

MÁSTER EN BIOINFORMÁTICA

**Uso de herramientas bioinformáticas en la respuesta
microbiana al cambio climático.**

**AUTOR: KAREN VIVIANA YAGUACHE VIVANCO
TUTOR: MARIANELA TIBISAY RAVELO VELÁSQUEZ**

CURSO 2024-2025

Agradecimientos

Dios ha sido mi fuente de mi inspiración todos estos años y hoy es a quién le debo todos mis logros.

A la mujer más valiosa de mi mundo, mi madre Itamar y al hombre que más me ha apoyado en todo, mi padre Milton, gracias por su amor y sacrificio. Sin ustedes, no hubiera podido llegar hasta aquí. Gracias por formarme de una manera tan única reflejando lo que soy ahora y por brindarme ese amor incondicional que todos necesitamos para salir adelante. Para mí es un orgullo y un privilegio ser su hija.

A mis hermanos Milton Daniel, Kelvin William, Gabriela Dayanna les agradezco, por estar siempre presente, por impulsarme a ser mejor cada día, por querer verme bien y feliz. Son el mejor regalo que me pudieron dar mis padres y estoy agradecida por ello.

Agradezco a todos los docentes de la Universidad Europea de Madrid, por impartir sus conocimientos en el camino de preparación de mi maestría y de manera especial a mi tutora de proyecto de fin de máster, Marianela Tibibsay Ravelo Velásquez, por su paciencia, por sus ideas, por su visión y sus conocimientos profesionales brindándome valiosos aportes para mi investigación.

INDICE

Resumen	5
Introducción	6
Microorganismos y cambio climático	7
Bioma marino	7
Herramientas Bioinformáticas en microorganismos	9
Hipótesis	9
Objetivo General	10
Objetivos Específicos	10
Metodología	10
Diseño de estudio	12
Variables	12
Criterios de inclusión y exclusión	12
Criterios de inclusión	12
Criterios de Exclusión	12
Resultados y Discusión	13
Visualización de Análisis Taxonómico	13
Análisis de diversidad Alfa	20
Análisis de diversidad Beta	21
Análisis Funcional	23
Conclusión	26
Bibliografía	28
1. ANEXOS	30
A. Selección y descarga de datasets, con metadatos ambientales (MGnify)	30
B. Preparación del entorno de trabajo	31
C. Descarga y procesamiento de datos crudos (Obtención de la ruta exacta de las muestras)	31
D. Verificación de archivos:	31
E. Control de calidad y limpieza de datos:	31
1. <i>Instalación FastQC</i>	32
2. <i>Instalación de MultiQC</i>	32
3. <i>Limpieza de datos con fastp</i>	32
4. <i>Verificación de limpieza</i>	33
F. Análisis Taxonómico / Diversidad	33
<i>Ejecución de análisis taxonómico</i>	33

	<i>Visualización de Análisis Taxonómico (Rstudio)</i>	35
G.	Análisis alfa y beta	36
	<i>Rstudio (vegan, phyloseq, ggplot2)</i>	38
H.	Análisis funcional	40
	<i>Ensamblaje con MEGAHIT para las nueve muestras</i>	40
	<i>Predicción de Genes con Prodigal</i>	41
	Anotación Funcional con eggNOG-mapper annotate	42
	Combinación de los Resultados en una Matriz de Abundancia	43

Resumen

Nos enfrentamos al cambio climático inducido por conductas humanas, lo que ha generado una alerta masiva en cuanto al incremento de temperatura, aumento de radiación solar, pérdida de nutrientes, estratificación oceánica, entre otros. Actualmente contamos con nuevas tecnologías, por lo que se ha podido evidenciar composiciones taxonómicas de microbiomas, sin embargo, aún existe poca investigación en cuanto a su impacto y funcionalidad de microorganismos frente a cada bioma. El presente estudio pretende analizar herramientas bioinformáticas para predecir adaptaciones funcionales en las comunidades microbianas frente a al cambio climático. Dentro del análisis taxonómico, se estableció un dominio entre las muestras, donde existe diversidad microbiana de arqueas y bacterias y su excelente adaptabilidad frente a entornos acuáticos extremos. Las bacterias resultaron ser sumamente abundantes en estos entornos acuáticos, sin embargo, las arqueas mantienen mayor prevalencia en cada muestra, a pesar de no ser un grupo predominante. Se analizaron la diversidad alfa y beta de cada una de las muestras, además se realizó el análisis funcional, dando como resultado un heatmap con las cincuenta funciones más comunes dentro de cada muestra, principalmente identificadores como K07012 y K00590 relacionados principalmente en el funcionamiento de defensa celular. K00525 relacionado a rutas metabólicas de purinas y pirimidinas de interés. K13993 destacando su funcionalidad y supervivencia frente a altas temperaturas ocasionadas por el estrés o choque ambiental.

Palabras claves

Microbioma acuático, altas temperaturas, meta-análisis, bacterias, arqueas, supervivencia.

Introducción

Debido a las múltiples actividades humanas, sus acciones hacen que aumente progresivamente gases de efecto invernadero en la atmósfera, dando como resultado que el clima a nivel mundial esté en constante cambio. Este fenómeno climático presenta diferentes factores ambientales como incremento de temperatura (retroalimentación positiva), modificaciones en los patrones de lluvia o precipitaciones, incremento de frecuencia, intensidad y duración de eventos extremos. Por otro lado, estos cambios están impactando tanto el funcionamiento como la estructura de la mayoría de los ecosistemas terrestres, afectando principalmente entornos naturales como también los que han sido manipulados por el ser humano. Este mecanismo de retroalimentación positiva está incorporado en los modelos climáticos del Panel Intergubernamental de Cambio Climático (IPCC), sin embargo, su capacidad para predecir su magnitud exacta es limitada. Esto se debe, en gran medida, a las incertidumbres sobre cómo se manifiestan procesos complejos microbianos del suelo en dichos modelos. (Dacal de Tomás, 2021)

Teniendo en cuenta, el efecto las actividades que realiza el ser humano día a día, causan problemas al clima, medio ambiente y a su vez, provocan extinciones de otros seres vivos como las plantas y animales, lo que trae como consecuencia una pérdida en su biodiversidad, poniendo en riesgo el reino animal y vegetal en el planeta donde habitamos. Los microorganismos constituyen a la conservación de un ecosistema sostenible y son los responsables de brindar soporte vital en la biosfera. Pese a que el impacto de estos microorganismos no es estudiado a profundidad, existe preocupación su capacidad para residir o adaptarse frente al cambio climático, o en el peor de los casos, deshabilitarse la biodiversidad que tenemos en la actualidad. (Cavicchioli et al., 2019)

Por otro lado, como es de esperarse, la tecnología ha venido avanzando drásticamente en los últimos años, por lo que se destaca la participación de herramientas bioinformáticas y su importancia de manejar grandes volúmenes de datos para la resolución de problemas a nivel

biológico, como la creación de programas, modelos, algoritmos de un estudio en específico. La bioinformática no solo combina la informática con la biología, sino también abarca otras ciencias como la matemática, química, estadística, entre otras. (Burguener & Germán F, 2017)

El presente trabajo de fin de máster destaca la aplicación de técnicas bioinformáticas en el microbioma acuático y sus efectos frente al cambio climático, corroborando su aplicación mediante la recopilación de artículos científicos que muestran estudios que avalen estas aplicaciones, a través del análisis de datos o datasets procedentes de repositorios públicos como PubMed y Scopus, así como de repositorios de la UEM.

Microorganismos y cambio climático

Los microorganismos han existido desde el inicio de la vida, como hace 3.800 millones de años y se estima que sobrevivirán a cualquier extinción futura. El papel de estos microbios es fundamental en el planeta tierra, ya que participan en el ciclo biogeoquímico de nutrientes y de flujos elementales como carbono, fósforo, nitrógeno, así también como en la flora, fauna, productividad agrícola y la red trófica universal. Además, consumen gases implicados en el calentamiento global. Se encuentran en todos los entornos naturales, como profundidades subterráneas, entornos extremos, etc. (Cavicchioli et al., 2019)

Durante décadas se ha estudiado la estrecha relación entre los microorganismos y sus hábitats. Las nuevas tecnologías como la metagenómica y secuenciación se ha descubierto la inmensa variedad de microorganismos en entornos terrestres, acuáticos y urbanos. sin embargo, a pesar de que se ha logrado identificar la composición taxonómica de estas comunidades, aún existe bastante desconocimiento respecto a sus funciones específicas y su respuesta ante el cambio climático.

Bioma marino

Los biomas marinos engloban aproximadamente el 70% de la superficie terrestre, desde ecosistemas costeros como estuarios, arrecifes de coral y manglares hasta las inmensas

profundidades de los océanos abiertos. Existen algunos microorganismos presentes en el bioma marino, como bacterias quimiolitotóxicas, que generan energía mediante un proceso de oxidación de compuestos inorgánicos, como por ejemplo bacterias sulfo-oxidantes, anammox, arqueas y bacterias nitrificantes. También podemos encontrar bacterias heterótrofas como las desnitrificantes canónicas y las reductoras de nitrato y sulfato. Por otro lado, se encuentran organismos vegetales como fitoplancton y cianobacterias adaptadas a la poca luz dentro de las zonas anóxicas marinas.

Los microorganismos fototróficos aprovechan la energía solar para sus procesos vitales a 200 metros superiores de la columna de agua. Sin embargo, en áreas más profundas la vida marina, se adapta a la ausencia de luz, obteniendo energía mediante químicos orgánicos e inorgánicos. (Cavicchioli et al., 2019)

El incremento de temperatura, radiación solar, aumento de agua dulce en capas superficiales del océano, es directamente proporcional en la estratificación oceánica, trayendo como consecuencia la reducción de paso de nutrientes, desde las profundidades del agua, hasta las superficiales. (Cavicchioli et al., 2019)

Las archaeas y bacterias son parte de diferentes grupos de microorganismos procariontes dentro de la comunidad microbiana. Ambos participan principalmente en las funciones ambientales, controlando o cambiando los componentes inorgánicos y orgánicos en diferentes ecosistemas, terrestres o acuáticos. Participan también en el consumo o producción de metano, oxidación, desnitrificación, impactando activamente en los grupos microbianos establecidos. (Silveira et al., 2021)

Aunque las archaeas y las bacterias son diferentes a nivel bioquímico y genético, no existe un parámetro como tal que las separe totalmente. Se estima que la capacidad de las archaeas para adaptarse al estrés energético crónico es lo que las distingue de las bacterias.

Los mecanismos que permiten a las archaeas vivir con poca energía incluyen membranas celulares con baja permeabilidad y vías metabólicas específicas. Por lo que estas

adaptaciones les han permitido evolucionar y prosperar en ambientes extremos. (Bravo, 2024)

Herramientas Bioinformáticas en microorganismos

Tabla 1

Tabla de Herramientas bioinformáticas.

