UNIVERSIDAD DE CONCEPCIÓN

Dirección de postgrado

Facultad De Ingeniería – Programa de Magister en Ciencias de la Ingeniería, Mención Ingeniería Civil

ANÁLISIS EXPLORATORIO DE LA COMPOSICIÓN Y DIVERSIDAD DE LAS COMUNIDADES BACTERIANAS PRESENTES EN RÍOS DE LA ZONA CENTRO SUR DE CHILE

Tesis presentada a la Facultad de Ingeniería de la Universidad de Concepción
para optar al grado académico de magíster en ciencias de la
ingeniería con mención en ingeniería civil

POR: JOAQUÍN ALFONSO AGONI AGONI

PROFESORES GUÍA: DR. MARCELO ANDRÉS AYBAR LAGOS

DR. OSCAR EDUARDO LINK LAZO

Marzo, 2024
Concepción (Chile)

A mi madre, Yasna Agoni Moreno, y mis abuelos,
Armando Agoni Uribe y Norma Moreno Guzmán

AGRADECIMIENTOS

El autor agradece a su profesor guía Marcelo Aybar, por su tiempo, apoyo y disposición, y al profesor Oscar Link, por su disposición y contribución en las etapas iniciales, en concreto, en la materialización de la campaña de terreno y la definición del tema de tesis.

De igual forma, considera necesario agradecer a Esteban Flores y René Iribarren, por la ayuda que brindaron durante la campaña de terreno, y a Pedro Arriagada, por su trabajo en el relleno de las estadísticas fluviométricas.

También agradece a Iván Ñancucheo, por la extracción del material genético desde las muestras de agua y de lecho, y a Danilo Pantoja y Felipe Melis, por la colaboración de estos en el procesamiento y en los análisis preliminares de los datos genéticos.

RESUMEN

Las bacterias son un componente fundamental de los ecosistemas acuáticos, estas participan activamente en los ciclos biogeoquímicos, forman parte de la cadena trófica e influyen significativamente en la calidad del agua. Por lo anterior, sumado a sus cortos tiempos de vida y alta sensibilidad con su entorno, varios grupos de investigación a nivel internacional han centrado sus esfuerzos en integrar las comunidades bacterianas en las campañas de monitoreo y seguimiento ambiental de ríos y lagos. Sin embargo, el conocimiento de líneas base con respecto a estas comunidades en ambientes naturales y sus interacciones con agentes externos es aún muy limitado, lo que dificulta su uso como bioindicadores. En esta tesis se analizaron las comunidades bacterianas planctónicas y sedimentarias presentes en siete ríos de la zona centro sur de Chile con el objetivo de conocer cómo se componen, qué diferencias existen entre comunidades de agua y de lecho, en qué condiciones ambientales habitan y qué influencia ejercen estas condiciones sobre las comunidades.

Se recolectaron muestras de agua y de lecho, en una única campaña de terreno, y se utilizó secuenciación genética de alto desempeño (HTS) para determinar la composición taxonómica de las comunidades bacterianas. Las condiciones ambientales en las que habitaban las bacterias se caracterizaron con datos de parámetros fisicoquímicos, de caudales y de usos de suelo, obtenidos de mediciones en terreno y también desde registros gubernamentales. El procesamiento de los datos genéticos, los análisis de la estructura de las comunidades, las comparaciones entre comunidades de agua y de lecho, la evaluación de la influencia de las condiciones ambientales y las pruebas estadísticas se realizaron en el entorno del lenguaje de programación libre R. La composición de las comunidades fue consistente con estudios previos y se reconocieron diferencias en composición y diversidad entre agua y lecho. Las comunidades de agua, pese a la variabilidad ambiental, fueron muy similares entre sí, la diversidad alfa de estas se correlacionó con el área de cuenca y la diversidad beta con variables espaciales y de calidad simultáneamente. Por el contrario, en lecho se observó gran disimilitud, lo que podría indicar mayor influencia de aspectos locales, pero no se detectaron correlaciones significativas con condiciones ambientales. Esta área de investigación muestra gran potencial de desarrollo, se espera que este trabajo pueda ser utilizado como sustento para futuros estudios.

SUMMARY

Bacteria are a key component of aquatic ecosystems, they actively participate in biogeochemical cycles, they are part of the trophic chain and significantly influence water quality. Furthermore, due to their short lifespans and high sensitivity to their environment, several international research groups have focused their efforts on integrating bacterial communities in environmental monitoring of rivers and lakes. However, the knowledge of a baseline regarding these communities in natural environments and their interactions with external agents is still very limited. This makes their use as bioindicators difficult. In this thesis, the planktonic and sedimentary bacterial communities present in seven rivers in the south-central zone of Chile were analyzed in their composition, differences between water and riverbed communities and environmental conditions they face.

Water and sediment samples were collected in a single field campaign, and high throughput sequencing (HTS) was used to determine the taxonomic composition of bacterial communities. The environmental conditions in which these communities thrived were characterized with data on physicochemical parameters, flows, and land use, obtained from field measurements and from government records. The processing of genetic data, analysis of the structure of the communities, comparisons between communities from water and riverbed, evaluation of the influence of environmental conditions and statistical tests were carried out in the R programming language environment. Observed composition of communities was consistent with previous studies and differences in composition and diversity between water and riverbed samples were identified. The water communities, despite the environmental variability, were very similar to each other, the alpha diversity of these was correlated with the basin area and the beta diversity with spatial and water quality variables simultaneously. On the contrary, great dissimilarity was observed in the riverbed samples, which could indicate a greater influence of local aspects, but no significant correlations with environmental conditions were detected. This area of research shows great potential for development, and it is expected that this work can support future lines of study.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO 1 INTRODUCCIÓN

1.1 Motivación

Los ecosistemas están deteriorándose a un ritmo alarmante producto de los impactos de la actividad humana. La degradación de los ambientes naturales ha generado una drástica reducción en la biodiversidad (variedad de genes, especies y rasgos presentes en los ecosistemas) a nivel mundial, siendo los ambientes de agua dulce aquellos que más se han visto afectados. Este declive ha repercutido negativamente en las funciones ecológicas de los ecosistemas, lo que ha reducido la resiliencia de estos frente a perturbaciones y degradado los servicios ecosistémicos que estos proveen. En el caso de los ecosistemas de agua dulce, como ríos, esta situación es preocupante, puesto que podría afectar la calidad de los recursos hídricos, vitales para la humanidad tanto por su consumo directo como para la producción de comida, con la agricultura y ganadería, actividades que requieren más del 70 % del agua dulce extraída a nivel mundial.

Los ríos son responsables del transporte y circulación de energía, nutrientes y materia orgánica entre continentes y océanos, lo que los torna piezas clave para los ciclos biogeoquímicos y responsables de sustentar la vida en las costas. Para la humanidad los ríos resultan de vital importancia, en consecuencia, la evaluación sistemática de la calidad de estos cuerpos de agua es primordial, lo que se ha traducido en la necesidad de conocer las características fisicoquímicas de estos y determinar la presencia de microorganismos patógenos. Este enfoque resulta suficiente para asegurar la calidad del recurso antes de su consumo, pero no permite evaluar la integridad ecológica de un río y, por tanto, asegurar su estabilidad en el tiempo, debido a que no proporciona información del estado de las diferentes comunidades biológicas presentes en el ambiente, ni con respecto a los procesos bioquímicos que están desarrollándose.

Las estrategias de monitoreo actualmente emplean componentes bióticos para evaluar el estado ecológico de los cuerpos de agua, denominados bioindicadores, los cuales son capaces de reflejar alteraciones en las condiciones naturales de un sistema mediante cambios en sus organismos o en sus poblaciones (Astudillo-García *et al.*, 2019; Cordier *et al.*, 2021). Peces, macroinvertebrados y algas han sido frecuentemente usados como bioindicadores y existen múltiples índices basados en estos (Fierro *et al.*, 2019). Sin embargo, hasta ahora la identificación de estos organismos se ha realizado únicamente en función de sus morfologías, lo que implica grandes consumos de tiempo y recursos, siendo necesario, además, contar con expertos para su realización. Más aún, este enfoque limita el rango de bioindicadores solo a aquellos cuya morfología sea fácilmente distinguible. En los últimos años, el auge de las tecnologías NGS (del inglés, *Next Generation Sequencing*) ha permitido secuenciar el material genético de muestras biológicas en menos tiempo (son capaces de procesar múltiples muestras simultáneamente) y con menores costos económicos (Pawloski *et al.*, 2021). Asimismo, el mayor uso de estas tecnologías ha enriquecido las bases de datos con nuevas secuencias genéticas. Además, debido al aumento generalizado en el poder de cómputo, el análisis bioinformático de las secuencias generadas puede realizarse incluso con una computadora personal. Las tecnologías NGS (también denominadas HTS, del inglés *High-Throughput Sequencing*) prometen potenciar las estrategias de monitoreo, esto al facilitar la identificación de bioindicadores, pasando de un enfoque sustentado en la morfología a uno en función de perfiles genéticos, ahorrando tiempo y recursos (Cordier *et al.*, 2021). También prometen ampliar el rango de bioindicadores disponibles y, con el análisis del ADN ambiental, prescindir de la recolección de especímenes para determinar la presencia de estos organismos en un ambiente. Estas tecnologías también posibilitan el estudio de una importante componente hasta ahora muy excluida en el monitoreo biológico: la vasta diversidad microbiana (Sagova-Mareckova *et al.*, 2020).. Dentro de esta se encuentran las bacterias, uno de los pilares de la cadena trófica en ambientes acuáticos y claves en los ciclos biogeoquímicos, motivos por los cuales estos microorganismos muestran un alto potencial para ser utilizados como bioindicadores.

