

ANÁLISIS BIOINFORMÁTICO DE SECUENCIAS NO CODIFICANTES EN METAGENOMAS: UNA REVISIÓN ACTUALIZADA

César Iván Cháves Garreta

*Universidad del Valle, Apartado Aereo 25360, Cali, Colombia.
correo electrónico: cesar.chaves@correounivalle.edu.com*

Pedro Antonio Moreno Tobar

*Universidad del Valle, Apartado Aereo 25360, Cali, Colombia.
correo electrónico: pedro.moreno@correounivalle.edu.com*

RESUMEN

Antecedentes. La metagenómica se apoya en la bioinformática para describir y clasificar un sin número de microorganismos. Muchos de los enfoques en esta área, centran su atención en las características funcionales derivadas de secuencias codificantes, descartando las secuencias no codificantes (*NCS*), aun conociendo su participación en la maquinaria regulatoria de los genes. Recientemente, un estudio desarrollado en la Universidad del Valle utiliza las *NCS* para caracterizar metagenomas y desarrollar un marco de trabajo bioinformático para el estudio de los microorganismos asociados a medio ambientes (*Ev* (Ambiental), *Ha* (Asociado a hospederos) y *Eg* (Diseñados)) (Tobar-Tosse *et al.* 2013).

Métodos. Con el objetivo de validar y verificar estos resultados, se propuso llevar a cabo un estudio más extenso, utilizando un mayor número de secuencias de metagenomas (90) y bajo la misma metodología. Aquí se estudiaron y determinaron la composición estructural y funcional de las secuencias, su clasificación en tres tipos de ecosistemas (*Ev*, *Ha* y *Eg*), y los perfiles estructurales y funcionales obtenidos.

Resultados. Inicialmente se encontró que los resultados y el enfoque metodológico propuesto por estos autores continúan revelando un comportamiento similar o parecido al estudio bajo referencia. También, se halló, una relación directa entre la cantidad de *NCS* y la composición particular de algunos metagenomas, destacando el potencial informativo de algunos patrones presentes en las *NCS* y las abundantes secuencias repetidas de trinucleótidos (*TRS*) encontrados en estas secuencias. Asimismo, se confirma la utilidad de los trinucleótidos para encontrar asociaciones entre los metagenomas.

Conclusión. Se comprobó que efectivamente, la información contenida en las *NCS* puede ayudar a clasificar metagenomas y describir sus características funcionales o estructurales dependientes del entorno, permitiendo validar gran parte del marco de trabajo bioinformático propuesto por Tobar-Tosse *et al.* (2013).

Palabras clave: bioinformática, Engineered, Environmental, Host-associated, metagenómica, *NCS*, trinucleótidos.

ABSTRACT

Background. The metagenomics is supports on bioinformatics to describe and classify a number of microorganisms. Many of the approaches in this area focus their attention on the functional characteristics derived from coding sequences, discarding non-coding sequences (*NCS*) even knowing their participation in the regulatory machinery of many genes. Recently, a study developed in Universidad Del Valle uses the *NCS* to characterize metagenomes and develop a bioinformatic framework for the study of microorganisms associated with environments (*Ev* (Environmental), *Ha* (Host-associated) and *Eg* (Engineered)) (Tobar-Tosse *et al.* 2013).

Methods. In order to validate and verify these results, it was proposed to carry out a more extensive study, using a greater number of sequences of metagenomas (90) and under the same methodology. Here we studied and determined the structural and functional composition of the sequences, their classification into three types of ecosystems (*Ev*, *Ha* and *Eg*), and the structural and functional profiles obtained.

Results. Initially it was found that the results and the methodological approach proposed by these authors continue to reveal a similar or similar behavior to the study under reference. Also, a direct relationship between the amount of *NCS* and the particular composition of some metagenomes was found, highlighting the informative potential of some patterns present in the *NCS* and the abundant trinucleotid repeated sequences (*TRS*) found in these sequences. Also, the utility of trinucleotides to find associations between metagenomes is confirmed.

Conclusion. It was verified that indeed, the information contained in *NCS* can help to classify metagenomas and describe their functional or structural characteristics depending on the environment, allowing to validate much of the bioinformatic framework proposed by Tobar-Tosse *et al.* (2013).

INTRODUCCIÓN

La metagenómica actualmente es una herramienta imprescindible para el estudio de microorganismos que con dificultad se pueden cultivar en el laboratorio (Stewart 2012, Vavourakis *et al.* 2016). Es un campo de investigación en crecimiento (Simon and Daniel 2011), donde se involucra el tratamiento del material genético que se extrae directamente de muestras ambientales (Gupta *et al.* 2016, Thomas *et al.* 2012). Se sabe que el 99 % del genoma microbiano no es posible cultivarse y que gran parte de las especies habitan ambientes mixtos y complejos (Jung *et al.* 2016, Tringe and Rubin 2005). Es por esta razón, que muchos estudios en este campo centran su atención en la estructura, función y dinámica de las comunidades microbianas (Fierer *et al.* 2012), además, permiten el descubrimiento de nuevos microorganismos, comprender la diversidad microbiana (Ventosa *et al.* 2015, Willner *et al.* 2011), y establecer hipótesis acerca de las interacciones que ocurren entre los miembros de la comunidad microbiana.

La metagenómica de la mano de la bioinformática han desarrollado un sin número de herramientas computacionales para el análisis de metagenomas, todas con el objetivo de descubrir nuevos conceptos biológicos en áreas del conocimiento microbiano (Kunin *et al.* 2008). Los avances en las tecnologías de secuenciación de última generación (NGS) han permitido extraer valiosa información metagenómica de hábitats cada vez más complejos, pero que a su vez han producido grandes cantidades de datos a un ritmo sin precedentes (Hiraoka *et al.* 2016). Las herramientas bioinformáticas de análisis de secuencias metagenómicas, en su mayoría utilizan sólo secuencias codificantes y son basadas en la predicción de marcos abiertos de lectura (ORF) (Teeling and Glockner 2012, Li 2009), haciendo a un lado las secuencias no codificantes (NCS) (Tobar-Tosse *et al.* 2013), lo que es preocupante si se sabe que estas secuencias, tanto en organismos procariotas y eucariotas unicelulares, pueden contener elementos reguladores de la expresión de los genes (Zou *et al.* 2014). De esta forma, si estas secuencias se desechan, se estaría perdiendo información reguladora o estructural importante, considerando que la proporción de NCS en procariotas corresponde al 18 % según Rogozin *et al.* (2002).

Estudios como el de Tobar-Tosse *et al.* (2013) han descubierto evidencia de la alta proporción de NCS que se descartan en los análisis de metagenomas, y la información relevante que estas secuencias podrían presentar. Estos autores encuentran una alta proporción de NCS con repeticiones de trinucleótidos (Tn) que tienen funciones regulatorias y adaptativas para las comunidades de metagenomas. Así mismo, descubren

asociaciones entre los valores de uso de Tn y la función de los genes, donde el agrupamiento de metagenomas está correlacionado con las adaptaciones de los microorganismos y tipos de metagenomas. De este modo proponen a las NCS como un recurso importante para la descripción de metagenomas, idea en la que se fundamenta este nuevo trabajo.

Los análisis para la búsqueda de información en genomas y metagenomas pueden tomar un enfoque apoyado en métodos basados en la secuencia o como la gran mayoría, en métodos basados en similitud de secuencias. Ej. BLAST (Culligan *et al.* 2014, Rai and Maiti 2015, Taft *et al.* 2007, Teeling and Glockner, 2012, Tobar-Tosse *et al.* 2013). Los métodos basados en la secuencia se interesan en definir patrones como el contenido de guanina y citosina (G+C), uso de codones (Cd) y frecuencias de di, tri, tetra o pentanucleótidos (Deschavanne *et al.* 1999 tomado de Tobar-Tosse *et al.* 2013), destacando el uso de Cd y tetranucleótidos para secuencias codificantes (Teeling *et al.* 2004) y uso de Tn para NCS (Tobar-Tosse *et al.* 2013), ofreciendo la posibilidad de identificar ciertas regularidades presentes entre los microorganismos.

De esta forma, esta investigación se sustenta en el análisis de metagenomas completos usando métodos basados en secuencias y similitud de secuencias, con el fin de examinar su composición y organización biológica. En este trabajo las secuencias codificantes y NCS de 90 metagenomas de diversos ecosistemas, se usaron para crear comparaciones de perfiles estructurales y funcionales, que confirmaron como efectivamente la información contenida en las NCS puede ayudar a clasificar metagenomas y describir características funcionales o estructurales dependientes del entorno, y además validaron en gran medida el marco de trabajo bioinformático propuesto Tobar-Tosse *et al.* (2013).

MÉTODOS

El desarrollo de este estudio está apoyado en los métodos de trabajo realizados por Tobar-Tosse *et al.* (2013) con algunas modificaciones, dichos métodos constan de 5 pasos:

1. Obtención del conjunto de datos metagenómicos y limpieza de secuencias.

Para este estudio se descargaron 90 metagenomas de la base de datos de *CAMERA project* (Seshadri *et al.* 2007), los cuales se clasificaron según el tipo de ecosistema, filogenia y función (Ivanova *et al.* 2010). Cada metagenoma se compone de secuencias de tipo “reads” sin ensamblar, las cuales fueron depuradas

con “scripts” basados en lenguaje *Python*, con el fin de eliminar nucleótidos desconocidos (X) y no especificados (N).

2. Identificación de secuencias no codificantes y aplicación de métodos basados en la secuencia.

Las secuencias codificantes y no codificantes fueron determinadas por predicciones de ORF, utilizando el algoritmo de *MetaGeneMark* (Zhu *et al.* 2010). Para cada metagenoma se definieron 3 conjuntos de datos que contienen las secuencias codificantes (predicción por ORF), secuencias completas (codificante y no codificante) y secuencias no codificantes (regiones por fuera de los ORF). Para los métodos basados en secuencias se definieron y analizaron tres patrones de secuencias: contenido G+C, Uso de Cd y Uso de Tn (Porceddu and Camiolo 2011). Cada uno de estos patrones se evaluaron en secuencias codificantes y secuencias completas, obteniendo patrones estructurales definidos así: i) Valores de Tn (Secuencias completas) o Cd (Secuencias codificantes) basados en la longitud $Tn(ls)$ o $Cd(ls)$ respectivamente, estos fueron definidos por la suma de las frecuencias del uso de Tn o frecuencias del uso de Cd, sobre la longitud de la secuencia (l): “ $Tn(ls)=(Tn)/l$ ” o “ $Cd(ls)=(Cd)/l$ ”, y ii) Valores de Tn o Cd sobre el número de Tn o Cd por secuencia, $Tn(ts)$ o $Cd(ts)$, respectivamente. Que se definen como la suma de las frecuencias del uso de Tn o frecuencias del uso de Cd, sobre el número de Tn o Cd ($n(Tn)$, $n(Cd)$): “ $Tn(ts)=(Tn)/n(Tn)$ ” ó “ $Cd(ts)=(Cd)/n(Cd)$ ”. Al final, estos valores se organizaron en una tabla comparativa llamada “Perfiles Estructurales”.