Herramientas	Descripción
bioinformaticas	
KRAKEN2	Kraken es una herramienta bioinformática, potencialmente utilizada en un sistema de clasificación taxonómica. En conjunto con otras herramientas como Bracken (mediante los resultados de kraken, bracken estudia la abundancia a nivel taxonómico) y para la visualización de los resultados taxonómicos, es compatible con Pavian , que permite la comparativa entre las clasificaciones de kraken 2 en todas las muestras a analizar. (Wood et al., 2022)
RStudio 4.1.0	Analiza de diversidad alfa y beta dentro de cada estudio individual. Para generar métricas de diversidad alfa, se realiza mediante el paquete phyloseq . Para el análisis de diversidad beta, utiliza el paquete vegan . (Li et al., 2023)
Otras herramientas especializadas	El ensamblaje metagenómico es un proceso bioinformático clave que permite recuperar catálogos de genes de microbiomas. Para el ensamblaje se emplea MEGAHIT v.1.1.2 . Finalmente, la predicción de genes en bacterias y arqueas se lleva a cabo mediante la herramienta bioinformática Prodigal v.2.6.3 . (Delgado & Andersson, 2022)

Hipótesis

El uso de herramientas bioinformáticas permite predecir adaptaciones funcionales en comunidades microbianas expuestas a variaciones ambientales asociadas al cambio climático.

Objetivo General

Analizar las herramientas bioinformáticas frente al estudio de la respuesta de las comunidades microbianas al cambio climático.

Objetivos Específicos

- ✚ Describir las herramientas más relevantes de la bioinformática utilizadas en el contexto del cambio climático

- ✚ Evaluar la capacidad de las herramientas bioinformáticas para predecir cambios funcionales y adaptativos con microorganismos en el cambio climático

- ✚ Identificar desafíos y limitaciones de la aplicación bioinformática en el cambio climático

Metodología

El presente trabajo se fundamenta en un enfoque bioinformático y de análisis de metadatos, aplicando herramientas computacionales para investigar la adaptación del microbioma acuático frente a condiciones asociadas al cambio climático. La metodología se estructura en las siguientes fases para garantizar la rigurosidad y que sea reproducible el análisis:

El análisis bioinformático comienza con la **preparación del entorno de trabajo**, que se realiza en un sistema operativo Linux por su compatibilidad con herramientas especializadas. En esta etapa se instalan los programas necesarios para el flujo de trabajo, como FastQC para control de calidad, Fastp para limpieza de datos, Samtools para manipulación de archivos, herramientas específicas para análisis taxonómico como Kraken2, análisis beta y alfa como Bracken, Rstudio para la visualización de datos. También se recomienda el uso de gestores de entornos como Conda para mantener las dependencias organizadas.

Una vez configurado el entorno, se procede a la **descarga de datos crudos** desde repositorios públicos como MGnify filtrando el bioma acuático a altas temperaturas, utilizando comandos como fastq-dump o wget, donde se descarga los paired-end en cada muestra. En

este caso, se escogieron nueve muestras con metadatos de un mismo estudio (MGYS00005690). Posteriormente, se realiza la **verificación de archivos** para garantizar su integridad mediante sumas de verificación (MD5) y la confirmación del formato correcto (FASTQ). Esta etapa es fundamental para evitar errores en los pasos posteriores.

El siguiente paso es el **control de calidad**, donde se evalúa la calidad de las secuencias mediante FastQC. Este análisis permite identificar problemas como adaptadores residuales, baja calidad en posiciones específicas y desviaciones en la distribución de GC. Con base en estos resultados, se guarda mediante MultiQC y se ejecuta la **limpieza de datos** utilizando fastp a través de un entorno de bioconda.

En el **análisis taxonómico** se ha empleado la herramienta Kraken2 para asignar lecturas a taxones y generar perfiles de composición microbiana. Y visualización utilizando Pavian en Rstudio. Además, se usa bracken para calcular la abundancia, misma que permite el análisis y visualización alfa y beta mediante Rstudio utilizando paqueres como vegan, phyloseq y ggplot2.

Finalmente, se lleva a cabo el **análisis funcional**, ensamblaje con MEGAHIT, predicción de genes con Prodigal y la anotación funcional con eggNOG-mapper annotate, para identificar procesos biológicos y rutas metabólicas enriquecidas. Este paso permite relacionar los patrones observados con funciones específicas y generar conclusiones relevantes para el estudio. Antes de concluir, se realiza una **validación final** para comprobar la consistencia de los resultados.

Diseño de estudio

Es un estudio observacional, principalmente de carácter descriptivo y analítico. Además, se enfoca en la aplicación de herramientas bioinformáticas (como Kraken, Rstudio y sus librerías, Prodigal, donde permite el análisis entre nueve muestras de un estudio de análisis metagenómico de las aguas termales de Kirishima, haciendo énfasis al microbioma acuático frente al cambio climático.

Variables

Se basa en variables cualitativas nominales, puesto que cualifica herramientas bioinformáticas en microbiomas, indistintamente de su entorno (biosfera, suelo, océano, entre otros) que no tienen un orden jerárquico intrínseco.

Criterios de inclusión y exclusión

Criterios de inclusión

Se incluyeron artículos científicos de alto impacto de análisis metagenómicos de microbioma acuático, donde apliquen herramientas bioinformáticas su análisis de datos frente al cambio climático.

Criterios de Exclusión

Se excluyeron artículos científicos donde hablen del microbioma terrestres y aéreos, sin relación directa con el cambio climático, donde no especifican las herramientas bioinformáticas utilizadas en el área de estudio.

Resultados y Discusión

Visualización de Análisis Taxonómico

Tabla 2

Pavian metagenomic data explorer (Dominio entre muestras)

Name	Rank	TID	Max	DRR163685	DRR163686	DRR163687	DRR163688	DRR163689	DRR163690	DRR163691	DRR163692	DRR163693
Bacteria	D	2	4068600	1359420	2903371	4068629	129940	54048	50562	55210	80301	122267
Archaea	D	2157	1543800	573158	64450	1242881	1543794	279516	159249	81324	311846	623517
Eukaryota	D	2759	17308	14000	14037	10429	4038	9088	17308	1252	4789	6499
Viruses	D	10239	1825	216	1825	612	121	136	209	186	151	226

Gracias a Pavian, se pudo visualizar la taxonomía de los datos metagenómicos de las nueve muestras en entornos marinos de este estudio, donde se analiza la presencia de microorganismos en todas las muestras, destacándose principalmente Bacterias y Archaeas y un leve grupo de Eucariotas. Se evidencia mayor prevalencia y/o dominio de bacterias en las muestras DRR163685, DRR163686 (casi 3 millones) y DRR163687 (más de 4 millones de lecturas). Sin embargo, a partir de la muestra DRR163688 existe mayor dominio de Archaeas, sin mencionar que, DRR163687 también presenta números considerables. Esta comparativa entre las bacterias y archaeas resulta interesante debido a una posible relación de exclusión o de nicho. Las bacterias son sumamente abundantes en muestras muy puntuales, sin embargo, las Archaeas, aunque presentes en todas las muestras, donde las bacterias tienen menor prevalencia, no se considera un grupo dominante. Esto infiere en que las condiciones de cada sitio favorecen a un dominio sobre el otro en el bioma marino, estableciendo una transición ecológica, donde puede influir algunos factores como temperatura, pH, salinidad, etc. llevando a una competencia entre comunidades microbianas para su supervivencia y/o prevalencia. Por otro lado, la presencia minoritaria de eucariotas y virus en todas las muestras en comparación con las bacterias y archaeas, podría representar cierta contaminación de bajo nivel o simplemente, en términos de biomasa o cantidad, son organismos minoritarios de este ecosistema en particular.

A pesar de que se evidencie poca presencia o adaptabilidad de los virus en este entorno; un estudio elaborado por (Anderson et al., 2014), destaca la evolución de estos microorganismos donde se evidencia la transferencia horizontal de genes de los virus hacia las arqueas y bacterias, formando virus archaeas y virus bacterianos en condiciones extremas de ventilación hidrotermal, destacando significativamente mayor abundancia y adaptabilidad en virus arqueales. Este estudio da un realce a los virus frente a diversas comunidades microbianas y su evolución en entornos totalmente extremos.

Figura 1

Rstudio: Diagrama de Sankey (DRR163685 - Kinyu)

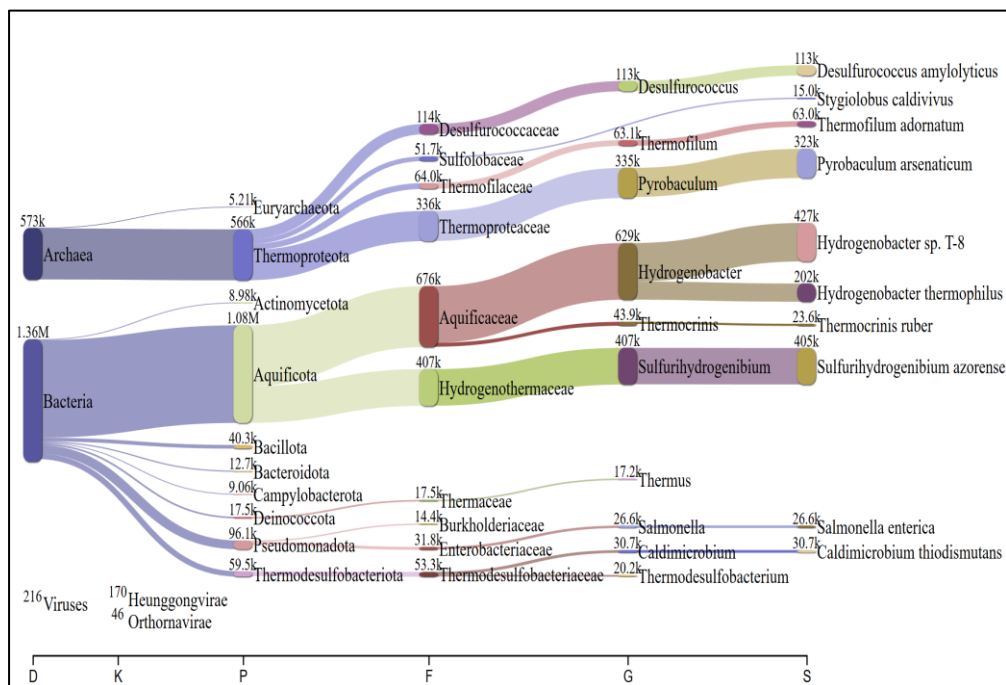


Figura 2

Rstudio: Diagrama de Sankey realizado por Pavian (DRR163686 - Yamanoshiro)