Las bacterias son uno de los conjuntos más abundantes del planeta y claves en múltiples ciclos biogeoquímicos. En ríos estas juegan un papel fundamental en la regulación de la calidad del agua, en la recirculación de nutrientes y la degradación de materia orgánica (Sigee, 2005; Battin *et al.*, 2016). El carácter ubicuo de estos microorganismos, sus bajos tiempos de generación, la sensibilidad de estos con su entorno y la rápida respuesta que muestran frente a alteraciones en este, les tornan candidatos para ser usados como bioindicadores (Washington *et al.*, 2013; Astudillo-García *et al.*, 2019). No obstante, esto exige contar con un vasto conocimiento con respecto a la respuesta de estas comunidades a múltiples factores de estrés, al menos a un nivel que permita asociar alteraciones funcionales y estructurales de una comunidad a cambios en su entorno, como la presencia de cierto contaminante o el aumento en la concentración de nutrientes. También exige contar con una línea base de conocimiento con respecto a las comunidades bacterianas propias de los ríos dentro de una región espacial definida, línea que se debería construir en base a comunidades desarrolladas en sistemas prístinos (o mínimamente perturbados) y que debiese considerar la variabilidad temporal de estas. No obstante, a nivel mundial existen regiones sin un mínimo de investigaciones que permitan la construcción de esta línea base, este es el caso de Chile (Habit *et al.*, 2019; Li *et al.*, 2021).

La zona central de Chile es reconocida a nivel mundial como un *hostspot* de biodiversidad, lamentablemente esta se encuentra bajo amenaza por el incremento de la presión antropogénica, derivada de la intensificación de la actividad agrícola en la zona y el creciente proceso de urbanización (Fierro *et al.*, 2019). A lo anterior debe agregarse que Chile podría verse especialmente afectado por el calentamiento global, consecuencias como el avance de la desertificación y la reducción de caudales ponen en una situación aún más delicada a los ecosistemas de la zona central del país (Habir *et al.*, 2019). Frente a esto, se tornará necesario implementar estrategias de monitoreo que permitan un control continuo del estado ecológico de los ecosistemas, especialmente de los acuáticos, antes de que los daños sean irremediables y los servicios ecosistémicos se vean significativamente perjudicados. El uso de las comunidades bacterianas en las estrategias de monitoreo

resulta atractivo dadas las posibles ventajas con respecto de los bioindicadores tradicionales, además, cada vez se ve más probable el surgimiento e implementación generalizada de índices alrededor del mundo. No obstante, en el caso de Chile la carencia de conocimiento en torno a la diversidad microbiana de sus ríos podría significar una gran limitante en la adaptación e implementación de estos dentro del territorio nacional.

El presente estudio busca dar los primeros pasos en la construcción de una línea base de conocimiento con respecto de la diversidad bacteriana de los ríos de Chile, la cual es crucial para la implementación de futuras estrategias de monitoreo que consideren a la componente bacteriana dentro de la evaluación de la integridad ecológica de ríos, alternativa que podría resultar más económica, rápida y fácil de ejecutar que las actualmente disponibles. De igual forma, se espera que la revisión efectuada en este estudio, junto con los métodos estadísticos y herramientas computacionales usadas puedan servir a quién esté iniciándose o busque profundizar dentro de esta línea de investigación.

1.2 Preguntas de investigación

- ¿Varían espacialmente la composición y diversidad de las comunidades bacterianas dentro de la zona de estudio?
- ¿Existen diferencias entre las comunidades bacterianas planctónicas (agua) y sedimentarias (lecho) en cuanto a composición y diversidad?
- ¿En qué condiciones ambientales se están desarrollando las comunidades bacterianas encontradas?

1.3 Objetivos

1.3.1 Objetivo General

Analizar cómo se componen las comunidades bacterianas presentes en ríos de la zona centro sur de Chile, los niveles de diversidad de estas, las diferencias entre comunidades de agua y de lecho, y las condiciones ambientales en las que estas habitan.

1.3.2 Objetivos Específicos

- Determinar la composición taxonómica de las comunidades bacterianas presentes en las muestras de agua y de lecho.
- Caracterizar puntos de muestreo mediante parámetros fisicoquímicos de calidad, caudales y usos de suelo en cuencas aportantes.
- Reconocer diferencias entre comunidades planctónicas y sedimentarias con respecto a sus composiciones y diversidad.
- Evaluar la influencia de las condiciones ambientales en la composición y diversidad de las comunidades bacterianas.

1.4 Metodología de trabajo

Esta investigación inició con la planificación y realización de una campaña de terreno, en esta se recolectaron muestras de agua y de lecho para análisis genéticos y químicos, y se midieron parámetros fisicoquímicos *in-situ*. Los análisis químicos fueron realizados en los laboratorios de la UdeC, en tanto que los análisis genéticos los realizó un laboratorio externo utilizando secuenciación genética de alto desempeño (HTS). Paralelamente, se realizó una revisión bibliográfica exhaustiva enfocada en cómo trabajar con datos genéticos de comunidades bacterianas y en investigaciones que hubiesen estudiado

comunidades planctónicas y/o sedimentarias de ríos. Sumado a lo anterior, los puntos de muestreo fueron caracterizados en función de lo medido en la campaña de terreno (posición geográfica, mediciones *in-situ* y resultados de laboratorio), de lo encontrado en los registros históricos DGA (datos fluviométricos y de calidad) y de los usos de suelo en las respectivas cuencas aportantes. Para el análisis de los datos genéticos se optó por usar el lenguaje de programación libre R, el trabajo inició con el procesamiento de las secuencias brutas entregadas por el laboratorio externo, con esto se determinaron la taxonomía de las bacterias detectadas y las abundancias relativas de estas en cada muestra de agua y de lecho. Posteriormente, usando paquetes de funciones para análisis ecológicos y estadísticos se analizó la estructura de las comunidades bacterianas (composición y diversidad), se compararon las comunidades de agua y de lecho, y se evaluó la influencia de las condiciones ambientales en las características estructurales de las comunidades.

1.5 Organización de la tesis

El presente documento de tesis está constituido por cinco capítulos donde el primero de estos, el actual, corresponde al capítulo introductorio. El segundo, en tanto, se centra en la revisión bibliográfica, en este se expone con respecto al delicado estado en el que se encuentra la biodiversidad de los ecosistemas fluviales y el potencial que tienen las comunidades bacterianas dentro del monitoreo del estado ecológico de estos ambientes. El siguiente capítulo, el tercero, trata lo relativo a los materiales y métodos utilizados en esta investigación, se detalla lo relacionado con la campaña de terreno, el procesamiento de las muestras, la recopilación de datos ambientales, el análisis de las comunidades bacterianas y la relación de las características de las comunidades con las condiciones ambientales del entorno que estas habitan. En el cuarto se exponen los resultados y se presenta la discusión en torno a estos. Por último, el quinto capítulo se encuentra reservado para las conclusiones en torno a este trabajo.

CAPÍTULO 2 REVISIÓN BIBLIOGRÁFICA

2.1 La integridad ecológica de los ríos está bajo amenaza

Los ríos conectan continentes y océanos, son responsables del transporte y circulación de energía, nutrientes, materia orgánica y otras sustancias desde los primeros hacia los segundos, esto los transforma en un componente fundamental para el desarrollo de los ciclos biogeoquímicos y en un elemento clave para la mantención de la vida en las costas (Aufdenkampe *et al.*, 2011; Masotti *et al.*, 2018; Zárate *et al.*, 2020). Pese a que constituyen cerca del 0.01 % de la hidrósfera y ocupan un 0.8 % de la superficie terrestre, los sistemas de agua dulce ostentan enormes niveles de biodiversidad y son capaces de sustentar un gran número de procesos ecológicos, lo que les torna particularmente vulnerables a los efectos de la actividad humana (Dudgeon *et al.*, 2006; Balian *et al.*, 2008; Cardinale *et al.*, 2012). Debido a los múltiples servicios ecosistémicos que proveen los ríos al ser humano, estos resultan esenciales para la existencia y el desarrollo de la humanidad, sin embargo, producto de los intensivos usos que se les han dado a estos recursos, los ecosistemas acuáticos se han visto sometidos a una constante y creciente presión (Pander y Geist, 2013; Arriagada *et al.*, 2019). Con el aumento de la población mundial y el consecuente proceso de urbanización, la presión antropogénica sobre los cauces se ha incrementado considerablemente, lo cual, en conjunto con la aceleración del cambio climático, han causado el deterioro de múltiples ecosistemas acuáticos, especialmente en países en vías de desarrollo. Dada la dependencia que existe de las funciones y los servicios ecosistémicos de un ecosistema hacia la biodiversidad de este, se teme que las pérdidas en la biodiversidad afecten negativamente dichas funciones y servicios, reduciendo la resiliencia de estos ecosistemas y menoscabando la capacidad de los sistemas acuáticos para mantener la calidad de sus aguas (Cardinale *et al*., 2012; Scheffers *et al.*, 2016; Alonso *et al.*, 2017; Feio *et al.*, 2021).