3. Determinación de asignaciones funcionales y aplicación de métodos basados en similitud de secuencias.

Los péptidos predichos por ORF a partir de las secuencias, se asignaron a una característica funcional utilizando BLASTP (Altschul 1997) (BLAST+ 2.2.30 release) como propone Li *et al.* (2008), usando un *e-value* de $\leq 1e/30$, e identidad de $\geq 95\%$. La base de datos local usada fue *Pfam-A* (Finn *et al.* 2014). Los resultados se agruparon en una tabla llamada “Perfiles Funcionales”, que contiene los valores de los modelos *Pfam*, definidos como la frecuencia de secuencias asignadas para cada modelo por metagenoma: “ $(Pfam)=(Pfam_{nq})/N(Pfam)$ ”. Donde (*Pfam*) es la frecuencia del modelo *Pfam* en cada metagenoma; (*Pfam_{nq}*) son el número de asignaciones encontradas por BLAST para el modelo, y *N(Pfam)* es el número total de modelos *Pfam* con asociaciones en el metagenoma. Adicionalmente, se aplicó otra estrategia, la cual está relacionada con una búsqueda BLAST de NCS en genomas bacterianos completos, con el fin de encontrar asociación para cualquier

función o taxonomía anotada.

4. Creación de perfiles funcionales y estructurales.

Se crearon cuatro perfiles estructurales, dos basados en secuencias codificantes ($Cd(ls)$, $Cd(ts)$), y dos basados en secuencias completas ($Tn(ls)$, $Tn(ts)$), además de un perfil funcional basado en las asociaciones funcionales *Pfam*. Estos perfiles se agruparon en tablas, que compararon los 90 metagenomas. Dichos perfiles funcionales y estructurales fueron analizados mediante árboles de agrupamiento jerárquico, usando Hierarchical Cluster Explore tool (Shneiderman 2002).

5. Creación de un marco de trabajo bioinformático.

Para cada metagenoma en cada clasificación ecosistémica, se calcularon y promediaron diez parámetros; tamaño, contenido de G+C, asignaciones funcionales, $Tn(ls)$, $Tn(ts)$, CGC, CCG, TTT, AAA, y porcentaje de NCS. Para algunos de estos parámetros se realizará una regresión lineal con el fin de calcular los coeficientes de correlación.

RESULTADOS

En esta investigación se utilizaron 90 metagenomas de diferentes tipos de ecosistema, a los cuales se les aplicaron cuatro métodos para el análisis de secuencias no codificantes propuestos por Tobar-Tosse *et al.* (2013).

Conjunto de datos metagenómicos y clasificación.

Los metagenomas obtenidos de la base de datos de *CAMERA Project* (Seshadri, 2007), fueron clasificados en tres grandes grupos ecosistémicos (*Environmental* (Ev), *Host-associated* (Ha) y *Engineered* (Eg)), tal como lo propone Ivanova *et al.* (2010). Para obtener una mejor comprensión de la biología del metagenoma, el grupo *Environmental* (Ev) se subdividió por categorías ecosistémicas (*Air* (Ai), *Aqua* (Aq) y *Terrestrial* (T)) y estos a su vez por el tipo de ecosistema al que pertenecen (*Freshwater* (f), *Hydrothermal* (h), *Marine* (m) y *soild* (s)). Además, todos los metagenomas fueron marcados con el tipo de comunidad (*bacteria* (b), *viruses* (v), *eukarya* (e) o *archaea* (a)) al que pertenecen las secuencias que contienen. Por citar un ejemplo, *Marine Metagenome from Line Islands* (MLI) es del grupo ecosistémico Ev, categoría ecosistémica Aq, tipo de ecosistema m y tipo de comunidad b, abreviándose EvAqmb_MLI para su fácil identificación. De esta manera, en la Tabla 1(anexo 1) se presenta la clasificación general de los datos metagenómicos, cada uno con una abreviación de su nombre, proyecto del que fue tomado, número de secuencias contenidas, y tamaño en Mpb. En total se analizaron 66.528.865 secuencias y 15.419 Mpb entre

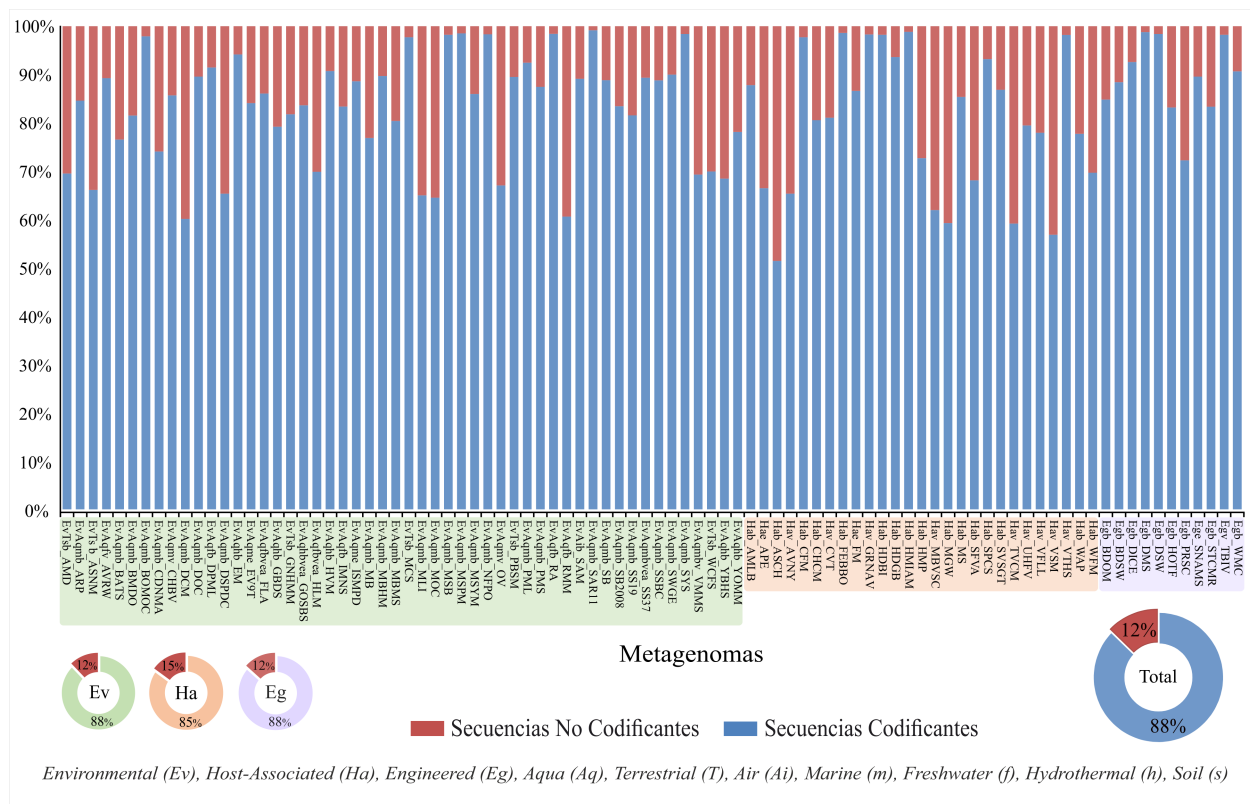


Figura 1. Proporción de secuencias codificantes y no codificantes basado en predicción ORF. Las barras azules representan el porcentaje de secuencias codificantes y las rojas el porcentaje de NCS en cada metagenoma. Los anillos contienen los porcentajes totales y para cada ecosistema de los dos tipos de secuencias.

todos los metagenomas.

Igualmente, la proporción de secuencias codificantes y no codificantes fueron determinadas para cada metagenoma (Fig. 1), encontrando en general para todos los metagenomas una baja proporción de NCS (12%) y una muy alta de secuencias codificantes (88%). En cuanto a los grupos, la más alta proporción

de NCS se encuentra en Ha con 15%, seguido de Ev (12%) y Eg (12%), que son comparables con los resultados presentados por Tobar-Tosse *et al.* (2013), relacionando así, el alto porcentaje de NCS como característica descriptiva del ecosistema Ha. Ej. Los metagenomas Hav_VSM, Hav_TVCM, Hab_ASCH o Hab_MGW.

Caracterización de metagenomas por métodos basados en la secuencia.

Los análisis utilizando métodos basados en la secuencia permitieron caracterizar y organizar las secuencias contenidas en cada metagenoma, evidenciando patrones que revelan características exclusivas para estos. El análisis de contenido de G+C mostro picos específicos (picos altos) y no específicos (picos bajos) para todos los metagenomas, distribuyéndose en un rango de 27-70% (Fig. 2A, línea punteada). Igualmente, se presentaron varios picos específicos para las secuencias codificantes (picos azules) en 56, 55, 52, 51, 50 y 49%. Por el contrario, las secuencias completas (picos rojos) solo presentan un pico específico en 52%. Destacando que estas diferencias entre las secuencias presentaron

un patrón similar al reportado por Tobar-Tosse *et al.* (2013). Por otra parte, se analizó el contenido de Cd (secuencias codificantes) y Tn (secuencias completas) en cada uno de los metagenomas. La Fig. 2B muestra la distribución radial del uso de Cd y Tn, evidenciando las diferencias y semejanzas entre los tipos de secuencias. Se encontraron tendencias en los patrones que relatan un uso similar (asterisco rojo) para GGC, GCG y GCC y dispar (asterisco verde) para CGC, CGG CCG, AAA y TTT. Estas dos tendencias están altamente correlacionadas con el trabajo de Tobar-Tosse *et al.* (2013) y dejan en evidencia la alta cantidad de secuencias de Tn presentes en las NCS de los metagenomas.

