species	High flexibility regions	
	Number and length of the sequence	Example
Methanococcus maripaludis chromosome	103-106(5); 141(1) 120(1) and 153(1) nucls.	-5' GTAATATTAATTTTAAT 3'-
Methanococcus maripaludis plasmid	130(1) nucls	-5' GATATTTTTTTATATAT 3'-
Sulfolobus islandicus Y.N.15.51 chromosome	101-110(7), 118(1) 128-133(2), 174(1) nucls.	-5' GATATATTTGGTGGTTA 3'-
Sulfolobus islandicus L.D.8.5 chromosome	32(1), 60(2), 101-107(3) 116(1) 125(2) 174(1) nucls.	-5' GTA AATATATGCATATA 3'-

Table 3. *High flexibility sequences identified in the chromosomes and plasmids in hyperthermophiles.*

In this table are only recognized all those sequence and regions with the default values of UGENE. All these sequences are distributed along in chromosome and plasmids not contiguously.

9.2 Figure captions

Figure 1. *Graphical representation of the profile values for the chromosomes and plasmids in hyperthermophiles.* The H and L values were used for X and Y coordinates and the V value were used as an absolute number for bubble size representation. The name of the species are pointed as follows: Aae: *Aquifex aeolicus* VF5; Sis: *Sulfolobus islandicus* (L.D. = L.D.8.5.; Y.N.= Y.N.15.51); Tpe: *Thermofilum pendens* Hrk5; Apro: *Archaeoglobus profundus* DSM 5631; Mma: *Methanococcus maripaludis* C5; Pab: *Pyrococcus abyssi* and Tbar: *Thermococcus barophilus* MP. According to the color is signaled the chromosome (gray), and the plasmid (blue) obtained values.

Figure 2. *Correlation between the codon usage in chromosomes and optimal growth temperature.* Here is represented the relation between the reported optimal growth temperature (X value) versus the percent value evaluated from the overall diversity of the genetic code. The according R^2 value of each resulting codon is shown below the corresponding quadrant.

Figure 3. *Correlation between the codon usage in plasmids and optimal growth temperature.* Here is represented the relation between the reported optimal growth temperature (X value) versus the percent value evaluated from the overall diversity of the genetic code. The according R^2 value of each resulting codon is shown below the corresponding quadrant.

9.3 Figures

Figure 1

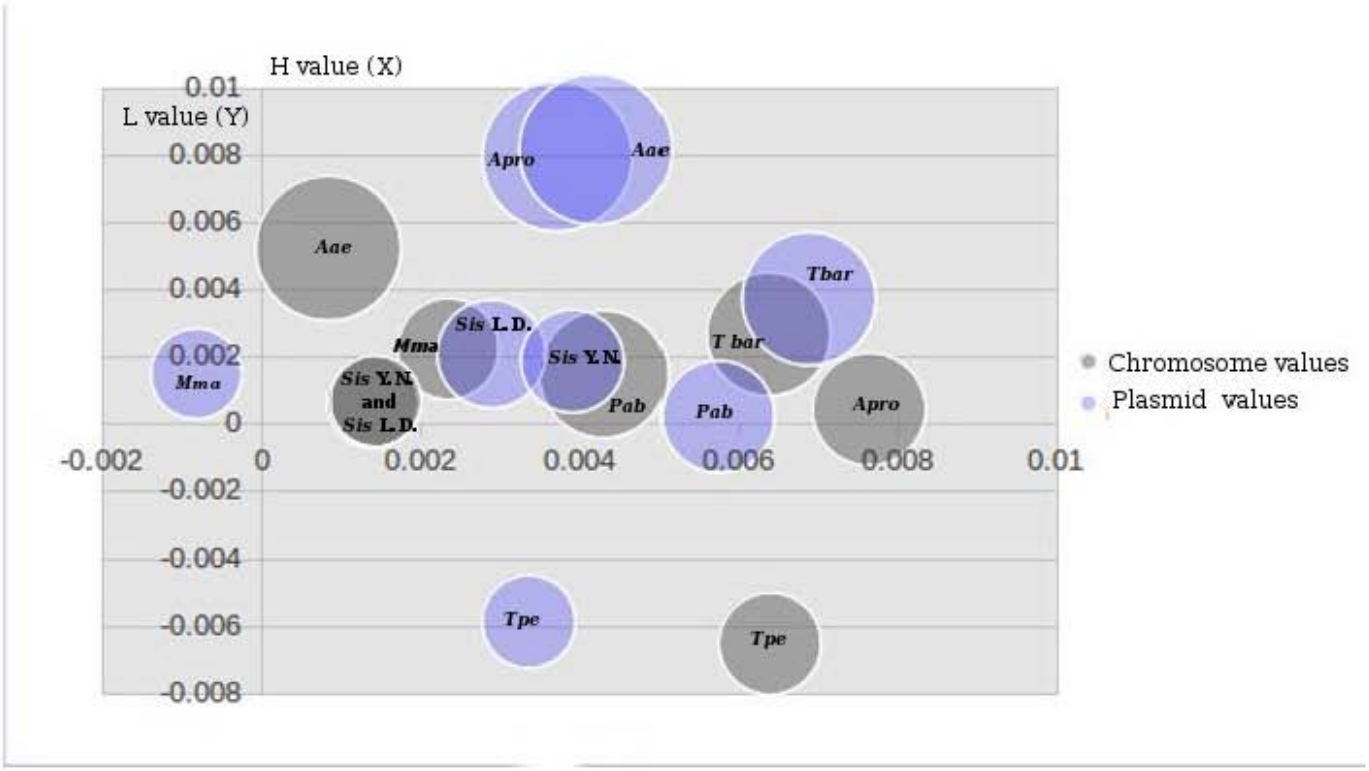


Figure 2

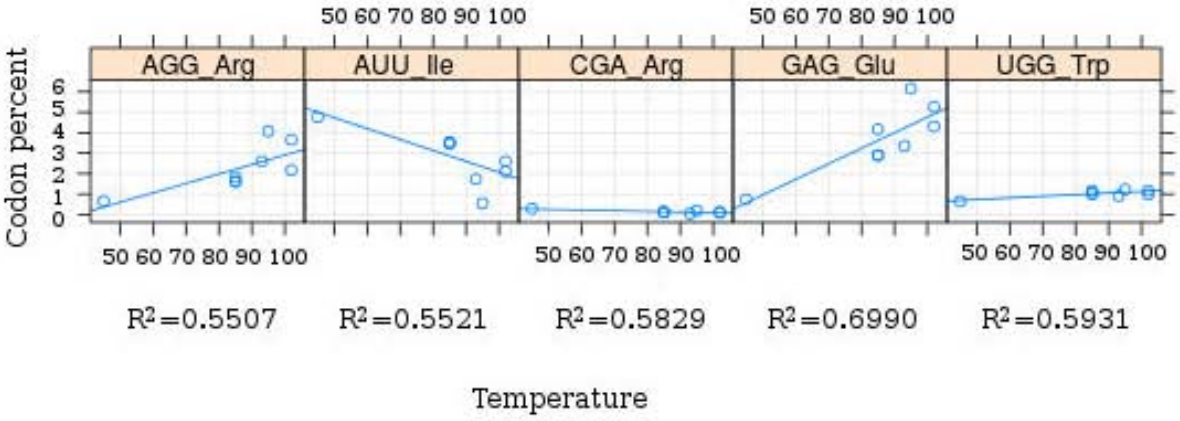
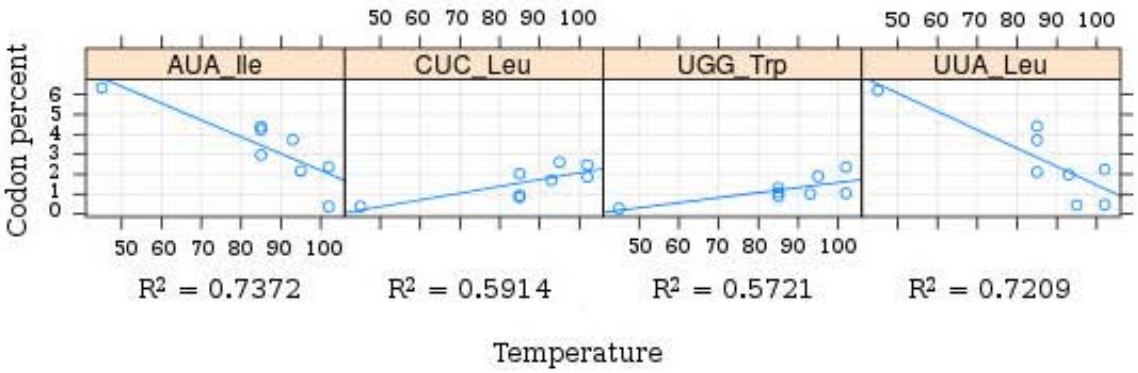

AGG_Arg
AUU_Ile
CGA_Arg
GAG_Glu
UGG_Trp
Codon percent
0
1
2
3
4
5
6
50 60 70 80 90 100
R²=0.5507
R²=0.5521
R²=0.5829
R²=0.6990
R²=0.5931
Temperature

Figure 3


AUA_Ile
CUC_Leu
UGG_Trp
UUA_Leu
Codon percent
Temperature
50 60 70 80 90 100
R² = 0.7372
R² = 0.5914
R² = 0.5721
R² = 0.7209

5 Capítulo III Resultados alternos al primer artículo

5.1 Resultados comparados dentro de las mediciones en diferentes estilos de vida

Dentro del desarrollo del doctorado fue posible el comparar los valores y resultados de los índices de Quintana y Miramontes e ntegrarlos en el tiempo de la candidatura. Dentro de la primera aproximaciòn y compilación fue posible reconocer que estos valores seguían un patrón relacionado a un estilo de vida mas que a una huella filogenética. El estudio inicial planteaba la comparación de organismos hipertermofílicos, frente a un control, de genomas pertenecientes a estilo de vida mesófilo y en otra comparativa frente a cromosomas de organismos tolerantes a alta salinidad.

Los primeros resultados de esta comparativa permitieron presentar una visión global de los arreglos que presentaban cada uno de los cromosomas de los diferentes estilos de vida, el número de organismos analizados se presenta a continuación:

Tabla 7: Tabla del número de cromosomas analizados dentro de la primera aproximación en el proyecto de doctorado. Sobre las especies analizadas se denotan en el anexo adjunto a este trabajo de tesis

Estilo de vida	Dominio Bacteria	Dominio Archaea
Hipertermófilos	11	41
Termófilos	65	8
Halófilos	11	12

Dentro del proyecto se buscaba identificar si alguno de los índices o valores establecidos por la propuesta de ambos autores (Quintana y Miramontes) se relacionaba con:

a) El tamaño de cromosoma dentro de la muestra en promedio del dominio Archaea y Bacteria se relaciona frente a algun arreglo o estructura del DNA.

b) Si el estilo de vida hipertermófilo, termófilo o halófilo presenta alguna relación frente a algún o alguno de los valores o índices establecidos en común, para alguno de los dos dominios

c) Por último se buscaba reconocer si había alguna relación frente a cierto aminoácido y estilo de vida que a nivel global no hubiese sido reportado

Cabe mencionar que para detectar si se presentaba una distribución normal para validar el valor obtenido del promedio o media, se usó el coeficiente de corelación de Pearson, el cual es una medida de la relación lineal entre dos variables aleatorias cuantitativas, se decidió utilizar esta prueba ya que esta no depende como la covarianza, de una misma escala de medidas. Esta medición permite establecer de una forma simple el establecer si dos variables se relacionan entre si.

Los resultados de este análisis estadístico se concentra dentro de la siguiente tabla:

Tabla 8: Comparativa de correlación entre los índices de Quintana y Miramontes y tamaño de genoma para organismos hipertermofilos - Dominio Archaea

Casos analizados		Índices de Heterogeneidad			Índices de Quintana			
Archaea HT	genoma	YRd-PROM	WSd-PROM	MKd-PROM	H-PROM	V-PROM	L-PROM	I-PROM
genoma	1							
YRd-PROM	-0.0403	1						
WSd-PROM	-0.0012	-0.2097	1					
MKd-PROM	0.0381	0.1433	0.4692	1				
H-PROM	-0.0602	0.6574	-0.6283	-0.3898	1			
V-PROM	0.0392	**-0.9998**	0.2102	-0.1374	-0.6587	1		
L-PROM	0.2617	0.4727	0.4872	0.6344	-0.0894	-0.4654	1	
I-PROM	0.0359	-0.8299	-0.1041	-0.3129	-0.4187	0.8292	-0.6748	1

Tabla 9: Comparativa de correlación entre los índices de Quintana y Miramontes y tamaño de genoma para organismos hipertermofilos - Dominio Bacteria

Casos analizados		Índices de Heterogeneidad			Índices de Quintana			
Bacteria HT	genoma	YRd-PROM	WSd-PROM	MKd-PROM	H-PROM	V-PROM	L-PROM	I-PROM
genoma	1							
YRd-PROM	-0.946	1						
WSd-PROM	0.2225	-0.3255	1					
MKd-PROM	-0.1907	0.0918	0.5196	1				
H-PROM	-0.5067	0.5530	-0.8910	-0.3651	1			
V-PROM	0.9470	**-0.9999**	0.3306	-0.08902	-0.5584	1		
L-PROM	-0.5356	0.6202	0.2916	0.5762	-0.0188	-0.6182	1	
I-PROM	0.5899	-0.6625	-0.2761	-0.5633	-0.0070	0.6605	**-0.9930**	1

Dentro de los primeros estudios, fue posible relacionar el valor de ambos valores V y YRd, no solo para el estilo de vida hipertermófilo sino para toda la muestra analizada, esto permitió reconocer un equivalente dentro de las dos escalas, esto hizo posible comprender que ambas mediciones evaluaban estructuras similares en los genomas.

Por otra parte la relación equivalente de índices de Quintana L e I resultante en los genomas de hipertermófilos del dominio Archaea hizo posible optimizar el artículo resultante a este estudio, con esto fue posible solamente evaluar solamente en comparaciones posteriores 3 y no 4 índices de Quintana.

Tabla 10: Comparativa de correlación entre los índices de Quintana y Miramontes y tamaño de genoma para organismos termófilos - Dominio Archaea

Casos analizados		Indices de Heterogeneidad			Indices de Quintana			
Archaea T	genoma	YRd-PROM	WSd-PROM	MKd-PROM	H-PROM	V-PROM	L-PROM	I-PROM
genoma	1							
YRd-PROM	0.6170	1						
WSd-PROM	-0.3340	-0.4912	1					
MKd-PROM	-0.3172	0.2625	0.3652	1				
H-PROM	0.2410	-0.2789	-0.4641	-0.9436	1			
V-PROM	-0.6187	**-0.9999**	0.4973	-0.2581	0.2766	1		
L-PROM	-0.0144	0.4046	0.5028	0.7255	-0.7475	-0.3954	1	
I-PROM	-0.4020	-0.4212	-0.2263	-0.1205	0.0981	0.4107	-0.7126	1

Tabla 11: Comparativa de correlación entre los índices de Quintana y Miramontes y tamaño de genoma para organismos termófilos - Dominio Bacteria

Casos analizados		Indices de Heterogeneidad			Indices de Quintana			
Bacteria T	genoma	YRd-PROM	WSd-PROM	MKd-PROM	H-PROM	V-PROM	L-PROM	I-PROM
genoma	1							
YRd-PROM	-0.8285	1						
WSd-PROM	-0.0383	0.0128	1					
MKd-PROM	-0.5854	0.5995	0.2737	1				
H-PROM	0.5444	-0.5988	-0.2124	-0.8006	1			
V-PROM	0.8295	**-0.9999**	-0.0174	-0.5990	0.6001	1		
L-PROM	-0.7796	0.8304	0.1530	0.7696	-0.5894	-0.8269	1	
I-PROM	0.7631	-0.9085	-0.1917	-0.6183	0.4997	0.9077	-0.9050	1

Para la comparativa realizada dentro de organismos termófilos para ambos dominios fue posible reconocer que únicamente se comportaba de forma similar la medición de V y YRd, del mismo modo, el que nos presentara únicamente esta señal compartida hizo posible descartar que debido a que los sistemas termófilos no denotan una señal compartida o única dentro de todos los genomas, hace suponer que los genomas de termófilos en ambos dominios, debido a tener formas de resolver y responder a cambios de temperatura menos drásticos que los hipertermófilos. Es gracias a esto que el proyecto descartó el estilo de vida termófilo frente a los índices. Sin embargo para proyectos posteriores se pensará en hacer también pangenómica o un estudio de rutas metabólicas comparadas que sean comunes o únicas a un grupo filogenético.

Tabla 12: Comparativa de correlación entre los índices de Quintana y Miramontes y tamaño de genoma para organismos halófilos- Dominio Archaea

Casos analizados		Indices de Heterogeneidad			Indices de Quintana			
Archaea Halófilos	genoma	YRd-PROM	WSd-PROM	MKd-PROM	H-PROM	V-PROM	L-PROM	I-PROM
genoma	1							
YRd-PROM	0.1233	1						
WSd-PROM	-0.4414	0.6026	1					
MKd-PROM	-0.4981	0.2905	0.6881	1				
H-PROM	0.5554	-0.2820	-0.7309	**-0.9809**	1			
V-PROM	-0.1269	**-0.9999**	-0.5994	-0.2860	0.2769	1		
L-PROM	-0.2750	0.6514	0.7926	0.9010	-0.8698	-0.6480	1	
I-PROM	0.2344	-0.7675	-0.8749	-0.7850	0.7730	0.7648	**-0.9653**	1

Discusión de Índices de Quintana y heterogeneidad frente a los primeros estudios de la muestra

La coincidencia de los valores entre los MKd y los valores H parecen tener una coherencia directa dentro de los genomas de los halófilos de ambos dominios. Esto permite comprender que la estructura de los radicales que ofrece el giro del DNA esta relacionado directamente por la incidencia de dímeros GC y GA. Esto es un arrreglo que se presenta

Tabla 13: Comparativa de correlación entre los índices de Quintana y Miramontes y tamaño de genoma para organismos halófilos- Dominio Bacteria

Casos analizados		Indices de Heterogeneidad			Indices de Quintana			
Bacteria Halófilos	genoma	YRd-PROM	WSd-PROM	MKd-PROM	H-PROM	V-PROM	L-PROM	I-PROM
genoma	1							
YRd-PROM	-0.6844	1						
WSd-PROM	-0.2351	0.7558	1					
MKd-PROM	-0.6534	0.8367	0.7386	1				
H-PROM	0.6998	-0.8930	-0.7656	**-0.9500**	1			
V-PROM	0.6863	**-0.9999**	-0.7569	-0.8387	0.8935	1		
L-PROM	-0.4980	0.9340	0.8153	0.7653	-0.8097	-0.9321	1	
I-PROM	0.2834	-0.7790	-0.8264	-0.5797	0.6422	0.7777	-0.9284	1

vinculando a ambos sistemas de índices y dando sentido al arreglo de ambos. Se cuenta con poca información para esclarecer a que se deban estos arreglos , sin embargo valdrá la pena continuar con este estudio, comparando genomas recientemente secuenciados, así como reconocer si este fenómeno se presenta dentro de los diferentes cromosomas y plásmidos que cuentan las especies arqueobacterianas.