2.2 Biodiversidad de un ecosistema

Esta es entendida como la variedad de genes, especies y rasgos funcionales presentes en un ecosistema, guarda estrecha relación con las funciones ecosistémicas de este y con los servicios ecosistémicos que puede proveer, generalmente un detrimento en la biodiversidad conlleva al deterioro de estas funciones y servicios (Cardinale *et al.*, 2012). En los últimos años se ha observado una drástica reducción en la biodiversidad a escala global, situación provocada por la acción humana y que ha afectado en mayor medida a los ecosistemas de agua dulce (Reid *et al.*, 2019; Tickner *et al.*, 2020). Producto del carácter sostenido y creciente, propio de las perturbaciones antropogénicas, se ha generado un efecto acumulativo en los daños que han sufrido estos ecosistemas, debilitándolos constantemente e impidiéndoles recuperarse (Arriagada *et al.* 2019). Los ecosistemas son capaces de lidiar con pérdidas menores de biodiversidad, sin embargo, a medida que avanza este proceso los impactos de los futuros detrimentos sobre el funcionamiento e integridad de los ecosistemas son cada vez más graves y difíciles de remediar (Cardinale *et al.*, 2012). Varios países han implementado estrategias para la rehabilitación de cuerpos de agua dañados, estas han resultado ser muy costosas y su éxito ha sido difícil de medir, principalmente, por el fracaso de las campañas de monitoreo destinadas a supervisar el proceso de recuperación (Pander y Geist, 2013). En vista de lo anterior, resulta necesario desarrollar estrategias capaces de detectar los daños ecológicos en ríos antes de que estos desemboquen en situaciones más complejas e irreversibles.

2.3 Monitoreo del estado ecológico de los ríos

Para evaluar el estado de afectación de un río no basta con conocer sus parámetros fisicoquímicos, dado que estos solo muestran una fotografía de una situación más compleja, se requiere analizar también componentes bióticos, denominados bioindicadores, capaces de reflejar el estado ecológico de un cauce mediante cambios en

sus organismos o poblaciones (Pawloski *et al.*, 2018; Astudillo-García *et al.*, 2019; Cordier *et al.*, 2021). Peces, algas bentónicas y macroinvertebrados han sido utilizados frecuentemente con este fin (Fierro *et al.*, 2019; Feio *et al.*, 2021), en base a estos se han definido diversos índices para detectar alteraciones en ríos, como *Hilsenhoff Biotic Index* (HBI) que utiliza macroinvertebrados para evaluar degradación química (Herman y Nejadhashemi, 2015; Fierro *et al.*, 2017; Fierro *et al.*, 2019). Sin embargo, la recolección e identificación de estos organismos implica generalmente un trabajo costoso, en tiempo y recursos, y que requiere de expertos para ser concretado (Pawlowski *et al.*, 2018; Cordier *et al.*, 2021). Además, esta visión del ecosistema es aún insuficiente para comprender su estado y funcionamiento holísticamente, esta carece de un importante componente: la diversidad microbiana (O'Brien et al., 2016; Sagova-Mareckova *et al.*, 2020). Con el auge de las tecnologías NGS (del inglés, *Next Generation Sequencing*), también conocidas como tecnologías HTS (del inglés, *High-Throughput Sequencing*), y la constante reducción en sus costos, se torna viable potenciar las estrategias de biomonitoreo con estas herramientas, las cuales permitirían entre otras cosas: simplificar el proceso de identificación (actualmente basado en la morfología); detectar la presencia de organismos utilizando rastros de ADN ambiental (sin necesidad de capturarles) y analizar comunidades microbianas, tanto en sus aspectos estructurales como también en funcionales (Clark *et al.*, 2018; Pawlowski *et al.*, 2018; Sagova-Mareckova *et al.*, 2020).

2.4 Bacterias y su potencial en el monitoreo de ríos

Las bacterias son ubicuas, en los ríos estas pueden subsistir en el lecho (como biopelículas sobre rocas, sedimentos u otras superficies sumergidas), o bien, en la columna de agua (adheridas a partículas o flotando libremente) (Mora-Gómez *et al*, 2016; Adhikari *et al.*, 2019). Son un componente fundamental dentro de estos ecosistemas debido a su participación en los ciclos biogeoquímicos, por formar parte de los pilares de la cadena trófica y por regular la calidad del agua, esto último al eliminar excesos de nutrientes y

contaminantes (Sabater, Guasch, Romaní y Muñoz, 2002; Sigee, 2005; Battin *et al.*, 2016; Roberto, Van Gray y Leff, 2018). Varios estudios han dado cuenta de que las características estructurales y funcionales de las comunidades bacterianas son afectadas por las condiciones fisicoquímicas de los cuerpos de agua que habitan (Romaní *et al.*, 2013; Mora-Gómez *et al*, 2016; Sagova-Mareckova *et al.*, 2020), asimismo, se han relacionado cambios en estas comunidades con impactos antropogénicos particulares, como descargas de plantas de tratamiento de aguas servidas (Ibekwe *et al.*, 2016), usos de suelo (Zhang *et al.*, 2016; Laperriere *et al*., 2020) y embalses (Wang *et al.*, 2018; Li *et al.*, 2018), y se ha hallado que algunas bacterias en estas comunidades responden a la presencia de contaminantes emergentes, como desechos farmacéuticos (Caracciolo, Topp y Grenni, 2015). Debido a la ubicuidad de estas comunidades, sus bajos tiempos de generación, la sensibilidad con su entorno y por responder rápidamente a las alteraciones en este, las comunidades bacterianas podrían ser utilizadas como bioindicadores, especialmente las biopelículas por su carácter sésil (Washington *et al.*, 2013; Borruso, Zerbe y Brusetti, 2015; Astudillo-García *et al.*, 2019).

Pese a que el conocimiento en torno a la biogeografía (distribución en el tiempo y espacio) de estas comunidades es aún muy limitado (Niño-García *et al.*, 2016), esto no imposibilita el uso de estos microorganismos como bioindicadores. Lau *et al.* (2015) desarrollaron el primer índice biótico basado en comunidades bacterianas para evaluar el estado ecológico de ríos, específicamente para cauces de Nueva Zelanda, el cual fue denominado *Bacterial Community Index* (BCI) y fue construido utilizando perfiles ARISA (del inglés *Automated Ribosomal Intergenetic Spacer* Analysis) de biopelículas. En concreto, utilizaron datos de perfiles genéticos (701 variables) para generar un modelo mediante regresión de mínimos cuadrados parciales, usando como variable de respuesta otro índice biótico validado para el área de estudio, el *Macroinvertebrate Community Index* (MCI). Este primer índice (BCI), pese a las limitaciones tecnológicas y a que solo consideró la composición taxonómica de las biopelículas, demostró la factibilidad de construir un indicador biótico utilizando características estructurales de las comunidades bacterianas, además, exhibió

correlaciones significativas con varios de los parámetros de calidad considerados. No obstante, tales correlaciones no resultaron ser fuertes y el BCI solo fue capaz de explicar un 35 % de la variabilidad observada en el MCI, por lo que no mostró suficiente potencial para ser integrado en campañas de monitoreo.

El ejemplo anterior es relativamente antiguo. Actualmente el estudio de microorganismos se realiza mediante las tecnologías NGS, las cuales ofrecen una serie de ventajas con respecto a sus predecesoras, entre estas: un mayor poder de secuenciación con menores consumos de tiempo (permite procesamiento de múltiples muestras en paralelo), costos económicos inferiores (y que siguen reduciéndose) y la capacidad de estudiar aspectos funcionales (Clark *et al.*, 2018), adicionalmente, con el aumento generalizado en la capacidad de cómputo, el procesamiento de las secuencias genéticas puede realizarse incluso con una computadora personal. Valiéndose de estas tecnologías Li *et al.* (2017) desarrollaron un índice de integridad biótica basado en comunidades bacterianas para la cuenca del río Qinhuai (China), el cual denominaron Ba – IBI (del inglés, *Bacteria – Index of Biotic Integrity*). La metodología utilizada en la construcción de Ba – IBI se asemeja a la que ha sido empleada en la confección de otros índices bióticos (Fierro *et al.*, 2018), a grandes rasgos, Li *et al.* (2017) primero establecieron métricas bióticas basadas en las características de las comunidades bacterianas (como la diversidad o abundancia de ciertos taxones) y luego seleccionaron aquellas que mejor discriminaron los sitios afectados de aquellos de referencia. El nivel de afectación de un punto de muestreo fue determinado mediante el IWQ (del inglés, *Index for the Water Quality*), un índice auxiliar generado por estos investigadores en función de la normativa china GB 3838 – 2002, la cual establece una clasificación de los cuerpos de agua según sus características fisicoquímicas. De acuerdo con los autores, el índice que desarrollaron logró clasificar los sitios estudiados según su nivel de afectación y mostró un nivel de correlación significativo y fuerte con el IWQ, así como también con el QHEI (del inglés, *Qualitative Habitat Evaluation Index*), adicionalmente, mostró una correlación significativa, fuerte y negativa con la densidad poblacional de las cercanías.