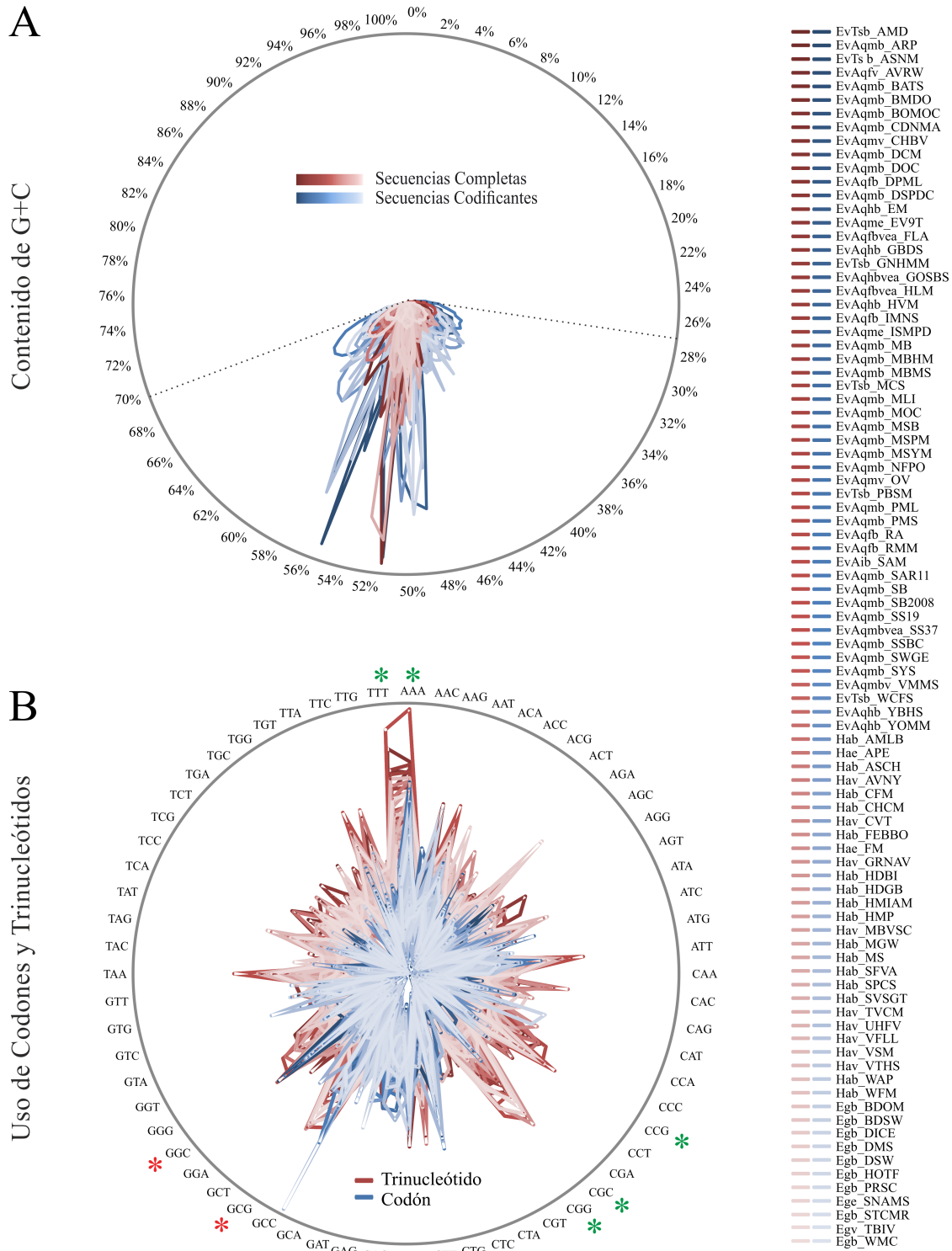


Figura 2. Patrones de secuencias analizados en cada metagenoma para secuencias codificantes (azul) y completas (rojo) A. Distribución del contenido de (G+C) B. Uso de codón y trinucleótidos.

Caracterización de metagenomas por métodos basados en similitud de secuencias.

Los métodos basados en similitud de secuencias permitieron comparar las características funcionales

de los metagenomas. El alineamiento de las secuencias codificantes con *Pfam* (Finn *et al.* 2014) identificó funciones importantes presentes en los metagenomas (“perfil funcional”). Sin embargo, la cantidad de

asignaciones *Pfam* encontradas por metagenoma fue muy baja (Fig. 3). De igual manera, la Fig. 4 contiene el análisis de agrupamiento jerárquico para el perfil funcional, donde se evidencia agrupaciones de metagenomas con conjuntos de funciones similares. Aquí se presenta claramente como los metagenomas formaron 6 subgrupos, donde característicamente, los subgrupos 1-5 están conformados en alta proporción por metagenomas de *Ev* (80 %), y el subgrupo 6 en su mayoría por metagenomas de *Ha* (58 %), permitiendo distinguir dos grandes grupos. Asimismo, para el grupo 1 se evidencian asociaciones de metagenomas en las categorías ecosistémicas, ejemplificado en los metagenomas EvAqmb_PMS, EvAqmb_SSBC, EvAqmb_NFPO y EvAqmb_ARP, que pertenecen a Aq. Igualmente, asociaciones entre tipos de ecosistema como EvAqfb_RA y EvAqfv_AVRW que pertenecen a f, o EvAqmb_SS19 y EvAqmbvea_SS37 relacionados con m. Por otro lado, el grupo 2 presenta asociaciones que hacen referencia al mismo tipo de comunidad, como en Hav_UHFV y Hav_GRNAV relacionados con

v, ó Hab_CFM y Hab_HDGB con b. En cuanto a los metagenomas del ecosistema *Eg*, estos no formaron grupos entre ellos, encontrándose asociados siempre a *Ev* o *Ha*. De igual forma, algunos subgrupos están compuestos por combinaciones de *Ev*, *Ha* y *Eg*, prueba de la existencia de funciones comunes entre los metagenomas, propuesta presentada por Tobar-Tosse *et al.* (2013), y que igualmente permitió seleccionar dichas funciones e identificar las más conservadas (Fig. 5). Aquí se destacan funciones relacionadas con reacciones fotosintéticas, factores de elongación en la síntesis de proteínas o producción de toxinas. Por otra parte, las *NCS* presentes en cada metagenoma fueron asociadas a 5.594 genomas bacterianos completos reportados en la base de datos del NCBI. La Fig.6 exhibe el porcentaje de *NCS* encontradas en las 30 clases taxonómicas más representativas descubiertas, destacando a las clases Gammaproteobacteria, Bacilli, Alphaproteobacteria, y Betaproteobacteria por su alta presencia de *NCS*.

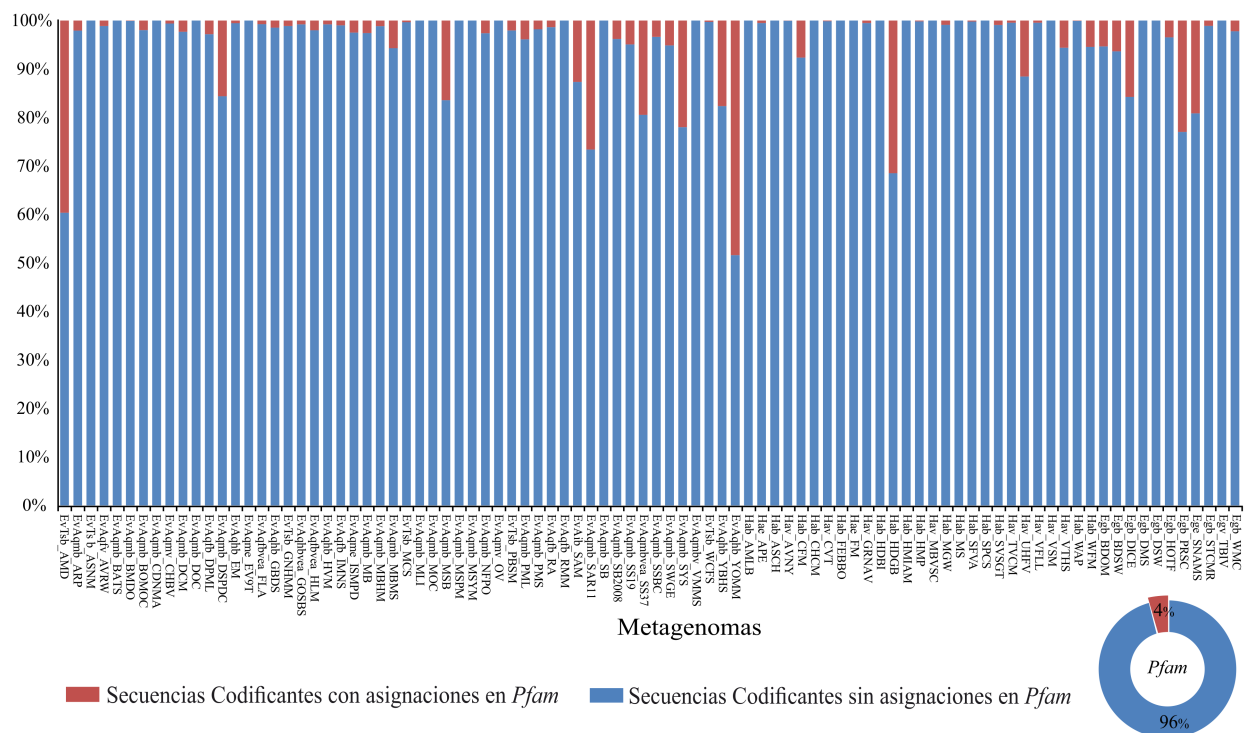


Figura 3. Proporción de secuencias codificantes con asignaciones en *Pfam* para cada metagenoma.