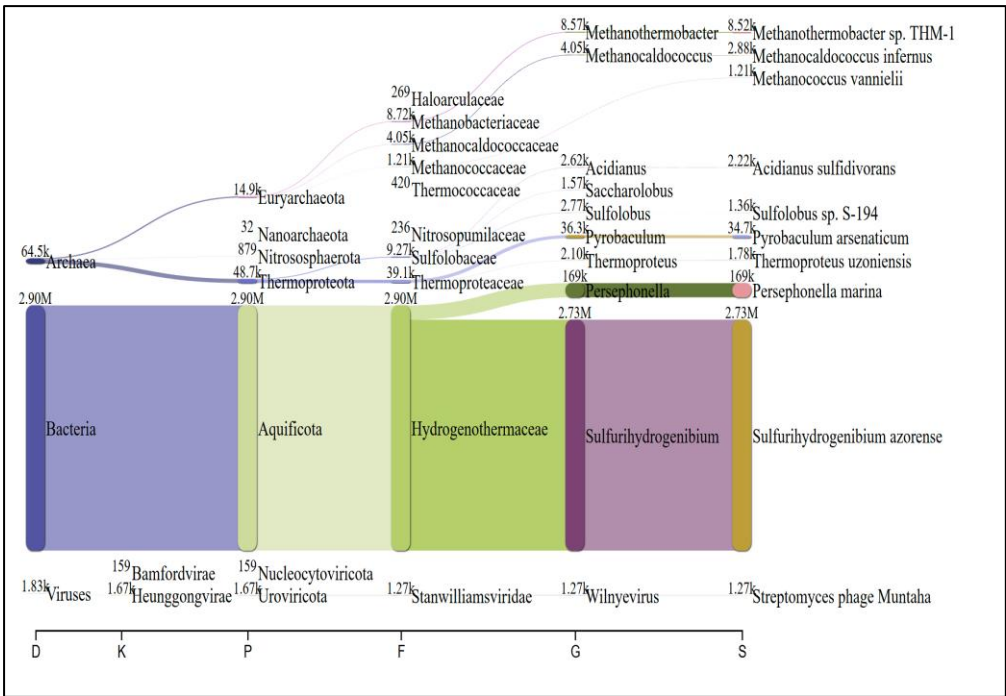


Figura 3

Rstudio: Diagrama de Sankey realizado por Pavian (DRR163687 - Nonoykoya)

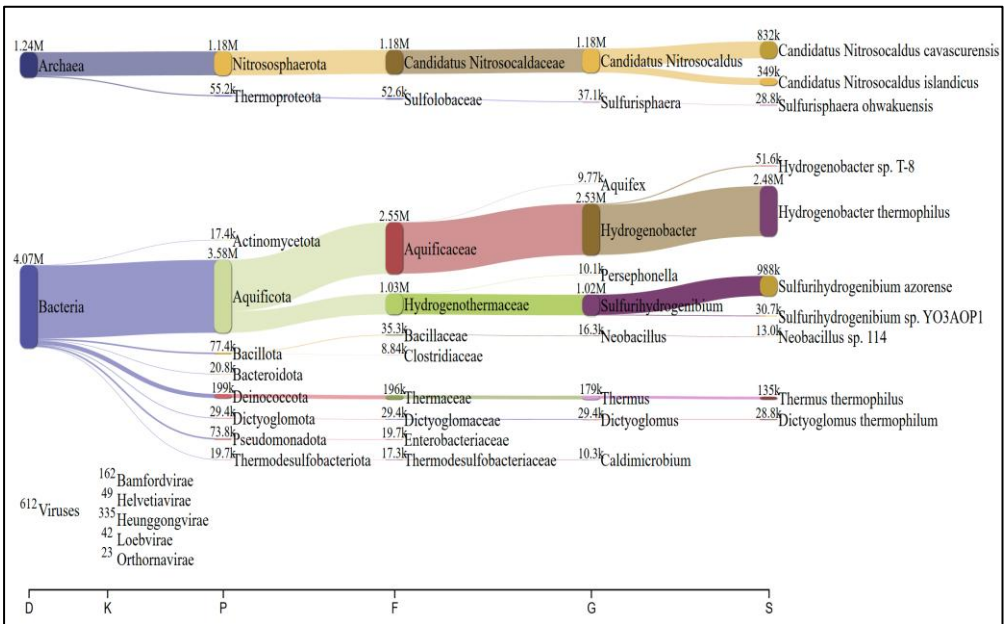


Figura 4

Rstudio: Diagrama de Sankey realizado por Pavian (DRR163688 - Miyama)

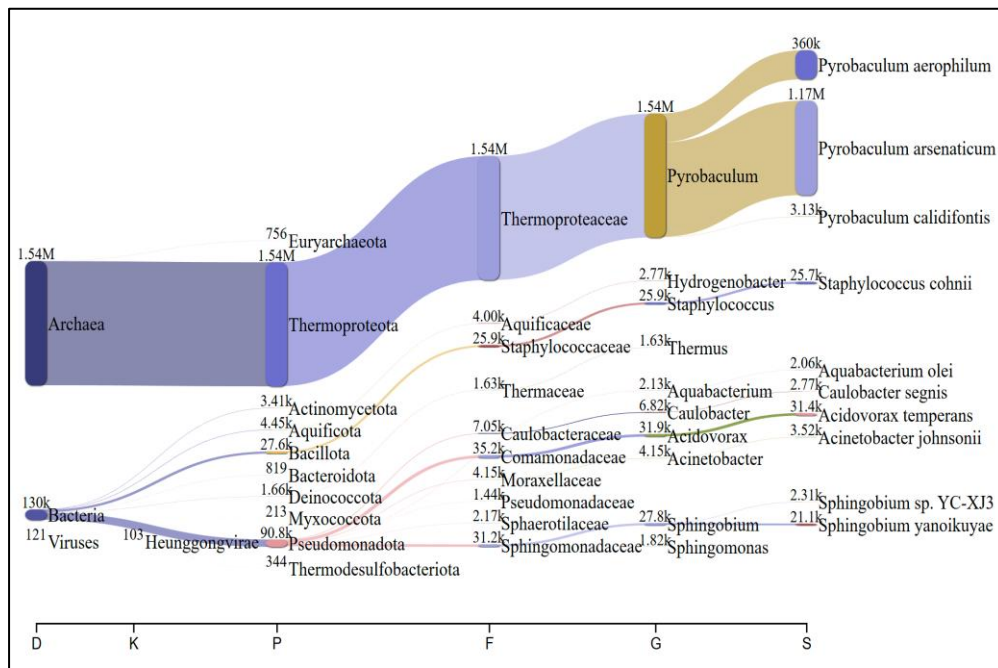


Figura 5

Rstudio: Diagrama de Sankey realizado por Pavian (DRR163689 - Kurinodake)

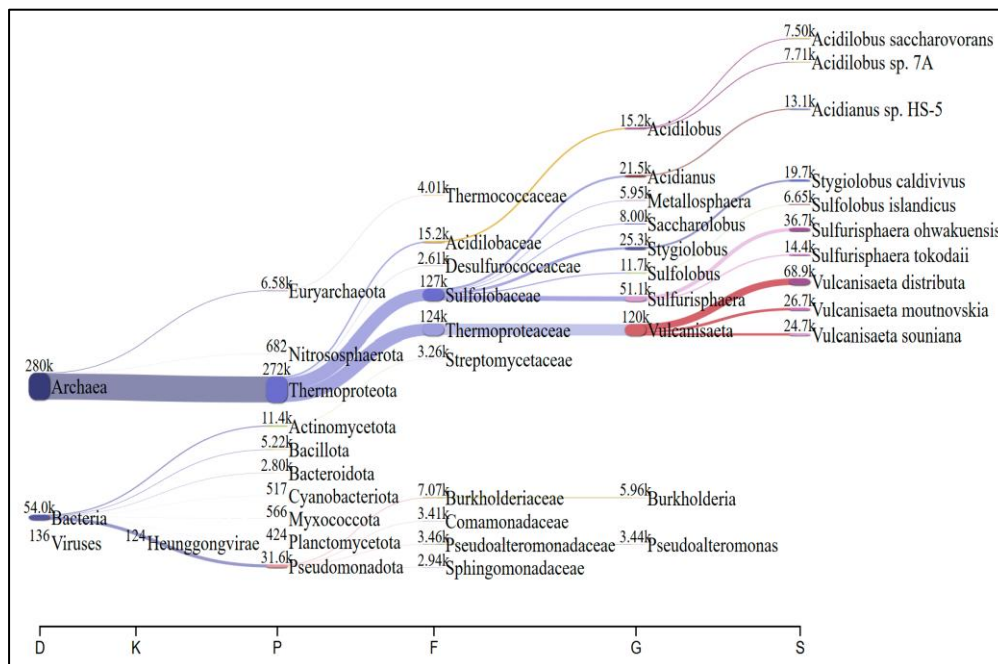


Figura 6

Rstudio: Diagrama de Sankey realizado por Pavian (DRR163690 - loudani)