Siguiendo un enfoque similar y motivados por las alteraciones ecológicas provocadas por la construcción y operación de la represa Tres Gargantas (China), Li *et al*. (2018) construyeron un índice para evaluar la integridad biótica de los ecosistemas acuáticos de este embalse según las características de sus comunidades bacterianas. Este índice también fue denominado Ba – IBI, no obstante, se consideraron otras métricas para su construcción, las cuales, de acuerdo señalan los autores, eran más apropiadas para el contexto del embalse. El Ba – IBI de estos investigadores mostró un buen desempeño para discriminar entre sitios afectados y de referencia, y dio cuenta de que el estado ecológico en distintos puntos del embalse dependía del nivel del agua, encontrándose que las peores condiciones se presentaron cuando el embalse se encontraba lleno. Niu *et al.* (2018) muestran otro ejemplo de la construcción de un índice para evaluar la integridad biótica de ambientes acuáticos, estos investigadores desarrollaron el MC-IBI, un índice basado en la información taxonómica y funcional de comunidades microbianas sedimentarias, el cual sería válido para los cauces de la cuenca del Río Taihu (China). Este índice tuvo un buen desempeño para discriminar entre sitios de referencia y sitios afectados, mostró un buen nivel de correlación con el IWQ y entregó un resultado consistente con un reporte realizado anteriormente en la zona, el cual utilizó índices multimétricos basados en macroinvertebrados para evaluar el estado de los ríos (aunque el índice MC-IBI entregó resultados menos conservadores). Y así como los casos anteriores, es posible encontrar más ejemplos de índices para la evaluación de la integridad biótica basados en microorganismos, sin embargo, el desarrollo de estos está actualmente concentrado en China y parece estar siendo llevado a cabo por un mismo grupo de trabajo.

Las tecnologías NGS prometen revolucionar el monitoreo ambiental, con respecto a esto Cordier *et al.* (2021) indican algunas estrategias para integrar estas herramientas en las actuales campañas de monitoreo, muchas de las cuales resultan compatibles con la incorporación del estudio de las comunidades bacterianas. Sin embargo, debido a la baja variedad de ambientes, escalas espaciales y, sobre todo, temporales que han abarcado las investigaciones con respecto a estas comunidades hasta ahora, así como también producto

de la falta de estandarización entre dichos estudios, actualmente no se dispone de una línea base de conocimiento que pueda usarse como referencia (Astudillo-García *et al.*, 2019), la cual resulta necesaria si se quieren desarrollar índices bióticos que cuenten con sentido ecológico. Para construir esta línea base es necesario conocer cómo son y se comportan las comunidades bacterianas presentes en sistemas de referencia, es decir, en ambientes capaces de conservar su integridad, mantener procesos ecológicos y cuyas comunidades se asemejen a las presentes en sistemas prístinos (o no perturbados) (Astudillo-García *et al.*, 2019). En caso de no contar con sistemas prístinos para usar de referencia, también es posible recurrir a sistemas poco perturbados y cuyo estado ecológico sea conocido, esto idealmente mediante el uso de bioindicadores tradicionales y considerando las características fisicoquímicas de los cuerpos de agua (Cordier *et al.*, 2021). Pese a que las investigaciones con respecto de la biodiversidad microbiana en sistemas de agua dulce se han incrementado en los últimos años, estas se encuentran hoy concentradas principalmente en Asia, Norte América y Europa (Li et al., 2021). Chile está dentro de las zonas con carencia de datos, exhibiendo poca o nula información con respecto de la diversidad microbiana de sus ríos (Habit *et al.*, 2019).

2.5 Contexto chileno y necesidades de investigación

Chile posee un amplio rango de condiciones climáticas y fisiográficas, producto de esto sus ríos exhiben un alto grado de diversidad morfológica e hidrológica (Andreoli *et al.*, 2012), lo cual permite la existencia de una gran variedad de ecosistemas acuáticos (Habit *et al.*, 2019) en su territorio. Particularmente, la zona centro sur del país se reconoce como una región con grandes niveles de biodiversidad, incluso a nivel mundial, y cuyas especies exhiben un alto grado de endemismo (Figueroa *et al.*, 2013; Fierro *et al.*, 2017). Lamentablemente, el carácter mediterráneo de esta zona le torna especialmente atractiva para el ser humano, tanto para el desarrollo de sus actividades como para vivir, lo que ha derivado en un fuerte proceso de urbanización (Pauchard *et al.*, 2006), que en conjunto

con los reemplazos de bosque nativo por plantaciones exóticas (o por terrenos agrícolas), la actividad de plantas industriales, la intensificación de la agricultura y otras actividades, han llevado a que los cuerpos de agua presentes en esta zona exhiban signos de deterioro (Karrasch *et al.*, 2006; Alonso *et al.*, 2017; Fierro *et al.*, 2017). Más aún, Chile es reconocido como un país particularmente vulnerable a las consecuencias adversas cambio climático (Habit *et al.*, 2019) y se espera que el avance de la desertificación, junto con la reducción de caudales, deriven en una mayor demanda por recursos hídricos y, simultáneamente, en una reducción en la oferta de estos. Lo anterior necesariamente incrementará la presión sobre los ecosistemas acuáticos, complicando aún más el escenario para estos, en vista de ello, urge disponer de mecanismos capaces de detectar tempranamente perturbaciones potencialmente peligrosas para la biodiversidad local, esto con el fin de poder tomar medidas antes de que los ecosistemas acuáticos se vean irremediablemente dañados y sus servicios ecosistémicos afectados.

El presente estudio explora la composición de las comunidades bacterianas sedimentarias y planctónicas presentes en distintos ríos de la zona centro sur de Chile, donde actualmente existe un gran vacío de conocimiento respecto de la diversidad microbiana presente (Habit *et al.*, 2019). El fin último es, desde luego, aportar en la construcción de una línea base de conocimiento, la cual se espera sirva de referencia para futuros estudios enfocados en el desarrollo y validación de índices para los ríos de Chile, cuyos cálculos estén basados en las características estructurales y/o funcionales de las comunidades bacterianas presentes. Asimismo, este estudio entrega nociones de cómo estudiar la estructura de las comunidades bacterianas y efectuar comparaciones entre muestras procedentes de distintos ambientes, así como también proporciona herramientas y métodos apropiados para lo anterior, todo esto dentro del entorno de R, un lenguaje de programación libre.

CAPÍTULO 3 MATERIALES Y MÉTODOS

3.1 Estructura del capítulo

El presente capítulo consta de seis secciones, donde la primera de estas, la actual, tiene como propósito entregar orientación. Las tres secciones que siguen a la presente se enfocan en la recopilación de los datos, en tanto, las dos restantes se centran cómo se efectuaron el procesamiento y el análisis de los datos recolectados.

En concreto, la segunda sección se centra en la campaña de terreno, en esta se explica la planificación en torno a la campaña (área de estudio, selección de los puntos de muestreo, entre otros) y las actividades que se realizaron en terreno. La tercera parte del capítulo trata el trabajo en laboratorio, en este se incluye el análisis químico de las muestras de agua y el procesamiento de las muestras biológicas por parte del laboratorio externo. En tanto, la cuarta sección expone lo relativo a la recopilación de datos desde fuentes gubernamentales (parámetros fisicoquímicos, caudales y de usos de suelo).

La quinta sección, la más extensa, se centra en el manejo de los datos genéticos de las comunidades bacterianas, esta inicia explicando los desafíos que supone trabajar con este tipo de datos y cómo se abordan estos. Posteriormente, se expone cómo se efectuó el procesamiento de las secuencias brutas y la asignación taxonómica. Luego, se explican en detalle los análisis estructurales realizados a las comunidades bacterianas, así como también se presentan las herramientas estadísticas utilizadas. Finalmente, la sexta sección contiene lo relativo al análisis de la relación entre las variables ambientales seleccionadas y las características estructurales de las distintas comunidades estudiadas.

3.2 Campaña de terreno

3.2.1 Área de estudio y puntos de muestreo

Se seleccionaron ríos ubicados entre la región del Libertador General Bernardo O'Higgins y la región del Biobío, el motivo detrás de esta elección territorial relativamente amplia fue capturar una mayor variabilidad en cuanto a condiciones ambientales, hidrológicas y de usos de suelo. Dadas las limitaciones de recursos y buscando aminorar los tiempos de viaje (para no comprometer la integridad de las muestras biológicas), la selección de los puntos de muestreo definitivos estuvo fuertemente influenciada por la facilidad de acceder a estos desde vías principales (como la Ruta 5) y la presencia de estaciones de monitoreo de la Dirección General de Aguas (DGA). Esto último para utilizar los registros históricos de caudales y de parámetros de calidad del agua en la caracterización de los ríos.