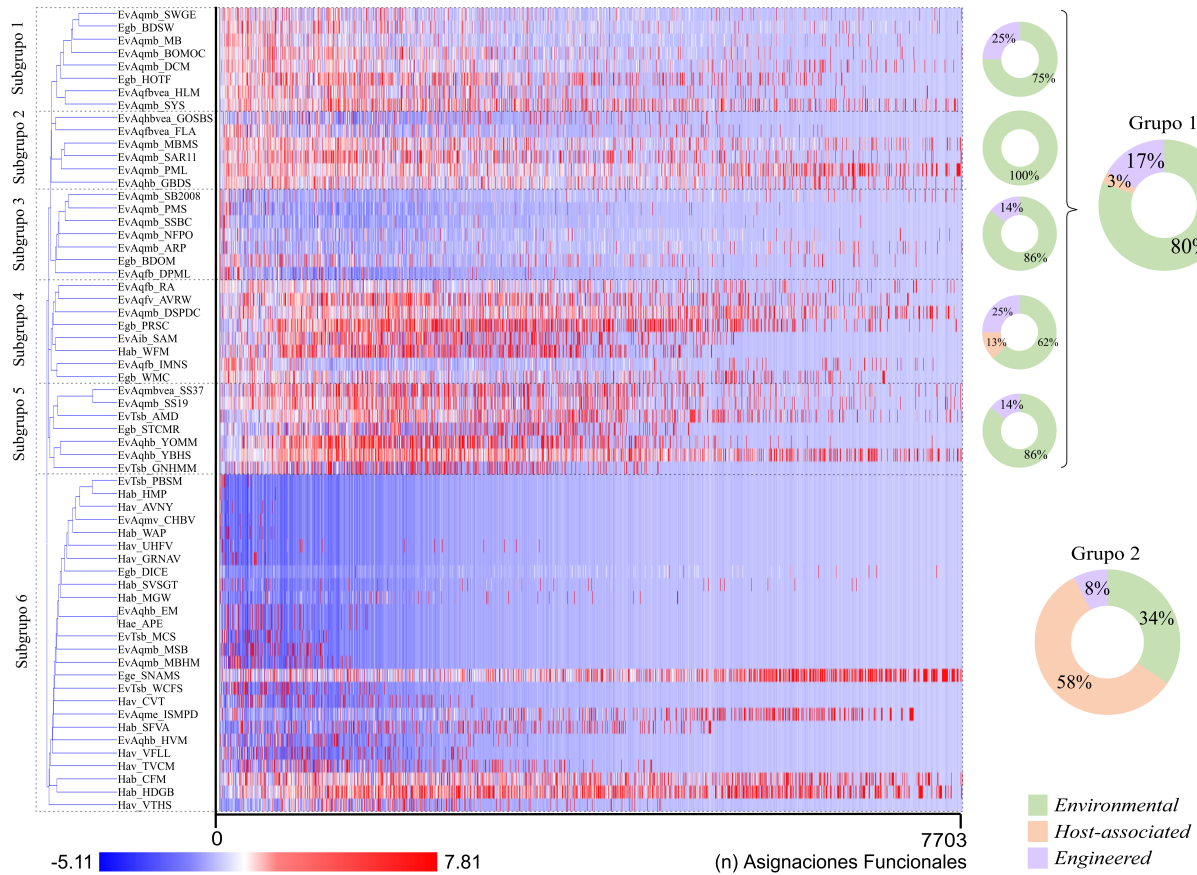


Figura 4. Análisis funcional de las secuencias codificantes para los metagenomas. Agrupamiento jerárquico de metagenomas basado en las asignaciones funcionales encontradas en *Pfam*. Las barras de color representan la frecuencia de la categoría funcional de bajo (azul) a alto (rojo). Los anillos representan el porcentaje de metagenomas de cada ecosistema para cada subgrupo en el agrupamiento.

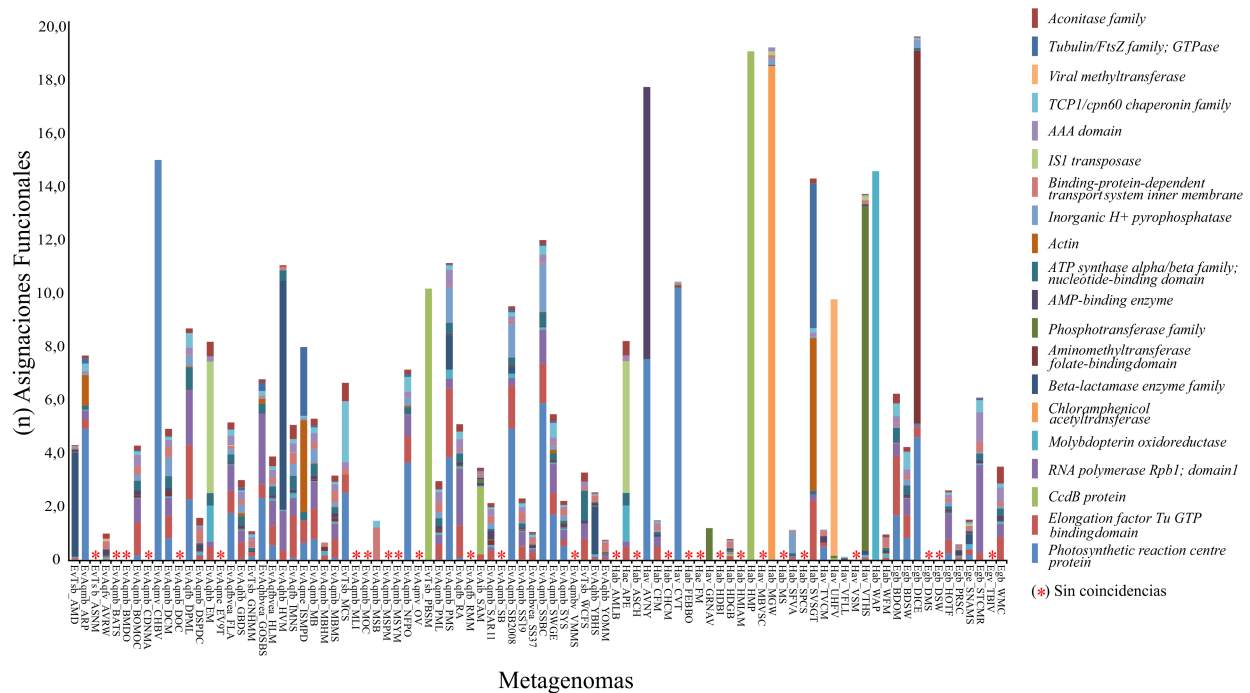


Figura 5. Asignaciones funcionales comunes entre los metagenomas. El tamaño de las barras representan el porcentaje de las categorías funcionales y el color indica el tipo de categoría.

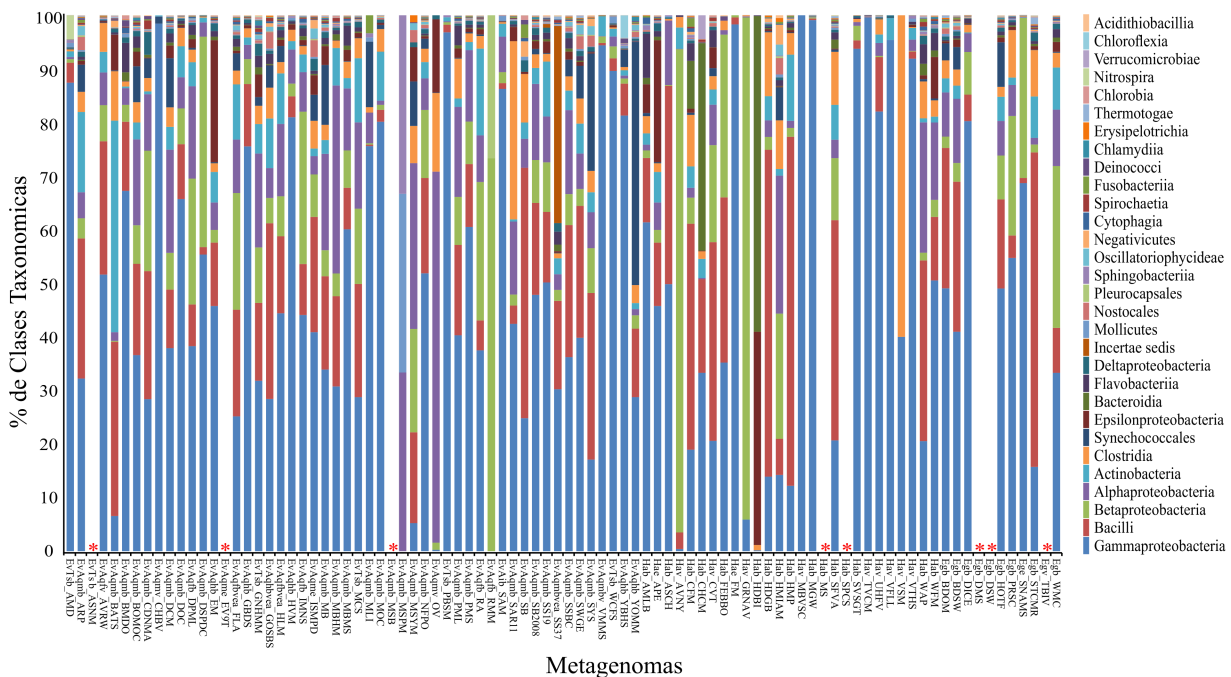


Figura 6. Proporción de *NCS* encontradas en genomas bacterianos completos para cada metagenoma. Los colores indican las clases taxonómicas encontradas y el tamaño de las barras el porcentaje. Los asteriscos rojos indican los metagenomas sin registros.

Perfiles Funcionales y Estructurales.

Las estimaciones de patrones estructurales de uso de Cd y Tn (*Cd(ls)*, *Cd(ts)*, *Tn(ls)*, y *Tn(ts)*), determinaron cuatro perfiles estructurales, al igual que las asignaciones de *Pfam* definieron un perfil funcional. Estos cinco perfiles fueron comparados por medio de árboles de agrupamiento jerárquico (Fig. 7). Existe un porcentaje de secuencias analizadas para cada árbol que está sujeto al método utilizado (secuencia o similitud); Los perfiles estructurales *Cd(ls)* y *Cd(ts)* tienen un 85 %, *Tn(ls)* y *Tn(ts)* tienen 100 %, y el perfil funcional 4 %. Las agrupaciones

de metagenomas encontradas en el árbol del perfil funcional, se compararon con los árboles de los perfiles estructurales (líneas punteadas). Aquí se revelan semejanzas (franjas sombreadas) en la conservación de los grupos ecosistémicos entre los perfiles estructurales y funcional, indicando la relevancia que presentan los patrones estructurales como los Tn en la clasificación de los metagenomas, concordando con una de las teorías importantes propuestas por Tobar-Tosse *et al.* (2013).

DISCUSIÓN

Evidencia de proporciones importantes de NCS en metagenomas.

Desafortunadamente en los estudios metagenómicos las NCS son muy poco usadas y se desechan (Tobar-Tosse *et al.* 2013). Estas pueden contener información importante, que puede estar involucrada en funciones regulatorias de la traducción como en el caso de los Riboswitch (Schrader *et al.* 2014) o estructurales como las secuencias T-box (Fu *et al.* 2013, Lassalle 2015). En este trabajo se descubre la proporción de NCS presente en 90 metagenomas de diferentes tipos de ecosistemas (Fig. 1). En general el 12 % de NCS encontradas, es casi la mitad de los valores reportados en otros estudios similares (20 %) como el de Tobar-Tosse *et al.* (2013). Esto puede estar relacionado con el gran tamaño de algunos metagenomas de *Eg* y *Ev*, además del elevado número de representantes de este último (Tabla 1), así como la corta longitud de algunas secuencias contenidas en varios metagenomas. Sin embargo, los reportes de NCS en procariotas muestran un porcentaje bajo (18 %) (Rogozin *et al.* 2002), que puede ser comparable con el que se presenta en este trabajo.

Relación directa entre la cantidad de NCS y la composición particular de metagenomas.