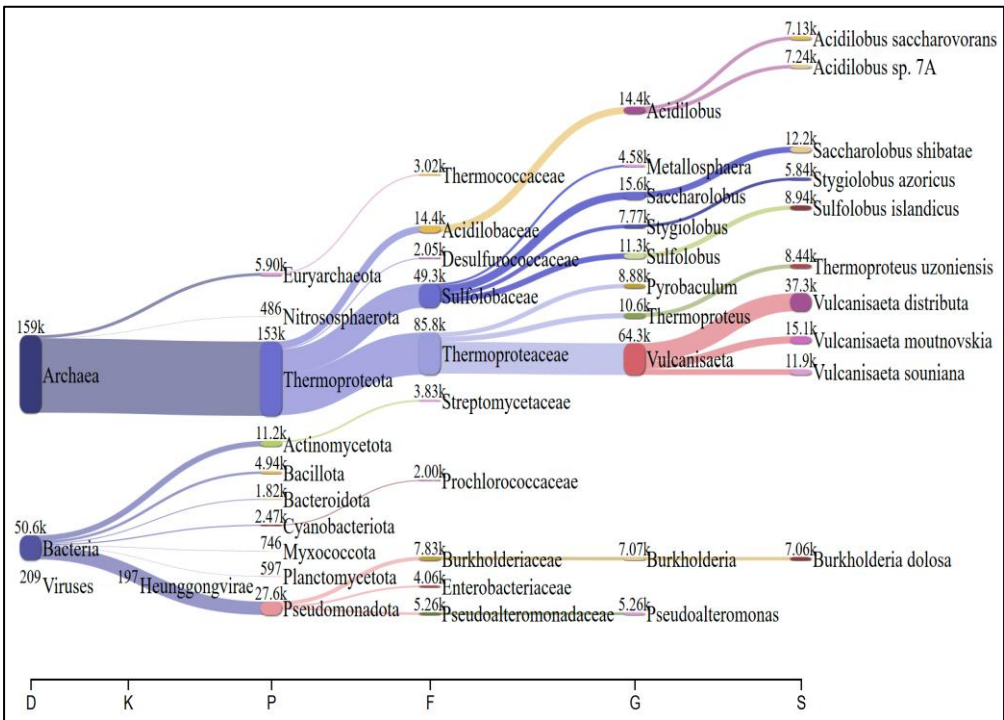
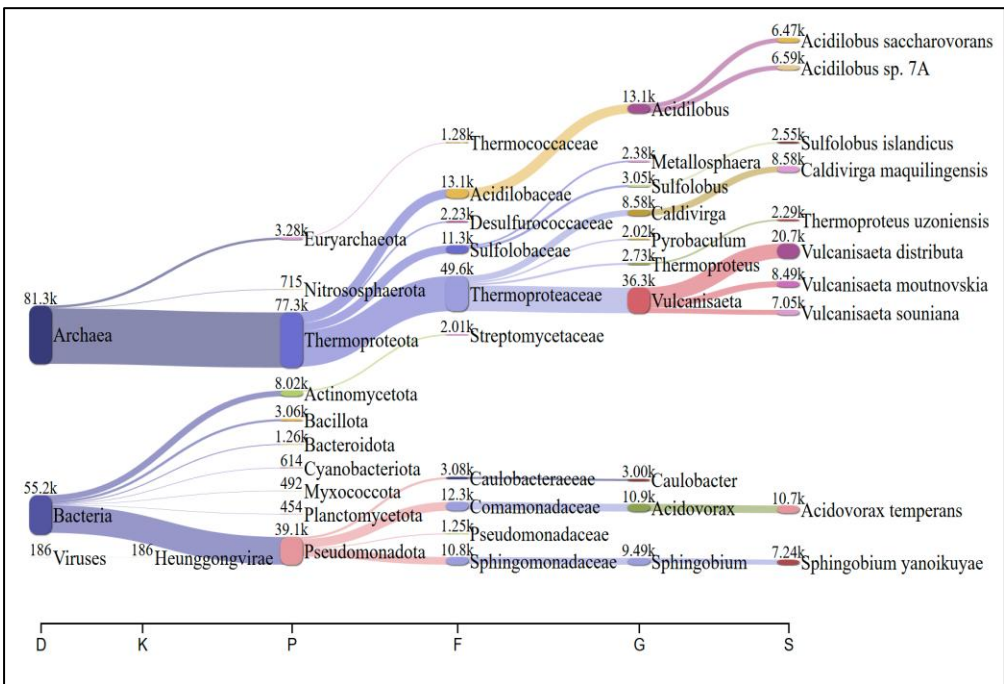


Figura 7

Rstudio: Diagrama de Sankey realizado por Pavian (DRR163691 - Torigigoku)



Rstudio: Diagrama de Sankey realizado por Pavian (DRR163692 – Ginyu)

[illegible]

Figura 1,2,3 (Dominio de Bacterias): En las primeras tres muestras, se evidencia predominio en comunidades bacterianas, donde el filo principal es Aquificotas, conocidas principalmente por su estabilidad y adaptación en temperaturas elevadas. Además, se evidencias dos géneros principales como *Hydrogenobacter* y *Sulfurihydrogenibium*

Figura 4-10 (Dominio Archaeas): En la muestra DRR163688, se observa que el filo dominante es Thermoproteota y el género dominante es Pyrobaculum, archaea mejor adaptada a un entorno aún más extremo de lo habitual. Se evidencia cierto tipo de evolución de archaeas a lo largo de las muestras restantes. Las muestras DRR163689 hasta DRR163692 dominadas por géneros *Vulcanisaeta*, *Thermoproteus* y *Caldivirga*. Sin embargo, en la figura 9, la muestra DRR163693 presenta un cambio en su composición, con géneros como *Metallosphaera* y *Sulfolobus* siendo los más prevalentes a este entorno.

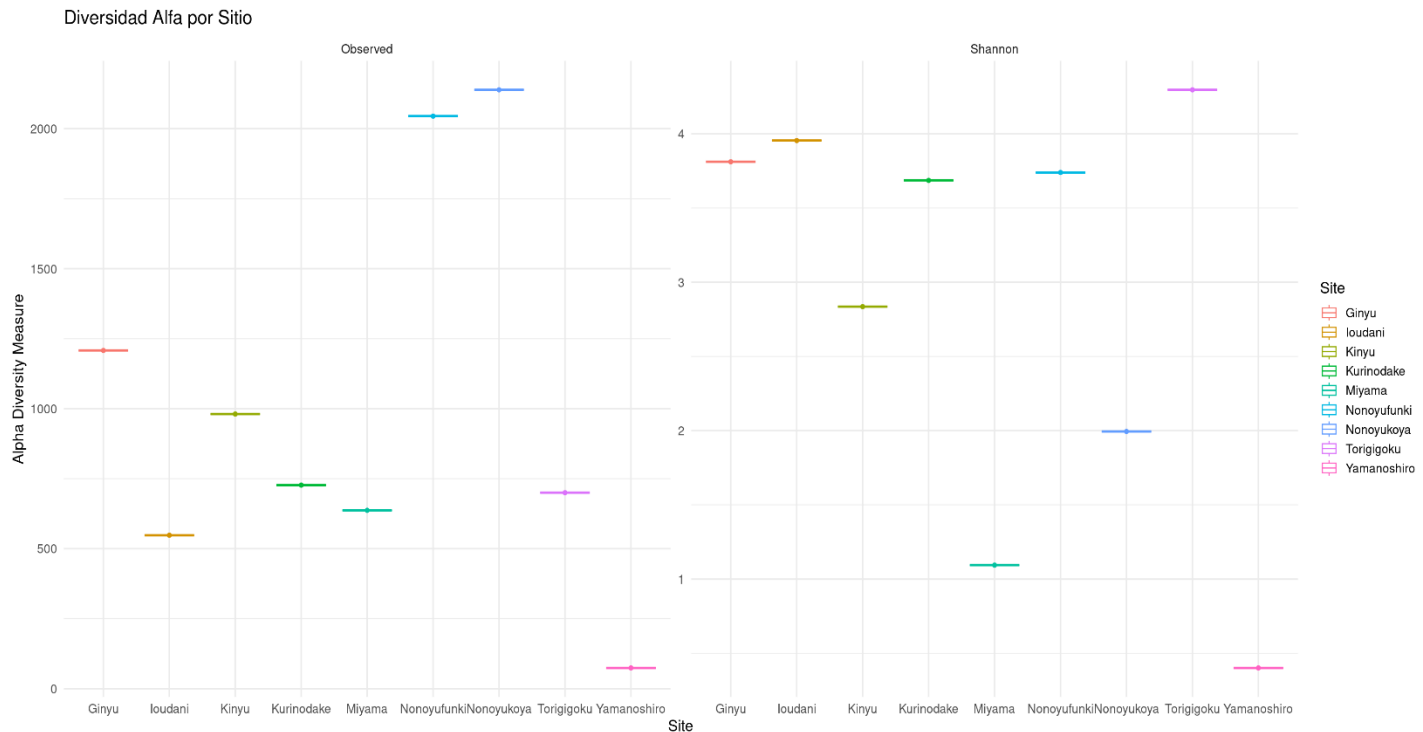
Un estudio en Kirishima, (Satoh et al., 2023) menciona que, las *Thermoproteus* o también llamadas *Crenarchaeotas*, son microorganismos aerobios, que prevalecen en aguas termales sulfurosas, mientras que, las *Aquificotas* al ser anaerobias pueden adaptarse a diferentes entornos, con un pH medio. Con esto da a entender que ambas comunidades pueden coexistir en un mismo entorno a pesar de sus diferencias, permitiendo contribuir en el ciclo de nutrientes y brindando estabilidad en el ecosistema.

Otro estudio en el desierto polar hiperárido, manifiesta ausencia de arqueas, por lo que sugieren que estos microorganismos no pueden tolerar el estrés térmico en ambientes xéridos extremos, eso puede explicar su incapacidad de sobrevivir en este entorno. Sin embargo, en suelos antárticos debido a su heterogeneidad ambiental, se puede presenciar un poco de arqueas, puesto que, en este entorno expresan menos estrés térmico debido a la humedad. En contraste con las bacterias, donde se evidencia una diversidad baja pero significativa en este tipo de suelos, sugiriendo mayor adaptabilidad que las arqueas.

Análisis de diversidad Alfa

Figura 10

Análisis de diversidad Alfa (Rstudio)



Este análisis comprende la diversidad de especies microbianas dentro de un mismo hábitat, donde evalúa dos parámetros importantes: La riqueza de las especies y el índice de Shannon. En la gráfica, quien presenta mayor riqueza de especies microbianas es en la muestra de Nonoyufunki y Nonoyukoya, con más de 2000 especies observadas. Ginyu no se queda atrás, con 1200 especies observadas. Por otro lado, quien tiene menor riqueza es la muestra de Yamanoshiro, con índices extremadamente bajos, menor a 100 especies observadas y el índice de Shannon también se presenta en menor escala.

Nonoyufunki en el índice de Shannon se mantiene, lo que indica que las comunidades están en perfecto equilibrio al igual que Ginyu.