Entre los cauces seleccionados se encuentra el río Biobío, perteneciente a la región homónima y uno de los principales ríos del país, el cual tiene una cuenca de 24 369 km^2, la tercera más grande a nivel nacional, y exhibe un caudal medio anual en torno a 900 m^3/s (Valdovinos y Parra, 2006; DGA, 2016). Sus aguas se utilizan para riego, acuicultura, generación hidroeléctrica, abastecimiento de agua potable e industrial, conservación de biodiversidad, como receptoras de efluentes urbanos e industriales, recreación y turismo, entre otros usos. Es necesario destacar que es una de las principales fuentes de agua para consumo humano en el Gran Concepción y otras ciudades ribereñas como Santa Juana y Hualqui (CENMA, 2017). Debido a la relevancia de este cauce a nivel nacional y regional, se decidió definir tres puntos de muestreo en este (Coihue, Nacimiento y Santa Juana).

En la región de Ñuble se seleccionaron los ríos Itata y Ñuble, antes de su confluencia por motivos prácticos como la cercanía entre dichos puntos y facilidad de acceso, pero también por la importancia que tienen estos cauces para la región, la cual destaca por su gran actividad agrícola (uno de los ejes de su economía) y consecuente dependencia hacia

los recursos hídricos con fines de riego (Uribe, 2020). A diferencia del río Biobío, solo fue posible definir un punto de muestreo en cada cauce, situación que se repitió para todos los otros ríos considerados. En la región del Maule se seleccionó el río homónimo cuya cuenca es una de las más grandes de Chile, con un poco más de 21 000 km^2, y cuyas aguas sustentan la agricultura y la generación hidroeléctrica en la zona, ambas actividades con niveles de producción significativos dentro del contexto nacional (Vicuña y Meza, 2012). En esta región también se visitaron los ríos Teno, tributario del río Mataquito, y Longaví, de la cuenca del Maule. Por último, en la región de O'Higgins se seleccionó el río Tinguiririca, uno de los principales afluentes del lago Rapel y cuyas aguas sustentan la producción agrícola en el valle de Colchagua, sector reconocido por su producción vitivinícola (Barbosa *et al.*, 2019).

La **Tabla 3.1** muestra el detalle de los puntos de muestreo (PMs), la asignación de los códigos se realizó en numeración creciente en el sentido Sur-Norte y, en el caso del río Biobío, en el sentido Este-Oeste.

Tabla 3.1 Descripción y coordenadas de los puntos de muestreo.

Código PM	Río	Sector	Coordenadas UTM		
			Huso	Norte (m)	Este (m)
P1	Biobío	Coihue en puente ferroviario	18H	5841184.30	712623.24
P2	Biobío	Aguas abajo de Nacimiento	18H	5858454.00	706182.00
P3	Biobío	Santa Juana	18H	5883895.40	683538.71
P4	Itata	Aguas arriba puente Ruta del Itata	18H	5934445.43	727775.11
P5	Ñuble	Puente Ruta N-62	18H	5941734.35	728274.74
P6	Longaví	Aguas arriba puente Ruta 5 km. 323	19H	6011970.36	254518.41
P7	Maule	Aguas abajo puente Ruta 5 km. 269	19H	6062095.35	255276.55
P8	Teno	Aguas debajo puente Ruta 5 km 176	19H	6135824.44	302030.10
P9	Tinguiririca	Aguas arriba puente Ruta 5 km. 142	19H	6167503.00	318452.70

Por otro lado, en la **Figura 3.1** se presenta el mapa con las posiciones geográficas de los puntos de muestreo visitados.

Figura 3.1 Ubicación de los puntos de muestreo.

3.2.2 Trabajo en terreno

La campaña de terreno se efectuó los días 8 y 9 de octubre de 2019, aunque cabe señalar que también se realizó una precampaña el 27 de septiembre del mismo año, en esta se visitó el río Carampangue (región del Biobío) y su finalidad fue establecer los protocolos para las actividades de terreno de la campaña principal.

En cada punto de muestreo la toma de muestras y las mediciones se realizaron siempre en el mismo sitio, generalmente en zonas con profundidades menores a 0.5 m y con flujo continuo de agua, condiciones encontradas habitualmente a distancias entre dos y ocho metros de la ribera. Cabe destacar que la seguridad del equipo humano fue siempre un factor decisivo en la definición de los puntos. En la **Figura 3.2** se resumen las actividades realizadas en cada punto de muestreo durante la campaña.

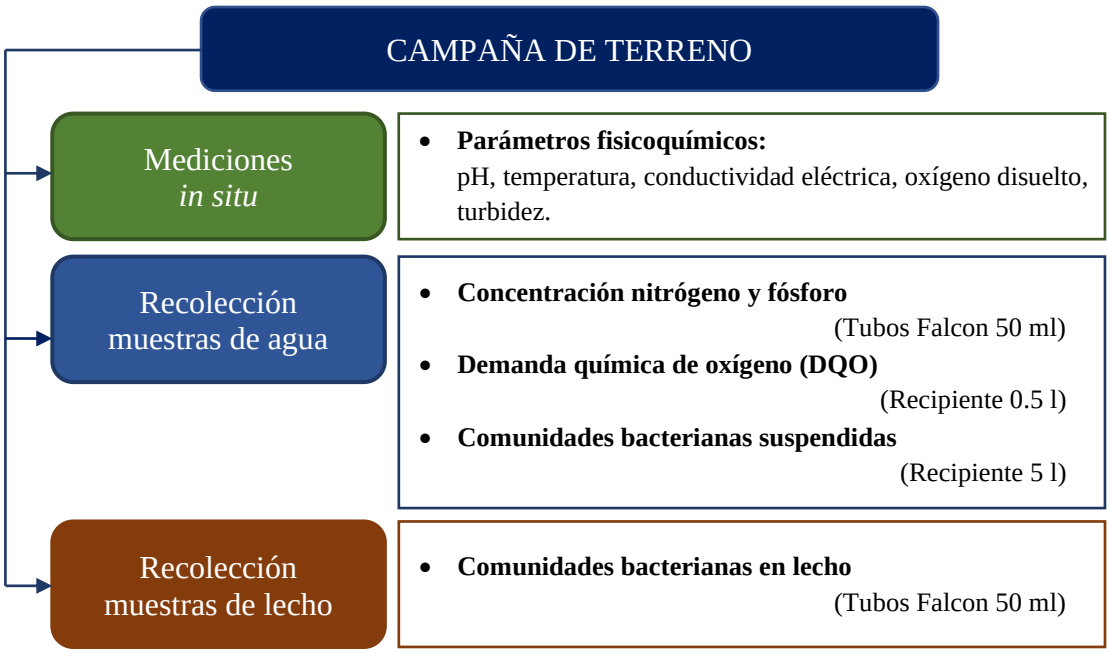

Figura 3.2 Actividades desarrolladas durante la campaña de terreno.

A. Mediciones de parámetros fisicoquímicos *in situ*

Se midieron cinco parámetros fisicoquímicos *in situ*: pH, temperatura, conductividad específica, oxígeno disuelto y turbidez, esto con la sonda multiparamétrica modelo U52 (Horiba, Kioto, Japón), la cual fue calibrada antes de la campaña de terreno de acuerdo con el manual del equipo. En cada punto de muestreo se realizó un mínimo de dos mediciones, en cada una se sumergió el dispositivo por un tiempo no menor a un minuto, esto para asegurar estabilidad en las lecturas.

B. Recolección de las muestras de agua

En cada punto de muestreo se recolectaron tres tipos de muestras, dos de estas destinadas a análisis químicos y una tercera para el estudio de las comunidades bacterianas plantónicas. Todas las muestras se obtuvieron desde el centro de la columna de agua, levemente por debajo de la superficie del río, y se almacenaron en un contenedor de plumavit provisto con hielo, esto para evitar el deterioro de las muestras debido a la alta temperatura ambiental. Una vez en el laboratorio, las muestras se mantuvieron en un refrigerador a 4°C.

Las muestras de agua para los análisis de fósforo y nitrógeno se recolectaron en tubos Falcon de 50 ml, mientras que aquellas para los análisis de DQO y las destinadas a los análisis genéticos fueron tomadas con envases plásticos de 5 l. En todos los casos los recipientes fueron sometidos a un proceso de acondicionamiento (tres lavados con agua de río) antes de la recolección de las muestras. Dada la imposibilidad de efectuar inmediatamente los análisis de fósforo y nitrógeno tras terminar la campaña, las muestras asociadas fueron sometidas a una reducción de pH con el fin de extender su tiempo de almacenamiento. Siguiendo lo señalado en el manual del espectrofotómetro, se utilizó una solución H_2SO_4 para disminuir el pH de las muestras a valores inferiores a dos, lo cual aseguraba la conservación de las muestras por un período máximo de 28 días.

C. Recolección de las muestras de lecho

Estas muestras se extrajeron utilizando dos métodos diferentes dependiendo de la naturaleza del lecho. Cuando este se encontraba constituido por sedimentos de gran tamaño (bolones) se optó por extraer aleatoriamente rocas desde el fondo y raspar la superficie de estas para obtener muestras de biopelículas, esto se realizó cuidadosamente y usando guantes de látex hasta conseguir un volumen de biomasa superior a 30 ml. En los casos donde el lecho estaba constituido en mayor medida por sedimentos de menor tamaño (arenas o finos), se optó por extraer la muestra de fondo perforando el lecho directamente con el envase de almacenamiento. El cerrado del recipiente en este último caso se efectuó bajo agua y tan pronto como se tomó la muestra, esto para evitar capturar una fracción significativa de bacterioplancton. Todas las muestras se guardaron en tubos Falcon de 50 ml y fueron transportadas en un contenedor de plumavit provisto de hielo. Una vez en el laboratorio, las muestras fueron reubicadas en un refrigerador a 4° C.