Por otra parte, el reconocer una señal de correlación compartida en ambos estilos de vida extremófilo, ha permitido el establecer que no importa que se mida el valor V o I, cualquiera de los dos evidenciará los arreglos entre ambos arreglos de material genético. Siendo esto comprobado, hace constar adicionalmente que la incidencia de regiones ricas en AT puede ser un rasgo compartido dentro de ambos estilos de vida extremo, siendo este un estado limitado y relacionado mas al estilo de vida que incluso al tamaño de genoma, y siendo esto una señal híbrida, ya que es posible solo encontrarlo en representantes del dominio Bacteria hipertermófilas y en arqueobacterias halófilas

Asimismo, el reconocer que no se presentan sesgos comunes para todos los estilos de vida frente a un tipo de codón o aminoácido en particular, frente al tamaño de genoma o incluso frente a la temperatura, permite esclarecer que el uso de codones presenta arreglos diferenciales. Además, como se reconoció dentro del artículo publicado, se requiere de establecer valores de corte menores al 0.95 para intentar buscar una correlación directa. Esto requiere revisarse a mas detalle dentro de estudios posteriores o identificar que para cierto tipo de proteínas se busca tener codones específicos que den termotolerancia o preferentes para estos estilos de vida extremos.

A continuación se anexan los controles en donde se logran corroborar que el valor de YRd y V presente en los dominios no solo es compartido en extremófilos, sino también frente a mesófilos, esclareciendo con esto que para estudios posteriores, la obtenión de estos valores parecen sinónimos.

Dentro de este trabajo, además, se planteó como premisa que si bien hay señales compartidas en la arquitectura para organismos extremófilos del dominio Archaea podían ser compartidas, se buscaba reconocer si se presentaban señales repetidas en tándem así como regiones de alta flexibilidad, prueba de los inicios de estos estudios se presentan dentro de los análisis presentados en el artículo publicado. Sin embargo se busca que se logre establecer un sistema automatizado para así corroborar la incidencia regionalizada de los diferentes motivos para asói sopesar la importancia

Tabla 14: Comparativa de correlación entre los índices de Quintana y Miramontes y tamaño de genoma para organismos mesófilos - Control de dominio Archaea

Casos analizados		Indices de Heterogeneidad			Indices de Quintana			
Archaea Mesófilos	genoma	YRd-PROM	WSd-PROM	MKd-PROM	H-PROM	V-PROM	L-PROM	I-PROM
genoma	1							
YRd-PROM	0.5380	1						
WSd-PROM	-0.3572	0.1460	1					
MKd-PROM	-0.2661	0.2968	0.6325	1				
H-PROM	0.3209	-0.0800	-0.6274	-0.8760	1			
V-PROM	-0.5460	**-0.9970**	-0.1300	-0.2782	0.0999	1		
L-PROM	-0.0550	0.4590	0.5635	0.9032	-0.7781	-0.4403	1	
I-PROM	-0.3396	-0.8271	-0.4377	-0.6085	0.3531	0.8157	-0.8001	1

Tabla 15: Comparativa de correlación entre los índices de Quintana y Miramontes y tamaño de genoma para organismos mesófilos - Control de dominio Bacteria

Casos analizados		Indices de Heterogeneidad			Indices de Quintana			
Bacteria Mesófilos	genoma	YRd-PROM	WSd-PROM	MKd-PROM	H-PROM	V-PROM	L-PROM	I-PROM
genoma	1							
YRd-PROM	-0.6187	1						
WSd-PROM	-0.2307	0.4936	1					
MKd-PROM	-0.6349	0.7644	0.5870	1				
H-PROM	0.5702	-0.5317	-0.4825	-0.9089	1			
V-PROM	0.6196	**-0.9999**	-0.4937	-0.7662	0.5336	1		
L-PROM	-0.6525	0.9420	0.5781	0.9006	-0.7512	-0.9430	1	
I-PROM	0.5814	-0.9370	-0.6375	0.4380	0.3531	0.9373	-0.8817	1

de estas secuencias frente al estilo de vida extremófilo.

Adicional a esta vertiente se habia buscado el identificar el pangenoma del estilo de vida hipertermófilo y halófilo para ambos dominios. en estos se intentó identificar ortólogos comunes y reconocer si existían una serie de duplicados, para así justificar y esclarecer desde la poliploidía de los genomas halófilos como alguna explicación del éxito de esta estrategia.

A continuación se da una breve introducción al concepto de du8plicación para así proponer las estrategias realizadas dentro de esta fase del doctorado

Pangenómica en estilos de vida extremo, el caso de los hipertermófilos y los halófilos

Dentro de los organismos extremófilos se pueden identificar grupos filogenéticos con estructura genómica compleja y con familias de proteínas afectadas en gran medida por eventos de diversificación vertical y transporte horizontal. Uno de los casos mas emblemáticos ha sido el de los Thermotogales, los cuales denotan una compleja historia evolutiva generada por estos procesos, dando así elementos de adaptación a la termofilia propios de un proceso evolutivo vertical, como elementos como solutos compatibles los cuales son resultantes de eventos de transporte vertical [Zhaxybayeva *et al.*, 2009] y diversificación interespecífica, resultante de procesos de transporte horizontal

[Santos *et al.*, 2002].

Es en estos estudios en donde la genómica ha podido dar propuestas de como dentro de todo un grupo filogenético, se pueden reconocer subgrupos y relacionarlos entre si, proponiendo incluso un estado ancestral a todos los genomas de un solo grupo filogenético [Conners *et al.*, 2006, Medrano-Soto *et al.*, 2004, Shockley *et al.*, 2005, Zhaxybayeva *et al.*, 2009] íntegro de la estructura genómica y la secuencia de sus diferentes proteínas, aunado a un estudio comparativo exhaustivo, da nuevas pautas en las propuestas evolutivas. Un ejemplo de esto, es el proponer a la temperatura ambiental como el principal determinante en la tasa de cambio dentro de los genomas del dominio Archaea [Groussin y Gouy, 2011], esta propuesta reciente permite reconocer que todavía no contamos con las herramientas o mediciones que nos hagan evidenciar mas elementos que sean presa de procesos evolutivos dinámicos. La ventaja con la que contamos es el diseño de nuevos algoritmos y herramientas de medición enfocadas al genoma. Una visión que recientemente ha permitido reconocer nuevos elementos dentro de diferentes grupos filogenéticos es el desarrollo de un pangenoma.

El definir un pangenoma es realizar una comparación entre genomas pertenecientes a una misma especie o género, esto con el objetivo de identificar la variabilidad, encontrando en la estructura y composición genómica la diversidad intraespecífica y proponer un repertorio genético único a cada taxón o variedad [Tettelin *et al.*, 2008]. Este tipo de estudios puede enfocarse a cualquier agrupación de seres vivos en donde se busque identificar la diversidad contenida de un grupo, así como evaluar la dinámica evolutiva en sus genes o proteínas, reconociendo sus elementos que se reportan de novo, así como los rasgos mas importantes conservados dentro de cada agrupación [Lapierre y Gogarten, 2009].

Si se desea aplicar el concepto de pangenoma para el estilo de vida hipertermofílico o halófilo, la metodología debe variar y los parámetros de comparación y valores de corte son los que hacen mas complicado la aproximación y el estudio.

Dentro de las limitantes de establecer un pangenoma de hipertermofílicos a los que nos enfrentamos son:

- El establecimiento de un pangenoma se ha desarrollado solamente frente a genes ortólogos o bases de datos similares al COG para establecer una estrutura común de genes [Klenk *et al.*, 2004], sin embargo no se ha logrado calcular

- La dispersión de la señal común es difícil de calcular debido a que los diferentes procesos de transferencia horizontal duplicación y mutación se presentan variados dentro de todos las familias de genes y grupos filogenéticos [Doolittle *et al.*, 1996, O'Brien y Fraser, 2005], los cuales el impacto de este proceso dentro de familias de genes en particular dentro de genomas arqueobacterianos no se conoce a la fecha con certeza.

- Solamente se ha buscado el reconocer una tendencia en el arreglo de todos los elementos que converja en un sesgo estrutural y mutacional presente en ambos dominios procariontes y a todos los organismos extremófilos [Zeldovich *et al.*, 2007].

Dentro de los primeros intentos para realizar un pangenoma de hipertermófilos nos enfrentamos con la problemática de reconocer elementos en donde se presentaban dentro de la comparación de únicamentes genes universales, es decir, el resultado de mas de una aproximación a estos modelos no era necesariamente el reconocer elementos comunes a un estilo de vida, sino a a la presencia de genes que se encontraban tanto comunes para los hipertermófilos como incluso dentro de los mesófilos.

Es en estos resultados previos (no mostrados) donde por el apoyo de los resultados en halófilos como de la diversidad limitada a unas cuantas ramas filogenéticas, que optamos por continuar únicanmente con el estilo de vida halófilo.

6 Capítulo IV Artículo en desarrollo. Pangenoma de organismos halófilos

Dentro de esta sección se realiza la descripción de algunos elementos del artículo que se esta preparando. Este lleva por título: *"Duplication bias in archaeal genes with halophilic lifestyle"*, Este trabajo se basa en la idea de reconocer eventos de duplicación, conversión y ecualización, presentes en genomas poliploides de bacterias Halobacteriales y halófilas del dominio Archaea, esto debido a su tendencia previamente reportada [Kapatai *et al.*, 2006, **?**]. Esto se desarrolló con el fin de localizar un sesgo en la duplicación de un grupo de genes que se correlacionaran al estilo de vida halofílico.

6.1 Introducción

El fenómeno de duplicación génica es uno de los mecanismos más importantes en desarrollo de la estructura y complejidad del genoma, dando pie al desarrollo de la diversidad génica y a la aparición de elementos parálogos dentro de la familia de genes [Lynch y Conery, 2003, Demuth y Hahn, 2009]. El fenómeno de la duplicación se ha estudiado a partir de reconcoer regiones duplicadas completas [Seoighe, 2003, Lefébure *et al.*, 2012], en regiones particulares como las islas de patogenia [Larsson *et al.*, 2009, Letek *et al.*, 2010], o en casos particulares para ciertos genes [Devos, 2010].

La divergencia de las secuencias duplicadas ha sido estudiada desde muchos enfoques, como el estudio de las redes génicas o "network genes" [Bhan *et al.*, 2002], y en particular en familias de genes particulares [Collins *et al.*, 2011]; [Demuth y Hahn, 2009] y rutas metabólicas [Becerra y Lazcano, 1998, Fani *et al.*, 1995, Marri *et al.*, 2006].

Algunos autores proponen que los procesos de duplicación han traído como consecuencia la adecuación y posterior adaptación ante condiciones ambientales cambiantes [Jarrous y Gopalan, 2010, Krylov *et al.*, 2003, Schmidt *et al.*, 2003, Weckwerth, 2010]. Incluso se ha sopesado el impacto en los cambios de ploidia dentro de ciertas especies, ya sea originado en eventos de especiación [Gerstein y Otto, 2009], como el inducido a partir de ciertas condiciones de estrés [Altermann, 2012]. Este tipo de eventos se ha logrado de manera selectiva dentro de algunas especies del dominio Bacteria [Batut *et al.*, 2004, Marri *et al.*, 2006], bacterias del orden de las Metanobacteriales [Hildenbrand *et al.*, 2011], y

dentro de diferentes especies en el orden de los halobacteriales como en los géneros *Halobacterium* [Soppa *et al.*, 2008] y *Haloferax* [Breuert *et al.*, 2006], los cuales entre otras especies del mismo orden presentan mecanismos de recombinación y duplicación específicos.

Es en el orden de las Halobacteriales en donde se ha logrado reconocer dos mecanismos particulares relacionados a la duplicación y a la poliploidía [Breuert *et al.*, 2006]. Por un lado se presenta el fenómeno de la ecualización, en donde ciertas regiones o copias del genoma se ven suprimidas por recombinación homóloga este proceso origina se encuentran en una fase S del ciclo celular o en un estado poliploide [Hildenbrand *et al.*, 2011]. El segundo proceso llamado conversión génica, es un fenómeno de regulación génica el cual se define como el flujo no reciproco de una molécula de DNA a otra [Pecoraro *et al.*, 2011].

Se ha postulado la incidencia de estos procesos y mas dentro de estructuras poliploides para originar arreglos heterólogos y la presencia de eventos de duplicación y recombinación específicos relacionándose incluso en algunos operones [Pei *et al.*, 2009]. Estos procesos aunque reportados ampliamente en los genomas de mamíferos e implicados en proceso meióticos [Cole *et al.*, 2012, Popa *et al.*, 2012], son limitados los conocimientos que se tienen acerca de los eventos de duplicación y recombinación relacionados en halobacterias, bacterias metanógenas, y otras especies pertenecientes a la división de los Euryarchaeota [Pecoraro *et al.*, 2011], sin embargo, sobre el impacto de estos procesos en la estructura génica en los genómicas procariontes y arqueobacterianos se conoce poco y su impacto en el número de duplicaciones génicas no está del todo caracterizado.

Es por estos mecanismos que proponemos que dentro de los genomas metanobacteriales y principalmente en halófilos arqueobacterianos se presenten eventos de duplicación que aún no se hayan reportado y que el resultado de estos eventos, junto con eventos de ecualización y conversión, sea un conjunto de genes que tenga importancia para el desarrollo de los organismos halófilos, los cuales tienen que responder ante condiciones adversas.

Al tener acceso a los genomas completos de halobacterias y metanógenos ha sido posible el buscar dentro de su genoma, la presencia de elementos más duplicados. Al comparar estas secuencias, podemos reconocer elementos relacionados a genes específicos, definiéndolos como elementos parálogos, por su similitud en secuencia primaria y suponiendo que estos dos grandes grupos de arqueobacterias, presentarán un mayor número de estas secuencias relacionadas que sus contrapartes con estilo de vida mesófilo.

Nuestro estudio ha permitido reconocer desde patrones específicos de genes y de secuencias relacionadas, con el estilo de vida halófilo, relacionando así de forma directa frente a especies que toleren o que dependan de una concentración de salinidad alta (¿1.8 M NaCl) dentro del medio, tanto del dominio Bacteria como del Archaea. Nuestra comparación de 24 genomas de procariontes halófilos, dió como resultado 199 genes de los cuales, la presencia de 85 genes solo dentro de Archaea y 18 genes de halófilos del dominio Bacteria, permitieron reconocer una relación y patrones relacionando solamente al primero de los dominios con un patrón común.

Al analizar de forma detallada este conjunto de genes es posible reconocer que se presente elementos relacionados a transportadores ABC, histidinacinasas y proteínas relacionadas en la oxido-reducción de ciertos aminoácidos como el glutamato.

Al comparar y buscar elementos comunes a ambos dominios, hemos identificado la presencia de genes que codifican para proteínas relacionadas a los procesos de segregación cromosómica, reparación, excisión del DNA y oxidorreducción de cofactores como la ubiquinona. La presencia de estos elementos nos permite reconocerlos como parte importante del estilo de vida halófilo en ambos dominios celulares

6.2 Material y Método

6.2.1 Grupo de control y experimental

Para definir nuestro grupo de estudio de los halófilos con genomas completos, utilizamos la base de datos ofrecida por KEGG [Kanehisa y Goto, 2000] en donde fueron reconocidos 24 genomas que iban desde halotolerantes hasta halófilos extremos. El criterio de ingreso era que, en base a lo investigado en las referencias consultadas, que presentara la presencia de las condiciones óptimas de crecimiento mayores o iguales de NaCl de 1.8 M.

Consultando el mismo KEGG se escogió de forma aleatoria la presencia de al menos un organismo no halófilo por cada halófilo integrado. Cuando fue posible se buscó el contar con representantes no halófilos lo más cercanos posibles para así hacer una comparación mas clara de los rasgos halófilos específicos a cada grupo filogenético. El listado de todas las especies manejadas como elementos control se anexa dentro de una tabla suplementaria.

Además, debido a que previamente se conoce que en el dominio Eukarya, se presentan eventos de duplicación a nivel cromosómico, se decidió integrar a 3 genomas de este dominio para tener un punto de comparación a este respecto, así se han integrado el genoma de dos levaduras: *Schizosaccharomyces pombe* y *Saccharomyces cerevisiae* y el genoma de una planta, *Zea mayz*, cuyo genoma presenta un gran número de copias de su genoma al ser poliploide.

6.2.2 Los genes "core" y la identificación de sus versiones duplicadas

Para identificar elementos genómicos comunes al estilo de vida halófilo se utilizó la paquetería BLASTP [Altschul *et al.*, 1990] para comparar los genomas del grupo control y el grupo de estudio. Se definió un valor de corte (cutoff) para la identificación de secuencias homólogas de e $\lesssim 10^{-7}$ para definir a un grupo de genes comunes al grupo de estudio y relacionar a un grupo de secuencias parálogas resultantes.

Los genes que forman el grupo común al grupo de estudio, los genes "core", son aquellos elementos que se encuentran comunes a todos los genomas halófilos. Ya que nos interesaba el poder reconocer casos en donde hubiese eventos de duplicación masiva, solamente se tomo en cuenta aquellos casos en donde tuviera más de dos secuencias relacionadas. Para evitar la presencia de falsos positivos, cada uno de los genes propuestos fue consultado en la base

del KEGG, en donde se identificaba si se reconocía la presencia de un grupo K0 relacionado. Dentro de esta misma página, se consulto la presencia de motivos Pfam (http://pfam.sanger.ac.uk) previamente reportados en la secuencia.

Para identificar la presencia de estos motivos dentro de cada uno de los genes de nuestro estudio se tomaron en cuenta todos aquellos motivos con un valor de e $\lesssim 1x10^{-10}$ siendo con esto más estrictos dentro del valor planteado dentro de las familias de proteínas propuesto por la base de datos de $1x10^{-3}$ a $1x10^{-6}$ [Finn *et al.*, 2010]. Del mismo modo fue utilizado el tamaño del producto del gen como otro factor a comparar entre dos genes que mostraran motivos similares.

Intentando correlacionar la función de cada uno de los productos de los diferentes genes aislados, se obtuvo el número de reacción EC, cuando se encontraba disponible. Esto permitió homologar los productos de los genes a partir de los dos primeros dígitos del número EC. Este número fue consultado también en la página del KEGG y en la página de nomenclatura enzimática (http://www.chem.qmul.ac.uk/iubmb/enzyme/).

6.2.3 Incidencia de duplicados y conocimiento de estadísticos

Para tener un valor comparativo del número de elementos duplicados identificados en genomas de halófilos y no halófilos, se utilizó la prueba de Mann-Whitmann U para comparar si cada uno de los genes contaba con un número de secuencias relacionadas significativamente diferentes. Esta prueba estadística fue realizada utilizando la paquetería SigmaPlot (Versión 11.0, de Systat Software Inc. San Jose California USA www.sigmaplot.com)

Para reconocer posibles sesgos únicos de cada dominio, se analizaron ambos grupos de genomas por separado. Posteriormente, se compararon ambas listas obtenidas y se buscaron elementos comunes a amabas listas. Para reconocer elementos comunes se utilizó los mismos elementos antes mencionados para la comparación.

Posteriormente en cada uno de los grupos de genes reconocidos, se propuso el índice de incidencia (IR) que permite comparar que tan alejado esta el valor del número de duplicados tomando como punto de comparación el numero de duplicados reconocidos en los genomas no halófilos. Utilizamos el valor 1 como el valor base presentado para el grupo control. Este número nos permitió reconocer patrones tanto individuales como compartidos en ambos dominios y ver en comparativa en donde se encontraba el mayor número de duplicaciones.

6.2.4 Heatmaps y análisis de cúmulos

Para analizar tanto el número como la presencia de secuencias duplicadas, se realizó una representación gráfica de los mismos. Se construyó un heatmap para representar a todos los elementos identificados como positivos para la prueba de Mann Whitmann U por cada dominio, del mismo modo en esta representación gráfica se incluyó un análisis de cúmulo jerárquico el cual permitía, el cual permitía evaluar la presencia de patrones comunes dentro de los resultados obtenidos. Este análisis se realizó utilizando la función heatmap.2 del lenguaje de programación R (Matloff, 2009). El

análisis fue realizado en cada uno de los dominios por separado, y en ambos análisis se incluyeron los datos comparados de los genomas eucariontes.