En los ecosistemas, *Ha* obtuvo el mayor porcentaje de NCS encontradas (15 %), sobrepasando el promedio general de 12 % hallado para todos los metagenomas, confirmando uno de los postulados propuestos por Tobar-Tosse *et al.* (2013), que relaciona el contenido elevado de NCS con metagenomas de tipo *Ha*. De igual manera, algunos metagenomas con comunidades de virus, como *Hav_VSM* o *Hav_TVCM* con contenidos altos de NCS por encima de 30 % (Fig. 1) permiten suponer la presencia de organismos con alta producción de *ncRNAs* (Tycowski *et al.* 2015). Además, se sabe que el genoma viral es de tamaño limitado, por lo tanto, estas NCS podrían ejercer papeles muy importantes en el ciclo del desarrollo del virus o en la infección viral (Alam *et al.* 2014, Tycowski *et al.*

2015, Miller *et al.* 2016). Asimismo, en metagenomas *Ha* de tipo bacteriano como *Hab_ASCH* ó *Hab_MGW* con 45 % y 42 % de NCS respectivamente, estas secuencias, podrían contener maquinaria regulatoria de la producción de antibióticos (Moody *et al.* 2013) o regiones intergénicas largas y especializadas como repeticiones CRISPR o pseudogenes (Koonin and Wolf 2008). Al final, estas inferencias permiten confirmar la relación directa que existe entre la cantidad de NCS y la composición particular de algunos metagenomas.

La variación en el contenido de G+C refleja la importancia que pueden tener las NCS en estudios de caracterización metagenómica.

En procariotas la composición nucleotídica está determinada por la filogenia y el ecosistema (Hildebrand *et al.* 2010). Igualmente, la variación en el contenido de G+C es mediada por el ambiente (Reichenberger *et al.* 2015) y por diferencias entre los patrones de mutación (Lassalle 2015). Esta variación se ve reflejada en los valores de contenido de G+C presentados en este trabajo (Fig. 2A), coincidiendo con los reportes de Tobar-Tosse *et al.* (2013). De igual manera, son comparables la distribución y el rango (27-70 %) del contenido de G+C encontrado en los metagenomas de este trabajo y los de estos autores. Este rango denominado “rango de vida” por Tobar-Tosse *et al.* (2013), donde sugieren estarían la mayoría de organismos simbiotes, intracelulares o aerobios, particularmente reconocidos por sus genomas con alto contenido de G+C (Bohlin 2015, Wu *et al.* 2012), deja entrever, como la variación en el contenido de G+C es apreciable incluso con la poca cantidad de ADN intergénico que se sabe presentan algunas bacterias endosimbiotes, por lo que se cree, que dicha variación también afectaría las NCS. Esto también es deducible en las diferencias marcadas de los contenidos de G+C de secuencias completas y codificantes aquí presentadas (Fig. 2A). De tal forma se confirma que las NCS, que ya se sabe median procesos regulatorios importantes, contienen información útil en sus patrones, que podrían dar un valor agregado a

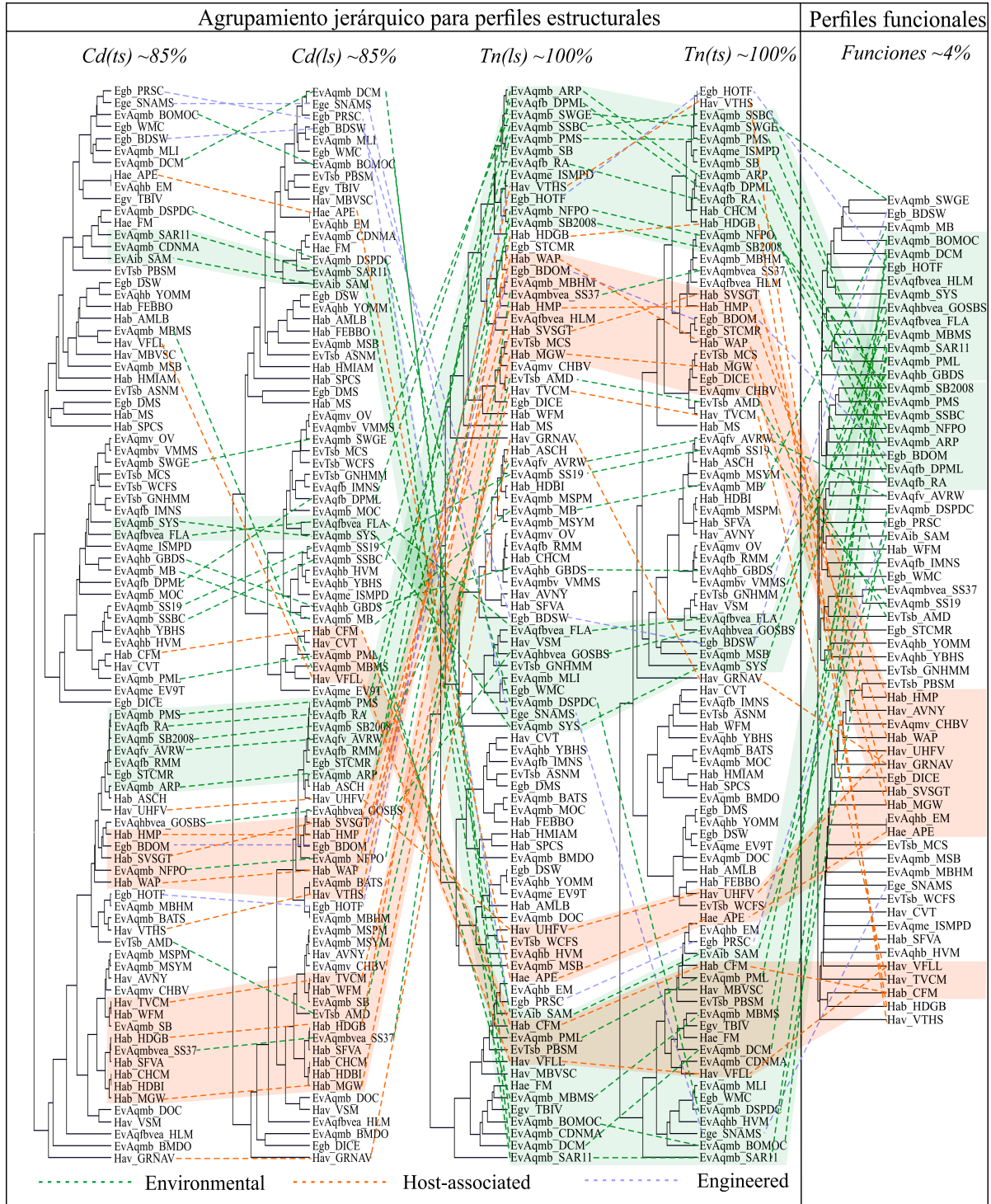


Figura 7. Comparación de árboles de agrupamiento jerárquico de los perfiles estructurales y funcionales. Las líneas punteadas comparan las agrupaciones de metagenomas entre los perfiles. Las franjas sombreadas representan los grupos mas conservados encontrados.

los análisis metagenómicos basados solo en predicción de genes (Kelley *et al.* 2012 tomado de Tobar-Tosse *et al.* 2013) y revelan el potencial que pueden tener en estudios de caracterización metagenómica.

Abundantes TRS encontrados en las NCS reafirman su importancia en la maquinaria reguladora de los microorganismos.

Por otra parte, las repeticiones de trinucleótidos contribuyen al polimorfismo del genoma y su variación afecta no solo la expresión génica, sino también, la adaptación a factores ambientales en procariotas (Lin and Kussell 2012, Li *et al.* 2004, Trivedi 2006, Zhou *et al.* 2014). El análisis de uso de Cd y Tn revelo semejanzas importantes con respecto a los resultados de Tobar-Tosse *et al.* (2013), observándose una relación similar en el uso de estos patrones en los metagenomas (Fig. 2A, asteriscos rojos y verdes). Aquí también se destaca la abundancia de TRS (trinucleotid repeat sequences) encontrados en las NCS, ratificando la relevancia de los TRS; CGC, CCG y CCG, y se hace hincapié en la participación de estos, en la maquinaria regulatoria de genes procariotas. Precisamente, Zou *et al.* (2014) presentan una lista de repeticiones de trinucleótidos con funciones regulatorias en genes de bacterias reportados en literatura, donde CCG, por ejemplo, hace parte de la secuencia reguladora del gen *modD* en *Neisseria meningitidis*, o Trivedi (2006) asegura que este TRS interviene en el plegamiento y solubilidad de proteínas, al igual que en las respuestas al estrés en arqueas.

Evidencia de perfiles funcionales discriminatorios entre los metagenomas y alta presencia de NCS en genomas bacterianos.

Las asignaciones funcionales encontradas a partir de secuencias codificantes en los metagenomas (Fig. 4) muestran asociaciones entre las distintas escalas de la clasificación ecosistémica. Estudios como el de Dinsdale *et al.* (2008), Roth-Schulze *et al.* (2016) y Wang and Sun (2016) reportan estas clases de asociaciones, demostrando que existen perfiles metabólicos discriminatorios en los ambientes y que son representados en las asociaciones de metagenomas con las mismas categorías ecosistémicas, tipos de ecosistema y tipo de comunidad de este trabajo. Sin embargo, las asociaciones de metagenomas compuestos por combinaciones de ecosistemas (*Ev*, *Ha* y *Eg*), son prueba de la existencia de funciones comunes entre los metagenomas (Fig. 5) (Tobar-Tosse *et al.* 2013). Por otra parte, la relación directa entre el tamaño del metagenoma y el número de asignaciones funcionales planteado por Tobar-Tosse *et al.* (2013), no es evidente en este trabajo, siendo esto reflejado en metagenomas como *EvAqmb_ARP* o *EvAqmb_NFPO*

de tamaños grandes, pero con pocas asignaciones funcionales (Fig. 3). Igualmente, se destaca que 27 metagenomas no tienen asignaciones funcionales en *Pfam* (Fig. 5, asteriscos rojos), sugiriendo la existencia de genes con funciones desconocidas (Culligan *et al.* 2014, Venter *et al.* 2004) o microorganismos no caracterizados aún (Menzel *et al.* 2015). En cuanto a los porcentajes de clases taxonómicas encontradas en los metagenomas a partir de NCS (Fig. 6), estos son comparables con los encontrados por Tobar-Tosse *et al.* (2013), permitiendo confirmar la alta presencia de NCS en genomas bacterianos y que según Matelska *et al.* (2016) pueden estar evolutivamente conservados en los microorganismos.

La importancia de los perfiles estructurales como trinucleótidos en la descripción de características comunes para los metagenomas y su caracterización ambiental.