3.3 Procesamiento y análisis de las muestras recolectadas

3.3.1 Muestras de agua: determinación de características químicas

Las muestras de agua recolectadas fueron analizadas en el laboratorio de calidad del agua perteneciente al departamento de Ingeniería Civil UdeC, esto con ayuda de un espectrofotómetro modelo DR 2800 (HACH, Loveland, Colorado, USA). No obstante, por exigencia de los protocolos del equipo, antes de efectuar cualquiera de los análisis fue necesario neutralizar el pH de las muestras, esto con una solución de NaOH. Tras corregir el pH de las muestras se continuó con los análisis químicos según los protocolos de la **Tabla 3.2**.

Tabla 3.2 Protocolos utilizados para los análisis químicos.

Parámetro Químico	Protocolo (manual HACH)
Demanda Química de Oxígeno (DQO)	Método 8000
Nitrógeno Total; TKN simplificado	TNT 880, Método 10210
Fósforo Total	TNT 843, Método 10242

3.3.2 Muestras biológicas: extracción, amplificación PCR y secuenciación

Las muestras de agua fueron filtradas mediante membranas de nitrato de celulosa con poros de tamaño 0.2 μm (diámetro 47 mm, Sartorious, Germany), estos filtros se almacenaron a -20 °C hasta la extracción. Para extraer el material genético de filtros y muestras de lecho se utilizó el kit Dneasy PorwerSoil Pro (Qiagen) siguiendo las instrucciones del fabricante. El material genético fue cuantificado para verificar la calidad de la extracción usando placas Take 3, que permiten medir 2 μl de muestra en un espectrofotómetro (Epoch, Biotek, Wakefiled, MA, USA). Luego, para llevar a cabo la secuenciación se amplificaron las secuencias objetivo con la técnica de reacción en cadena de la polimerasa (PCR, del inglés *Polymerase Chain Reaction*). Las amplificaciones por PCR se realizaron con 2.5 μl de una solución tampon MasterMix que contiene GoTaq Flexi DNA Polimerase de concentración final 1X con 2 nM de MgCl2, 0.3 μM de dNTPs, 0.3 μl de cada partidor y 4 ng del ADN extraído previamente. Se amplificó la región V4 hipervariable del gen 16s del ARN ribosomal (ARNr) usando los partidores 515Fseq (5′- GTGCCAGCMGCCGCGGTAA) y 806rcbc (5′- GACTACHVGGGTWTCTAAT) (Parada *et al*., 2016). La amplificación por PCR se realizó bajo las siguientes condiciones: primero, una etapa de desnaturalización de 3 min a 94 °C; segundo, 28 ciclos de la combinación 94 °C por 30 s, 57 °C durante 1 min y 72 °C por 1.5 min; y, por último, un intervalo final de 10 min a una temperatura de 72 °C.

Los partidores Illumina se obtuvieron desde el *Earth Microbiome Project*. El conjunto de amplicones se cuantificó mediante un ensayo qPCR estándar utilizando un kit Library

Quant Illumina (Kapa) siguiendo las instrucciones del fabricante. La biblioteca genética amplificada se analizó con el equipo Bioanalyzer 2100 (Agilent Technologies, California, USA) utilizando un DNA 1000 chip (Agilent Technologies, California, USA) siguiendo las instrucciones del fabricante. La biblioteca genética se cargó en la plataforma Illumina MiSeq (Illumina Inc., San Diego, CA, USA) para la generación de clústeres, siendo sometida a secuenciación de 150 bp paired – end. Cabe señalar que la extracción del material genético se llevó a cabo en el Laboratorio de Biominería y Microbiología de Extremófilos de la Facultad de Ingeniería USS (Concepción, Chile), mientras que la secuenciación de muestras fue efectuada por Genoma Mayor (Santiago, Chile).

3.4 Recolección datos de caudales, históricos de calidad y de usos de suelo

3.4.1 Registros históricos DGA de caudal y relleno de estadísticas

Se consultó la base de datos de la Dirección General de Aguas (DGA) para obtener registros históricos de caudales en las cercanías de los puntos de muestreo. Un número importante de las estaciones fluviométricas próximas a los puntos de muestreo no contaban con registros de caudal aceptables e incluso algunas se encontraban descontinuadas desde hace años. En total, se obtuvieron registros válidos solo para cinco puntos, para los restantes fue necesario efectuar un relleno de estadísticas. Como un paso previo al relleno, aquellas estaciones sin información fluviométrica fueron reemplazadas por otras ubicadas dentro del mismo cauce, aunque no tan cercanas a los puntos de muestreo correspondientes. El relleno de los registros se realizó utilizando el algoritmo de machine *learning* MissForest siguiendo la metodología de Arriagada, Karelovic y Link (2021). Una vez construidos los registros fluviométricos, estos fueron utilizados para calcular el caudal medio anual asociado a cada punto de muestreo.

3.4.2 Registros históricos DGA de parámetros fisicoquímicos

La mayoría de los puntos contaba con estaciones de calidad en sus cercanías, las excepciones fueron P2 (Nacimiento), P8 (Teno) y P9 (Tinguiririca). Para el primero no se encontró una estación próxima que fuese válida, ya que las existentes ya estaban asociadas a otros puntos (P1 - Coihue y P3 – Santa Juana) y se optó por evitar repetir valores entre PMs, dado que podría tener repercusiones negativas en los análisis posteriores. Por el contrario, para P8 y P9 si se encontraron estaciones alternativas, aunque no tan cercanas. En la **Figura A 3.4.1** se muestra la ubicación de las estaciones DGA usadas y en la **Tabla A** 3.4.1 se presentan las coordenadas de estas. Se estableció un horizonte temporal de 30 años para los registros de calidad, los cuales, en los mejores casos, contaron con 3 o 4 mediciones anuales. Del universo total de parámetros de calidad se mantuvieron aquellos comparables con lo medido en la campaña (*in situ* o laboratorio) y aquellos con registros relativamente completos, de esta forma se conservaron 13 parámetros, estos se muestran en la **Tabla A 3.4.2** junto con sus métodos de medición y unidades correspondientes.

3.4.3 Usos de suelo en cuencas aportantes

Para determinar los porcentajes de usos de suelo dentro de cada cuenca aportante, se utilizaron los catastros de usos de suelo y vegetación elaborados por la Corporación Nacional Forestal (CONAF), los que se encuentran disponibles como archivos en formato SHP dentro la Infraestructura de Datos Geoespaciales (https://www.ide.cl). Para trabajar con estos archivos se usó el programa SIG de libre acceso QGIS (QGIS.org, 2022) en su versión A Coruña 3.10.12. Se usó una nomenclatura basada en la utilizada por los catastros de usos de suelo, así, se trabajó con ocho categorías: usos de suelo: agrícola, bosque nativo, plantación forestal, praderas y matorrales, nieve y glaciares, superficie de agua, sin vegetación y, por último, urbano. Cabe señalar que estos datos son composicionales, lo que obliga a tener ciertas precauciones al trabajar con ellos.

3.5 Análisis de la estructura de las comunidades bacterianas (CBs)

Esta sección está dividida en tres subcapítulos: el primero se centra en las cualidades de los datos genéticos de estas comunidades y los enfoques que pueden adoptarse al trabajar con estos; el segundo, en tanto, explica cómo se realizó el procesamiento de las secuencias brutas entregadas por Genoma Mayor y la consiguiente asignación taxonómica; y, por último, el tercero detalla y explica todos los procedimientos y herramientas utilizadas en los análisis de las características estructurales de las comunidades estudiadas.

3.5.1 Consideraciones previas para trabajar con datos genéticos de CBs

A. Características de los datos

La información genética asociada a comunidades bacterianas cuenta con características que dificultan y hasta imposibilitan su análisis utilizando métodos estadísticos tradicionales (Xia, Sun y Chen, 2018), estos se explicarán a continuación. Debe señalarse que las herramientas utilizadas durante esta investigación consideran estas cualidades y son capaces de lidiar con ellas, asimismo, se utilizaron ciertas prácticas (explicadas más adelante) para facilitar el manejo de estos datos por dichas herramientas.

Primero que nada, las matrices de lecturas, que es como son organizados estos datos, se caracterizan por tener un número de columnas (muestras) considerablemente menor al de filas (especies), mientras las primeras rondan en las decenas, las segundas lo hacen en las decenas de miles. Esta condición, se conoce como alta dimensionalidad e implica que estos datos no pueden ser analizados por medio de métodos estadísticos clásicos y se requiere de técnicas para reducir el número de variables, las cuales deben contemplar las relaciones taxonómicas que existen entre especies. Sumado a lo anterior, los datos genéticos también presentan sobredispersión, esto es, para cada especie se observa una enorme variabilidad en el número de lecturas detectadas entre distintas muestras, la cual

puede ser de origen técnico o biológico. Del trabajo con replicados se ha encontrado que el proceso de secuenciación puede ser modelado bajo una distribución Poisson, sin embargo, esta no logra explicar la sobredispersión observada cuando se trabaja con muestras de diferente origen. En su reemplazo se ha sugerido la distribución Binomial Negativa para modelar estos datos, ya que esta guarda gran similitud con la distribución Poisson, pero incluye un parámetro para la sobredispersión (McMurdie, 2018).