Posteriormente para evidenciar los casos más significativos de los genes propuestos, se analizó la varianza en cada uno de los grupos de genes analizados. Utilizando como base la varianza de los datos se identificaron los valores que sobrepasaban el setenta y cincoava unidad percentil, así, solamente se reconocerán el 25% de los valores más variables de cada muestra. Este análisis se llevo a cabo en el mismo estilo de gráficas antes propuesto y se basó en un protocolo propuesto para resolver señales de micro arreglos.

6.3 Resultados

6.3.1 Los grupos control y estudio resultantes

Encontramos dentro de la base de datos del KEGG 12 genomas pertenecientes a 12 bacterias con estilo de vida halófilo: tres Bacteroidetes, una Delta-proteobacteria, tres Firmicutes y cinco Gamma-proteobacteria del mismo modo identificamos 12 arqueobacterias halófilas, 11 del orden de las Halobacteriales y una perteneciente a las Methanobacteriales.

La muestra se muestra en la siguiente tabla:

Dentro de la tabla es posible ver que los halófilos extremos se encuentran más ampliamente distribuidos dentro del dominio Archaea, a diferencia de los halófilos bacterianos que presentan una amplia diversidad de respuesta a la concentración de sal al presentarse más halófilos moderados y "borderline". A pesar de ello las respuestas a la temperatura son similares con la excepción de la especie de *H. lacusprofundum* el cual presenta una respuesta psicrófila a la temperatura. Sin embargo un factor importante a reconocerse dentro del grupo es la presencia de un valor de tamaño de genoma similar en todas las especies.

El grupo control obtenido dentro de este estudio está conformado por seis Bacteroidetes, siete Gamma proteobacteria, tres Delta proteobacteria y seis Firmicutes. Del dominio Archaea se seleccionaron siete Euryarchaeota, 11 Crenarchaeota, una especie Korarchaeota y una Nanoarchaeota. Como se había escrito anteriormente, se han incluido 3 genomas eucariontes en el estudio, dos Ascomycetes y una planta angiosperma. Cada uno de los genomas del grupo control se seleccionó a partir de su baja tolerancia a la salinidad.

6.3.2 Los genes "core" o genes comunes al estilo de vida halófilo

Nuestro estudio del BLASTP no nos reveló genes únicos al estilo de vida halófilo en ambos dominios, sin embargo al intentar reconocer la presencia de genes compartidos a los genomas del lote de estudio e identificando su presencia variable dentro de los genomas del lote de control, fue posible seleccionar un grupo particular de 195 genes.

Después de reconocer y eliminar aquellos que compartían dominio y tamaño o estaban implicados en reacciones enzimáticas similares, fue posible seleccionar a 174 genes y eliminar 21 genes de este grupo debido a una posible ho-

Tabla 16: En esta tabla se compilaron todas las características ambientales del lote de estudio del dominio Bacteria. La clasificación en respuesta a la salinidad, está dada por la clasificación previa de [Mesbah y Wiegel, 2008]. La información fue obtenida a partir de referencias localizadas en la página del NCBI.

Subdivisión	Especies	Tamaño del Cromosoma (Mb)	Salinidad (NaCl M)	Clasificación por temperatura	Referencia
Delta proteobacteria	*Haliangium ochraceum*	9.45	Halófilo moderado (0.03–0.87)	Mesófilo	[Ivanova *et al.*, 2010]
Gamma proteobacteria	*Nitrosococcus halophilus*	4.15	Halófilo moderado (0.7–1.77)	Mesófilo	[Campbell *et al.*, 2011]
	Halorhodospira halophila	4.15	Borderline (2.2)	Mesófilo	[Tsuihiji *et al.*, 2006]
	Thioalkalivibrio sulfidophilus HL-EbGR7	3.46	Extremófilo (4.5)	Mesófilo	[Muyzer *et al.*, 2011]
	Thioalkalivibrio sp.	2.99	Extremófilo (4.5)	Mesófilo	[Sorokin y Kuenen, 2005]
	Halomonas elongata DSM2581	4.3	Extremófilo (4.3)	Mesófilo	[Sorokin y Kuenen, 2005]
Firmicutes	*Bacillus halodurans*	4.2	Halófilo moderado (0.35-2.33)	Mesófilo	[Takami *et al.*, 2000]
	Natranaerobius thermophilus	3.19	Extremófilo (3.3-3.9)	Mesófilo	[Mesbah y Wiegel, 2008]
	Halothermothrix orenii	2.58	Halófilo moderado (1.89)	Termófilo	[Mavromatis *et al.*, 2009]
	Acetohalobium arabaticum	2.47	Halófilo moderado (3.7)	Mesófilo	[Sikorski *et al.*, 2010]
Bacteroidetes	*Salinibacter ruber* DSM13855	3.59	Extremófilo (2.5-3.9)	Mesófilo	[Antón *et al.*, 2002]
	Salinibacter ruber M8	3.83	Extremófilo (2.5-3-9)	Mesófilo	[Mongodin *et al.*, 2005]

mología. Al analizar los dos primeros dígitos del número EC en estos genes, hga sido posible caracterizar únicamente 25 funciones diferentes, de los cuales, 11 de estas reacciones fueron agrupaciones realizadas por su estructura y funciones, pero que no presentaban un número EC.

La representación gráfica de estos resultados así como el número identificado a cada función es representada en la Figura 8. Los tres grupos de genes mejor representados dentro de este conjunto son los transportadores ABC, los genes que codifican para proteínas sintasas y aquellas relacionadas a reacciones de transferasas.

Tabla 17: En esta tabla se compilaron todas las características ambientales del lote de estudio del dominio Archaea. La clasificación en respuesta a la salinidad, está dada por la clasificación previa de [Mesbah y Wiegel, 2008]. La información fue obtenida a partir de referencias localizadas en la página del NCBI.

Subdivisión	Especies	Tamaño del Cromosoma (Mb)	Salinidad (NaCl M)	Clasificación por temperatura	Referencia
Halobacteriales	*Haloarcula marismortui* ATCC43049	4.27	Extremófilo (3)	Termófilo	[Baliga *et al.*, 2004]
	Halomicrobium mukohataei	3.33	Halófilo moderado (1.8)	Mesófilo	[Cui *et al.*, 2009]
	Halobacterium salinarum R1	2.67	Extremófilo (3.9)	Termófilo	[Zeng *et al.*, 2006]
	Halobacterium salinarum NRC-1	2.57	Extremófilo (4.3)	Mesófilo	[Zeng *et al.*, 2006]
	Haloterrigena turkmenica	5.44	Extremófilo (3.5)	Mesófilo	[Saunders *et al.*, 2010]
	Halorhabdus utahensis	2.67	Halófilo moderado (1.8)	Mesófilo	[Siddaramappa *et al.*, 2012]
	Haloquadratum walsbyi	3.18	Extremófilo (3.3)	Mesófilo	[Bolhuis *et al.*, 2006]
	Natrialba magadii	4.44	Extremófilo (3.5)	Mesófilo	[Siddaramappa *et al.*, 2012]
	Halorubrum lacusprofundi	3.69	Extremófilo (3.4)	Psicrófilo	[Gibson *et al.*, 2005]
	Halalkalicoccus jeotgali	3.7	Halófilo moderado (1.8)	Mesófilo	[Roh *et al.*, 2007]
	Natronomonas pharaonis	2.75	Extremófilo (3.5)	Mesófilo	[Falb *et al.*, 2008]
Methanobacteriales	*Methanohalobium evestigatum*	2.41	Extremófilo (3.5)	Termotolerante	[Zhilina y Zavarzin, 1987]

6.3.3 Índice de incidencia de duplicación

Después de analizar los 174 genes utilizando la prueba de Mann Whitney U en cada uno de los dominios por separado, ha sido posible el limitar el conjunto de genes a solo aquellos que presentan un número significativamente diferente de secuencias duplicadas seleccionadas entre organismos halófilos y no halófilos. Así fue posible reconocer 18 genes en el dominio Bacteria cuyas secuencias duplicadas eran mayores en número frente al grupo control. Mientras que en el dominio Archaea fue posible reconocer 85 genes cuyas secuencias relacionadas eran mayores que en los no halófilos. Cada uno de estos genes es copilado en el material adicional anexo de la tesis.

El índice de incidencia de duplicación (IR) bacteriano se reconoció considerablemente mas bajo que el identificado en todos los casos de Archaea. Solamente dos de los elementos bacterianos sobrepasan el doble de secuencias relacionadas: el gen caa (VNG038G) con un índice de 1:3.67 y el gen aup (VNG0136G) con un índice de 1:3.5 y solo

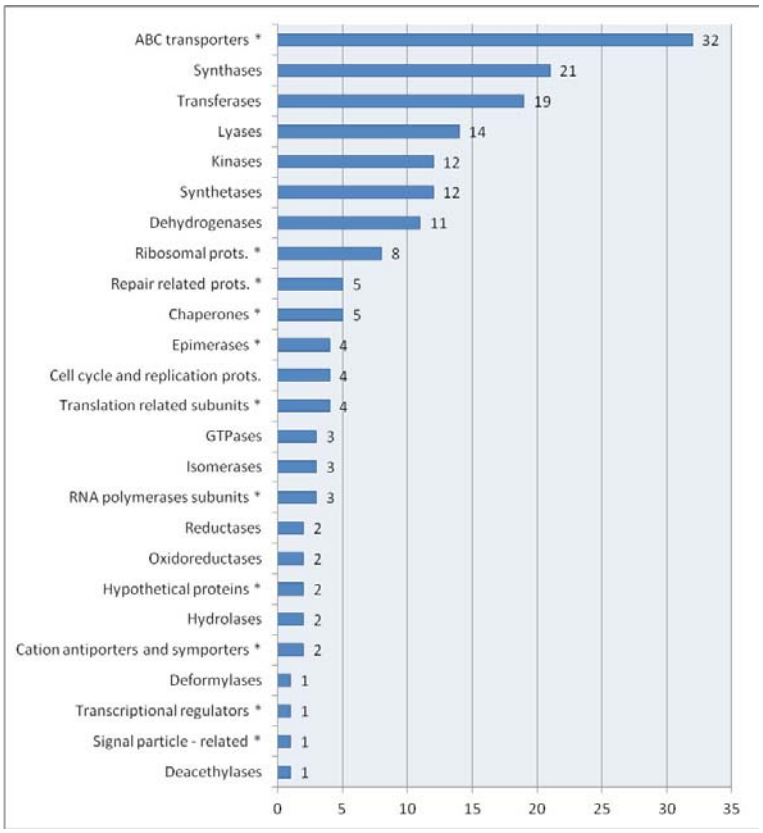


Figura 7: Representación de la clasificación de los 174 genes implicados en el estudio. Esta gráfica es el resultado del primer grupo de genes identificados en el estudio. Los grupos marcados con un asterisco son una agrupación artificial basada en la función reportada en la base de datos del KEGG. La lista de genes se anexa en el material digital

seis de los 18 genes superan el valor de IR de 1:2. Dentro de las funciones mas importantes del dominio Bacteria se presentan los transportadores de cationes, desacetilasas, deshidrogenasas, liasas y oxidorreductasas.

Para el caso del dominio Archaea, podemos reconocer la presencia de los mayores valores de IR relacionados a tres elementos, los tres más altos, relacionados a histidina cinasas, los valores que muestras estos tres genes en base a su IR corresponden a un valor que va de 1:11.74 a 1:9.26. Posterior a este valor se presentan elementos con un valor de IR cercano al 1:6.0, en este grupo se pueden reconocer una deshidrogenasa, un gen que codifica para una chaperona, una proteína histidina cinasa y dos genes que codifican para proteínas implicadas en la reparación de DNA. Es en este caso donde se logra reconocer la presencia de valores en su mayoría que superan el doble de secuencias relacionadas o duplicados entre los halófilos y no halófilos.

6.3.4 Heatmaps y análisis de cúmulo

Utilizando el número de secuencias relacionadas para los 85 genes resultantes del dominio Archaea y 18 genes para el dominio Bacteria (estos se anexan en el material digital adicional) se realizó una gráfica en donde se combinaba el número de copias o secuencias relacionadas para cada gen en cada uno de los organismos del grupo control y de

estudio. El resultado de este proceso son las Figuras 9 y 10 que se presentan a continuación.

Al analizar ambas gráficas ha sido posible solamente el identificar dos sesgos específicos únicamente para las arqueobacterias halófilas. Por una parte , dentro del primer grupo se presenta la presencia de la mayoría de los halófilos junto con especies de metanobacterias y dos especies externas a ellos: *T. pendens* y *T. kodakaraensis*. Dentro de un segundo grupo es posible ver una huella relacionando a un mayor número de secuencias parálogas, relacionando así a solamente tres especies: *N.magadii*, *H. marismortui* y *H. turkmenica*. Las únicas dos especies que no se relacionan junto con estos dos grupos son las pertenecientes al género *Halobacterium*.

Para el caso de Bacteria, es posible reconocer la diversidad de los genomas halófilos relacionada en un solo grupo subdividido en dos subgrupos, esta agrupación esta definida primordialmente gracias a tres genes principales: un gen de transferasa (VNG0060G), la proteína relacionada a transportadores ABC (VNG0645) y un gen que codifica para una oxidorreductasa ($Hhal_0811$). El primer subgrupo compuesto por 7 especies halófilas y adicionalmente a tres especies que no están relacionadas con la tolerancia a ambientes salinos. Estas especies forman parte de los mismos órdenes que los halófilos incluidos en este estudio, siendo solamente especies relacionadas a las divisiones de Delta y Gamma-proteobacteria.

El segundo subgrupo de halófilos está representado por especies pertenecientes a las divisiones de Firmicutes, Bacteroidetes y una especie perteneciente a las Gamma proteobacteria, y es en este grupo en donde solo se incluyen especies del grupo control y a las Gamma proteobacterias. La presencia de estos arreglos permiten incluir el comportamiento de secuencias duplicadas únicamente del gen de la transferasa y el transportador ABC, lo que limita la presencia de eventos de duplicación presentes en halófilos.

El análisis de los sistemas bacterianos permite incluso diferenciar rasgos compartidos dentro de los genomas eucariontes establecidos como elementos control, siendo estos 5 genes analizados en este conjunto de los genes y de los cuales presentan un sesgo completamente independiente de las bacterias halófilas y las especies cercanas filogenéticamente a los mismos.

Al reconocer en los heatmaps a los elementos más variantes dentro del estudio es posible reconocer para el dominio Archaea, que los genes relacionados a los transportadores ABC presentan un patrón independiente de duplicación al de tres cinasas, una deshidrogenasa y una oxidasa. Esto correlaciona únicamente con las tres especies más variables previamente descritas.

Para el caso de los elementos variantes del dominio Bacteria, se permite reconocer que el transportador ABC y la transferasa antes mencionadas dentro del cluster de halófilos, sin embargo el numero de secuencias duplicadas para el gen de la oxidorreductasa se relaciona de manera mas directamente con las especies pertenecientes al grupo de las Gamma proteobacterias halófilas y no halófilas y a una especie del grupo de las Delta proteobacteria, *Geobacter*sp. M18.

Dentro de ambos análisis es posible reconocer eventos de duplicación particulares dentro de los genomas eucari-

ontes incluidos la presencia de una chaperona, una epimerasa y una liasa, evidenciada dentro del heatmap arqueobacteriano, así como un partron compartido por parte de una proteína hipotética, una desacetilasa y una transferasa dentro del heatmap bacteriano.

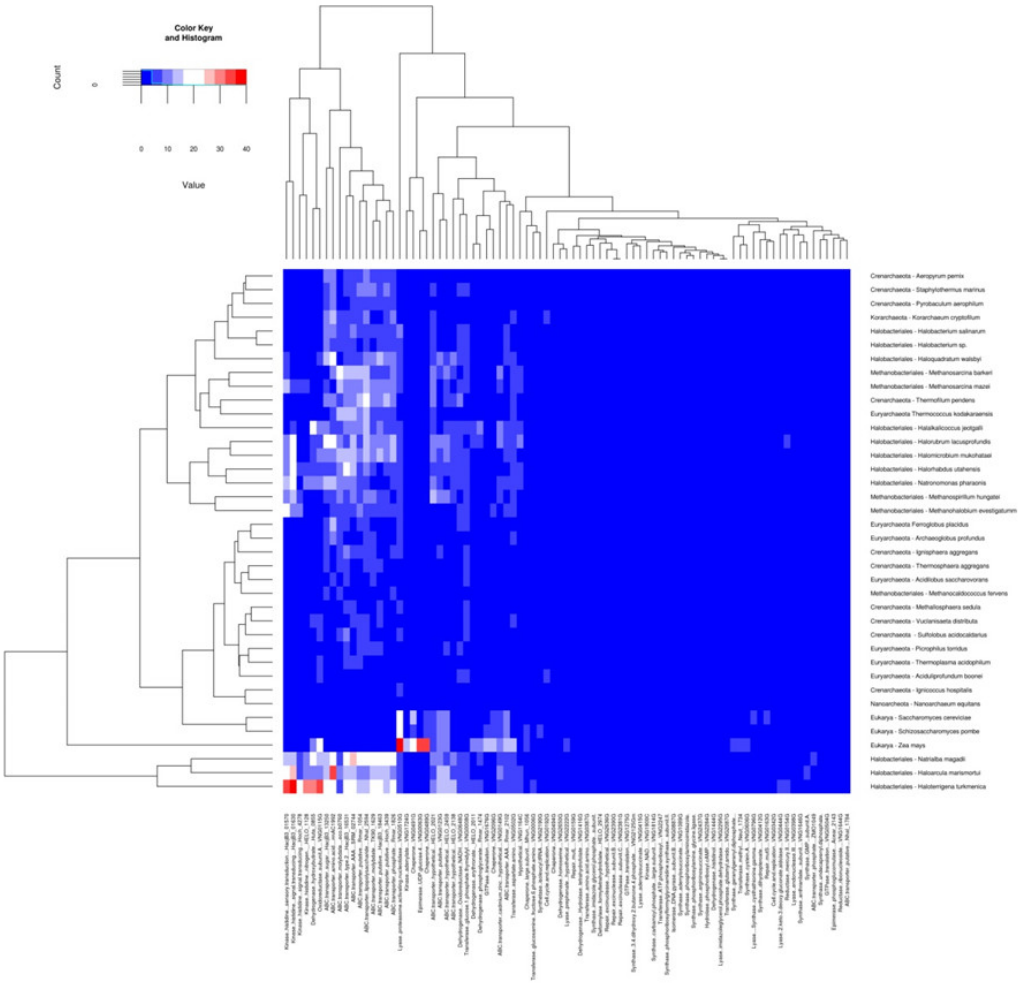


Figura 8: Análisis de cúmulo y gráfica de calor (heatmap) correlacionando las especies con las secuencias duplicadas identificadas en el dominio Archaea. La clave de color es señalada como una guía de menor a mayor de azul a rojo y con valores intermedios con tendencia al blanco. Cada uno de estos análisis es el resultado de integrar un análisis jerárquico de los valores de secuencias duplicadas frente a cada uno de los genomas