Nonoyukoya de mayor riqueza, pasa a un índice de 2.0, lo que significa poca diversidad en el entorno. Sin embargo, resulta interesante que, loudani, Kurinodake, y Torigigoku

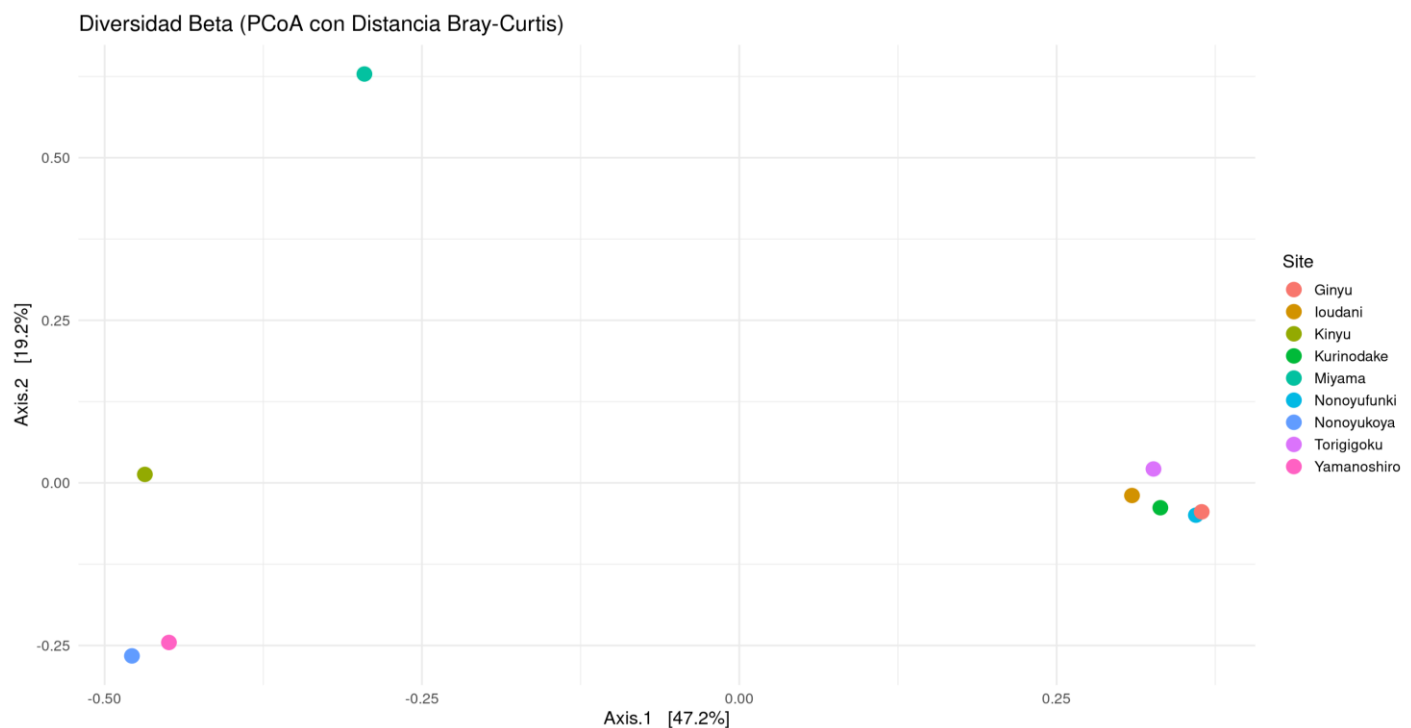
presenten una riqueza entre 500 a 750 especies y su índice de Shannon sea relativamente alto, siendo estos lo más cercanos a 4.0. lo que indica mayor diversidad en el entorno.

Lo interesante de este análisis es que a pesar de que Nonoyukoya represente muchísimas especies superando a las demás muestras, son pocas las especies dominantes en el entorno y el resto de ellas pueden ser raras, por tanto, su diversidad es baja. A diferencia de loudani y Kurinodake y Torigigoku, las especies dentro de estos entornos se encuentran distribuidas de una manera más uniforme, lo que sugiere mayor equilibrio en el entorno. En conclusión, este análisis de diversidad alfa nos indica que los sitios varían enormemente en su riqueza y estructura comunitaria. Algunos son ricos pero dominados por pocas especies, mientras que otros son más pobres, pero ecológicamente más equilibrados.

Análisis de diversidad Beta

Figura 11

Análisis de diversidad Beta (Rstudio)



Nonoyukoya y Yamanoshiro son entornos con microorganismos muy similares entre sí

que comparten las mismas especies, siendo prevalentes las bacterias. A pesar de que Nonoyukoya mostró mayor riqueza, ambas muestras tienen baja diversidad, lo que corrobora su similitud entre especies.

Miyama se presenta totalmente distinto al resto de muestras, además, mantiene baja diversidad, lo que sugiere que los microorganismos en este entorno no pueden mantenerse en condiciones extremas, o pueden influir factores ambientales que no les permita a las arqueas adaptarse con facilidad.

Ioudani, Kurinodake, Torigigoku, Nonoyufunki y Ginyu son entornos que tienen alta diversidad, por lo que están en condiciones equilibradas. Aunque no tienen la mayor cantidad de especies, las que están presentes coexisten en abundancias similares, lo que podría indicar un ecosistema más estable.

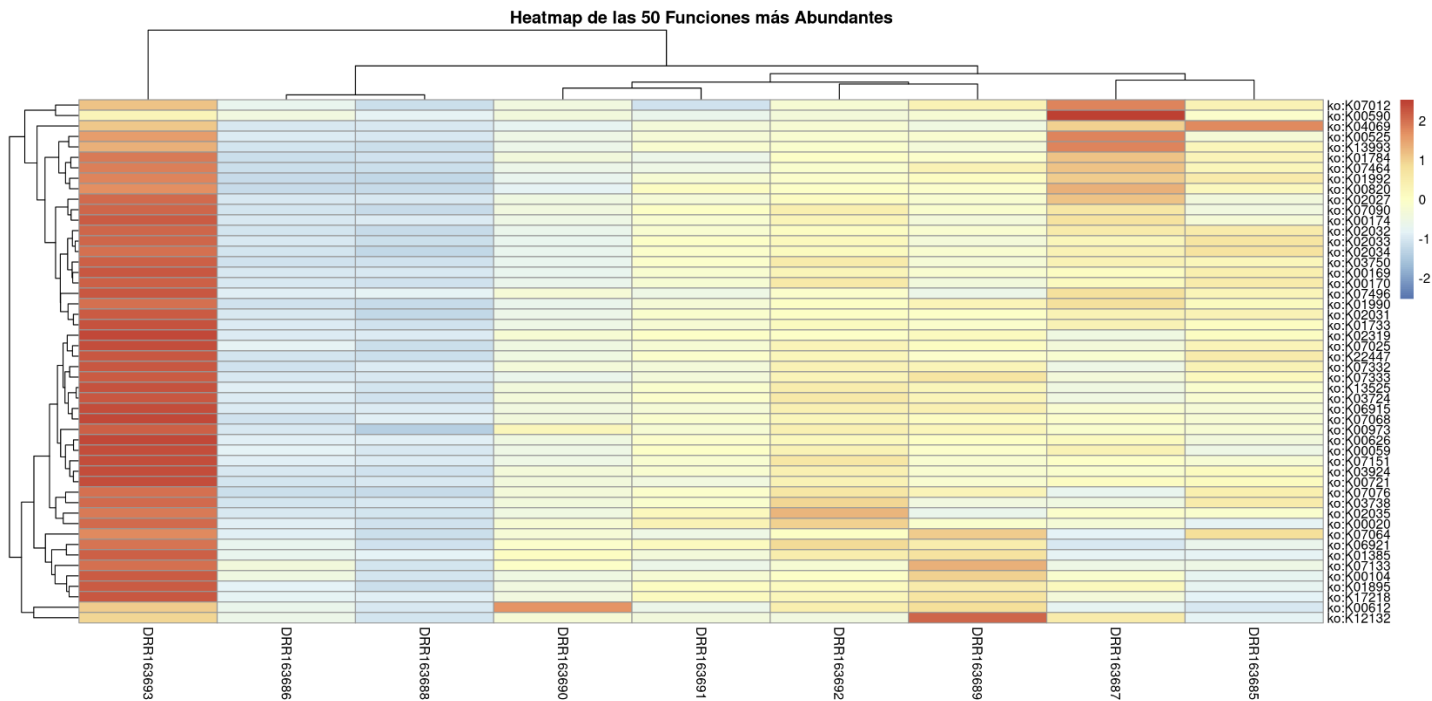
En conclusión, los análisis presentados evidencian que las diferencias entre los entornos de muestreo no se limitan a variaciones en la riqueza de especies. Más bien, las diferencias fundamentales radican en la composición taxonómica y la estructura comunitaria, incluyendo la abundancia relativa de los taxones. Todo indica que estos factores son un reflejo directo de las condiciones ambientales de cada entorno como su pH, temperatura o nutrientes. Cada entorno, con sus características únicas, da forma a un tipo particular de comunidad biológica, explicando porqué son tan distintas entre sí.

(Jurdzinski et al., 2023) estudiaron distintos biomas acuáticos, como entornos de agua dulce, salada y salobre. Analizaron la presencia de comunidades bacterianas, genéticamente distintas entre sí y separadas en distintos entornos. Los autores escatiman que se debe a la salinidad y geografía del bioma acuático, evidenciando mayor adaptabilidad de bacterias en el entorno salobre, es decir, que las bacterias se adaptan con facilidad a entornos neutros, ni tan dulces, ni tan salados. A su vez, nivel geográfico evidencian diversidad en distintas partes del mundo, por lo que impiden unirse entre sí.

Análisis Funcional

Figura 12

Análisis Funcional (Heatmap - Rstudio)



Mediante el presente mapa de calor o heatmap, se logra visualizar las 50 funciones más abundantes y comunes entre las 9 muestras. Los KEGG Orthology (KO) son genes asociados a rutas metabólicas, por lo que explica qué tipo de función tienen en cada uno de ellos. Dentro de la escala de colores, se evidencia el rojo intenso (que comprende una sobreexpresión o alta abundancia), amarillo (cuando la abundancia es promedio o cercana a cero), azul (indica baja abundancia).

Se puede evidenciar que el clúster jerárquico agrupa por columnas según su similitud. Como por ejemplo la muestra DRR163686 y DRR163688 se encuentran agrupadas como baja abundancia, de igual forma la muestra DRR163690 y DRR163691. DRR163692 con la DRR163689 (con una abundancia promedio) y finalmente la DRR163687 con la DRR163685 (relacionada con sobreexpresión en ciertas zonas)

Se evidencia que, el identificador K07012 está involucrado en la mayoría de las muestras, especialmente en las muestras DRR163693 (Nonoyufunki) y DRR163687 (Nonoykoya). Su ortología se relaciona a un sistema de defensa en procariontes, asociado a CRISPR Cas3. Un estudio de (Moody et al.,2024) revela su presencia en LUCA (último ancestro común universal), haciendo énfasis en su capacidad para generar energía a nivel celular, es decir, LUCA poseía características metabólicas complejas, lo que generó una supervivencia en el entorno primitivo.