Otro rasgo distintivo y problemático de las matrices de lecturas es que estas son altamente dispersas, es decir, gran parte de sus valores son ceros (generalmente más del 70 %). Esta característica genera que los modelos paramétricos de la estadística tradicional fallen en estimar con precisión la varianza de los datos y, en su reemplazo, se deben utilizar modelos estadísticos *Zero-Inflated* (modelos para datos con alta presencia de ceros). Otro inconveniente asociado a los ceros es el origen de estos: algunos de ellos pueden producirse porque una especie efectivamente no está presente en una muestra (cero estructural), pero otros pueden ocurrir porque la profundidad de secuenciación no fue suficiente para detectar una especie que sí estaba presente (Tsilimigras *et al*. 2016). Puesto que cada tipo de cero lleva asociado un significado diferente, se ha propuesto trabajar con distribuciones diferentes para modelar los ceros según su origen para lograr una correcta interpretación de los datos (Xia, Sun y Chen, 2018). Algunas prácticas simples para evitar trabajar con este exceso de ceros es, por un lado, filtrar especies o ASVs en función del número de lecturas (o abundancias relativas) para descartar aquellas con baja presencia en las muestras y, por otro, utilizar niveles taxonómicos superiores en los análisis (como Filo), lo cual puede disminuir notoriamente la presencia de ceros, pero también significa sacrificar información taxonómica relevante (podría ser interesante y de valor conocer la presencia de un determinado género o especie).

Por último, y no menos importante, estos datos son composicionales, esto quiere decir que para cada muestra sus partes (entendido como el número de lecturas por especie) están ligadas entre sí por una suma predefinida (Gloor *et al.*, 2016). El valor de esta suma está

determinado por el proceso de secuenciación y no proporciona información útil con respecto a la comunidad, a diferencia de las relaciones (proporciones) entre las partes (Quinn *et al.*, 2019). La estadística clásica requiere que los datos estén en el espacio euclidiano para poder ser aplicada, no obstante, los datos composicionales residen en un subespacio denominado *simplex*. Si se hace caso omiso de lo anterior pueden obtener resultados erróneos, por ejemplo, producto de la restricción presente en la suma estos datos sufren de correlaciones espurias, las cuales podrían equivocadamente considerarse como verdaderas. Para solucionar esto se requiere aplicar técnicas que liberen estos datos de la restricción, sacándolos del *simplex* y llevándolos al espacio euclidiano, de forma que los análisis estadísticos tradicionales puedan ser utilizados.

B. Enfoque adoptado al analizar los datos: tradicional o composicional

Según señala McMurdie (2018) no existe una única forma de trabajar con los datos genéticos de comunidades bacterianas, qué normalización utilizar o qué análisis efectuar dependerán de lo que se está estudiando y de las características de los datos. Arbitrariamente, Gloor *et al.* (2017) clasificaron las formas de trabajar con estos datos en dos categorías según si es o no considerado el carácter composicional de estos, de esta forma reconocen dos enfoques: tradicional y composicional. Por simplicidad en este estudio se adoptó esta clasificación para diferenciar aquellas ocasiones en las que trabajó considerando el carácter composicional de los datos de las que no.

Con respecto a la primera, estos investigadores incluyen todas aquellas prácticas que se han utilizado comúnmente para el análisis de microbiomas y que hacen caso omiso con respecto a lo composicional. Un aspecto central que tratar bajo este enfoque es la normalización de los datos, puesto que esto afecta todos los análisis posteriores. McMurdie (2018) señala que normalizar puede entenderse como una estandarización de los datos a una misma escala de medida, o bien, a una escala de incertidumbre similar (variabilidad). Esta práctica busca que distintas muestras sean comparables entre sí, así

como también corregir sesgos presentes en los datos originales, lo cual se asocia con las características mencionadas en el **punto 3.5.1A**, de tal forma que puedan obtenerse resultados significativos (Weiss *et al.*, 2017). No existe una única forma de normalizar, en la literatura es posible encontrar múltiples alternativas (Weiss *et al.*, 2017). Tras normalizar los datos es posible realizar análisis posteriores, como aplicar técnicas de ordenación para visualizar la disimilitud entre muestras o efectuar comparaciones entre muestras para identificar taxones con abundancias significativamente diferentes entre estas. Cabe destacar que para esto último (también conocido como análisis diferencial de abundancias) se han desarrollado programas y algoritmos en R, los cuales normalizan de forma automática los datos antes de realizar el análisis central.

Con respecto al segundo enfoque (composicional), este utiliza y adapta las técnicas propuestas por Aitchison (1982) para el análisis de datos composicionales. Aitchison fue pionero en esta materia y propuso transformaciones matemáticas para remover la restricción inherente a estos datos, denominadas transformaciones $log - ratio$. Básicamente, estas consisten en primero dividir cada parte de un vector de composiciones por un valor de referencia y, posteriormente, calcular el logaritmo del cociente obtenido. El nuevo vector conserva las relaciones originales entre partes, pero no está sujeto a la restricción de la suma, por lo cual es posible aplicar métodos estadísticos tradicionales. Estrictamente, la información genética de comunidades bacterianas no cumple con la definición de datos composicionales, según esta un vector de composiciones debe tener todos sus valores mayores a cero, lo que se contrapone con lo que sucede en los datos genéticos de estas comunidades (Quinn *et al.*, 2018). Lo anterior se debe a que las transformaciones *log-ratio* no pueden ser aplicadas cuando existen valores nulos. Como solución se ha planteado efectuar un reemplazo de los valores ceros por pseudo – conteos, esto es, por números en el rango $(0, DL]$, donde DL es el límite de detección del instrumento (1 en el caso del secuenciador). Rigurosamente, esto solo es válido para ceros redondeados, sin embargo, debido a la imposibilidad de reconocer la naturaleza de los ceros, no suele realizarse distinción al momento de aplicar el tratamiento.

3.5.2 Procesamiento bioinformático y asignación taxonómica

El análisis bioinformático de las secuencias genéticas se realizó en R v4.1.3 (R Core Team, 2022) utilizando el protocolo DADA2 disponible en el paquete de funciones homónimo (Callahan *et al.*, 2016), cabe señalar que todas las funciones mencionadas en esta sección pertenecen a este paquete. Las secuencias brutas entregadas por Genoma Mayor estaban demultiplexadas, además, tanto los cebadores como los adaptadores ya habían sido removidos, por tanto, se procedió con la inspección de la calidad de estas lecturas, esto mediante la función *plotQualityProfile*. En base a los resultados proporcionados por esta función, se decidió recortar las lecturas *Forward* en 145 bp y las *Reverse* en 140 bp, esto mediante la función *filterAndTrim* con parámetros trunLen = (145, 140) y maxEE = 2, donde este segundo parámetro indica el máximo número aceptable de errores esperados (para los otros argumentos se utilizaron los valores predeterminados). Posteriormente, se usó la función *learnErrors* para ajustar un modelo paramétrico para cada muestra de los errores en las secuencias de amplicones (según calidad y composición de secuencias de cada muestra) y se utilizó la función *derepFastq* para derreplicar las lecturas de secuencias (se detectan aquellas secuencias que son iguales y se deja solo una, indicando cuántas veces se repite esta). Los resultados de las funciones anteriores se utilizaron como argumentos para la función *dada*, la cual usa el modelo de errores previamente ajustado para remover errores de secuenciación e inferir las ASVs (del inglés, *Amplicon Sequence Variants*), las cuales corresponden a lecturas de secuencias que difieren en solo un nucleótido. Cabe destacar que en la función *dada* se usó el argumento *pool* = TRUE, lo cual implica que el algoritmo cruzó información entre muestras, incrementando su capacidad para detectar secuencias raras o poco abundantes (*dada* con *pool* = FALSE analiza las muestras separadamente y remueve las secuencias que tengan solo una observación). Tras aplicar la función *dada* las lecturas *Forward* y *Reverse* resultantes fueron fusionadas en una única secuencia mediante la función *mergePairs*, luego se utilizó la función *makeSequenceTable* para generar una tabla de secuencias y, por último, se usó la función *removeBimeraDenovo* para remover las secuencias quimera (artefactos del

proceso de secuenciación). Finalmente, la asignación taxonómica de las ASVs se realizó con las funciones *assignTaxonomy* (Dominio a Género) y *addSpecies* (Especie) utilizando archivos *fasta* de entrenamiento (Callahan, 2018) construidos de acuerdo con la base de datos genéticos SILVA 132 (Quast *et al.*, 2013).

3.5.3 Análisis estructural de las comunidades en R

Todas las gráficas, análisis estructurales de las comunidades y análisis estadísticos se realizaron con R v.4.1.3 (R Core Team, 2022), utilizando los paquetes de la **Tabla 3.3**.