La variación en los valores de *Tn(ls)* y *Tn(ts)* y *Cd(ls)* y *Cd(ts)* como medida de comparación entre metagenomas propuesta por Tobar-Tosse *et al.* (2013) fue comprobada. Las asociaciones entre metagenomas expuestas por estos perfiles estructurales en comparación con las del perfil funcional, muestran semejanzas en la conservación de los grupos ecosistémicos (Fig.7, áreas sombreadas). Así mismo los porcentajes altos de secuencias usadas de los perfiles de trinucleótidos (100 %) con respecto a los otros perfiles (codones 85 % y funcionales 4 %) son destacables. Al igual que para Tobar-Tosse *et al.* (2013) aquí se ratifica la importancia de los trinucleótidos en la comparación de metagenomas y caracterización ambiental (Fig.7), pero para nosotros es indiferente de si se logra con *Tn(ts)* o *Tn(ls)*. Esto es evidente en las agrupaciones que forman los metagenomas *EvAqmb_SSBC*, *EvAqmb_PMS*, *EvAqmb_ARP*, *EvAqfb_DPL* y *EvAqfb_RA* que son equiparables entre los tres tipos de perfiles (*Tn(ts)*, *Tn(ls)* y funcional) de este trabajo. Biológicamente, las semejanzas encontradas entre estos perfiles sugieren mecanismos conservados de regulación para genes específicos presentes en las secuencias de codificación (Tobar-Tosse *et al.* 2013). Estudios como el de Marchais *et al.* (2009) exponen evidencia de firmas específicas presentes en secuencias de *ncRNAs*, capaces de agrupar genomas bacterianos. De igual forma, las diferencias entre estos perfiles, no son más que la presencia de mecanismos conservados de regulación no compartidos con el perfil funcional (Fu *et al.* 2013, Tobar-Tosse *et al.* 2013), que pueden estar relacionados con elementos no codificantes como *YRNAs* y *sbRNAs* que Kowalski (2015) encuentra conservados tanto en procariotas como eucariotas.

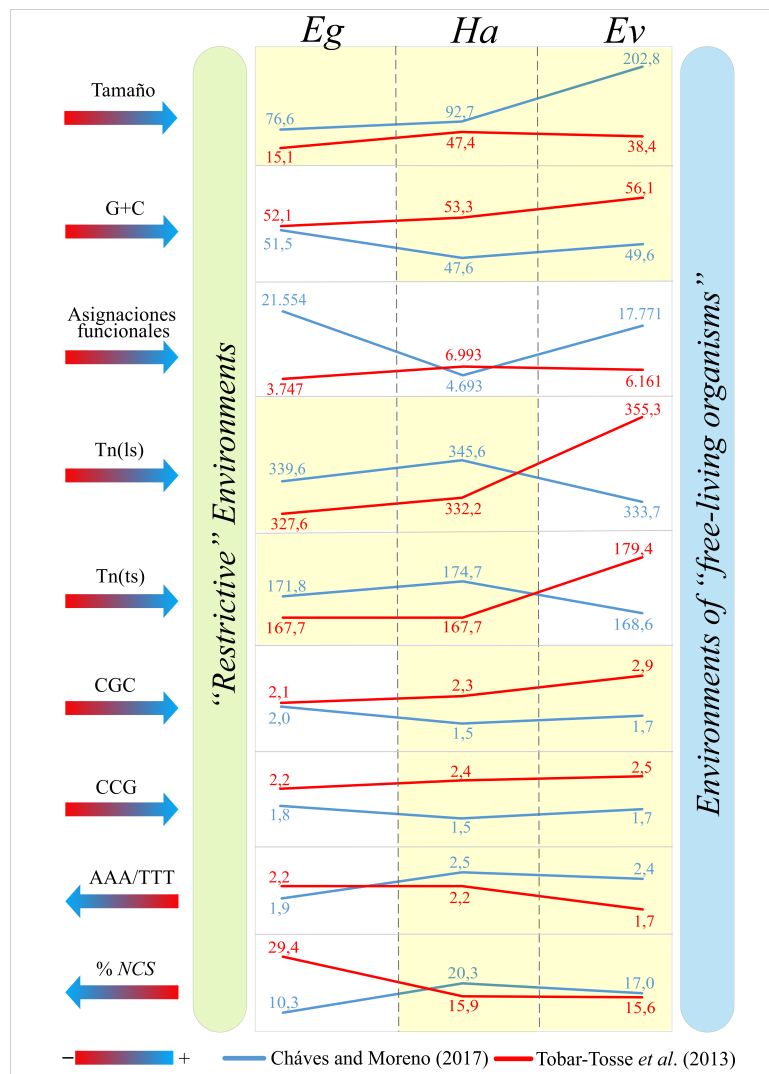


Figura 8. Análisis comparativo del marco de trabajo bioinformático de Tobar-Tosse *et al.* (2013). Las líneas azules representan los resultados de este trabajo y las rojas los de Tobar-Tosse *et al.* (2013) con sus respectivos valores para cada categoría. Las flechas indican la tendencia en la que deben aumentar los valores según el modelo en cada categoría. Las regiones amarillas denotan los tipos de ecosistemas que cumplen las tendencias de cada categoría.

Se valida gran parte del marco de trabajo bioinformático propuesto por Tobar-Tosse *et al.* (2013).

El marco de trabajo bioinformático para el estudio de metagenomas ambientales propuesto por Tobar-Tosse *et al.* (2013) establece 10 parámetros que pueden correlacionarse e identificar las tendencias $Eg \rightarrow Ha \rightarrow Ev$ entre los ecosistemas, sugiriendo el movimiento de estos parámetros metagenómicos desde ambientes restrictivos.^a ambientes de “organismos de vida libre”. Este marco conceptual fue evaluado sobre los 90 metagenomas (Fig. 8, líneas azules) y sugiere que un 67 % de los resultados, son compatibles con el trabajo de Tobar-Tosse *et al.* (2013) (Líneas rojas). Esto es destacable, si se tiene en cuenta la variabilidad en el tamaño y tipo de comunidad de los metagenomas analizados (Tabla 1). Es de resaltar que

siete parámetros (contenido de G+C, asignaciones funcionales, CGC, CCG, AAA, TTT y %NCS) exhiben solo la tendencia $Ha \rightarrow Ev$, demostrando la afinidad del marco de trabajo con la mayoría de metagenomas aquí analizados, pero que excluye a los metagenomas Eg , tal vez por el poco número de estos. De esta manera, las correlaciones entre los parámetros encontrados por Tobar-Tosse *et al.* 2013, solo son reproducibles entre los ecosistemas Ha y Ev , especialmente para los trinucleótidos: CGC, CCG, TTT y AAA. Aquí no es evidente la relación inversa entre el número de asignaciones funcionales y el porcentaje de NCS, que puede estar influenciado por la cantidad considerable de metagenomas que no obtuvieron coincidencias en *Pfam* (Fig. 5). Como tampoco se observa una correlación entre el contenido de G+C frente a $Tn(ts)$ y $Tn(ls)$. Todo esto permite deducir que el marco de

trabajo bioinformático propuesto por Tobar-Tosse *et al.* (2013) si es aplicable a los metagenomas *Ha* y *Ev* usados aquí, permitiendo observar las tendencias que siguen ciertos patrones desde ambientes “restrictivos” a ambientes de “organismos de vida libre”. Además, proponemos que los patrones que no exhiben las tendencias esperadas deben completarse con otros enfoques que sustenten mejor esta propuesta o aumentando el número de muestras de metagenomas *Eg.*

La existencia de una relación directa entre la cantidad de *NCS* y la composición particular de algunos metagenomas, así como el potencial informativo de algunos patrones presentes en las *NCS* o los abundantes *TRS* encontrados en estas secuencias, permiten concluir que efectivamente las *NCS* presentan información suficiente para describir metagenomas tanto estructural como funcionalmente, destacando a los patrones estructurales como trinucleótidos por su capacidad de encontrar asociaciones entre los metagenomas.

Igualmente se concluye que el marco de trabajo bioinformático propuesto por Tobar-Tosse *et al.* (2013) es válido solo para metagenomas *Ha* y *Ev* de este estudio, observándose las tendencias esperadas que siguen ciertos patrones desde ambientes “restrictivos” a ambientes de “organismos de vida libre”.

AGRADECIMIENTOS

Agradezco a la Universidad del Valle, Departamento de Biología, Escuela de Ingeniería de Sistemas y Computación, en especial al Ph.D. Pedro Moreno por su apoyo y oportuna gestión para la culminación de este trabajo, además al Ph.D. Fabian Tobar por apoyar el proceso investigativo, así como al Ph.D. Luis Garreta por apoyar la parte computacional y al candidato a doctor Miguel Guevara por la optimización del “Cluster HPC” para el análisis de los datos, al igual que a todos los integrantes del grupo de investigación en bioinformática por su ayuda y colaboración.

LITERATURA CITADA

- Alam, C. M., Singh, A. K., Sharfuddin, C. and Ali, S. (2014) ‘Genome-wide scan for analysis of simple and imperfect microsatellites in diverse carlaviruses’, *Infection, Genetics and Evolution*, 21, pp. 287–294. doi: 10.1016/j.meegid.2013.11.018.
- Bohlin, J. (2015) ‘Genome expansion in bacteria: the curious case of *Chlamydia trachomatis*’, *BMC Research Notes*. BioMed Central, 8(1), p. 512. doi: 10.1186/s13104-015-1464-6.
- Carr, R. and Borenstein, E. (2014) ‘Comparative analysis of functional metagenomic annotation and the mappability of short reads.’, *PLoS ONE*. Edited by C. Gibas. Public Library of Science, 9(8), p. e105776. doi: 10.1371/journal.pone.0105776.
- Culligan, E. P., Sleator, R. D., Marchesi, J. R. and Hill, C. (2014) ‘Metagenomics and novel gene discovery’, *Virulence*. Taylor & Francis, 5(3), pp. 399–412. doi: 10.4161/viru.27208.
- Dinsdale, E. a, Edwards, R. a, Hall, D., Angly, F., Breitbart, M., Brulc, J. M., Furlan, M., Desnues, C., Haynes, M., Li, L., McDaniel, L., Moran, M. A., Nelson, K. E., Nilsson, C., Olson, R., Paul, J., Brito, B. R., Ruan, Y., Swan, B. K., Stevens, R., Valentine, D. L., Thurber, R. V., Wegley, L., White, B. a and Rohwer, F. (2008) ‘Functional metagenomic profiling of nine biomes.’, *Nature*. Nature Publishing Group, 452(7187), pp. 629–632. doi: 10.1038/nature07346.
- Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A. and Knight, R. (2012) ‘Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients.’, *The ISME Journal*. Nature Publishing Group, 6(5), pp. 1007–17. doi: 10.1038/ismej.2011.159.
- Finn, R. D., Bateman, A., Clements, J., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J. and Punta, M. (2014) ‘Pfam: The protein families database’, *Nucleic Acids Research*. Oxford University Press, 42(D1), pp. D222–30. doi: 10.1093/nar/gkt1223.
- Fu, Y., Deiorio-Haggar, K., Anthony, J. and Meyer, M. M. (2013) ‘Most RNAs regulating ribosomal protein biosynthesis in *Escherichia coli* are narrowly distributed to Gammaproteobacteria’, *Nucleic Acids Research*. Oxford University Press, 41(6), pp. 3491–3503. doi: 10.1093/nar/gkt055.
- Gupta, P., Manjula, A., Rajendhran, J., Gunasekaran, P. and Vakhlu, J. (2016) ‘Comparison of metagenomic DNA extraction methods for soil sediments of high elevation Puga hot spring in Ladakh, India to explore bacterial diversity’, *Geomicrobiology Journal*, doi: 10.1080/01490451.2015.1128995.
- Hildebrand, F., Meyer, A. and Eyre-Walker, A. (2010) ‘Evidence of selection upon genomic GC-content in bacteria’, *PLoS Genetics*. Edited by M. W. Nachman. Public Library of Science, 6(9), p. e1001107. doi: 10.1371/journal.pgen.1001107.
- Hiraoka, S., Yang, C. and Iwasaki, W. (2016) ‘Metagenomics and bioinformatics in microbial ecology: current status and beyond’, *Microbes and Environments*. Japanese Society of Microbial Ecology · The Japanese Society of Soil Microbiology, 31(3), pp. 204–212. doi: 10.1264/jsme2.ME16024.
- Jung, J., Philippot, L. and Park, W. (2016) ‘Metagenomic and functional analyses of the consequences of reduction of bacterial diversity on soil functions and bioremediation in diesel-contaminated microcosms’, *Scientific Reports*. Nature Publishing Group, 6(October 2015), pp. 1–10. doi: 10.1038/srep23012.