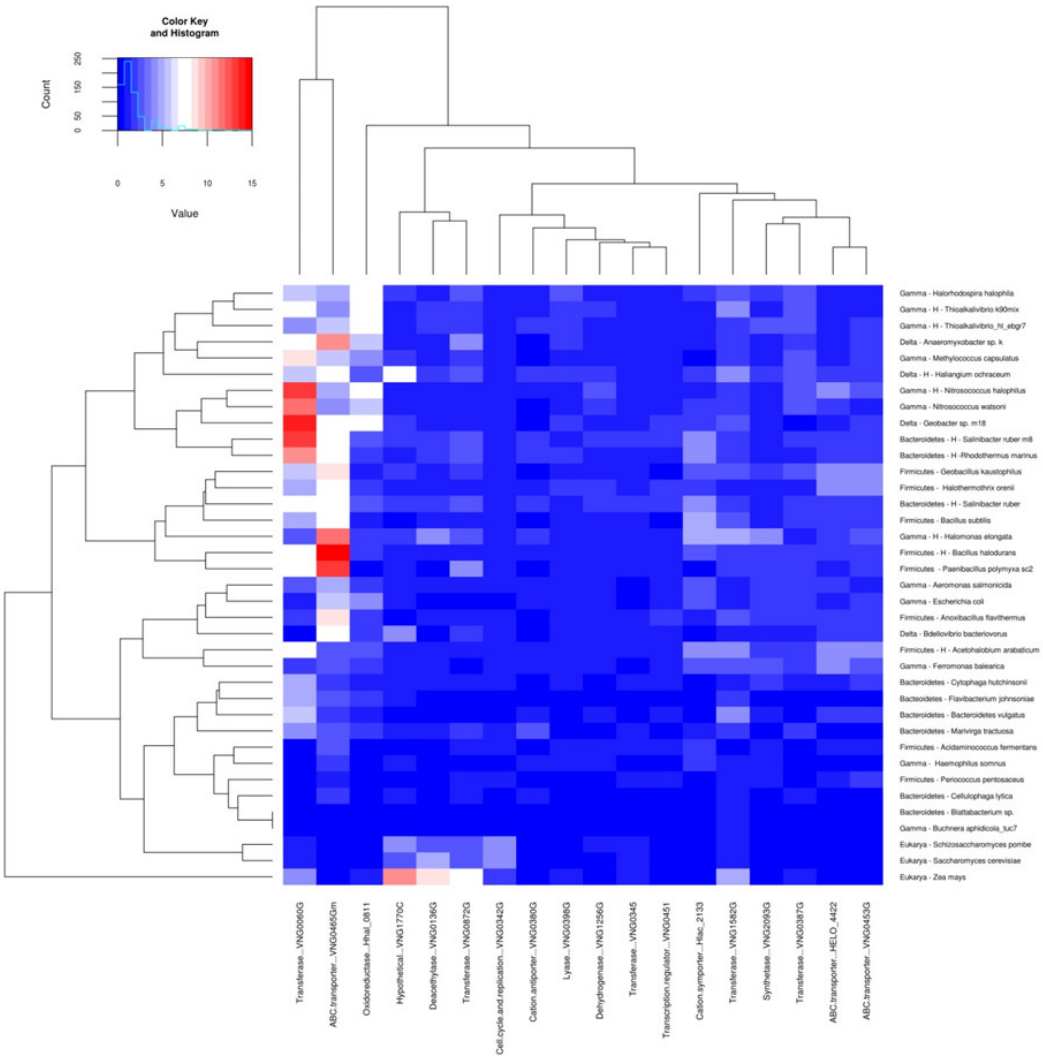


Figura 9: Análisis de cúmulo y gráfica de calor (heatmap) correlacionando las especies con las secuencias duplicadas identificadas en el dominio Bacteria. La clave de color es señalada como una guía de menor a mayor de azul a rojo y con valores intermedios con tendencia al blanco. Cada uno de estos análisis es el resultado de integrar un análisis jerárquico de los valores de secuencias duplicadas frente a cada uno de los genomas

Cuando se comparan los genes resultantes de ambos dominios, es posible reconocer y correlacionar usando como huella la presencia de los motivos de Pfam y la reacción que cataliza el producto de estos genes, la presencia de solo 4 genes, estos genes codifican para una proteína endonucleasa tipo III (VNG0398G), una aminotransferasa (VNG0387G), una proteína relacionada en la segregación de cromosomas (VNG0342G) y una oxidorreductasa / deshidrogenasa (Hhal-0811). Los valores IR en estos genes comunes a ambos dominios no sobrepasan en promedio el valor IR de 1:2.11, y con una prueba de normalidad positiva (P=0.094).

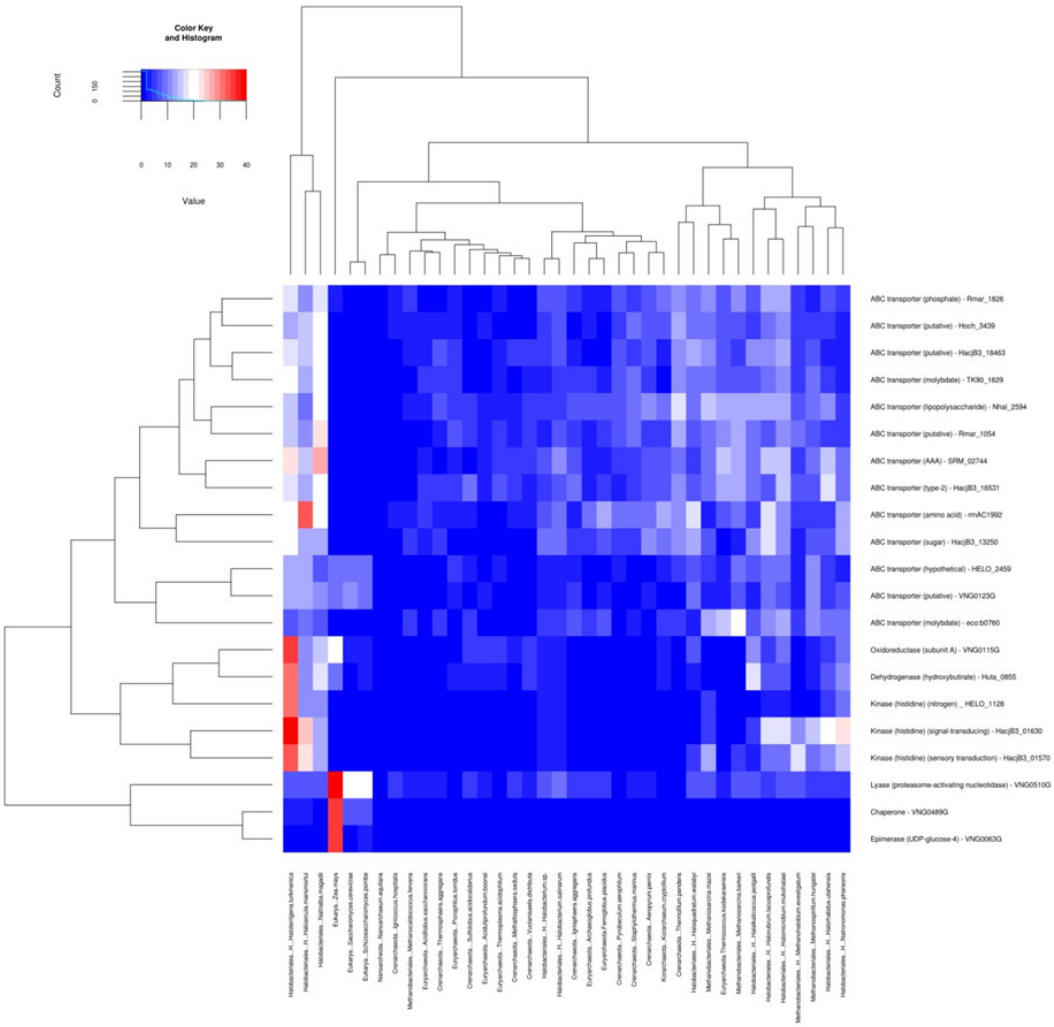


Figura 10: Elementos génicos mas variables dentro de los genes resultantes del “core gene” arqueobacteriano. Comparando los valores con la más alta varianza exclusivamente para todas las especies se puede comprobar la incidencia y la estabilidad de cada uno de las agrupaciones previamente propuestas. Nuevamente el análisis se realizó calculando la varianza de cada uno de los genes por separado y calculando nuevamente la relación entre ellos con el análisis jerárquico.

6.4 Discusión y Conclusión

La estructura de los elementos duplicados es un tópico coplicado de abordar, mas en elcaso de la pangenómica comparada. En este intento de etablecer una constante de duplicidad ha sido una verteinte inicial ya que las primeras críticas que se pueden resentar son los valores de corte de los homólogos a los cuales nos fundamentamos dentro de la familia de genes y estudios previos en donde se establece frente a la familia de genes [Collins *et al.*, 2011]. Sin embargo es un hecho que cierts familias sesgan de una forma diferencial la presencia de un alto índice de genes relacionados. Esta familia es la de los transportadores ABC.

La presencia de elementos en el dominio Archaea como resultado principal hace pensar que la diversificación de

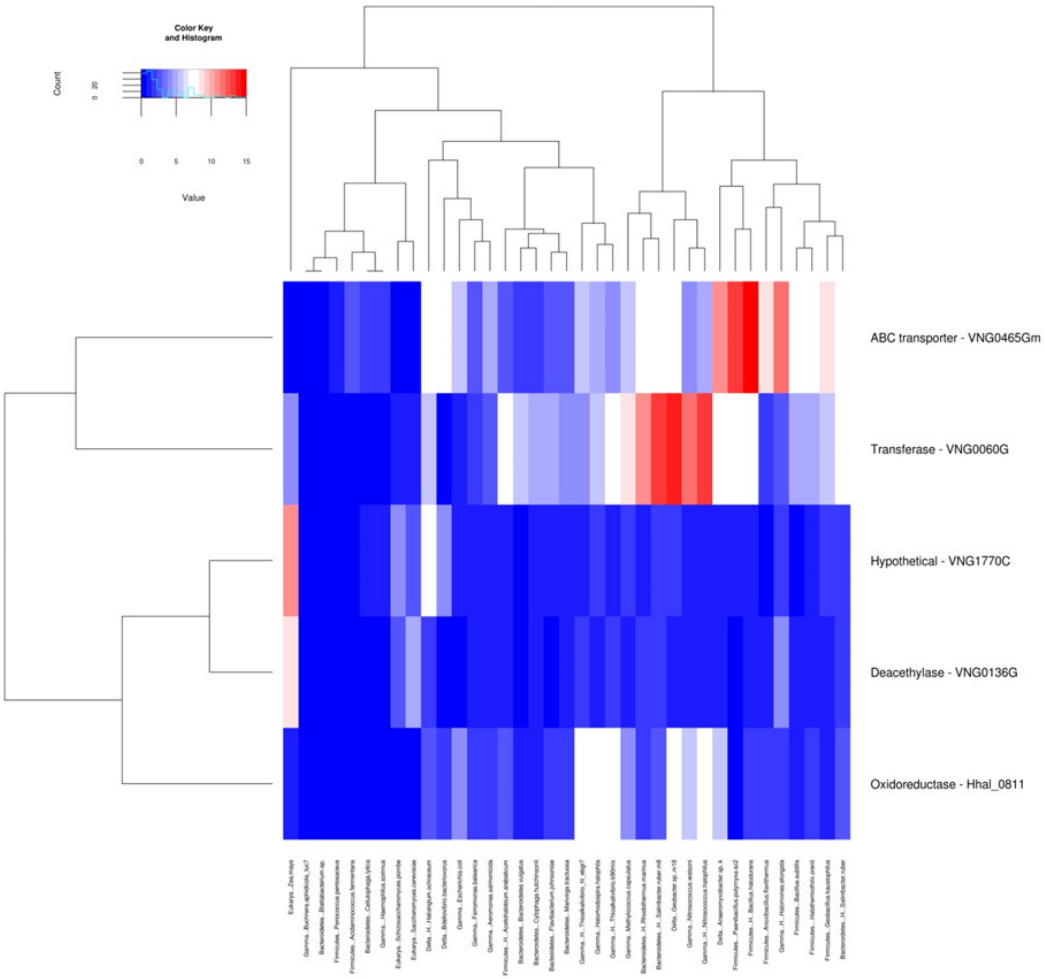


Figura 11: Elementos génicos mas variables dentro de los genes resultantes del “core gene” bacteriano. Comparando los valores con la más alta varianza exclusivamente para todas las especies se puede comprobar la incidencia y la estabilidad de cada uno de las agrupaciones previamente propuestas. Nuevamente el análisis se realizó calculando la varianza de cada uno de los genes por separado y calculando nuevamente la relación entre ellos con el análisis jerárquico.

esta familia es una constante mayor incluso en este dominio, dando con ello mas preguntas de las que puede resolver, dado que la importancia de esta familia puede relacionarse en flujos de iones, los cuales pueden er utilizados para poder regular de manera dinámica el interior celular [Oren, 2008], sin embargo, es un sistema que es también usado en sistemas de metabolismo central y adaptación a estilos de vida extremófilos [Ahmed *et al.*, 2005], es por esto ue valoe la pena reevaluar este estudio a una muestra mas grande y de manera alterna relacionar esta familia de genes (ABC) y reconocer su estado de arte dentro del dominio arqueobacteriano mismo, un tratamiento posterior se debe de hacer para el caso de las kinasas.

El reconocer una incidencia marcada dentro de las arqueobacterias ppodría tratarse de una forma simplista de

búsueda de homólogos, pero para este caso particular es posible intentar encontrar tanto homólogos como la incidencia de estudios a nivel de network, para así reconocer su papel metabólico de uno o varios genes e identificar el impacto de uno o varios procesos y rutas metabólicas que están regulando mas de una respuesta de genes, se desea continuar el trabajo de una forma similar al desarrollado en *Haloarcula marismortui* [Baliga *et al.*, 2004], en donde se reconocen redes metabólicas e identidades de redes conocidas y así extrapolando este resultado en mas de un genoma.

Para el caso de los resultados identificados en los genomas de halófilos bacterianos se presenta solamente un representantes de transportadores ABC una transferasa y una oxidorreductasa, esto permite discutir que al menos por un gen identificado, se hace constar la diferencia de las estrategias de tolerancia general del dominio, ya que la incidencia de transportadores de azúcares y aminoácidos [Elevi Bardavid y Oren, 2012, Empadinhas y Costa, 2008] darían una señal común al proceso de síntesis de solutos compatibles, sin embargo, se denota que es mas complicado el poder establecer un patrón común en estrategias por la polifilia del grupo y elementos diversos de respuesta a variación osmótica. Será necesario el hacer una aproximación similar al estudio e impacto de uno o mas genes dentro de las redes metabólicas y ver si es posible establecer un consenso en todos los genomas.

7 Anexo II

Manuscrito en preparación: **Differential gene duplication in archaeal halophilic genomes**, este trabajo será sometido a una revista indizada del área. Adicional a esto se anexan las listas de genes de los pangenomas de hipertermófilos y halófilos

Differential gene duplication in archaeal halophilic genomes

Héctor Vázquez[1,2], Fabián Reyes-Prieto[1], Arturo Becerra[1]

[1]Facultad de Ciencias, UNAM, Apdo. Postal 70-407, Cd. Universitaria, 04510 México D.F., México

[2]Posgrado en Ciencias Biológicas, UNAM, Av. Universidad 3000, Coyoacán, 04510 México D.F., México

H.V., hectorgilbertovazquez@gmail.com

F.R-P., fabian.reyes@ciencias.unam.mx

A.B., abb@ciencias.unam.mx

Correspondence should be addressed to Arturo Becerra, abb@ciencias.unam.mx

Key words: archaea, halophilic, gene duplication

Abstract

The gene duplication event is a critical process in the genomic evolution of prokaryotic organisms. In certain cases, high-scale gene duplications have an important impact on the genome structure and metabolic repertoire of those organisms. One unexplored aspect of gene duplication is the identification of the selection process leading to gene conversion in halophilic organisms. Here we used a comparative genomic approach to analyze 24 genomes of halophilic prokaryotes to evaluate the impact of gene duplication over groups of core genes possibly involved in traits related to halophilic strategies for lifestyle. We identified 85 and 18 genes in Archaea and Bacteria respectively which are involved in ion transport, environmental sensing and DNA repair, and which are affected by gene duplication events. Our results suggest that some genes closely associated with the halophilic lifestyle have evolved via duplication events.

1. Introduction

The process of gene duplication is a key mechanism that affects genomic complexity and intrinsic genetic diversity [1], and is ultimately a major source of evolutionary innovation. Duplications may involve entire genomic copies [2-3], clusters [4], gene islands or just single genes [5-6]. The sequence divergence of duplicated genes has been thoroughly studied in gene networks [7], the evolution of gene families [8–10] and in metabolic pathways [11–13].

The fact that gene duplication has had an important role in species-specific adaptations with regard to particular lifestyles is well documented [14]. This phenomenon occurs in biological groups that can tolerate changes in their ploidy as the result of adaptation to environmental conditions [15-16]. Such cases include different Bacteria species [17–19], some Methanobacteriales species [2] and the Halobacteriales species, like *Halobacterium* sp. NRC-1 and *Haloferax volcanii* [20-21].

Haloarchaea share two main genomic mechanisms related to gene duplication: a) change of ploidy [20] associated with external stimuli, and b) events of equalization [21-22]. Although diverse genomic studies of halophilic Archaea have been carried out [23-24], the role and impact of gene duplication in halophilic prokaryotes is still not well understood.

Using the information provided by complete genome sequences, we can identify; 1) genetic elements constantly present in our study group, and 2) the number of copies in each case. The aim of this article is to identify genetic composition patterns and shared traits present in the genomes of halophilic prokaryotes.

Our comparison of the 24 complete genomes of halophilic prokaryotes revealed the existence of 103 genes which are always present, each one with a significant number of duplicates. After analyzing a particular pattern for both cellular domains, it is possible to recognize 85 genes in Archaea and 18 genes in Bacteria. These genes are associated with ion and metabolite transport, oxido-reduction reactions, synthetases, transferases and DNA repair enzymes. We propose that this common duplication pattern is related to the halophilic lifestyle.

2. Material and Methods

2.1. Study and control groups

In order to define our study group, we used the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [25-26] (http://www.genome.jp/kegg) as a reference. We selected 24 genome sequences of prokaryotes that are extreme (>3.0M), borderline extreme (1.4 – 4.0 M), and moderately halophilic (0.5 – 2.0 M) organisms (Table 1).

The control group included complete genomes of non-halophilic taxa closely related to the selected halophiles. We have chosen at least one genome member from the same taxonomic group of each genome included in the study group (Supplementary Material, Table 1).

We also added the complete nuclear genomes from the unicellular Ascomycetes *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, and the polyploidic plant *Zea mayz* to this analysis. We used them to

compare their high gene duplication incidence [27-28], and contrasted them against the archaeal and bacterial cases.

2.2. The core genes and their duplicated versions

In order to identify common genomic elements related with halophilic lifestyle, we used BLASTP [29] searches to compare *study* and *control* groups. We defined a BLASTP cutoff value for homologous identification of e $\leq 10^{-7}$ to define a group of genes and recognize the number of paralogous related to them. This value was based on similar approaches used in the study of protein families and in genomic comparisons [30–33].

The core genes were defined using the common elements shared by the genomes from the study group. Because our goal was to recognize duplication events in core genes, we focused on sequences with more than two copies. To avoid redundancy, each proposed gene was verified using the KEGG Orthology (KO) database and their motifs comprised in the Pfam database (http://pfam.sanger.ac.uk). We only considered motifs with e-value of 1×10^{-10} or lower [34]. We based identification of the obtained motifs and the approximate size as the gene product. We used the EC number to highlight the catalytic reaction of each enzyme in order to associate the function of the gene products with duplication events. We used the reported function of the gene products in the KEGG database for the enzymes with no EC number.

2.3. Duplication incidence, statistical recognition

In order to have a comparison value for the duplicated elements in halophilic and non-halophilic genomes, we used the Mann–Whitney U test. This analysis was performed using SigmaPlot software (version 11.0, from Systat Software, Inc., San Jose California USA, www.sigmaplot.com). Looking for particular trends, we examined the two prokaryotic domains separately. This enabled us to identify two different sets of genes and recognize their shared elements.

An incidence ratio (IR) was estimated by using the mean value of the paralogous sequences in the control group. We assumed an IR baseline value of one (IR=1:1) when genes are in a single-copy status and compared it with the mean value of duplicated sequences recognized in the genomes of the study group.

In order to identify the most variable genes in the average sample, we designed a test to recognize those genes whose values exceed the cut off by more than 75%. The values of those genes were provided in Table 2.