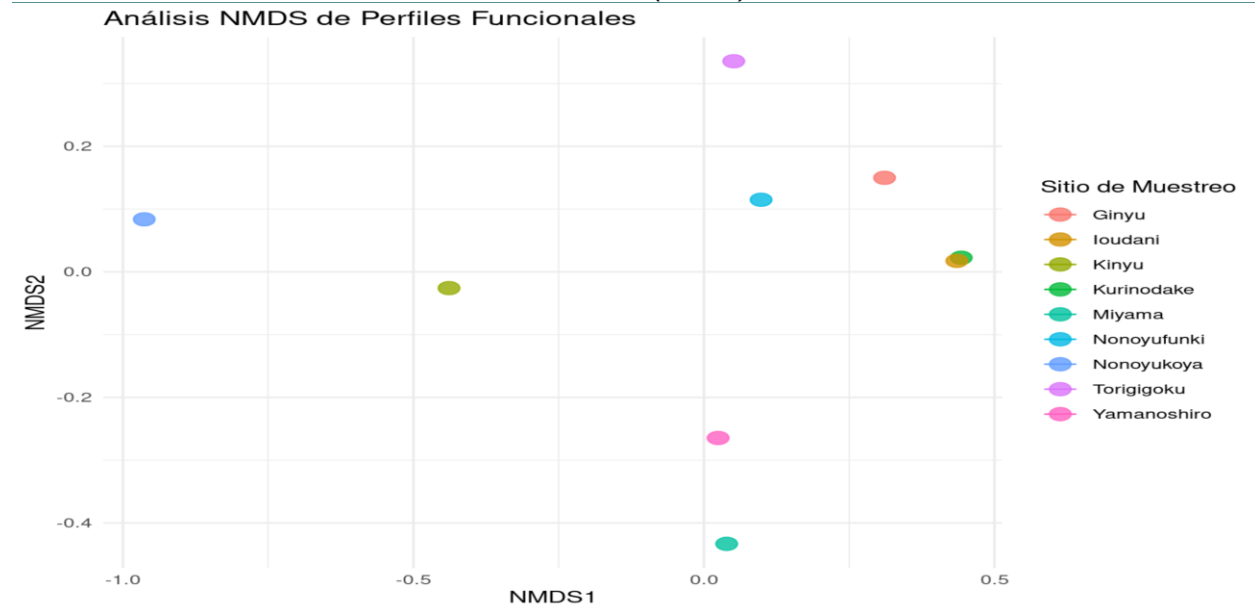
Al igual que el K00590, también se ve involucrado en el sistema de defensa celular. (Seong et al., 2022) sugiere que este identificador participa activamente en procesos de metilación celular, ayudando a la regulación del ciclo celular y combatir contra virus.

Otro identificador de importancia es el K00525, puesto que comprende rutas metabólicas de purinas y pirimidinas, esenciales para la replicación y función celular bacteriana. A pesar de esto, un análisis metagenómico de (Enav et al., 2014) pone en relevancia al virus como invasor, beneficiándose de estas rutas metabólicas bacterianas, convirtiendo a las bacterias en excelentes hospedadoras.

No podía faltar el identificador K13993, responsables de proteger las células del choque o estrés térmico. Una investigación por (Xie et al., 2024) corrobora la información resaltando a los miembros de la familia HSP20 en cuanto a la supervivencia frente a altas temperaturas. El K13993 no necesita un activador, para mantener resistencia térmica de los microorganismos frente a condiciones extremas.

Figura 13

Análisis de Escalamiento Multidimensional no Métrico (NMDS)



El análisis de escalamiento multidimensional no métrico permite visualizar que muestras son similares o diferentes entre sí. Los puntos más cercanos, son los más similares entre sí, sin embargo, los puntos más lejanos son los más diferentes entre las muestras.

Analizando la gráfica, se logra evidenciar que, Ioudani y Kurinodake son funcionalmente casi idénticos entre sí, están relativamente superpuestos.

Se puede evidenciar que, Ginyu y Nonoyufunki muestran una alta similitud funcional entre sí y con el par Ioudani/Kurinodake, formando un grupo compacto en el cuadrante superior derecho.

Nonoyukoya es el más diferente de todos, por lo que, su posición aislada en el extremo izquierdo del gráfico indica que tiene un perfil funcional único y muy distinto al resto de las muestras. Por otro lado, Yamanoshiro y Torigigoku también se separan del grupo principal, aunque no de forma tan extrema como Nonoyukoya, lo que también sugiere que tienen perfiles funcionales particulares.

Conclusión

El uso de herramientas bioinformáticas da un salto a la ciencia, ya que permite un análisis más profundo y robusto de metadatos al momento de estudiar microorganismos frente al cambio climático. Actualmente hay muchas herramientas que permiten realizar este tipo de análisis. En este estudio se utilizó FastQC, para el control de calidad de los metadatos, MultiQC para guardar los resultados, Fastq para limpieza de datos, Kakren2 para el análisis taxonómico con visualización en Pavian, Bracken para calcular la abundancia y generar el análisis beta y alfa a través de Rstudio con paquetes como vengam, phyloseq y ggplot2 para la visualización. En el análisis funcional, se realizó el ensamblaje con MEGAHIT, predicción de genes con Prodigal y la anotación funcional con eggNOG-mapper annotate para identificar procesos biológicos y rutas metabólicas de interés

En base a los resultados, se pudo analizar la diversidad de microorganismos como bacterias, arqueas y virus. En el análisis taxonómico, se pudo evidenciar abundancia y predominio en bacterias, en tres muestras: DRR163685 (Kinyu), DRR163686 (Yamanoshiro) y DRR163687 (Nonoykoya). Sin embargo, en el resto de muestras se evidenció mayor adaptabilidad en arqueas. Resulta de interés el hecho de que a pesar que las arqueas presentaron mayor adaptabilidad en la mayoría de las muestras, las bacterias siguen siendo un grupo dominante. Se infiere que puede deberse a un estado de competencia entre comunidades microbianas para sobrevivir, mismos que pueden verse afectados por factores ambientales como temperatura, pH, salinidad, transición ecología, entre otros.

En el análisis de diversidad alfa y beta se destaca principalmente Nonoyufunki, ya que presentó alta riqueza con más de 2000 especies y además, se evidenció un equilibrio estable en el entorno. Finalmente, el análisis funcional dio un plus para conocer qué identificadores de KEGG Orthology se encuentran en cada una de las muestras, destacándose funciones relacionadas la defensa celular frente a las condiciones ambientales, identificadores relacionado en procesos de metilación celular, ayudando a la regulación del ciclo celular y

combatir virus, identificadores con rutas metabólicas de purinas y pirimidinas de interés frente a virus e identificadores resistentes al choque y estrés termino ocasionado por el cambio climático.

Bibliografía

- Anderson, R. E., Sogin, M. L., & Baross, J. A. (2014). Evolutionary strategies of viruses, bacteria and archaea in hydrothermal vent ecosystems revealed through metagenomics. *PLoS ONE*, 9(10). <https://doi.org/10.1371/journal.pone.0109696>
- Bravo, E. (2024). *Los microorganismos como sujeto de derechos. Su importancia en el origen de la vida y continuidad de los ciclos vitales*. ASTM.
- Burguener, & Germán F. (2017). *Desarrollo de herramientas bioinformáticas para la anotación de genomas* [Universidad de Buenos Aires].
https://hdl.handle.net/20.500.12110/tesis_n6264_Burguener
- Cavicchioli, R., Ripple, W. J., Timmis, K. N., Azam, F., Bakken, L. R., Baylis, M., Behrenfeld, M. J., Boetius, A., Boyd, P. W., Classen, A. T., Crowther, T. W., Danovaro, R., Foreman, C. M., Huisman, J., Hutchins, D. A., Jansson, J. K., Karl, D. M., Koskella, B., Mark Welch, D. B., ... Webster, N. S. (2019). Scientists' warning to humanity: microorganisms and climate change. In *Nature Reviews Microbiology* (Vol. 17, Issue 9, pp. 569–586). Nature Publishing Group. <https://doi.org/10.1038/s41579-019-0222-5>
- Dacal de Tomás, M. (2021). *Regulación microbiana de los efectos del cambio climático en los ecosistemas a distintas escalas espaciotemporales*. Universidad Rey Juan Carlos.
- Delgado, L. F., & Andersson, A. F. (2022). Evaluating metagenomic assembly approaches for biome-specific gene catalogues. *Microbiome*, 10(1). <https://doi.org/10.1186/s40168-022-01259-2>
- Enav, H., Mandel-Gutfreund, Y., & Béjà, O. (2014). Comparative metagenomic analyses reveal viral-induced shifts of host metabolism towards nucleotide biosynthesis. *Microbiome*, 2(1). <https://doi.org/10.1186/2049-2618-2-9>
- Jurdzinski, K. T., Mehrshad, M., Fernando Delgado, L., Deng, Z., Bertilsson, S., & Andersson, A. F. (2023). *M I C R O B I O L O G Y Large-scale phylogenomics of aquatic bacteria reveal molecular mechanisms for adaptation to salinity*. <https://www.science.org>

- Li, J., Bates, K. A., Hoang, K. L., Hector, T. E., Knowles, S. C. L., & King, K. C. (2023). Experimental temperatures shape host microbiome diversity and composition. *Global Change Biology*, 29(1), 41–56. <https://doi.org/10.1111/gcb.16429>
- Satoh, S., Tanaka, R., Yokono, M., Endoh, D., Yabuki, T., & Tanaka, A. (2023). Phylogeny analysis of whole protein-coding genes in metagenomic data detected an environmental gradient for the microbiota. *PLoS ONE*, 18(2 February). <https://doi.org/10.1371/journal.pone.0281288>
- Seong, H. J., Roux, S., Hwang, C. Y., & Sul, W. J. (2022). Marine DNA methylation patterns are associated with microbial community composition and inform virus-host dynamics. *Microbiome*, 10(1). <https://doi.org/10.1186/s40168-022-01340-w>
- Silveira, R., Silva, M. R. S. S., de Roure Bandeira de Mello, T., Alvim, E. A. C. C., Marques, N. C. S., Kruger, R. H., & da Cunha Bustamante, M. M. (2021). Bacteria and Archaea Communities in Cerrado Natural Pond Sediments. *Microbial Ecology*, 81(3), 563–578. <https://doi.org/10.1007/s00248-020-01574-x>
- Wood, D., Lu, J., & Langmead, B. (2022). *KRAKEN 2*. <https://ccb.jhu.edu/software/kraken2/>
- Xie, X., Deng, X., Chen, J., Chen, L., Yuan, J., Chen, H., Wei, C., Liu, X., & Qiu, G. (2024). Two new clades recovered at high temperatures provide novel phylogenetic and genomic insights into Candidatus Accumulibacter. *ISME Communications*, 4(1). <https://doi.org/10.1093/ismeco/ycae049>

1. ANEXOS

A. Selección y descarga de datasets, con metadatos ambientales (MGnify)

MGnify, es una plataforma bioinformática de acceso libre facilitando el análisis, la visualización y el descubrimiento de diversos tipos de datos de microbioma incluyendo la metagenómica, metatranscriptómica, amplicones y ensamblaje.

En el presente estudio, se identifican y descargan conjuntos de datos metagenómicos de biomas marinos, a través de la plataforma MGnify (EMBL-EBI), utilizando descriptores de búsqueda específicos como temperatura extrema, entorno acuático, identificando estudios que contengan muestras de ambientes en estas condiciones.

1. Ir a la página de MGnify <https://www.ebi.ac.uk/metagenomics>
2. Clic en Browse data
3. Clic en samples. Aquí encontraras un filtro de biomas. seleccionar bioma acuático.
4. En el filter biome selecciona aquatic, luego donde dice filter, escribe hot spring (zona de altas temperaturas) y selecciona las muestras que contengan metadatos. La estrategia se centró en identificar un único estudio de alta calidad (Estudio MGYS00005690) que contuviera múltiples muestras de réplica, con el fin de minimizar el batch effects y asegurar una comparación robusta.