- Koonin, E. V. and Wolf, Y. I. (2008) 'Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world', *Nucleic Acids Research*. Oxford University Press, 36(21), pp. 6688–6719. doi: 10.1093/nar/gkn668.
- Kowalski, M. P. and Krude, T. (2015) 'Functional roles of non-coding Y RNAs.', *The international journal of biochemistry & cell biology*, 66, pp. 20–9. doi: 10.1016/j.biocel.2015.07.003.
- Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. and Hugenholtz, P. (2008) 'A bioinformatician's guide to metagenomics.', *Microbiology and molecular biology reviews* : MMBR. American Society for Microbiology (ASM), 72(4), pp. 557–78, Table of Contents. doi: 10.1128/MMBR.00009-08.
- Lassalle, F., Périan, S., Bataillon, T., Nesme, X., Duret, L. and Daubin, V. (2015) 'GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands', *PLoS genetics*. Edited by D. A. Petrov. Public Library of Science, 11(2), p. e1004941. doi: 10.1371/journal.pgen.1004941.
- Li, W. (2009) 'Analysis and comparison of very large metagenomes with fast clustering and functional annotation.', *BMC bioinformatics*. BioMed Central, 10(1), p. 359. doi: 10.1186/1471-2105-10-359.
- Li, W., Wooley, J. C. and Godzik, A. (2008) 'Probing Metagenomics by Rapid Cluster Analysis of Very Large Datasets', *PLoS ONE*. Edited by D. Jones. Public Library of Science, 3(10), p. e3375. doi: 10.1371/journal.pone.0003375.
- Li, Y. C., Korol, A. B., Fahima, T. and Nevo, E. (2004) 'Microsatellites within genes: Structure, function, and evolution', *Molecular Biology and Evolution*. Oxford University Press, 21(6), pp. 991–1007. doi: 10.1093/molbev/msh073.
- Lin, W. H. and Kussell, E. (2012) 'Evolutionary pressures on simple sequence repeats in prokaryotic coding regions', *Nucleic Acids Research*. Oxford University Press, 40(6), pp. 2399–2413. doi: 10.1093/nar/gkr1078.
- Marchais, A., Naville, M., Bohn, C., Boulloc, P. and Gautheret, D. (2009) 'Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles', *Genome Research*. Cold Spring Harbor Laboratory Press, 19(6), pp. 1084–1092. doi: 10.1101/gr.089714.108.
- Matelska, D., Kurkowska, M., Purta, E., Bujnicki, J. M. and Dunin-Horkawicz, S. (2016) 'Loss of conserved non-coding RNAs in genomes of bacterial endosymbionts.', *Genome biology and evolution*. Oxford University Press, 8(2), p. evw007–. doi: 10.1093/gbe/evw007.
- Menzel, P., Gudbergssdóttir, S. R., Rike, A. G., Lin, L., Zhang, Q., Contursi, P., Moracci, M., Kristjansson, J. K., Bolduc, B., Gavrilov, S., Ravin, N., Mardanov, A., Bonch-Osmolovskaya, E., Young, M., Krogh, A. and Peng, X. (2015) 'Comparative Metagenomics of Eight Geographically Remote Terrestrial Hot Springs', *Microbial Ecology*. Springer US, 70(2), pp. 411–424. doi: 10.1007/s00248-015-0576-9.
- Miller, W. A., Shen, R., Staplin, W. and Kanodia, P. (2016) 'Noncoding RNAs of Plant Viruses and Viroids: Sponges of Host Translation and RNA Interference Machinery.', *Molecular Plant-Microbe Interactions*. Molecular Plant-Microbe Interactions, X(X), p. MPMI10150226FI. doi: 10.1094/MPMI-10-15-0226-FI.
- Moody, M. J., Young, R. A., Jones, S. E. and Elliot, M. A. (2013) 'Comparative analysis of non-coding RNAs in the antibiotic-producing *Streptomyces* bacteria', *BMC Genomics*. BioMed Central, 14(1), p. 558. doi: 10.1186/1471-2164-14-558.
- Reichenberger, E. R., Rosen, G., Hershberg, U. and Hershberg, R. (2015) 'Prokaryotic nucleotide composition is shaped by both phylogeny and the environment', *Genome Biology and Evolution*, 7(5), pp. 1380–1389. doi: 10.1093/gbe/evv063.
- Rogozin, I. B., Makarova, K. S., Natale, D. a, Spiridonov, A. N., Tatusov, R. L., Wolf, Y. I., Yin, J. and Koonin, E. V. (2002) 'Congruent evolution of different classes of non-coding DNA in prokaryotic genomes.', *Nucleic acids research*. Oxford University Press, 30(19), pp. 4264–4271. doi: 10.1093/nar/gkf549.
- Schrader, J. M., Zhou, B., Li, G. W., Lasker, K., Childers, W. S., Williams, B., Long, T., Crosson, S., McAdams, H. H., Weissman, J. S. and Shapiro, L. (2014) 'The Coding and Noncoding Architecture of the *Caulobacter crescentus* Genome', *PLoS Genetics*, 10(7). doi: 10.1371/journal.pgen.1004463.
- Simon, C. and Daniel, R. (2011) 'Metagenomic analyses: Past and future trends', *Applied and Environmental Microbiology*. American Society for Microbiology, 77(4), pp. 1153–1161. doi: 10.1128/AEM.02345-10.
- Stewart, E. J. (2012) 'Growing unculturable bacteria', *Journal of Bacteriology*. American Society for Microbiology, 194(16), pp. 4151–4160. doi: 10.1128/JB.00345-12.
- Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., Heroux, M. A., Ioannidis, J. P. A. and Tauber, M. (2016) 'Enhancing reproducibility for computational methods', *Science*, 354(6317).
- Teeling, H. and Glöckner, F. O. (2012) 'Current opportunities and challenges in microbial metagenome analysis-A bioinformatic perspective', *Briefings in Bioinformatics*, 13(6), pp. 728–742. doi: 10.1093/bib/bbs039.
- Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. and Glockner, F. O. (2004) 'Application of tetranucleotide frequencies for the assignment of genomic fragments', *Environmental Microbiology*. Blackwell Science Ltd, 6(9), pp. 938–947. doi: 10.1111/j.1462-2920.2004.00624.x.
- Thomas, T., Gilbert, J. and Meyer, F. (2012) 'Metagenomics - a guide from sampling to data analysis', *Microbial Informatics and Experimentation*. *BioMed Central*, 2(1), p. 3. doi: 10.1186/2042-5783-2-3.
- Tobar-Tosse, F., Rodríguez, A. C., Vélez, P. E., Zambrano, M. M. and Moreno, P. A. (2013) 'Exploration of Noncoding Sequences in Metagenomes', *PLoS ONE*. Edited by A. A.-J. Golden. *Public Library of Science*, 8(3), p. e59488. doi: 10.1371/journal.pone.

- Tringe, S. G. and Rubin, E. M. (2005) 'Metagenomics: DNA sequencing of environmental samples.', *Nature reviews. Genetics*. Nature Publishing Group, 6(11), pp. 805–14. doi: 10.1038/nrg1709.
- Trivedi, S. (2006) 'Comparison of simple sequence repeats in 19 Archaea.', *Genetics and Molecular Research : GMR*, 5(4), pp. 741–72. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17183484> (Accessed: 13 February 2017).
- Tycowski, K. T., Guo, Y. E., Lee, N., Moss, W. N., Vallery, T. K., Xie, M. and Steitz, J. A. (2015) 'Viral noncoding RNAs: More surprises', *Genes and Development*. Cold Spring Harbor Laboratory Press, 29(6), pp. 567–584. doi: 10.1101/gad.259077.115.
- Vavourakis, C. D., Ghai, R., Rodriguez-Valera, F., Sorokin, D. Y., Tringe, S. G., Hugenholtz, P. and Muyzer, G. (2016) 'Metagenomic insights into the uncultured diversity and physiology of microbes in four hypersaline soda lake brines', *Frontiers in Microbiology*. Frontiers Media SA, 7(FEB), p. 211. doi: 10.3389/fmicb.2016.00211.
- Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H. and Smith, H. O. (2004) 'Environmental Genome Shotgun Sequencing of the Sargasso Sea', *Science*, 304(5667), pp. 66–74. doi: 10.1126/science.1093857.
- Wang, H. and Sun, L. (2016) 'Comparative metagenomic analysis of the microbial communities in the surroundings of Iheya north and Iheya ridge hydrothermal fields reveals insights into the survival strategy of microorganisms in deep-sea environments', *Journal of Marine Systems*. doi: 10.1016/j.jmarsys.2016.10.006.
- Wu, H., Zhang, Z., Hu, S. and Yu, J. (2012) 'On the molecular mechanism of GC content variation among eubacterial genomes.', *Biology direct*. BioMed Central, 7(1), p. 2. doi: 10.1186/1745-6150-7-2.
- Zhou, K., Aertsen, A. and Michiels, C. W. (2014) 'The role of variable DNA tandem repeats in bacterial adaptation', *FEMS Microbiology Reviews*, 38(1), pp. 119–141. doi: 10.1111/1574-6976.12036.