2.4. Heatmaps and cluster analysis

A heatmap representation of duplicates and a clustering of the results with a statistical value were constructed. The representation was plotted using the heatmap.2 package of the R programming language [35]. We used the number of duplicated sequences and correlated them with a cluster hierarchical analysis to evaluate the presence of shared patterns and create groups with common behavior. The analysis was performed separately in Archaea and Bacteria domains. The data from the eukaryotic genomes was included in both cases.

3. Results and Discussion

3.1. Study and control groups

We retrieved twelve genomes from halophilic Bacteria (three Bacteroidetes, one Delta-proteobacteria, three Firmicutes, and five Gamma-proteobacteria) as well as twelve genomes from halophilic Archaea (eleven Halobacteriales, and one Methanobacterial). Genomic general data and optimal environmental growth conditions (pH, temperature) of the selected taxa are presented in Table 1.

Extreme halophilic lifestyle is broadly distributed in Archaea. In Bacteria, we can identify the most diverse response to salinity, from moderate to extreme. However, genome size and even poliextremophilic response are presented in a similar distribution for both prokaryote domains.

The control group is comprised of complete genomes from six Bacteroidetes, seven Gamma-proteobacteria, three Delta-proteobacteria, six Firmicutes, three Methanobacteriales, seven Euryarchaeota, eleven Crenarchaeota, one Korarchaeota, one Nanoarchaeota, two Ascomycetes and one nuclear plant genome.

3.2. The core genes

Our BLASTP search did not reveal unique genes exclusively present in the genomes of the selected halophilic prokaryotes. In contrast, we identified a set of shared genes composed of 195 genes. This group of genes is made up of all the elements always present in the halophilic genomes (File 1. Supplementary Material).

To avoid redundancy in the core genes, we compared the size and the motifs from Pfam and KO groups, and we could identify 21 identical genes from the sample, after suppressing these repeated

elements, we obtained only 174 genes (the core genes). Using the EC number and reaction reported in KEGG, we analyzed the reaction for the whole group of core genes. It was then possible to classify this set of 25 gene groups according to their enzymatic mechanisms. Each one of these genes is reported in the Supplementary Material (File 2) and their graphical representations are shown in Figure 1. The three functions most frequently identified in these core genes are ABC transporters, synthases, and transferases with a representation of greater than 15 instances per group.

We could recognize particular patterns from the onset of our investigation of the core genes. These patterns are similar to those conserved elements which comprise the universal set of the last common ancestor [13], like ABC transporters, ribosomal proteins, cell cycle, transcription and translation regulators.

However, we found additional gene families which do not correspond with this trend, such as epimerase, chaperones and diverse elements implicit in transferase and synthase reactions. This provided us with a particular set of elements which follow different pathways and involve properties possibly implicated in the halophilic lifestyle strategy.

3.3. Duplication incidence ratio and normal distribution

From the 174 core genes, we identified a total set of 103 genes: 18 genes for Bacteria and 85 for Archaea. These genes, have a significant difference in the number of duplications between halophilic and non-halophilic species. The genes, with their corresponding functions and reactions are presented in Table 2. We noted a greater IR value in archaeal genomes than in the bacterial genomes.

The highest duplication ratio for Archaea was 1:11.74 from the signal-transducing histidine kinase-like protein (HacjB3_01630). We were able to group these 85 genes into 18 gene families (Figure 2). The highest family ratios are generated by chaperones, dehydrogenases, kinases and proteins involved in DNA repair.

Histidine kinases and chaperones in Archaea could play a vital role in the halophilic lifestyle, because they are involved in the changing environmental conditions of pH, temperature and internal ion concentration. The significance of having found a wide representation in histidine kinases is that they could be involved in signal transduction pathways, which participate in sensing the environment. The particular role of these proteins composed of a transmembrane and an intracellular domain [36], has been previously reported in bacterial models, implicating them in processes from chemotaxis to motility, secretion and cross-talking [37]. Their actual recognition as part of archaeal genes could implicate homologous regulation processes. Chaperones could play a major role in supporting stability and the folding of elements involved even in the central metabolism [38]. This proposal is supported by the work developed by Lahav *et al.* where the activity and expression of chaperones is modified by saline stress [39].

The prevalence of some epimerases and transferases in the core genes can be indirectly associated with shared elements of carbohydrate metabolism, in the synthesis of some carbohydrate-based compatible solutes [40-41]. However, further studies are needed to recognize their impact on specific metabolic routes.

Bacteria IR has a considerably lower number of representatives when compared with Archaea. Only two elements reach values of 1:3.67 and 1:3.5; these genes are the cation antiporter (caa) gene (VNG0380G) and the acetoin utilization (aup) gene (VNG0136G) respectively. Because of the low number of genes, it is not possible to correlate a particular function with the number of duplicated sequences. Only six of the 18 bacterial sets of genes have duplicated sequences that exceed the IR of 1:2.

The identification of a cation antiporter and the deacethylase genes in bacterial genomes could suggest two features of their biology. On one hand, the cation antiporters as proposed by Krulwich are the main regulators of pH, mostly in alkaline conditions [42-43]. This sign although it is the first approach, indicates a possible shared trend, where these genes and their paralogous elements are the result of common duplication events. The other hand, the paralogous sequences of aup acetoin-deacethylases, have been implicated in transcriptional downregulation because of their relation to the histone-like proteins presented in Archaea [44]. However, this result might make reference to horizontal gene transfer mechanism or a particular homolog event shared only by halophilic Bacteria. This aspect needs further study. The presence of histone-like proteins in Bacteria was not found in the core set of duplicated genes. Moreover, proteins could be regulating elements still not recognized or reported.

The genes with larger variability in the number of duplications are shown in Table 2. Those genes are found in greater numbers in Archaea than in Bacteria. In the archaeal domain, the family with the most representatives is the ABC transporters. The second largest group with the most variable genes is the histidine kinase genes. Bacteria is represented by only 5 genes, these genes are an ABC transporter, a transferase, one hypothetical gene, one deacetylase, and one oxidorreductase gene.

3.4. Heatmap and cluster analysis

The set of 85 genes in Archaea and 18 genes in Bacteria are plotted separately in Figures 3 and 4. We noted a remarkable bias in the halophilic Archaea, which involved at least two important groups with a similar pattern (Figure 3). The first group is composed of six halobacterial Archaea: *H. lacusprofundis, H. mukohataei, H. utahensis, H. walsbyi, N. pharaonis, H. jeotgali* and the four Methanobacteriales species included in this study. The second group, involving the highest values of related sequences, includes *N. magadii*, *H. marismortui* and *H. turkmenica*. Only two species of the first group are not involved with the Halobacteriales or Methanobacteriales: *T. pendens* and *T. kodakaraensis*. Two halobacterial species remain external to these groups: the species of the genera *Halobacterium.*

The results from the heatmap graphics for Archaea have allowed us to confirm the importance of histidine kinases, as well as their relation to certain ABC transporters. The first duplication bias involving Methanobacteriales and Halobacteriales could be explained by the shared halotolerance implied in both groups. The halotolerance of Methanobacterias is well documented and has been studied, identifying unique and specific ion transporter proteins [45-46] and particular compatible solutes [47]. However, it has been proposed that Methanobacteriales could have an incidence of universal compatible solutes in their structures [48]. The presence of these universal compatible solutes and their activation, which is regulated by changes due to intracellular ion concentration, could explain the shared pattern with Halobacteriales. A possible example of this proposal is the case of glycine betaine, and how it is regulated by an increase in potassium by the methanoarchaeon, *Methanohalophilus portucalensis* [48].

Although the direct relation of ion transporters with their compatible solutes or transduction signals is still not specified, we might infer that some paralogous found for ABC transporters could be important as sensors or as ionic transporters and regulators [49-50].

The proposal of elements shared between halophilic and hyperthermophilic species was previously referenced in *Thermofilum pendens* [51], showing a convergent pattern for this particular group of genes.

The high frequency of proteins related to sensory environment in our results, such as histidine kinases and ABC transporters, suggests that for both groups of archaeas, these genes are important in polyextremophily conditions [52-54].

Finally, the location of *Halobacterium* genera could be explained by their low K+ input activity, which was reported [55], and the possibility that *Halobacterium* genera could have a different duplication pattern, not focused on ion transport. This supports the possibility that the paralogous elements that we recognized might be common for archaeal species with regard to the salt-in strategy and the metabolism of some compatible solutes related to ion flux.

In the case of halophilic bacteria, we noted two particular sets of genes in which the whole of the halophilic species was included, with the exception of *A. arabaticum*. However, the structure and composition of these groups is diverse due to the presence of seven different species not related by their halotolerance.

We propose that these genes correlate with the importance of ABC transporters for the archaeal genomes, supporting the case that ion transporters spur the halophilic lifestyle in both cellular domains.

Conclusion:

Halophilic microorganisms show a great diversity in their metabolic strategies and pathways in order to respond to the stress produced by temperature, pH, and changes in salinity. By trying to recognize common elements, we were able to identify two particular traits: 1) a set (the core genes) with a high incidence of paralogous sequences in Bacteria (15 genes) but even more in Archaea, which have higher IR values (85 genes). This approach led us to consider that, in halophilic genomes, it is possible to recognize that duplication is one of the most important features needed to generate diversity in genomic evolution; and 2) in archaeal and bacterial domains, we found duplication events of genes involved in ion transporter proteins and genes involved in DNA repair.

A remarkable shared function recognized in both sets of genes is the ion transporter proteins; the ABC transporter subunits in Archaea and the caa antiporters in Bacteria. These abundant gene duplications can be explained in archaeal genomes, taking into account their roles in salt-in strategy and their activation of compatible solute metabolisms. However, it is necessary to identify which elements could couple with the ion transport and regulation of K+ in Archaea and which compatible solutes could have a direct relation with the ion flux in Bacteria.

Despite the bias of complete genome databases for halophilic organisms, it was possible to identify a group of duplicated genes which remain preserved in their genomes, as proposed by Bratlie *et al,* [14]. As we have previously noted, the differential paralogous sequence arrangements presented throughout this study have prompted us to divide all archaeal diversity into two main groups, and to omit the only archaeal genera with salt-out strategy: *Halobacterium* [54]. We propose that the core genes recognized by this study, mainly composed of ABC ion transporter subunits and histidine kinase paralogous genes, could represent a common feature of the salt-in strategy.

Acknowledgments

This paper constitutes a partial fulfillment of the Graduate Program in Biological Sciences of the National Autonomous University of México (UNAM). H.V. acknowledges the scholarship and financial support provided by the National Council of Science and Technology (CONACYT), and UNAM. F.R-P. was supported by the Programa de Becas Posdoctorales, UNAM. The support of CONACYT (100199) to A.B. is gratefully acknowledged. Part of the work reported here was completed during a sabbatical leave of absence of AB, with support of DGAPA-UNAM, where he enjoyed the hospitality of Prof. Juli Peretó at the Instituto Cavanilles (Valencia, Spain). We are indebted to Dr. Adrian Reyes-Prieto for many useful suggestions.

References:

[1] S.-G. Kong, W.-L. Fan, H.-D. Chen, Z.-T. Hsu, N. Zhou, B. Zheng, and H.C. Lee. 2009. "Inverse symmetry in complete genomes and whole-genome inverse duplication.," *PloS one*, vol. 4, no. 11, p. e7553

[2] C. Hildenbrand, T. Stock, C. Lange, M. Rother, and J. Soppa, 2011. "Genome copy numbers and gene conversion in methanogenic archaea.," *Journal of bacteriology*, vol. 193, no. 3, pp. 734–43.

[3] V. Pecoraro, K. Zerulla, C. Lange, and J. Soppa, 2011. "Quantification of ploidy in proteobacteria revealed the existence of monoploid, (mero-)oligoploid and polyploid species.," *PloS one*, vol. 6, no. 1, p. e16392, Jan. 2011.

[4] I. K. Jordan, K. S. Makarova, J. L. Spouge, Y. I. Wolf, and E. V Koonin, 2001. "Lineage-specific gene expansions in bacterial and archaeal genomes.," *Genome research*, vol. 11, no. 4, pp. 555–65.

[5] J. N. Davidson, K. C. Chen, R. S. Jamison, L. A. Musmanno, and C. B. Kern, 1993. "The evolutionary history of the first three enzymes in pyrimidine biosynthesis.," *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 15, no. 3, pp. 157–64.

[6] K. M. Devos, "Grass genome organization and evolution. 2010. " *Current opinion in plant biology*, vol. 13, no. 2, pp. 139–45, Apr. 2010.

[7] A. Bhan, D. J. Galas, and T. G. Dewey, 2002. "A duplication growth model of gene expression networks.," *Bioinformatics (Oxford, England)*, vol. 18, no. 11, pp. 1486–93.

[8] P. R. Marri, J. P. Bannantine, and G. B. Golding, 2006. "Comparative genomics of metabolic pathways in Mycobacterium species: gene duplication, gene decay and lateral gene transfer.," *FEMS microbiology reviews*, vol. 30, no. 6, pp. 906–25.

[9] J. P. Demuth and M. W. Hahn, 2009. "The life and death of gene families.," *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 31, no. 1, pp. 29–39.

[10] R. E. Collins, H. Merz, and P. G. Higgs, 2011. "Origin and evolution of gene families in Bacteria and Archaea.," *BMC bioinformatics*, vol. 12 Suppl 9, p. S14.

[11] A. Lazcano, E. Díaz-Villagómez, T. Mills, and J. Oró, 1995. "On the levels of enzymatic substrate specificity: implications for the early evolution of metabolic pathways.," *Advances in space research : the official journal of the Committee on Space Research (COSPAR)*, vol. 15, no. 3, pp. 345–56.

[12] A. Becerra and A. Lazcano. 1998. "The role of gene duplication in the evolution of purine nucleotide salvage pathways.," *Origins of life and evolution of the biosphere : the journal of the International Society for the Study of the Origin of Life*, vol. 28, no. 4–6, pp. 539–53, Oct. 1998.

[13] L. Delaye, A. Becerra, and A. Lazcano. 2005. "The last common ancestor: what's in a name?," *Origins of life and evolution of the biosphere : the journal of the International Society for the Study of the Origin of Life*, vol. 35, no. 6, pp. 537–54.

[14] M. S. Bratlie, J. Johansen, B. T. Sherman, D. W. Huang, R. A. Lempicki, and F. Drabløs. 2010. "Gene duplications in prokaryotes can be associated with environmental adaptation.," *BMC genomics*, vol. 11, p. 588.

[15] W. B. Watt and M. Dean. 2000. "Molecular-functional studies of adaptive genetic variation in prokaryotes and eukaryotes.," *Annual review of genetics*, vol. 34, pp. 593–622.

[16] A. C. Gerstein and S. P. Otto. 2009 "Ploidy and the causes of genomic evolution.," *The Journal of heredity*, vol. 100, no. 5, pp. 571–81.

[17] R. Maldonado, J. Jimenez, and J. Casadesus. 1994. "Changes of ploidy during the Azotobacter vinelandii growth cycle.," *J. Bacteriol.*, vol. 176, no. 13, pp. 3911–3919.

[18] N. J. Trun and S. Gottesman, 1991. “Characterization of Escherichia coli mutants with altered ploidy.” *Research in microbiology*, vol. 142, no. 2–3, pp. 195–200.

[19] M. Griese, C. Lange, and J. Soppa. 2011. “Ploidy in cyanobacteria.,” *FEMS microbiology letters*, vol. 323, no. 2, pp. 124–31.

[20] S. Breuert, T. Allers, G. Spohn, and J. Soppa. 2006. “Regulated polyploidy in halophilic archaea.,” *PloS one*, vol. 1, no. 1, p. e92.

[21] C. Lange, K. Zerulla, S. Breuert, and J. Soppa. 2011.“Gene conversion results in the equalization of genome copies in the polyploid haloarchaeon Haloferax volcanii.,” *Molecular microbiology*, vol. 80, no. 3, pp. 666–77.

[22] J. Soppa, 2011“Ploidy and gene conversion in Archaea.,” *Biochemical Society transactions*, vol. 39, no. 1, pp. 150–4.

[23] J. Soppa. 2005. “From replication to cultivation: hot news from Haloarchaea.,” *Current opinion in microbiology*, vol. 8, no. 6, pp. 737–44.

[24] J. A. Leigh, S.-V. V Albers, H. Atomi, and T. Allers. 2011. “Model organisms for genetics in the domain archaea: methanogens, halophiles, thermococcales and sulfolobales.,” *FEMS microbiology reviews*, vol. 35, no.4, pp. 577-608

[25] M. Kanehisa and S. Goto. 2000. “KEGG: kyoto encyclopedia of genes and genomes.,” *Nucleic acids research*, vol. 28, no. 1, pp. 27–30.

[26] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. 2004. “The KEGG resource for deciphering the genome.,” *Nucleic acids research*, vol. 32, no. Database issue, pp. D277–80.

[27] C. Seoighe. 2003. “Turning the clock back on ancient genome duplication.,” *Current opinion in genetics & development*, vol. 13, no. 6, pp. 636–43.

[28] A. Lawton-Rauh. 2003. “Evolutionary dynamics of duplicated genes in plants.,” *Molecular Phylogenetics and Evolution*, vol. 29, no. 3, pp. 396–409.

[29] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic local alignment search tool.,” *Journal of molecular biology*, vol. 215, no. 3, pp. 403–10.

[30] E. V Koonin, 2005. “Orthologs, paralogs, and evolutionary genomics.,” *Annual review of genetics*, vol. 39, pp. 309–38.

[31] J. P. Gogarten and L. Olendzenski, 1999. “Orthologs, paralogs and genome comparisons.,” *Current opinion in genetics & development*, vol. 9, no. 6, pp. 630–6.

[32] A. M. Altenhoff and C. Dessimoz, 2012. “Inferring orthology and paralogy.,” *Methods in molecular biology (Clifton, N.J.)*, vol. 855, pp. 259–79.

[33] R. L. Tatusov, A. R. Mushegian, P. Bork, N. P. Brown, W. S. Hayes, M. Borodovsky, K. E. Rudd, and E. V Koonin, 1996. “Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coli.,” *Current biology : CB*, vol. 6, no. 3, pp. 279–91.

[34] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, and A. Bateman, 2010. “The Pfam protein families database.,” *Nucleic acids research*, vol. 38, no. Database issue, pp. D211–22.

[35] N. Matloff, “The Art of R Programming,” 2009. San Francisco CA. USA. No Starch Press. 179 pp.

[36] Z. Zhang and W. A. Hendrickson, 2010. “Structural characterization of the predominant family of histidine kinase sensor domains.,” *Journal of molecular biology*, vol. 400, no. 3, pp. 335–53.

[37] X. Sheng, M. Huvet, J. W. Pinney, and M. P. H. Stumpf, 2012. “Evolutionary characteristics of bacterial two-component systems.,” *Advances in experimental medicine and biology*, vol. 751, pp. 121–37.

[38] R. Jaenicke, “Protein stability and molecular adaptation to extreme conditions.,” *European journal of biochemistry / FEBS*, vol. 202, no. 3, pp. 715–28, Dec. 1991.

[39] R. Lahav, A. Nejidat, and A. Abeliovich, 2004. “Alterations in protein synthesis and levels of heat shock 70 proteins in response to salt stress of the halotolerant yeast Rhodotorula mucilaginosa.,” *Antonie van Leeuwenhoek*, vol. 85, no. 4, pp. 259–69.

[40] M. S. da Costa, H. Santos, and E. A. Galinski, 1998. “An overview of the role and diversity of compatible solutes in Bacteria and Archaea.,” *Advances in biochemical engineering/biotechnology*, vol. 61, pp. 117–53.

[41] N. Empadinhas and M. S. da Costa, 2010.“Diversity, biological roles and biosynthetic pathways for sugar-glycerate containing compatible solutes in bacteria and archaea.,” *Environmental microbiology*, vol. 13, pp. 2056–2077.

[42] T. A. Krulwich, 1983. “Na+/H+ antiporters.,” *Biochimica et biophysica acta*, vol. 726, no. 4, pp. 245–64.