DRS082842: DRR163693 (Nonoyufunki)

DRS082835: DRR163686 (Yamanoshiro)

DRS082836: DRR163687 (Nonoykoya)

DRS082838: DRR163689 (Kurinodake)

DRS082834: DRR163685 (Kinyu)

DRS082841: DRR163692 (Ginyu)

DRS082840: DRR163691 (Torigigoku)

DRS082839: DRR163690 (Ioudani)

DRS082837: DRR163688 (Miyama)

B. Preparación del entorno de trabajo

Se ha trabajado mediante la terminal de Linux, . Luego, se ha creado un directorio principal para el proyecto para mantener todos los archivos y resultados organizados, mediante los procesos: `Bash:mkdir tfmconda install -c bioconda sra-tools`

C. Descarga y procesamiento de datos crudos (Obtención de la ruta exacta de las muestras)

`curl"https://www.ebi.ac.uk/ena/portal/api/filereportaccession=DRR163693&result=read_r
un&fields=fastq_ftp"22. Descargar el primer archivo de los pares.`

Se requiere descargar ambos archivos porque la técnica de secuenciación utilizada se llama "Paired-End" (lectura de extremos pareados). En esta técnica, cada fragmento de ADN se lee por ambos lados.

```
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/DRR163/DRR163693/DRR163693_1.fastq.gz -P  
tfm/raw_reads/aquatic/
```

```
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/DRR163/DRR163693/DRR163693_2.fastq.gz -P  
tfm/raw_reads/aquatic/
```

Así se realiza con el resto de las muestras.

D. Verificación de archivos:

```
ls -lh tfm/raw_reads/aquatic/
```

E. Control de calidad y limpieza de datos:

La calidad de los datos crudos de secuenciación (archivos FASTQ) fue evaluada inicialmente con el programa FastQC(v0.11.9). Posteriormente, los reportes de calidad de todas las muestras fueron agregados y visualizados de forma conjunta utilizando MultiQC(v1.28) para obtener una visión general y detectar posibles anomalías o efectos de lote. Basado en los resultados del control de calidad, presentó que la calidad general de las 18 muestras era alta, con una puntuación Phred promedio >30 (figura 1) en la mayoría de las secuencias. Sin

embargo, se detectó la presencia de secuencias de adaptadores en el extremo 3' de las lecturas en un 14% de las secuencias de algunas muestras (figura2). Por ello, se procedió a un paso de limpieza y recorte.

Control de Calidad:

1.Instalación FastQC

```
conda install -c bioconda fastqc
```

Creación de carpeta para los resultados: `mkdir -p tfm/quality_control/aquatic`

Ejecutar FastQC: `fastqc tfm/raw_reads/aquatic/*.fastq.gz -o tfm/quality_control/aquatic/`

para un control de calidad de archivos de secuenciación de ADN

2.Instalación de MultiQC

Es un programa que toma los resultados de FastQC y guarda el reporte resumido en esa misma carpeta.

```
Conda install -c bioconda -c conda-forge multiqc
```

```
multiqc tfm/quality_control/aquatic/ -o tfm/quality_control/aquatic/
```

3.Limpieza de datos con fastp

```
conda install -c bioconda fastp
```

```
mkdir -p tfm/cleaned_reads/aquatic
```

```
nano clean_reads.sh
```

```
#!/bin/bash
```

Crear directorio de salida si no existe

```
mkdir -p tfm/cleaned_reads/aquatic
```

Bucle para procesar cada par de archivos

```
for file1 in tfm/raw_reads/aquatic/*_1.fastq.gz do base=$(basename $file1 _1.fastq.gz)
```

```
file2="tfm/raw_reads/aquatic/${base}_2.fastq.gz"
```

```
output1="tfm/cleaned_reads/aquatic/${base}_cleaned_1.fastq.gz"
```

```
output2="tfm/cleaned_reads/aquatic/${base}_cleaned_2.fastq.gz"
```

```
echo "Procesando muestra: $base"
```

```
fastp -i $file1 -l $file2 -o $output1 -O $output2 done
```

```
echo "Proceso de limpieza completado."
```

Script ejecutable:

```
chmod +x clean_reads.sh
```

```
./clean_reads.sh
```

4.Verificación de limpieza

```
mkdir -p tfm/quality_control/aquatic_cleaned
```

```
fastqc tfm/cleaned_reads/aquatic/*.fastq.gz -o tfm/quality_control/aquatic_cleaned/
```

```
multiqc tfm/quality_control/aquatic_cleaned/ -o tfm/quality_control/aquatic_cleaned/
```

F. Análisis Taxonómico / Diversidad

```
mkdir ~/kraken_db
```

```
cd ~/kraken_db
```

```
wget -c https://genome-
```

```
idx.s3.amazonaws.com/kraken/k2_standard_08gb_20240112.tar.gz
```

```
tar -zxvf k2_standard_08gb_20240112.tar.gz
```

Para activar el entorno: conda activate kraken_env

Ejecución de análisis taxonómico

```
nano run_kraken_pipeline.sh
```

```
#!/bin/bash
```

```
# Script para correr Kraken2 y Bracken en un lote de muestras pareadas
```

```
# Ruta a la base de datos de Kraken2
```

```
DB_PATH="/kraken_db"
```

```
# Ruta al directorio que contiene tus archivos FASTQ limpios.
```

```
INPUT_DIR="cleaned_reads"
```

```

# Número de hilos/procesadores para el análisis.

THREADS=10

# Longitud de lecturas

READ_LEN=150

# --- CREACIÓN DE DIRECTORIOS DE SALIDA ---

KRAKEN_REPORTS="results/kraken_reports"

BRACKEN_OUTPUTS="results/bracken_outputs"

mkdir -p $KRAKEN_REPORTS

mkdir -p $BRACKEN_OUTPUTS

echo "Directorios de resultados creados en la carpeta 'results/'."

echo "---"

# --- BUCLE PRINCIPAL DE ANÁLISIS ---

# El bucle buscará todos los archivos que terminan en _1.fastq.gz

# y los procesará como el archivo "forward" (R1) de un par.

for R1 in ${INPUT_DIR}/*_1.fastq.gz; do

    # Extraer el nombre base de la muestra (ej. DRR163685_cleaned)

    SAMPLE_ID=$(basename ${R1} _1.fastq.gz)

    # Definir el nombre del archivo "reverse" (R2) correspondiente

    R2="${INPUT_DIR}/${SAMPLE_ID}_2.fastq.gz"

    echo "Procesando muestra: ${SAMPLE_ID}"

#Paso 1: Ejecutar Kraken2

    echo "  Running Kraken2..."

    kraken2 --db $DB_PATH \

        --threads $THREADS \

        --paired \

        --gzip-compressed \

```

```

--report "${KRAKEN_REPORTS}/${SAMPLE_ID}.report" \
--output /dev/null \

$R1 $R2

# Usamos --output /dev/null para descartar la clasificación por lectura
# ya que solo necesitamos el archivo de reporte (--report).

#Paso 2: Ejecutar Bracken

echo "  Running Bracken..."

bracken -d $DB_PATH \

-i "${KRAKEN_REPORTS}/${SAMPLE_ID}.report" \
-o "${BRACKEN_OUTPUTS}/${SAMPLE_ID}.S.bracken" \
-w "${BRACKEN_OUTPUTS}/${SAMPLE_ID}.S.kreport" \
-r $READ_LEN \

-l S # 'S' para nivel de especie

echo "Muestra ${SAMPLE_ID} finalizada."

echo "---"

done

echo "¡Análisis completado para todas las muestras!"

Script ejecutable:

chmod +x run_kraken_pipeline.sh

./run_kraken_pipeline.sh

```

Visualización de Análisis Taxonómico (Rstudio)

Se usa Pavian, para la visualización y análisis taxonómico con metagenomas.

Instalación:

```

install.packages("remotes")

remotes::install_github("fbreitwieser/pavian")

pavian::runApp()

```

automáticamente se abrirá una ventana Pavian corriendo, listo para analizar los datos. Para ello, da clic en Use data on server. Posteriormente coloca la ruta donde se encuentran los archivos en formato S.kreport de todas las muestras. Por ejemplo (TFM\results\kraken_reports) selecciona todos tus archivos y da clic en Read selected directories.

Una vez cargado los datos, dar clic en View results. En la parte izquierda hay una sección que dice Sample. Se da click y automáticamente aparece la visualización Sankey, en table o text. Este diagrama o visualización de Sankey es una de las mejores formas de visualizar la composición taxonómica de una muestra. Te muestra cómo se distribuyen las secuencias de ADN a través de los diferentes niveles de la jerarquía biológica. De izquierda a derecha Dominio (D) à Filo (P) à Familia (F) à Género (G) à Especie (S). Además, en la sección comparison, puedes visualizar todas las muestras mediante una tabla taxonómica comparativa.

G. Análisis alfa y beta

Se debe combinar todos tus archivos *S.bracken* en una única matriz de abundancia y crear un archivo de metadatos. Para ello se realiza lo siguiente en la terminal de Ubuntu:

```
nano combine_bracken.py.  
  
#!/usr/bin/env python  
  
import os  
  
import sys  
  
import pandas as pd  
  
# --- Configuración ---  
  
# Directorio donde están tus archivos .S.bracken  
  
bracken_dir = 'results/bracken_outputs/'  
  
# Nombre del archivo de salida  
  
output_file = 'abundance_matrix.tsv'  
  
# --- Fin de la Configuración ---  
  
# Diccionario para guardar los datos: {especie: {muestra: conteo}}
```

```

abundance_data = {}

print(f"Buscando archivos .S.bracken en: {bracken_dir}")

# Recorrer cada archivo en el directorio
for filename in os.listdir(bracken_dir):

    if filename.endswith(".S.bracken"):

        sample_id = filename.replace(".S.bracken", "")

        print(f"Procesando muestra: {sample_id}")

        file_path = os.path.join(bracken_dir, filename)

        with open(file_path, 'r') as f:

            next(f) # Saltar la primera línea (el encabezado)

            for line in f:

                parts = line.strip().split('\t')

                species_name = parts[0]

                read_count = int(parts[5]) # Columna 'new_est_reads'

                # Añadir los datos al diccionario

                if species_name not in abundance_data:

                    abundance_data[species_name] = {}

                abundance_data[species_name][sample_id] = read_count

# Convertir el diccionario a un DataFrame de pandas
df = pd.DataFrame(abundance_data).fillna(0).astype(int)

# Reorganizar las columnas para que las muestras estén en el orden correcto
df = df.reindex(sorted(df.columns), axis=1)

# Guardar la matriz en un archivo separado por tabuladores
df.to_csv(output_file, sep='\t')

print(f"\n¡Listo! Matriz de abundancia guardada en: {output_file}")

```

script ejecutable:

```
chmod +x combine_bracken.py
```

```
./combine_bracken.py
```

Rstudio (vegan, phyloseq, ggplot2)

Subir los archivos obtenidos en Ubuntu (abundance_matrix) y la creación de la tabla metadata (donde van las características principales como pH, Salinidad, Sitio, etc).