Anexos

Tabla 1. Clasificación de metagenomas y cantidad de secuencias analizadas

Ecosistema	Categoría Ecosistémica	Tipo de Ecosistema	Tipo de comunidad	Abreviación	Proyecto Metagenómico	Tamaño (Mpb)	# Secuencias
Environmental	Air	Air	Bacteria	EvAqib_SAM	Singapore Air Metagenome	16,0	16.539
				EvAqfb_DPML	Diel gene expression profiles of a phosphorus limited mountain lake using metatranscriptomics	25,0	137.204
	Aquatic	Freshwater	Bacteria	EvAqfb_IMNS	Ice Metagenome of the Northern Schneeferner	230,3	1.035.051
			Bacteria	EvAqfb_RA	River Amazon Metagenome	374,8	1.150.910
			Bacteria	EvAqfb_RMM	Microbial metagenome from Rios Mesquites microbiolite	47,4	454.430
			Bacteria, Virus, Eukarya, Archaea	EvAqfbvea_FLA	Freshwater Lagoon Albufera	237,4	651.327
			Bacteria, Virus, Eukarya, Archaea	EvAqfbvea_HLM	Mediterranean Hypersaline Coastal Lagoon	231,9	698.585
			Virus	EvAqfv_AVRW	Metagenomic analysis of viruses in reclaimed water	323,0	1.476.389
			Bacteria	EvAqhb_EM	Epibiont Metagenome	30,0	36.429
			Bacteria	EvAqhb_GBDS	Guaymas Basin deep-sea Metagenome	1.026,8	4.721.230
			Bacteria	EvAqhb_HVM	Hydrothermal Vent Metagenome	20,6	25.392
			Bacteria	EvAqhb_YBHS	Metagenome from Yellowstone Bison Hot Spring	244,9	284.067
	Marine	Marine	Bacteria	EvAqhb_YOMM	Yellowstone National Park Octopus/Mushroom Hot Springs Metagenome	29,5	29.975
			Bacteria, Virus, Eukarya, Archaea	EvAqhbvea_GOSBS	Global Ocean Sampling Baltic Sea	93,5	318.206
			Bacteria	EvAqmb_ARP	Microbial community gene expression across a productivity gradient of the Amazon River plume	789,5	3.756.679
			Bacteria	EvAqmb_BATS	BATS 16S rRNA pyrosequencing time- and depth-series	722,2	2.744.286
			Bacteria	EvAqmb_BMDO	Transcriptional Response of Bathypelagic Marine Bacteria to the Deepwater Horizon Oil Spill	175,9	359.042
			Bacteria	EvAqmb_BOMOC	Bermuda Oceanic Microbial Observatory Course	492,1	1.180.976
			Bacteria	EvAqmb_CDNMA	Comparative day/night metatranscriptomic analysis of microbial communities	27,3	277.518
			Bacteria	EvAqmb_DCM	Deep Chlorophyll Maximum Metagenome	267,3	1.043.403
			Bacteria	EvAqmb_DOC	Community succession in DOC utilization experiment reflects functional adaptation in marine bacterioplankton	157,6	613.726
			Bacteria	EvAqmb_DSPDC	Diverse syntrophic partnerships by direct cell capture and metagenomics	92,6	485.989

Ecosistema	Categoría Ecosistémica	Tipo de Ecosistema	Tipo de comunidad	Abreviación	Proyecto Metagenómico	Tamaño (Mpb)	# Secuencias
			Bacteria	EvAqmb_MB	Marine Bacterioplankton Metagenomes	285.7	1,249,363
			Bacteria	EvAqmb_MBHM	Mediterranean Bathypelagic Habitat Metagenome	5.5	6,905
			Bacteria	EvAqmb_MBMS	Monterey Bay Microbial Study	267.1	1,123,695
			Bacteria	EvAqmb_MLI	Marine Metagenome from Line Islands	217.1	2,078,504
			Bacteria	EvAqmb_MOC	Microbial Oceanography Course	26.9	93,470
			Bacteria	EvAqmb_MSB	Metagenome of Marine NaCl-Saturated Brine	0.3	445
			Bacteria	EvAqmb_MSPM	Metagenomic signatures of the Peru Margin	60.3	578,317
			Bacteria	EvAqmb_MSVM	Marine <i>Synechococcus</i> metagenome experiment	74.7	726,942
			Bacteria	EvAqmb_NFPO	Influence of nitrogen-fixation on microbial community gene expression in the oligotrophic Southwest Pacific Ocean	765.0	3,845,744
			Bacteria	EvAqmb_PML	Marine Metagenome from Coastal Waters project at Plymouth Marine Laboratory	295.2	1,375,026
			Bacteria	EvAqmb_PMS	Polyamine Metatranscriptomic Study	103.3	672,764
			Bacteria	EvAqmb_SAR11	AAP gene operon study of SAR11 clades	47.8	208,654
			Bacteria	EvAqmb_SB	Sapelo Island Bacterioplankton Metagenome	32.9	350,324
			Bacteria	EvAqmb_SB2008	Sapelo Island Summer 2008 Bacterioplankton Metatranscriptome	445.5	2,128,215
			Bacteria	EvAqmb_SS19	Santa Pola Saltern 19 Percent	421.3	1,166,606
			Bacteria	EvAqmb_SSBC	Sargasso Sea Bacterioplankton Community	119.4	572,204
			Bacteria	EvAqmb_SWGE	Surface Water Marine Microbial Community Gene Expression	276.2	1,491,285
			Bacteria	EvAqmb_SYS	<i>Synechococcus</i> Study	93.4	451,497
			Bacteria, Virus	EvAqmbv_VMMS	Viral and Microbial Metagenomes from Salterns of Differing Salinities	406.6	4,034,831
			Bacteria, Virus, Eukarya, Archaea	EvAqmbvea_SS37	Santa Pola Saltern 37 percent	282.5	674,018
Host-associated	Host-associated	Host-associated	Eukarya	EvAqme_EV9T	Eukaryal V9 Tag Sequencing Methodology	14.7	93,397
			Eukarya	EvAqme_JSMPD	The use and regulation of ferritin for iron sequestration in marine pennate diatoms	115.5	527,646
			Virus	EvAqmvv_CHBV	Chesapeake Bay Virioplankton Metagenome	26.7	39,231
			Virus	EvAqmvv_OV	Ocean Viruses	262.4	2,597,495
			Bacteria	EvTs_b_ASNM	Bacterial Diversity in Soils under <i>Artemisia sieberi</i> and <i>Noaea mucronata</i> plants in the Negev Desert	0.4	617
			Bacteria	EvTsb_AMD	Acid Mine Drainage Metagenome	89.3	92,494
			Bacteria	EvTsb_GNHMM	Metagenome isolated from Guerrero Negro hypersaline microbial mat in Mexico	75.4	117,519
			Bacteria	EvTsb_MCS	Metatranscriptomics of contaminated soil	38.9	115,698
			Bacteria	EvTsb_PBSM	Pacific Beach Sand Metagenome	0.9	732
			Bacteria	EvTsb_WCFS	Waseca County Farm Soil Metagenome	39.5	35,850
			Bacteria	Hab_AMLB	Association <i>M. leiyi</i> and bacteria	14.3	34,546

Ecosistema	Categoría Ecosistémica	Tipo de Ecosistema	Tipo de comunidad	Abreviación	Proyecto Metagenómico	Tamaño (Mpb)	# Secuencias
			<i>Bacteria</i>	Hab_ASCH	Metagenomic analysis of stressed coral holobionts	106,1	983.499
				Hab_CFM	Canine fecal metagenome	178,3	466.196
				Hab_CHCM	Chicken Cecum Microbiome	57,0	548.986
				Hab_FEBBO	Fluorescent epibiotic biofilm on Bahamian ostracods	10,0	75.682
				Hab_HDBI	Metagenomic Diagnosis of Bacterial Infections	20,5	200.396
				Hab_HDGB	Human Distal Gut Biome project	164,0	156.114
				Hab_HMIAM	Host-microbe interactions in an Atriplex microbiome	119,2	313.705
				Hab_HMP	Human Microbiome Project	438,5	421.584
				Hab_MGW	Mediterranean Gutless Worm Metagenome	109,3	111.366
				Hab_MS	Mouse Skin Metagenome	0,6	532
				Hab_SFVA	Swine fecal viromes with and without in-feed antibiotics	366,2	1.074.567
				Hab_SPCS	Symbiotic prokaryotic communities in sponges	0,2	631
				Hab_SVSGT	Solemya velum symbiont gill transcriptome	445,8	1.422.242
				Hab_WAP	White pox in Acropora palmata	174,4	826.116
				Hab_WFM	Whale Fall Metagenome	28,1	27.737
				Hae_APE	Alvinella pompejana Epibiont Metagenome	30,0	36.429
				Hae_FM	Fossil Metagenome	6,1	54.217
				Hav_AVNY	Allochthonous viruses in upstate NY freshwater resources	11,3	35.168
				Hav_CVT	Population structure of wild ocean cyanobacterial viruses revealed by viral-tagged metagenomics	51,1	159.753
				Hav_GRNAV	GorgonianRNAViruses	3,0	32.655
				Hav_MBVSC	Analysis of a viral metagenomic library from 200 m depth in Monterey Bay, California constructed by direct shotgun cloning	0,4	878
				Hav_TVCM	Trichodesmium Viral Community Metagenome	58,7	148.268
				Hav_UHFV	Uncultured Human Fecal Virus Metagenome	8,6	6.694
				Hav_VFLL	Viral Metagenome from the Freshwater Lake Limnopolar	18,8	81.234
				Hav_VSM	Viral Stromatolite Metagenome	71,6	693.171
				Hav_VTHS	Viral Metagenomes from Terrestrial Hot Springs	12,1	12.402
Engineered	Engineered	Engineered	<i>Bacteria</i>	Egb_BDOM	Bloomer DOM addition experiment	665,4	2.720.826
			<i>Bacteria</i>	Egb_BDSW	Bloomer DSW addition experiment	1.014,3	5.008.547
			<i>Bacteria</i>	Egb_DICE	Dauphin Island Cubitainer Experiment (DICE)	134,4	767.735
			<i>Bacteria</i>	Egb_DMS	Microbial controls on DMS degradation and DMS formation in the Sargasso Sea	1,9	17.789
			<i>Bacteria</i>	Egb_DSW	Diversity and population structure of sewage-derived microorganisms in wastewater treatment plant influent	27,6	268.930
			<i>Bacteria</i>	Egb_HOTF	Two HOT Fosmid end depth profiles (HOT179 and HOT186)	91,8	194.570

Ecosistema	Categoría Ecosistémica	Tipo de Ecosistema	Tipo de comunidad	Abreviación	Proyecto Metagenómico	Tamaño (Mpb)	# Secuencias
			Bacteria	Egb_PRSC	Phosphorus Removing (EBPR) Sludge Community	65,6	67.900
			Bacteria	Egb_STCMR	Metagenomic sequencing of terephthalate-degrading consortia from mother reactor (TA_mother_DNA)	62,9	163.266
			Bacteria	Egb_WMC	Wastewater Metagenome Mallard Creek	100,1	414.014
			Eukarya	Ege_SNAMS	Study of a non-axenic Micromonas strain (CCMP1764)	174,8	744.820
			Virus	Egv_TBIV	Tampa Bay Induced Viral Metagenome	30,0	288.859
Total general						15.419,1	66.528.865