[43] T. A. Krulwich, G. Sachs, and E. Padan, 2011. “Molecular aspects of bacterial pH sensing and homeostasis.,” *Nature reviews. Microbiology*, vol. 9, no. 5, pp. 330–43.

[44] N.C. Kyrpides, C.A. Ouzounis, 1999. “Transcription in archaea.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 15. pp. 8545-50.

[45] R. Ciulla, C. Clougherty, N. Belay, S. Krishnan, C. Zhou, D. Byrd, and M. F. Roberts, 1994. “Halotolerance of Methanobacterium thermoautotrophicum delta H and Marburg.,” *Journal of bacteriology*, vol. 176, no. 11, pp. 3177–87.

[46] L. V Parfenova, B. M. Crane, and B. S. Rothberg, 2006. “Modulation of MthK potassium channel activity at the intracellular entrance to the pore.,” *The Journal of biological chemistry*, vol. 281, no. 30, pp. 21131–8.

[47] M. C. Lai, K. R. Sowers, D. E. Robertson, M. F. Roberts, and R. P. Gunsalus, 1991. “Distribution of compatible solutes in the halophilic methanogenic archaebacteria.,” *J. Bacteriol.*, vol. 173, no. 17, pp. 5352–5358.

[48] M.C. Lai, D.R.Yang, M.J. Chuang, 1999. “Regulatory factors associated with synthesis of the osmolyte glycine betaine in the halophilic methanoarchaeon Methanohalophilus portucalensis.” *Applied and environmental microbiology.* vol. 65. no. 2, pp. 828-33.

[49] M. F. Roberts, 2005.“Organic compatible solutes of halotolerant and halophilic microorganisms.,” *Saline systems*, vol. 1, p. 5.

[50] E. Dassa, 2011. “Natural history of ABC systems: not only transporters.,” *Essays in biochemistry*, vol. 50, no. 1, pp. 19–42.

[51] I. Anderson, J. Rodriguez, D. Susanti, I. Porat, C. Reich, L. E. Ulrich, J. G. Elkins, K. Mavromatis, A. Lykidis, E. Kim, L. S. Thompson, M. Nolan, M. Land, A. Copeland, A. Lapidus, S. Lucas, C. Detter, I. B. Zhulin, G. J. Olsen, W. Whitman, B. Mukhopadhyay, J. Bristow, and N. Kyrpides, 2008. “Genome sequence of Thermofilum pendens reveals an exceptional loss of biosynthetic pathways without genome reduction.,” *Journal of bacteriology*, vol. 190, no. 8, pp. 2957–65.

[52] M. D’Antonio and F. D. Ciccarelli, 2011“Modification of gene duplicability during the evolution of protein interaction network.,” *PLoS computational biology*, vol. 7, no. 4, p. e1002029.

[53] X. Zhang, M. Kupiec, U. Gophna, and T. Tuller, 2011. “Analysis of coevolving gene families using mutually exclusive orthologous modules.,” *Genome biology and evolution*, vol. 3, pp. 413–23.

[54] S. E. Giuliani, A. M. Frank, D. M. Corgliano, C. Seifert, L. Hauser, and F. R. Collart, 2011. “Environment sensing and response mediated by ABC transporters.,” *BMC genomics*, vol. 12 Suppl 1, no. Suppl 1, p. S8.

[55] S. P. Kennedy, W. V Ng, S. L. Salzberg, L. Hood, and S. DasSarma, 2001."Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence.," *Genome research*, vol. 11, no. 10, pp. 1641–50.

[56] H. Ivanova, N., Daum, C., Lang, E., Abt, B., Kopitz, M., Saunders, E., Lapidus, A., Lucas, S., Glavina Del Rio, T., Nolan, M., Tice, H., Copeland, A., Cheng, J-F., Chen, F., Bruce, D., Goodwin, L., Pitluck, S., Mavromatis, K., Pati, A., Mikhailova, N., Chen, 2010. "Complete genome sequence of Haliangium ochraceum type strain (SMP-2T)," *Standards in Genomic Sciences*, vol. 2, no. 1, pp. 96–106.

[57] M. a Campbell, P. S. G. Chain, H. Dang, A. F. El Sheikh, J. M. Norton, N. L. Ward, B. B. Ward, and M. G. Klotz, 2011."Nitrosococcus watsonii sp. nov., a new species of marine obligate ammonia-oxidizing bacteria that is not omnipresent in the world's oceans: calls to validate the names 'Nitrosococcus halophilus' and 'Nitrosomonas mobilis'.," *FEMS microbiology ecology*, vol. 76, no. 1, pp. 39–48.

[58] H. Tsuihiji, Y. Yamazaki, H. Kamikubo, Y. Imamoto, and M. Kataoka, 2006. "Cloning and characterization of nif structural and regulatory genes in the purple sulfur bacterium, *Halorhodospira halophila*.," *Journal of bioscience and bioengineering*, vol. 101, no. 3, pp. 263–70.

[59] D. Y. Sorokin and J. G. Kuenen, 2005. "Chemolithotrophic haloalkaliphiles from soda lakes.," *FEMS microbiology ecology*, vol. 52, no. 3, pp. 287–95.

[60] R. H. Vreeland, C. D. Litchfield, E. L. Martin, and E. Elliot, 1980. "*Halomonas elongata*, a New Genus and Species of Extremely Salt-Tolerant Bacteria," *International Journal of Systematic Bacteriology*, vol. 30, no. 2, pp. 485–495.

[61] H. Takami, K. Nakasone, Y. Takaki, G. Maeno, R. Sasaki, N. Masui, F. Fuji, C. Hirama, Y. Nakamura, N. Ogasawara, S. Kuhara, K. Horikoshi. 2000. "Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparidon with *Bacillus subtilis*." *Nucleic Acids Research*. vol. 28. no. 21. pp. 4317-31.

[62] B. Zhao, N. M. Mesbah, E. Dalin, L. Goodwin, M. Nolan, S. Pitluck, O. Chertkov, T. S. Brettin, J. Han, F. W. Larimer, M. L. Land, L. Hauser, N. Kyrpides, and J. Wiegel. 2011. "Complete Genome Sequence of the Anaerobic, Halophilic Alkalithermophile *Natranaerobius thermophilus* JW/NM-WN-LF.," *Journal of bacteriology*, vol. 193, no. 15, pp. 4023–4.

[63] K. Mavromatis, N. Ivanova, I. Anderson, A. Lykidis, S. D. Hooper, H. Sun, V. Kunin, A. Lapidus, P. Hugenholtz, B. Patel, and N. C. Kyrpides, 2009. "Genome analysis of the anaerobic thermohalophilic bacterium *Halothermothrix orenii*.," *PloS one*, vol. 4, no. 1, p. e4192.

[64] H.P. Sikorski, J. Lapidus, A. Chertkov, O. Lucas, S. Copeland, A. Glavina Del Rio, T. Nolan, M. Tice, H. Cheng, J.F. Han, C. Brambilla, E. Pitluck, S. Liolios, K. Ivanova, N. Mavromatis, K. Mikhailova, N. Pati, A., Bruce, D., Detter, C., Tapia, 2010. "Complete genome sequence of *Acetohalobium arabaticum* type strain (Z-7288T)," *Standards in Genomic Sciences*, vol. 3, no. 1, pp. 57–65.

[65] E. F. Mongodin, K. E. Nelson, S. Daugherty, R. T. Deboy, J. Wister, H. Khouri, J. Weidman, D. A. Walsh, R. T. Papke, G. Sanchez Perez, A. K. Sharma, C. L. Nesbø, D. MacLeod, E. Bapteste, W. F. Doolittle, R. L. Charlebois, B. Legault, and F. Rodriguez-Valera, 2005."The genome of *Salinibacter ruber*: convergence and gene exchange among hyperhalophilic bacteria and archaea.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 50, pp. 18147–52.

[66] J. Brito-Echeverría, M. Lucio, A. López-López, J. Antón, P. Schmitt-Kopplin, and R. Rosselló-Móra, 2011 "Response to adverse conditions in two strains of the extremely halophilic species *Salinibacter ruber*.," *Extremophiles : life under extreme conditions*, vol. 15, no. 3, pp. 379–89.

[67] N. S. Baliga, R. Bonneau, M. T. Facciotti, M. Pan, G. Glusman, E. W. Deutsch, P. Shannon, Y. Chiu, R. S. Weng, R. R. Gan, P. Hung, S. V Date, E. Marcotte, L. Hood, and W. V. Ng, 2004. "Genome sequence of

Haloarcula marismortui: a halophilic archaeon from the Dead Sea.," *Genome research*, vol. 14, no. 11, pp. 2221–34.

[68] H.L. Cui, P.J. Zhou, A. Oren, and S.J. Liu, 2009. "Intraspecific polymorphism of 16S rRNA genes in two halophilic archaeal genera, *Haloarcula* and *Halomicrobium*.," *Extremophiles : life under extreme conditions*, vol. 13, no. 1, pp. 31–7.

[69] S. Leuko, M. J. Raftery, B. P. Burns, M. R. Walter, and B. A. Neilan, 2009. "Global protein-level responses of Halobacterium salinarum NRC-1 to prolonged changes in external sodium chloride concentrations.," *Journal of proteome research*, vol. 8, no. 5, pp. 2218–25.

[70] E. Saunders, B. J. Tindall, R. Fähnrich, A. Lapidus, A. Copeland, T. G. Del Rio, S. Lucas, F. Chen, H. Tice, J.-F. Cheng, C. Han, J. C. Detter, D. Bruce, L. Goodwin, P. Chain, S. Pitluck, A. Pati, N. Ivanova, K. Mavromatis, A. Chen, K. Palaniappan, M. Land, L. Hauser, Y.-J. Chang, C. D. Jeffries, T. Brettin, M. Rohde, M. Göker, J. Bristow, J. A. Eisen, V. Markowitz, P. Hugenholtz, H.P. Klenk, and N. C. Kyrpides, 2010. "Complete genome sequence of *Haloterrigena turkmenica* type strain (4k).," *Standards in genomic sciences*, vol. 2, no. 1, pp. 107–16.

[71] I. Anderson, B. J. Tindall, H. Pomrenke, M. Göker, A. Lapidus, M. Nolan, A. Copeland, T. Glavina Del Rio, F. Chen, H. Tice, J.-F. Cheng, S. Lucas, O. Chertkov, D. Bruce, T. Brettin, J. C. Detter, C. Han, L. Goodwin, M. Land, L. Hauser, Y.-J. Chang, C. D. Jeffries, S. Pitluck, A. Pati, K. Mavromatis, N. Ivanova, G. Ovchinnikova, A. Chen, K. Palaniappan, P. Chain, M. Rohde, J. Bristow, J. A. Eisen, V. Markowitz, P. Hugenholtz, N. C. Kyrpides, and H.-P. Klenk, 2009. "Complete genome sequence of *Halorhabdus utahensis* type strain (AX-2).," *Standards in genomic sciences*, vol. 1, no. 3, pp. 218–25.

[72] H. Bolhuis, P. Palm, A. Wende, M. Falb, M. Rampp, F. Rodriguez-Valera, F. Pfeiffer, and D. Oesterhelt. 2006. "The genome of the square archaeon *Haloquadratum walsbyi* : life at the limits of water activity.," *BMC genomics*, vol. 7, p. 169.

[73] S. Siddaramappa, J.F. Challacombe, R.E. De Castro, F. Pfeiffer, S. Friedhelm, M.I. Giménez,R.A. Paggi, J.C. Detter, K.W. Davenport, L.A. Goodwin, N. Kyrpides, R. Tapia, S. Pitluck, S. Lucas, T. Woyke, J.A. Maupin-Furlow, 2012. "A comparative genomics perspective on the genetic content of the alkaliphilic haloarchaeon *Natrialba magadii* ATCC 43099T. *BMC Genomics*.vol.13, no. 1. pp. 165.

[74] S. W. Roh, Y.-D. Nam, H.-W. Chang, Y. Sung, K.-H. Kim, H.-M. Oh, and J.-W. Bae, 2007. "*Halalkalicoccus jeotgali* sp. nov., a halophilic archaeon from shrimp jeotgal, a traditional Korean fermented seafood.," *International journal of systematic and evolutionary microbiology*, vol. 57, no. Pt 10, pp. 2296–8.

[75] J. A. E. Gibson, M. R. Miller, N. W. Davies, G. P. Neill, D. S. Nichols, and J. K. Volkman, 2005. "Unsaturated diether lipids in the psychrotrophic archaeon *Halorubrum lacusprofundi*.," *Systematic and applied microbiology*, vol. 28, no. 1, pp. 19–26.

[76] M. Falb, F. Pfeiffer, P. Palm, K. Rodewald, V. Hickmann, J. Tittor, and D. Oesterhelt, 2005. "Living with two extremes: conclusions from the genome sequence of Natronomonas pharaonis.," *Genome research*, vol. 15, no. 10, pp. 1336–43.

[77] T. N. Zhilina and G. A. Zavarzin, 1987. "*Methanohalobium evestigatus*, n. gen., n. sp., the extremely halophilic methanogenic Archaebacterium," *Dokl. Akad. Nauk USSR*, vol. 293, pp. 464–468.

[78] N.M. Mesbah, J. Wiegel, 2008. "Life at extreme limits: the anaerobic halophilic alkalithermophiles." *Annals of the New York Academy of Sciences*. vol.1125, pp. 44-57.

8 Capítulo V. Discusión final

El estudio de los organismos extremófilos, siempre ha dado una buena pauta para conocer mas acerca del variedades a nivel metabólico [Stetter, 2006], para obtener enzimas que pueden ser utilizadas en el campo de la biotecnología [de Champdoré *et al.*, 2007] e incluso para cuestionarnos acerca de las diferentes ramas filogenéticas, las cuales se les ha una incidencia dentro de etapas tempranas de la vida en la tierra, aunque no primigenias [Islas *et al.*, 2003].

Actualmente, uno de los elementos que ahora se puede analizar, incluso a nivel in silico, son los genomas en su totalidad y la composición de los mismos, reconociendo cada una de sus regiones y genes. Aunque ya se han reportado y desmentido varias características que se relacionaban para el estilo de vida hipertermofílico como la presencia de un contenido de GC que sobrepase el 60 %, o la presencia de más de un sesgo de esta concentración a través de todo el genoma, todavia se desconocen diferentes tipos de regulación y señales de transducción.

La teoría de que el genoma esta sesgado y limitado en su composición es una propuesta compartida por varios grupos de investigación, y que si es analizada dentro de diferentes estilos de vida extremófilo, es posible reconocer patrones que aún no se han reportado, sin embargo su estructura se basa en regiones tándem, de alta flexibilidad, mas esto requiere de una comparativa en ambientes variados, grupos filogenéticos, para así intentar reconocer y extrapolar elementos propios a un estilo de vida o a un grupo filogenético.

La aproximación presentada dentro del artículo aceptado, permite detectar que aún hay variantes que se pueden analizar como continuación a este proyecto: si bien las características del transporte horizontal las estamos relacionando directamente a las regiones en tándem y de alta flexibilidad, estas podrían darnos una pauta adicional de como responden los diferentes genomas y no implicarse en todos los casos a este proceso.

El pensamiento de Conant y Cordero [Conant y Wolfe, 2008, Cordero y Polz, 2014] ha dado pautas nuevas e innovaciones dentro del estudio de la pangenómica, asi, vale la pena sopesar e intentar vincular la historia ambiental y evolutiva de mas de un organismo extremófilo y mesófilo. es necesario reevaluar incluso la manera que se esta analizando en un inicio este trabajo ya que cabe la posibilidad de que una señal en plásmidos y cromosomas variados puede estarse pasando por alto.

Esto nos hizo posible integrarla incluso ideas de varios autores, que daban por alto que los sesgos eran estables y que no se presentaba algo diferente a adicionar al esquema de respuesta del genoma [Singer y Hickey, 2003, Van der Linden y de Farias, 2006]; y pensar que el genoma se pueda ver afectado en su estructura y composición por incluso, por variables ambientales, y que su remodelación puede interpretarse por elementos, variables e índices diferenciales. Si bien los estudios previos nos dan una pauta de que los genes ortólogos nos dan una pauta clara, la comparativa entre el cromosoma y el plásmido nos podría estar otorgando una visión mas detallada de como cambia una secuencia, codificante o no codificantes ante un sesgo mutacional local o un arreglo resultante de una presión de selección ambiental, esto de todas formas vale la pena reevaluarlo y establecerlo en una propuestas mas amplia y

diferente.

En trabajos posteriores, valdría la pena también el reconocer la importancia del manosilglicerato, único soluto compatible relacionado de forma independiente con el estilo de vida hipertermófilo, ya que aunque hay mas solutos compatibles en hipertermófilos y en especies que resisten tanto una alta salinidad como una temperatura mayor a la mesófila, cavbría la posibilidad de que esta ruta metabólica podría estudiarse a manera de reloj molecular en diferentes especies.

Dentro de los mismos resultados parciales para hipertermófilos, parece ser que los índices estructurales postulados por Miramontes et al.[Miramontes *et al.*, 1995], dan la posibilidad de que a nivel genómico, se encuentra un valor parcial para los hipertermófilos del dominio Archaea. Estos índices muestran una señal muy debil de correlación a nivel global, sin embargo deben de analizarse con detenimiento y corroborar la incidencia común entre el valor V y el Yrd, para el caso de grupos divesos y eucariontes.

Un resultado que vale la pena recalcar es la incidencia de una colección de genes que buscando un set único de genes relacionados a un estilo de vida hipertermófilo, se logró solo obtener un conjunto de genes universales. Esta propuesta, del mismo modo, no dio un un resultado positivo que permitiera corelacionar un conjunto de secuencias con el estilo de vida. Esto es coherente con los eventos de disminución de genoma y las diferentes estrategias y componentes reportados previamente, tales como la presencia parcial de la reverso girasa [Atomi *et al.*, 2004], o la incidencia de los diferentes elementos moleculares que permiten resolver y evitar los procesos de desnaturalización como son chaperonas [Sterner y Liebl, 2001], sistemas de reparación únicos y homólogos al Bacteria y Eukarya [DiRuggiero *et al.*, 1999], scorde a esto, se ha postulado que el estilo de vida termófilo parece haberse originado tanto en las etapas tempranas de la vida, [Islas *et al.*, 2003] y que al menos el ser termófilo se ha originado de forma independiente mas de una vez a nivel bacteriano.

Es por esta posibilidad y por eventos de transporte horizontal, que se dieron en diferentes tiempos evolutivos, que sus estrategias pudieron haberse enriquecido, para así originar el estilo de vida hipertermófilo en varias subdivisiones bacterianas y en tiempos diferentes, no solo por la reverso girasa, sino por regiones codificantes y componentes que se encontraban cercanos al mismo gen, y que por comparación de secuencias, se han identificado como elementos de origen arqueobacteriano, permitiendo así la colonización de ambientes hipertermófilos [Brochier y Philippe, 2002].

El otro punto a discutir son los diferentes eventos de duplicación hipotetizados que pudieron dar origen a toda la abundancia de secuencias parálogas que inciden de manera preferente en genomas de halófilos. Es en esto casos en donde se postula que estos extremófilos dependan de este acervo y configuración particular les permita: *a)* responder ante condiciones cambiantes e intervalos variados y de forma consecuente, *b)* contar con un grupo de secuencias las cuales fungen como elementos que a la larga, podrían dar origen a elementos nuevos para mas de una familia de genes implementando tasa de mutación y eventos de divergencia.