Una vez subido los archivos, se realiza lo siguiente:

```
# Paso 1: Instalar y Cargar los Paquetes Necesarios
```

```
if (!requireNamespace("BiocManager", quietly = TRUE))
```

```
  install.packages("BiocManager")
```

```
if (!requireNamespace("phyloseq", quietly = TRUE))
```

```
  BiocManager::install("phyloseq")
```

```
if (!requireNamespace("ggplot2", quietly = TRUE))
```

```
  install.packages("ggplot2")
```

```
if (!requireNamespace("vegan", quietly = TRUE))
```

```
  install.packages("vegan")
```

```
cargar librerías:
```

```
library(phyloseq)
```

```
library(ggplot2)
```

```
library(vegan)
```

```
# Paso 2: Cargar tus Archivos de Datos
```

```
# Cargar la matriz de abundancia (formato: especies en filas, muestras en columnas).
```

```
abundance_table <- read.delim("abundance_matrix.tsv", sep = "\t", row.names = 1)
```

```
# Cargar la tabla de metadatos ("metadata.txt" o "metadata.tsv")
```

```
metadata_table <- read.delim("metadata.txt", sep = "\t", row.names = 1)
```

```

# Paso 3: Crear el Objeto phyloseq

# Convertimos los datos al formato que phyloseq requiere.
OTU <- otu_table(abundance_table, taxa_are_rows = TRUE)
META <- sample_data(metadata_table)

# Combinamos todo en un solo objeto phyloseq.
ps_object <- phyloseq(OTU, META)
print("Objeto phyloseq creado exitosamente:")
print(ps_object)

# Paso 4: Análisis de Diversidad Alfa

# Calculamos y graficamos la riqueza de especies (Observed) y el índice de Shannon.
# El gráfico agrupa y colorea las muestras según la columna "Site" de los metadatos.
plot_richness(ps_object, x = "Site", measures = c("Observed", "Shannon"), color =
"Site") +
  geom_boxplot() +
  theme_minimal(base_size = 14) +
  ggtitle("Diversidad Alfa por Sitio")

# Paso 5: Análisis de Diversidad Beta

# Transformamos los conteos a abundancia relativa (porcentajes).
ps_relative <- transform_sample_counts(ps_object, function(x) x / sum(x) * 100)

# Calculamos la ordenación PCoA usando la distancia de Bray-Curtis.
pcoa_bray <- ordinate(ps_relative, method = "PCoA", distance = "bray")

# Creamos el gráfico de PCoA. Cada punto es una muestra.

```

```
plot_ordination(ps_relative, pcoa_bray, color = "Site") +
  geom_point(size = 5) +
  theme_minimal(base_size = 14) +
  ggtitle("Diversidad Beta (PCoA con Distancia Bray-Curtis)")
```

H. Análisis funcional

Ensamblaje con MEGAHIT para las nueve muestras

1. Crear entorno:

```
conda create -n functional_analysis -c conda-forge -c bioconda megahit prodigal
eggnog-mapper -y
```

2. activar entorno:

```
conda activate functional_analysis
```

```
crear: mkdir -p resultados_funcionales/1_ensamblaje
```

3. crear: nano run_assembly.sh

```
#!/bin/bash
```

```
# Bucle para ensamblar cada muestra desde la carpeta cleaned_reads
```

```
for r1 in cleaned_reads/*_cleaned_1.fastq.gz; do
```

```
# Extrae el nombre base de la muestra
```

```
base=$(basename ${r1} _cleaned_1.fastq.gz)
```

```
# Define el archivo R2 correspondiente
```

```
r2=cleaned_reads/${base}_cleaned_2.fastq.gz
```

```
echo ""
```

```
echo "--- Ensamblando ${base} ---"
```

```
# Ejecuta MEGAHIT y guarda los resultados en la carpeta que ya creamos
```

```
megahit -1 ${r1} -2 ${r2} -o resultados_funcionales/1_ensamblaje/${base}_megahit --
```

```
min-contig-len 1000
```

```
done

echo ""

echo "Proceso de ensamblaje completado para todas las muestras"

script ejecutable:

chmod +x run_assembly.sh

./run_assembly.sh
```

Predicción de Genes con Prodigal

```
nano run_prodigal.sh

#Bucle para predecir genes en cada muestra ensamblada

for contigs in resultados_funcionales/1_ensamblaje/*_megahit/final.contigs.faa; do #
Extrae el nombre base de la muestra base=$(basename $(dirname ${contigs}) _megahit)

echo ""

echo "Prediciendo genes para ${base} "


# Creación del directorio de salida para este paso

mkdir -p resultados_funcionales/2_prediccion_genes


# Ejecutar Prodigal en modo metagenoma

prodigal -i ${contigs} -a resultados_funcionales/2_prediccion_genes/${base}_proteinas.faa -p
meta

done

echo "" echo "--- ¡Predicción de genes completada para todas las muestras! ---"


Script ejecutable:

chmod +x run_prodigal.sh

./run_prodigal.sh
```

Anotación Funcional con eggNOG-mapper annotate

conda activate functional_analysis

nano run_annotation.sh

```
#!/bin/bash
```

```
# Bucle para anotar los archivos de proteínas de cada muestra
```

```
for proteinas in resultados_funcionales/2_prediccion_genes/*_proteinas.faa; do
```

```
    # Extrae el nombre base de la muestra
```

```
    base=$(basename ${proteinas} _proteinas.faa)
```

```
    echo ""
```

```
    echo "--- Anotando funciones para ${base} ---"
```

```
    # Crea el directorio de salida para este paso
```

```
    mkdir -p resultados_funcionales/3_anotacion_eggnog
```

```
    # Ejecuta eggNOG-mapper
```

```
    emapper.py -i ${proteinas} \
```

```
        -m diamond \
```

```
        --data_dir ~/db/eggnog/ \
```

```
        --output resultados_funcionales/3_anotacion_eggnog/${base} \
```

```
        --cpu 12 # Usa 12 hilos, ajústalo según tu PC
```

```
done
```

```
echo ""
```

```
echo "--- Anotación funcional completada para todas las muestras ---"
```

Script ejecutable:

```
chmod +x run_annotation.sh
```

```
./run_annotation.sh
```

Combinación de los Resultados en una Matriz de Abundancia

Instalación de pandas: `pip install pandas`

```
nano crear_matriz.py
```

```
import os
```

```
import pandas as pd
```

```
from collections import Counter
```

```
# Ruta a la carpeta con los resultados de eggNOG
```

```
annotation_dir = "resultados_funcionales/3_annotacion_eggnog/"
```

```
# Diccionario para guardar los conteos de cada muestra
```

```
all_ko_counts = {}
```

```
print("Leyendo archivos de anotación...")
```

```
# Itera sobre cada archivo en el directorio
```

```
for filename in os.listdir(annotation_dir):
```

```
    if filename.endswith(".emapper.annotations"):
```

```
        sample_name = filename.split('.')[0]
```

```
        print(f"Procesando muestra: {sample_name}")
```

```
        filepath = os.path.join(annotation_dir, filename)
```

```
        # Lista para guardar los KOs de esta muestra
```

```
        sample_kos = []
```

```
        with open(filepath, 'r') as f:
```

```
            for line in f:
```

```

# Ignora las líneas de comentario que empiezan con '#'
if not line.startswith('#'):
    columns = line.strip().split('\t')

    # La columna de KEGG KOs es la número 11 (índice 10)
    if len(columns) > 11 and columns[11] != '-':
        # Un gen puede tener varios KOs, separados por comas
        kos = columns[11].split(',')
        sample_kos.extend(kos)

# Cuenta la frecuencia de cada KO en la muestra
all_ko_counts[sample_name] = Counter(sample_kos)

# Convierte el diccionario de conteos en una tabla (DataFrame) de pandas
df = pd.DataFrame(all_ko_counts).fillna(0).astype(int)

# Guarda la matriz final en un archivo CSV
output_file = "ko_abundance_matrix.csv"
df.to_csv(output_file)

print(f"\n¡Listo! Matriz de abundancia guardada en: {output_file}")
print(f"La matriz tiene {df.shape[0]} funciones (KOs) y {df.shape[1]} muestras.")

```

python3 crear_matriz.py

Análisis y Visualización en R

Paso 1: Cargar librerías

```
library(vegan)
```

```
library(ggplot2)
```

Paso 2: Cargar datos

```
ko_matrix <- read.csv("ko_abundance_matrix.csv", row.names = 1)
```

```
metadata <- read.table("metadata.txt", header = TRUE, row.names = 1, sep = "\t")
```

```

# Paso 3: Preparar datos

ko_matrix_t <- t(ko_matrix)

ko_matrix_t <- ko_matrix_t[rownames(metadata), ]

# Paso 4: Realizar NMDS

set.seed(123)

nmDS <- metaMDS(ko_matrix_t, distance = "bray")

# Extraer coordenadas

nmDS_points <- as.data.frame(scores(nmDS)$sites)

nmDS_points <- merge(nmDS_points, metadata, by = "row.names")


# Paso 5: Crear el gráfico

ggplot(nmDS_points, aes(x = NMDS1, y = NMDS2, color = Site)) +

  geom_point(size = 4, alpha = 0.8) +

  labs(title = "Análisis NMDS de Perfiles Funcionales",

        x = "NMDS1",

        y = "NMDS2",

        color = "Sitio de Muestreo") +

  theme_minimal() +

  stat_ellipse(aes(group = Site), type = 't') # Agrupa las elipses por 'Site'

```

Creación de heatmap

```

if (!require("pheatmap")) install.packages("pheatmap")

library(pheatmap)

# Selecciona las 50 funciones (KOs) más abundantes para visualizarlas

top50_kos <- names(sort(rowSums(ko_matrix), decreasing = TRUE)[1:50])

heatmap_matrix <- ko_matrix[top50_kos, ]

# Crea el heatmap

```

```

pheatmap(heatmap_matrix,

        scale = "row", # Escala los valores por fila (función) para ver patrones

        cluster_cols = TRUE, # Agrupa las muestras por similitud

        cluster_rows = TRUE, # Agrupa las funciones por similitud

        main = "Heatmap de las 50 Funciones más Abundantes")

# Guarda el gráfico

ggsave("Heatmap_funcional.png", width = 10, height = 12)

```

Gráfico 1.

FastQC. Puntuaciones de calidad por secuencia