Este conjunto de secuencias parálogas del mismo modo ha hecho posible tener un factor predictivo sobre una de las

dos estrategias identificadas en las estrategias de los halófilos. Si bien la mayoría de los halófilos del dominio Archaea presentan la estrategia salt – in, solamente el género de *Halobacterium* tiene un mayor uso de solutos compatibles que de transportadores de iones. Esto es coherente con el resultado obtenido dentro de los análisis del capítulo 2, en donde fue posible el reconocer y el dejar claro que el conjunto de elementos parálogos, mayoriamente presentado por histidin cinasas y por transportadores de iones,.permite incluso identificar y diferenciar este rasgo aunado a reconocer 3 genomas (*Natrialba magadii, Haloterrigena turkmenica y Haloarcula marismortui*) que hacen diferenciar a este comportamiento al menos en dos niveles diferentes incluso dentro del dominio Archaea.

Esto nos ha permitido sospechar que la esencia de este descubrimiento es el poder reconocer que ciertos elementos de la estrategia salt – in 1) dependen de eventos de divergencia, 2) estos procesos de divergencia permiten que se de una respuesta a mas de un factor ambiental, esto, por los rasgos de poliextremofilia de las tres especies con mas secuencias parálogas y 3) es un sistema de mas de una estrategia , hay elementos adicionales que permiten suponer procesos de oxidorreducciń y metabolismo de solutos compatibles aun no descritos.

Este estudio comparativo ha permitido el reconocer una tendencia propia de los genomas de los organismos halófilos, definir un conjunto de genes vitales, posiblemente para su respuesta ante los cambios de salinidad e intervalos de tolerancia. La realidad es que el genoma de los microrganismos y mas aquellos ue habitan en condiciones de vida extrema, parecen darnos una pauta en donde la adecuación y adaptación podria romper el esquema inicial de analizar la estabilidad enzimática como principal modelador de la secuencia de los genes. Mas aún es dentro de este análisis donde es posible el ubicar que gracias a los motivos conservados de pFam y al valor IR incluso es posible el ubicar elementos comunes y la presencia de homólogos que sobrepasan el sesgo del dominio.

La presencia de una oxidorreductasa putativa en ambos dominios procariontes hace pensar que incluso la presencia de estas tres variables como la actividad el valor IR y la presencia de motivos específicos conservados son variables que podrían utilizarse para reconocer desde arreglos en genes ortólogos hasta la identificación de parálogos.

Esta nueva metodología valdría ponerla a prueba dentro de rasgos conservados y familias de ortólogos conocidas para asi comprender si estos nos pueden evindenciar elementos conservados que todavía son cripticos por la tasa de mutación. Es recomendable continuar evaluando con el valor IR nos puede incluso dat nociones de como un evento de duplicación y divergencia nos da cavida para proponer nuevos modelos de secuencias ancestrales y homólogas.

Vale la pena realizar una actualización cuando ya se cuente con una muestra mayor de genes, del mismo modo quedará pendiente analizar este fenómeno en intervalos menores dentro de los genomas bacterianos, ideitificar la secuencia de los genes parálogos, así como intentar a nivel parcial la presencia de sesgos en sus aminoácidos.

Referencias

[Agarwal y Grover, 2008] Agarwal, S. W. y Grover, A. (2008). Nucleotide composition and amino acid usage in AT-rich hyperthermophilic species. *Open Bioinform J*, 2:11–19.

[Ahmed *et al.*, 2005] Ahmed, H., Ettema, T. J. G., Tjaden, B., Geerling, A. C. M., van der Oost, J., y Siebers, B. (2005). The semi-phosphorylative Entner-Doudoroff pathway in hyperthermophilic archaea: a re-evaluation. *The Biochemical journal*, 390(Pt 2):529–40.

[Allers y Mevarech, 2005] Allers, T. y Mevarech, M. (2005). Archaeal genetics - the third way. *Nature reviews. Genetics*, 6(1):58–73.

[Altermann, 2012] Altermann, E. (2012). Tracing lifestyle adaptation in prokaryotic genomes. *Frontiers in microbiology*, 3:48.

[Altschul *et al.*, 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., y Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–10.

[Anderson *et al.*, 2009] Anderson, I., Ulrich, L. E., Lupa, B., Susanti, D., Porat, I., Hooper, S. D., Lykidis, A., Sieprawska-Lupa, M., Dharmarajan, L., Goltsman, E., Lapidus, A., Saunders, E., Han, C., Land, M., Lucas, S., Mukhopadhyay, B., Whitman, W. B., Woese, C., Bristow, J., y Kyrpides, N. (2009). Genomic characterization of methanomicrobiales reveals three classes of methanogens. *PloS one*, 4(6):e5797.

[Antón *et al.*, 2002] Antón, J., Oren, A., Benlloch, S., Rodríguez-Valera, F., Amann, R., y Rosselló-Mora, R. (2002). Salinibacter ruber gen. nov., sp. nov., a novel, extremely halophilic member of the Bacteria from saltern crystallizer ponds. *International journal of systematic and evolutionary microbiology*, 52(Pt 2):485–91.

[Archibald y Roger, 2002] Archibald, J. M. y Roger, A. J. (2002). Gene duplication and gene conversion shape the evolution of archaeal chaperonins. *Journal of molecular biology*, 316(5):1041–50.

[Atomi *et al.*, 2004] Atomi, H., Matsumi, R., y Imanaka, T. (2004). Reverse gyrase is not a prerequisite for hyperthermophilic life. *Journal of Bacteriology*, 186(14):4829–33.

[Baliga *et al.*, 2004] Baliga, N. S., Bonneau, R., Facciotti, M. T., Pan, M., Glusman, G., Deutsch, E. W., Shannon, P., Chiu, Y., Weng, R. S., Gan, R. R., Hung, P., Date, S. V., Marcotte, E., Hood, L., y Ng, W. V. (2004). Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea. *Genome research*, 14(11):2221–34.

[Batut *et al.*, 2004] Batut, J., Andersson, S. G. E., y O'Callaghan, D. (2004). The evolution of chronic infection strategies in the alpha-proteobacteria. *Nature reviews. Microbiology*, 2(12):933–45.

[Becerra y Lazcano, 1998] Becerra, A. y Lazcano, A. (1998). The role of gene duplication in the evolution of purine nucleotide salvage pathways. *Origins of life and evolution of the biosphere : the journal of the International Society for the Study of the Origin of Life*, 28(4-6):539–53.

[Bhan *et al.*, 2002] Bhan, A., Galas, D. J., y Dewey, T. G. (2002). A duplication growth model of gene expression networks. *Bioinformatics (Oxford, England)*, 18(11):1486–93.

[Bolhuis *et al.*, 2006] Bolhuis, H., Palm, P., Wende, A., Falb, M., Rampp, M., Rodriguez-Valera, F., Pfeiffer, F., y Oesterhelt, D. (2006). The genome of the square archaeon Haloquadratum walsbyi : life at the limits of water activity. *BMC genomics*, 7:169.

[Boussau *et al.*, 2008] Boussau, B., Blanquart, S., Necsulea, A., Lartillot, N., y Gouy, M. (2008). Parallel adaptations to high temperatures in the Archaean eon. *Nature*, 456(7224):942–5.

[Breuert *et al.*, 2006] Breuert, S., Allers, T., Spohn, G., y Soppa, J. (2006). Regulated polyploidy in halophilic archaea. *PloS one*, 1(1):e92.

[Brochier y Philippe, 2002] Brochier, C. y Philippe, H. (2002). Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature*, 417(6886):244.

[Cai *et al.*, 2009] Cai, Y., Patel, D. J., Geacintov, N. E., y Broyde, S. (2009). Differential nucleotide excision repair susceptibility of bulky DNA adducts in different sequence contexts: hierarchies of recognition signals. *Journal of molecular biology*, 385(1):30–44.

[Cambillau y Claverie, 2000] Cambillau, C. y Claverie, J. M. (2000). Structural and genomic correlates of hyperthermostability. *The Journal of biological chemistry*, 275(42):32383–6.

[Campbell *et al.*, 2011] Campbell, M. a., Chain, P. S. G., Dang, H., El Sheikh, A. F., Norton, J. M., Ward, N. L., Ward, B. B., y Klotz, M. G. (2011). Nitrosococcus watsonii sp. nov., a new species of marine obligate ammonia-oxidizing bacteria that is not omnipresent in the world's oceans: calls to validate the names 'Nitrosococcus halophilus' and 'Nitrosomonas mobilis'. *FEMS microbiology ecology*, 76(1):39–48.

[Chan *et al.*, 2011] Chan, P. P., Cozen, A. E., y Lowe, T. M. (2011). Discovery of permuted and recently split transfer RNAs in Archaea. *Genome biology*, 12(4):R38.

[Cole *et al.*, 2012] Cole, F., Keeney, S., y Jasin, M. (2012). Preaching about the converted: how meiotic gene conversion influences genomic diversity. *Annals of the New York Academy of Sciences*, 1267:95–102.

[Coleman *et al.*, 2006] Coleman, M. L., Sullivan, M. B., Martiny, A. C., Steglich, C., Barry, K., Delong, E. F., y Chisholm, S. W. (2006). Genomic islands and the ecology and evolution of Prochlorococcus. *Science (New York, N.Y.)*, 311(5768):1768–70.

[Collins *et al.*, 2011] Collins, R. E., Merz, H., y Higgs, P. G. (2011). Origin and evolution of gene families in Bacteria and Archaea. *BMC bioinformatics*, 12 Suppl 9:S14.

[Conant y Wolfe, 2008] Conant, G. C. y Wolfe, K. H. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nature reviews. Genetics*, 9(12):938–50.

[Conners *et al.*, 2006] Conners, S. B., Mongodin, E. F., Johnson, M. R., Montero, C. I., Nelson, K. E., y Kelly, R. M. (2006). Microbial biochemistry, physiology, and biotechnology of hyperthermophilic Thermotoga species. *FEMS microbiology reviews*, 30(6):872–905.

[Cordero y Polz, 2014] Cordero, O. X. y Polz, M. F. (2014). Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Publishing Group*, 12(4):263–273.

[Cui *et al.*, 2009] Cui, H.-L., Zhou, P.-J., Oren, A., y Liu, S.-J. (2009). Intraspecific polymorphism of 16S rRNA genes in two halophilic archaeal genera, Haloarcula and Halomicrobium. *Extremophiles : life under extreme conditions*, 13(1):31–7.

[Das *et al.*, 2006] Das, S., Paul, S., Bag, S. K., y Dutta, C. (2006). Analysis of Nanoarchaeum equitans genome and proteome composition: indications for hyperthermophilic and parasitic adaptation. *BMC genomics*, 7:186.

[de Champdoré *et al.*, 2007] de Champdoré, M., Staiano, M., Rossi, M., y D'Auria, S. (2007). Proteins from extremophiles as stable tools for advanced biotechnological applications of high social interest. *Journal of the Royal Society, Interface / the Royal Society*, 4(13):183–91.

[Deckert *et al.*, 1998] Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E., Overbeek, R., Snead, M. A., Keller, M., Aujay, M., Huber, R., Feldman, R. A., Short, J. M., Olsen, G. J., y Swanson, R. V. (1998). The complete genome of the hyperthermophilic bacterium Aquifex aeolicus. *Nature*, 392(6674):353–8.

[Demuth y Hahn, 2009] Demuth, J. P. y Hahn, M. W. (2009). The life and death of gene families. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 31(1):29–39.

[Devos, 2010] Devos, K. M. (2010). Grass genome organization and evolution. *Current opinion in plant biology*, 13(2):139–45.

[Dickerson y Ng, 2001] Dickerson, R. E. y Ng, H. L. (2001). DNA structure from A to B. *Proceedings of the National Academy of Sciences of the United States of America*, 98(13):6986–8.

[Diekmann, 1989] Diekmann, S. (1989). Definitions and nomenclature of nucleic acid structure parameters. *Journal of Molecular Biology*, 205(4):787–791.

[DiRuggiero *et al.*, 1999] DiRuggiero, J., Brown, J. R., Bogert, A. P., y Robb, F. T. (1999). DNA repair systems in archaea: mementos from the last universal common ancestor? *Journal of molecular evolution*, 49(4):474–84.

[Doolittle *et al.*, 1996] Doolittle, R. F., Feng, D. F., Tsang, S., Cho, G., y Little, E. (1996). Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science (New York, N.Y.)*, 271(5248):470–7.

[Doolittle y Brown, 1994] Doolittle, W. F. y Brown, J. R. (1994). Tempo, mode, the progenote, and the universal root. *Proceedings of the National Academy of Sciences of the United States of America*, 91(15):6721–8.

[Dutta y Paul, 2012] Dutta, C. y Paul, S. (2012). Microbial lifestyle and genome signatures. *Current genomics*, 13(2):153–62.

[Elevi Bardavid y Oren, 2012] Elevi Bardavid, R. y Oren, A. (2012). The amino acid composition of proteins from anaerobic halophilic bacteria of the order Halanaerobiales. *Extremophiles : life under extreme conditions*, 16(3):567–72.

[Empadinhas y Costa, 2008] Empadinhas, N. y Costa, M. S. (2008). Osmoadaptation mechanisms in prokaryotes : distribution of compatible solutes. *International Microbiology*, 11:151–161.

[Falb *et al.*, 2008] Falb, M., Müller, K., Königsmaier, L., Oberwinkler, T., Horn, P., von Gronau, S., Gonzalez, O., Pfeiffer, F., Bornberg-Bauer, E., y Oesterhelt, D. (2008). Metabolism of halophilic archaea. *Extremophiles : life under extreme conditions*, 12(2):177–96.

[Fani *et al.*, 1995] Fani, R., Liò, P., y Lazcano, A. (1995). Molecular evolution of the histidine biosynthetic pathway. *Journal of molecular evolution*, 41(6):760–74.

[Finn *et al.*, 2010] Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., y Bateman, A. (2010). The Pfam protein families database. *Nucleic acids research*, 38(Database issue):D211–22.

[Galtier y Duret, 2007] Galtier, N. y Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in genetics : TIG*, 23(6):273–7.

[Gerstein y Otto, 2009] Gerstein, A. C. y Otto, S. P. (2009). Ploidy and the causes of genomic evolution. *The Journal of heredity*, 100(5):571–81.

[Gibson *et al.*, 2005] Gibson, J. A. E., Miller, M. R., Davies, N. W., Neill, G. P., Nichols, D. S., y Volkman, J. K. (2005). Unsaturated diether lipids in the psychrotrophic archaeon Halorubrum lacusprofundi. *Systematic and applied microbiology*, 28(1):19–26.

[Goodsell *et al.*, 1993] Goodsell, D., Kopka, M., Cascio, D., y Dickerson, R. (1993). Crystal structure of CATGGC-CATG and its implications for A-tract bending models. *Proceedings of the National Academy of Sciences of the United States of America*, 90(7):2930–4.

[Greaves y Warwicker, 2007] Greaves, R. B. y Warwicker, J. (2007). Mechanisms for stabilisation and the maintenance of solubility in proteins from thermophiles. *BMC structural biology*, 7:18.

[Gribaldo y Brochier-Armanet, 2006] Gribaldo, S. y Brochier-Armanet, C. (2006). The origin and evolution of Archaea: a state of the art. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 361(1470):1007–1022.

[Groussin y Gouy, 2011] Groussin, M. y Gouy, M. (2011). Adaptation to Environmental Temperature is a Major Determinant of Molecular Evolutionary Rates in Archaea. *Molecular Biology and Evolution*, 28(9):1–42.

[Hildenbrand *et al.*, 2011] Hildenbrand, C., Stock, T., Lange, C., Rother, M., y Soppa, J. (2011). Genome Copy Numbers and Gene Conversion in Methanogenic Archaea. *Journal of bacteriology*, 193(3):734–743.

[Hong *et al.*, 2008] Hong, M., Fitzgerald, M. X., Harper, S., Luo, C., Speicher, D. W., y Marmorstein, R. (2008). Structural basis for dimerization in DNA recognition by Gal4. *Structure (London, England : 1993)*, 16(7):1019–26.

[Hua-Van *et al.*, 2011] Hua-Van, A., Le Rouzic, A., Boutin, T. S., Filée, J., y Capy, P. (2011). The struggle for life of the genome's selfish architects. *Biology direct*, 6:19.

[Islas *et al.*, 2003] Islas, S., Velasco, A., Becerra, A., Delaye, L., y Lazcano, A. (2003). Hyperthermophily and the origin and earliest evolution of life. *International microbiology : the official journal of the Spanish Society for Microbiology*, 6(2):87–94.

[Ivanova *et al.*, 2010] Ivanova, N., Daum, C., Lang, E., Abt, B., Kopitz, M., Saunders, E., Lapidus, A., Lucas, S., Glavina Del Rio, T., Nolan, M., Tice, H., Copeland, A., Cheng, J.-F., Chen, F., Bruce, D., Goodwin, L., Pitluck, S., Mavromatis, K., Pati, A., Mikhailova, N., Chen, A., Palaniappan, K., Land, M., Hauser, L., Chang, Y.-J., Jeffries, C. D., Detter, J. C., Brettin, T., Rohde, M., Göker, M., Bristow, J., Markowitz, V., Eisen, J. A., Hugenholtz, P., Kyrpides, N. C., y Klenk, H.-P. (2010). Complete genome sequence of Haliangium ochraceum type strain (SMP-2). *Standards in genomic sciences*, 2(1):96–106.

[Jaenicke y Böhm, 1998] Jaenicke, R. y Böhm, G. (1998). The stability of proteins in extreme environments. *Current opinion in structural biology*, 8(6):738–48.

[Jarrous y Gopalan, 2010] Jarrous, N. y Gopalan, V. (2010). Archaeal/eukaryal RNase P: subunits, functions and RNA diversification. *Nucleic acids research*, 38(22):7885–94.

[Johnson *et al.*, 2013] Johnson, S., Chen, Y.-J., y Phillips, R. (2013). Poly(dA:dT)-rich DNAs are highly flexible in the context of DNA looping. *PloS one*, 8(10):e75799.

[Kanehisa y Goto, 2000] Kanehisa, M. y Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

[Kapatai *et al.*, 2006] Kapatai, G., Large, A., Benesch, J. L. P., Robinson, C. V., Carrascosa, J. L., Valpuesta, J. M., Gowrinathan, P., y Lund, P. a. (2006). All three chaperonin genes in the archaeon Haloferax volcanii are individually dispensable. *Molecular microbiology*, 61(6):1583–97.

[Karlin y Mrázek, 1997] Karlin, S. y Mrázek, J. (1997). Compositional differences within and between eukaryotic genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 94(19):10227–32.

[Klein *et al.*, 2002] Klein, R. J., Misulovin, Z., y Eddy, S. R. (2002). Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proceedings of the National Academy of Sciences of the United States of America*, 99(11):7542–7.

[Klenk *et al.*, 2004] Klenk, H.-P., Spitzer, M., Ochsenreiter, T., y Fuellen, G. (2004). Phylogenomics of hyperthermophilic Archaea and Bacteria. *Biochemical Society transactions*, 32(Pt 2):175–8.

[Klipcan *et al.*, 2006] Klipcan, L., Safro, I., Temkin, B., y Safro, M. (2006). Optimal growth temperature of prokaryotes correlates with class II amino acid composition. *FEBS letters*, 580(6):1672–6.

[Konings *et al.*, 2002] Konings, W. N., Albers, S.-V., Koning, S., y Driessen, A. J. M. (2002). The cell membrane plays a crucial role in survival of bacteria and archaea in extreme environments. *Antonie van Leeuwenhoek*, 81(1-4):61–72.

[Krupovic *et al.*, 2013] Krupovic, M., Gonnet, M., Hania, W. B., Forterre, P., y Erauso, G. (2013). Insights into dynamics of mobile genetic elements in hyperthermophilic environments from five new Thermococcus plasmids. *PloS one*, 8(1):e49044.

[Krylov *et al.*, 2003] Krylov, D. M., Nasmyth, K., y Koonin, E. V. (2003). Evolution of eukaryotic cell cycle regulation: stepwise addition of regulatory kinases and late advent of the CDKs. *Current biology : CB*, 13(2):173–7.

[Kumar y Nussinov, 2001] Kumar, S. y Nussinov, R. (2001). How do thermophilic proteins deal with heat? *Cellular and molecular life sciences : CMLS*, 58(9):1216–33.

[Kurz, 2008] Kurz, M. (2008). Compatible solute influence on nucleic acids: Many questions but few answers. *Saline Systems*, 4(6):1–14.

[Lapierre y Gogarten, 2009] Lapierre, P. y Gogarten, J. P. (2009). Estimating the size of the bacterial pan-genome. *Trends in genetics : TIG*, 25(3):107–10.

[Larsson *et al.*, 2009] Larsson, P., Elfsmark, D., Svensson, K., Wikström, P., Forsman, M., Brettin, T., Keim, P., y Johansson, A. (2009). Molecular evolutionary consequences of niche restriction in Francisella tularensis, a facultative intracellular pathogen. *PLoS pathogens*, 5(6):e1000472.

[Lefébure *et al.*, 2012] Lefébure, T., Richards, V. P., Lang, P., Pavinski-Bitar, P., y Stanhope, M. J. (2012). Gene repertoire evolution of Streptococcus pyogenes inferred from phylogenomic analysis with Streptococcus canis and Streptococcus dysgalactiae. *PloS one*, 7(5):e37607.

[Letek *et al.*, 2010] Letek, M., González, P., Macarthur, I., Rodríguez, H., Freeman, T. C., Valero-Rello, A., Blanco, M., Buckley, T., Cherevach, I., Fahey, R., Hapeshi, A., Holdstock, J., Leadon, D., Navas, J., Ocampo, A., Quail, M. A., Sanders, M., Scortti, M. M., Prescott, J. F., Fogarty, U., Meijer, W. G., Parkhill, J., Bentley, S. D., y Vázquez-Boland, J. A. (2010). The genome of a pathogenic rhodococcus: cooptive virulence underpinned by key gene acquisitions. *PLoS genetics*, 6(9).

[Li *et al.*, 2007] Li, W., Zou, H., y Tao, M. (2007). Sequences downstream of the start codon and their relations to G + C content and optimal growth temperature in prokaryotic genomes. *Antonie van Leeuwenhoek*, 92(4):417–27.

[Lightfield *et al.*, 2011] Lightfield, J., Fram, N. R., y Ely, B. (2011). Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. *PloS one*, 6(3):e17677.

[Lynch y Conery, 2003] Lynch, M. y Conery, J. S. (2003). The origins of genome complexity. *Science (New York, N.Y.)*, 302(5649):1401–4.

[Mallick *et al.*, 2002] Mallick, P., Boutz, D. R., Eisenberg, D., y Yeates, T. O. (2002). Genomic evidence that the intracellular proteins of archaeal microbes contain disulfide bonds. *Proceedings of the National Academy of Sciences of the United States of America*, 99(15):9679–84.

[Marathe y Bansal, 2011] Marathe, A. y Bansal, M. (2011). An ensemble of B-DNA dinucleotide geometries lead to characteristic nucleosomal DNA structure and provide plasticity required for gene expression. *BMC structural biology*, 11(1):1.

[Marck y Grosjean, 2002] Marck, C. y Grosjean, H. (2002). tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA (New York, N.Y.)*, 8(10):1189–232.

[Marri *et al.*, 2006] Marri, P. R., Bannantine, J. P., y Golding, G. B. (2006). Comparative genomics of metabolic pathways in Mycobacterium species: gene duplication, gene decay and lateral gene transfer. *FEMS microbiology reviews*, 30(6):906–25.

[Martins *et al.*, 1996] Martins, L. O., Carreto, L. S., Da Costa, M. S., y Santos, H. (1996). New compatible solutes related to Di-myo-inositol-phosphate in members of the order Thermotogales. *Journal of bacteriology*, 178(19):5644–51.

[Martínez-Cano *et al.*, 2015] Martínez-Cano, D. J., Reyes-Prieto, M., Martinez-Romero, E., Partida-Martinez, L. P., Latorre, A., Moya, A., y Delaye, L. (2015). Evolution of small prokaryotic genomes. *Frontiers in Microbiology*, 5:742.

[Mavromatis *et al.*, 2009] Mavromatis, K., Ivanova, N., Anderson, I., Lykidis, A., Hooper, S. D., Sun, H., Kunin, V., Lapidus, A., Hugenholtz, P., Patel, B., y Kyrpides, N. C. (2009). Genome analysis of the anaerobic thermohalophilic bacterium Halothermothrix orenii. *PloS one*, 4(1):e4192.

[McCall *et al.*, 1986] McCall, M., Brown, T., Hunter, W., y Kennard, O. (1986). The crystal structure of d(GGATGGGAG): an essential part of the binding site for transcription factor IIIA. *Nature*, 322(6080):661–4.

[Medrano-Soto *et al.*, 2004] Medrano-Soto, A., Moreno-Hagelsieb, G., Vinuesa, P., Christen, J. A., y Collado-Vides, J. (2004). Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. *Molecular biology and evolution*, 21(10):1884–94.

[Mesbah y Wiegel, 2008] Mesbah, N. M. y Wiegel, J. (2008). Life at extreme limits: the anaerobic halophilic alkalithermophiles. *Annals of the New York Academy of Sciences*, 1125:44–57.

[Miramontes *et al.*, 1995] Miramontes, P., Medrano, L., Cerpa, C., Cedergren, R., Ferbeyre, G., y Cocho, G. (1995). Structural and thermodynamic properties of DNA uncover different evolutionary histories. *Journal of Molecular Evolution*, 40:698–704.

[Mongodin *et al.*, 2005] Mongodin, E. F., Nelson, K. E., Daugherty, S., Deboy, R. T., Wister, J., Khouri, H., Weidman, J., Walsh, D. a., Papke, R. T., Sanchez Perez, G., Sharma, a. K., Nesbø, C. L., MacLeod, D., Bapteste, E., Doolittle, W. F., Charlebois, R. L., Legault, B., y Rodriguez-Valera, F. (2005). The genome of Salinibacter ruber: convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proceedings of the National Academy of Sciences of the United States of America*, 102(50):18147–52.

[Murrell *et al.*, 2005] Murrell, A., Rakyan, V. K., y Beck, S. (2005). From genome to epigenome. *Human molecular genetics*, 14 Spec No(1):R3–R10.

[Muyzer *et al.*, 2011] Muyzer, G., Sorokin, D. Y., Mavromatis, K., Lapidus, A., Clum, A., Ivanova, N., Pati, A., D'Haeseleer, P., Woyke, T., y Kyrpides, N. C. (2011). Complete genome sequence of "Thioalkalivibrio sulfidophilus" HL-EbGr7. *Standards in genomic sciences*, 4(1):23–35.

[Ng y Dickerson, 2002] Ng, H.-L. y Dickerson, R. E. (2002). Mediation of the A/B-DNA helix transition by G-tracts in the crystal structure of duplex CATGGGCCCATG. *Nucleic acids research*, 30(18):4061–7.

[O'Brien y Fraser, 2005] O'Brien, S. J. y Fraser, C. M. (2005). Genomes and evolution: the power of comparative genomics. *Current opinion in genetics & development*, 15(6):569–71.

[Ochman y Jones, 2000] Ochman, H. y Jones, I. B. (2000). Evolutionary dynamics of full genome content in Escherichia coli. *The EMBO journal*, 19(24):6637–43.

[Okonechnikov *et al.*, 2012] Okonechnikov, K., Golosova, O., y Fursov, M. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics (Oxford, England)*, 28(8):1166–7.

[Ollivier *et al.*, 1994] Ollivier, B., Caumette, P., Garcia, J. L., y Mah, R. A. (1994). Anaerobic bacteria from hypersaline environments. *Microbiological reviews*, 58(1):27–38.

[Oren, 2008] Oren, A. (2008). Microbial life at high salt concentrations: phylogenetic and metabolic diversity. *Saline systems*, 4:2.

[Pecoraro *et al.*, 2011] Pecoraro, V., Zerulla, K., Lange, C., y J., S. (2011). Quantification of ploidy in proteobacteria revealed the existence of monoploid, (mero-)oligoploid and polyploid species. *PloS one*, 6:e16392.

[Pei *et al.*, 2009] Pei, A., Nossa, C. W., Chokshi, P., Blaser, M. J., Yang, L., Rosmarin, D. M., y Pei, Z. (2009). Diversity of 23S rRNA genes within individual prokaryotic genomes. *PloS one*, 4(5):e5437.

[Pereira y Reeve, 1998] Pereira, S. L. y Reeve, J. N. (1998). Histones and nucleosomes in Archaea and Eukarya: a comparative analysis. *Extremophiles : life under extreme conditions*, 2(3):141–8.

[Pikuta *et al.*, 2007] Pikuta, E. V., Hoover, R. B., y Tang, J. (2007). Microbial extremophiles at the limits of life. *Critical reviews in microbiology*, 33(3):183–209.

[Popa *et al.*, 2012] Popa, A., Samollow, P., Gautier, C., y Mouchiroud, D. (2012). The sex-specific impact of meiotic recombination on nucleotide composition. *Genome biology and evolution*, 4(3):412–22.

[Quintana *et al.*, 1992] Quintana, J., Grzeskowiak, K., Yanagi, K., y Dickerson, R. (1992). Structure of a B-DNA decamer with a central T-A step: C-G-A-T-T-A-A-T-C-G. *Journal of Molecular Evolution*, 225(1/2):379–95.

[Rhodes y Klug, 1986] Rhodes, D. y Klug, A. (1986). An underlying repeat in some transcriptional control sequences corresponding to half a double helical turn of DNA. *Cell*, 46(1):123–32.

[Roh *et al.*, 2007] Roh, S. W., Nam, Y.-D., Chang, H.-W., Sung, Y., Kim, K.-H., Oh, H.-M., y Bae, J.-W. (2007). Halalkalicoccus jeotgali sp. nov., a halophilic archaeon from shrimp jeotgal, a traditional Korean fermented seafood. *International journal of systematic and evolutionary microbiology*, 57(Pt 10):2296–8.

[Ronimus y Musgrave, 1996] Ronimus, R. S. y Musgrave, D. R. (1996). Purification and characterization of a histone-like protein from the Archaeal isolate AN1, a member of the Thermococcales. *Molecular microbiology*, 20(1):77–86.

[Santos *et al.*, 2002] Santos, H., da Costa, M. S., y Costa, M. S. (2002). Compatible solutes of organisms that live in hot saline environments. *Environmental microbiology*, 4(9):501–9.

[Saunders *et al.*, 2010] Saunders, E., Tindall, B., Fähnrich, R., Lapidus, A., Copeland, A., Glavina Del Rio, T., Lucas, S., Chen, F., Tice, H., Cheng, J., Han, C., Detter, C., y Bruce, D.; Goodwin, L. C. P. P. S. P. A. I. N. M. K. C. A. P. N. (2010). Complete genome sequence of Haloterrigena turkmenica type strain (4kT). *Standards in Genomic Sciences*.

[Schmidt *et al.*, 2003] Schmidt, S., Sunyaev, S., Bork, P., y Dandekar, T. (2003). Metabolites: a helping hand for pathway evolution? *Trends in biochemical sciences*, 28(6):336–41.

[Schofield y Hsieh, 2003] Schofield, M. J. y Hsieh, P. (2003). DNA mismatch repair: molecular mechanisms and biological function. *Annual review of microbiology*, 57:579–608.

[Schuerman y van Meervelt, 2000] Schuerman, G. S. y van Meervelt, L. (2000). Conformational Flexibility of the DNA Backbone. *Journal of the American Chemical Society*, 122(2):232–240.

[Seoighe, 2003] Seoighe, C. (2003). Turning the clock back on ancient genome duplication. *Current opinion in genetics & development*, 13(6):636–43.

[She *et al.*, 2001] She, Q., Singh, R. K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M. J., Chan-Weiher, C. C., Clausen, I. G., Curtis, B. A., De Moors, A., Erauso, G., Fletcher, C., Gordon, P. M., Heikamp-de Jong, I., Jeffries, A. C., Kozera, C. J., Medina, N., Peng, X., Thi-Ngoc, H. P., Redder, P., Schenk, M. E., Theriault, C., Tolstrup, N., Charlebois, R. L., Doolittle, W. F., Duguet, M., Gaasterland, T., Garrett, R. A., Ragan, M. A., Sensen, C. W., y Van der Oost, J. (2001). The complete genome of the crenarchaeon Sulfolobus solfataricus P2. *Proceedings of the National Academy of Sciences of the United States of America*, 98(14):7835–40.

[Shockley *et al.*, 2005] Shockley, K. R., Scott, K. L., Pysz, M. A., Conners, S. B., Johnson, M. R., Montero, C. I., Wolfinger, R. D., y Kelly, R. M. (2005). Genome-wide transcriptional variation within and between steady states for continuous growth of the hyperthermophile Thermotoga Maritima. *Applied and environmental microbiology*, 71(9):5572–6.

[Siddaramappa *et al.*, 2012] Siddaramappa, S., Challacombe, J. F., De Castro, R. E., Pfeiffer, F., Sastre, D. E., Giménez, M. I., Paggi, R. A., Detter, J. C., Davenport, K. W., Goodwin, L. A., Kyrpides, N., Tapia, R., Pitluck, S., Lucas, S., Woyke, T., y Maupin-Furlow, J. A. (2012). A comparative genomics perspective on the genetic content of the alkaliphilic haloarchaeon Natrialba magadii ATCC 43099T. *BMC genomics*, 13(1):165.

[Sikorski *et al.*, 2010] Sikorski, J., Lapidus, A., Chertkov, O., Lucas, S., Copeland, A., Glavina Del Rio, T., Nolan, M.and Tice, H., Cheng, J.-F., Han, C., Brambilla, E., Pitluck, S., Liolios, K., Ivanova, N., Mavromatis, K., Mikhailova, N., Pati, A., Bruce, D., Detter, C., y Tapia, H.-P. (2010). Complete genome sequence of Acetohalobium arabaticum type strain (Z-7288T). *Standards in Genomic Sciences*, 3(1):57–65.

[Singer y Hickey, 2003] Singer, G. A. C. y Hickey, D. A. (2003). Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene*, 317(1-2):39–47.

[Soppa *et al.*, 2008] Soppa, J., Baumann, a., Brenneis, M., Dambeck, M., Hering, O., y Lange, C. (2008). Genomics and functional genomics with haloarchaea. *Archives of microbiology*, 190(3):197–215.

[Sorokin y Kuenen, 2005] Sorokin, D. Y. y Kuenen, J. G. (2005). Chemolithotrophic haloalkaliphiles from soda lakes. *FEMS microbiology ecology*, 52(3):287–95.

[Sterner y Liebl, 2001] Sterner, R. y Liebl, W. (2001). Thermophilic adaptation of proteins. *Critical reviews in biochemistry and molecular biology*, 36(1):39–106.

[Stetter, 1996] Stetter, K. (1996). Hyperthermophilic prokaryotes. *FEMS Microbiology Reviews*, 18(2-3):149–158.

[Stetter, 2006] Stetter, K. O. (2006). History of discovery of the first hyperthermophiles. *Extremophiles : life under extreme conditions*, 10(5):357–362.

[Su *et al.*, 2013] Su, A. A. H., Tripp, V., y Randau, L. (2013). RNA-Seq analyses reveal the order of tRNA processing events and the maturation of C/D box and CRISPR RNAs in the hyperthermophile Methanopyrus kandleri. *Nucleic Acids Research*, 41(12):6250–6258.

[Sved y Bird, 1990] Sved, J. y Bird, A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proceedings of the National Academy of Sciences of the United States of America*, 87(12):4692–6.

[Takami *et al.*, 2000] Takami, H., Nakasone, K., Takaki, Y., Maeno, G., Sasaki, R., Masui, N., Fuji, F., Hirama, C., Nakamura, Y., Ogasawara, N., Kuhara, S., y Horikoshi, K. (2000). Complete genome sequence of the alkaliphilic bacterium Bacillus halodurans and genomic sequence comparidon with Bacillus subtilis. *Nucleic Acids Research*, 28(21):4317–4331.

[Tekaia y Yeramian, 2006] Tekaia, F. y Yeramian, E. (2006). Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. *BMC genomics*, 7:307.

[Tereshko *et al.*, 1999] Tereshko, V., Minasov, G., y Egli, M. (1999). The Dickerson-Drew B-DNA Dodecamer Revisited at Atomic Resolution. *Communications*, 121:470–471.

[Tettelin *et al.*, 2008] Tettelin, H., Riley, D., Cattuto, C., y Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology*, 11(5):472–477.

[Trent, 2000] Trent, J. (2000). Extremophiles in astrobiology: per Ardua ad Astra. *Gravitational and space biology bulletin : publication of the American Society for Gravitational and Space Biology*, 13(2):5–11.

[Tsuihiji *et al.*, 2006] Tsuihiji, H., Yamazaki, Y., Kamikubo, H., Imamoto, Y., y Kataoka, M. (2006). Cloning and characterization of nif structural and regulatory genes in the purple sulfur bacterium, Halorhodospira halophila. *Journal of bioscience and bioengineering*, 101(3):263–70.

[Van der Linden y de Farias, 2006] Van der Linden, M. G. y de Farias, S. T. (2006). Correlation between codon usage and thermostability. *Extremophiles : life under extreme conditions*, 10(5):479–81.

[van der Oost *et al.*, 2014] van der Oost, J., Westra, E. R., Jackson, R. N., y Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nature reviews. Microbiology*, 12(7):479–92.

[Vieille y Zeikus, 2001] Vieille, C. y Zeikus, G. J. (2001). Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiology and molecular biology reviews : MMBR*, 65(1):1–43.

[Weckwerth, 2010] Weckwerth, W. (2010). Metabolomics: an integral technique in systems biology. *Bioanalysis*, 2(4):829–36.

[Wiezer y Merkl, 2005] Wiezer, A. y Merkl, R. (2005). A comparative categorization of gene flux in diverse microbial species. *Genomics*, 86(4):462–75.

[Wojciechowski *et al.*, 2013] Wojciechowski, M., Czapinska, H., y Bochtler, M. (2013). CpG underrepresentation and the bacterial CpG-specific DNA methyltransferase M.MpeI. *Proceedings of the National Academy of Sciences of the United States of America*, 110(1):105–10.

[Yavartanoo y Choi, 2013] Yavartanoo, M. y Choi, J. K. (2013). ENCODE: A Sourcebook of Epigenomes and Chromatin Language. *Genomics & informatics*, 11(1):2–6.

[Zeldovich *et al.*, 2007] Zeldovich, K., Berezovsky, I., y Shakhnovich, E. (2007). Protein and DNA sequence determinants of thermophilic adaptation. *PLoS computational biology*, 3(1):e5.

[Zeng *et al.*, 2006] Zeng, C., Zhu, J.-C., Liu, Y., Yang, Y., Zhu, J.-Y., Huang, Y.-P., y Shen, P. (2006). Investigation of the influence of NaCl concentration on Halobacterium salinarum growth. *Journal of Thermal Analysis and Calorimetry*, 84(3):6.

[Zhaxybayeva *et al.*, 2009] Zhaxybayeva, O., Swithers, Kristen S, a. L. P., Fournier, G. P., Bickhart, D. M., DeBoy, R. T., Nelson, K. E., Nesbø, C. L., Doolittle, W. F., Gogarten, J. P., y Noll, K. (2009). On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14):5865–70.

[Zhilina y Zavarzin, 1987] Zhilina, T. N. y Zavarzin, G. (1987). Methanohalobium evestigatus, n. gen., n. sp., the extremely halophilic methanogenic Archaebacterium. *Dokl. Akad. Nauk USSR*, 293:464–468.

[Zhu *et al.*, 2013] Zhu, Y., Lin, Z., y Nakhleh, L. (2013). Evolution after whole-genome duplication: a network perspective. *G3 (Bethesda, Md.)*, 3(11):2049–57.

[Zuckerkand y Pauling, 1965] Zuckerkand, E. y Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, 8(2):357–366.