Tabla 3.3 Paquetes de funciones utilizados en R.

Paquete y versión	Referencia
ALDEx2 v1.26.0	Fernandes *et al.* (2013); Fernandes *et al.* (2014); Gloor, Macklaim y Fernandes (2016)
corrplot v0.92	Wei y Simko (2021)
cowplot v1.1.1	Wilke (2020)
DESeq2 v1.34.0	Love, Huber y Anders (2014)
ggdendro v0.1.23	de Vries y Ripley (2022)
ggforce v0.3.3	Pedersen (2021)
ggpmisc v0.4.6	Aphalo (2022)
ggrepel v0.9.1	Slowikowski (2021)
ggtext v0.1.1	Wilke (2021)
paleetter v1.4.0	Hvitfeldt (2021)
patchwork v1.1.1	Pedersen (2020)
phyloseq v1.38.0	McMurdie y Holmes (2013).
RcolorBrewer v1.1-3	Neuwirth (2022)
tidyverse v1.3.1	Wickham *et al.* (2019)
vegan v2.6-2	Oksanen *et al.* (2022)
VennDiagram 1.7.3	Chen (2022)
zCompositions 1.4.0-1	Palarea-Albaladejo y Martin-Fernandez (2015)

Como texto guía principal se usó el libro de Xia, Sun y Chen (2018), dentro de este fue posible encontrar la teoría detrás de varios de los análisis, recomendaciones, descripciones y ejemplos de funciones y paquetes usados para trabajar con este tipo de datos, así como también herramientas para desarrollar diversos análisis estadísticos. Adicionalmente, se

recurrió, aunque en menor medida, a los libros de Legendre y Legendre (1998), Borcard, Gillet y Legendre (2011) y Beiko, Hsiao y Parkinson (2018).

Del procesamiento bioinformático de las secuencias encontradas entre todas las muestras se obtuvieron una matriz de lecturas totales (m_{ij} indica el número de lecturas encontradas para la ASV_i en la $muestra_j$) y una tabla con la taxonomía de cada una de las ASVs, esto para los niveles Dominio, Filo, Clase, Orden, Familia, Género y Especie. Utilizando estas dos tablas y una tercera con la información asociada a los puntos de muestreo (características ambientales, espaciales, etc.) se construyó un objeto phyloseq para poder trabajar con las funciones del paquete Phyloseq, las cuales facilitan los análisis estadísticos y estructurales de microbiomas. Dicho objeto phyloseq fue sometido al siguiente filtro según taxonomía, con la función *filter_taxa* (paquete Phyloseq), se mantuvieron todas aquellas ASVs que cumpliesen simultáneamente las siguientes condiciones: primero, que perteneciesen al Dominio Bacteria; segundo, que tuviesen asignación taxonómica al menos hasta el nivel Filo; tercero, que no correspondiesen al Orden Cloroplasto; y, cuarto, que no pertenecieran a la Familia Mitocondria.

El objeto phyloseq obtenido tras el filtro anterior se utilizó como base para efectuar los análisis estructurales posteriores, no obstante, antes de efectuar algunos de estos fue necesario aplicar un segundo filtro en función de la abundancia relativa y prevalencia de las ASVs (en cuántas muestras se observa una ASVs). Este segundo filtro fue utilizado con el propósito de reducir la proporción de ceros en la matriz de lecturas, esto para mejorar el desempeño de las herramientas utilizadas. Los detalles con respecto a este filtro abundancia/prevalencia serán explicados más adelante dentro de esta sección.

La **Figura 3.3** exhibe un diagrama indicando los procesos desarrollados en R para llevar cabo los análisis estructurales. Con tal de favorecer la comprensión del diagrama, la **Tabla 3.4** muestra los significados de siglas utilizadas.

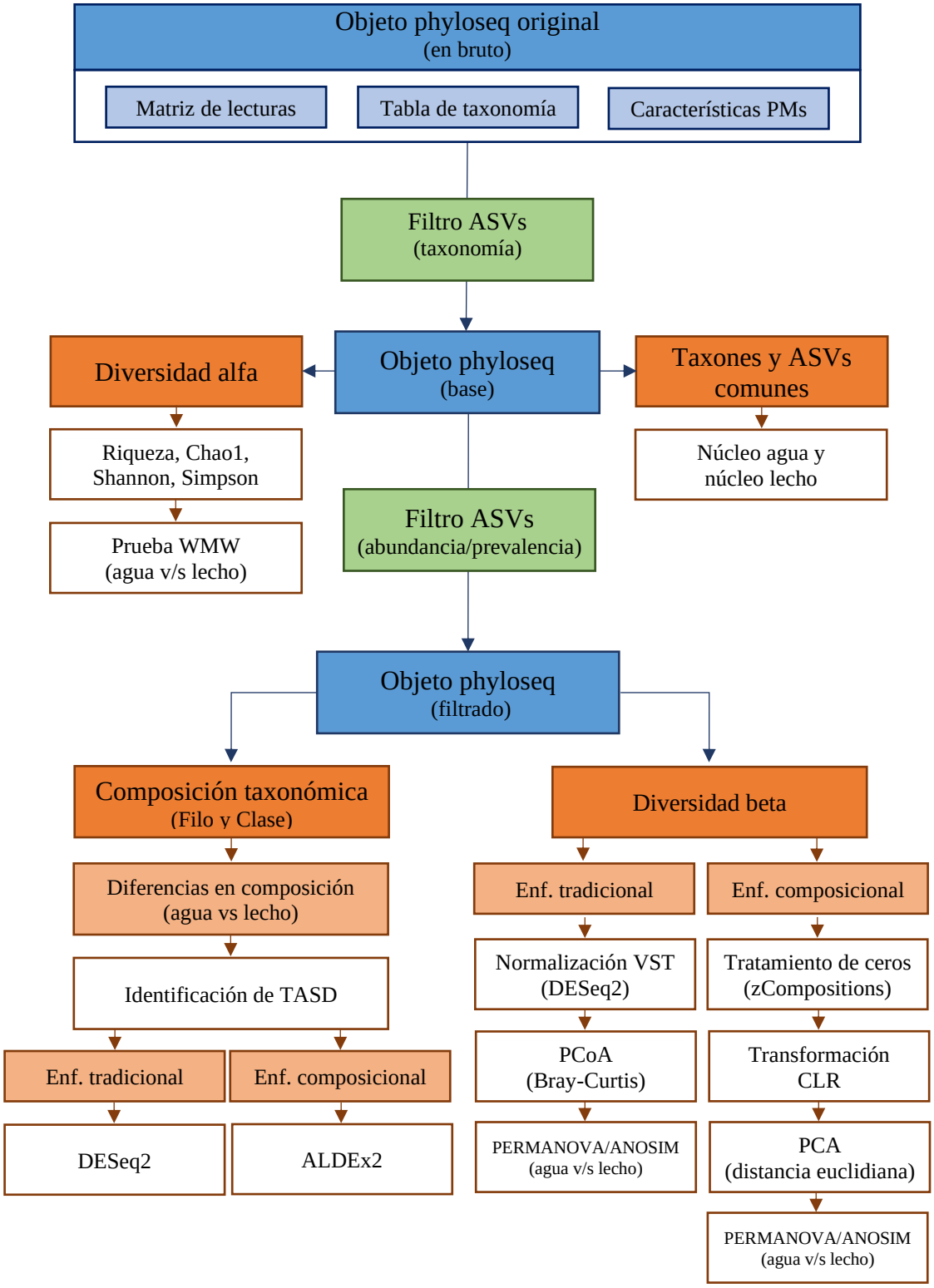

Figura 3.3 Secuencia de trabajo para el análisis estructural de las comunidades en R.

Tabla 3.4 Significado de siglas utilizadas en **Figura 3.3**

Sigla	Significado
PM	Punto de Muestreo
ASV	*Amplicon Sequence Variant*: secuencias que varían en solo un nucléotido.
WMW	Wilcoxon – Mann – Whitney
TASD	Taxones con Abundancias Significativamente Diferentes
VST	*Variance Stabilizing Transformation*: método de normalización disponible en DESeq2
CLR	*Centred Log-Ratio*: transformación utilizada para trabajar con datos composicionales
PCA	Principal Components Analysis: análisis de componentes principales (técnica de ordenación)
PCoA	Principal Coordinates Analysis: análisis de coordendas principales (técnica de ordenación)

Continuando, en los siguientes párrafos se explicarán en detalle todos los análisis estructurales efectuados a las comunidades bacterianas de agua y de lecho:

A. Diversidad alfa

La diversidad alfa es la que se observa en un espacio geográfico reducido, lo que en el contexto de este estudio se tradujo en la diversidad de una muestra. Comúnmente, para estimar esta propiedad se utilizan especies, no obstante, también es posible usar otro nivel como Filo o Clase. En este estudio se estimó la diversidad en base a las ASVs, pero sin aplicar ningún filtro a las matrices de abundancias, puesto que esto afectaría el cálculo de algunos de los índices. La diversidad alfa fue estimada mediante cuatro índices: Riqueza observada, Riqueza estimada (Chao1), Shannon – Wiener y Simpson; y se utilizó la función *estimate_richness* (paquete Phyloseq) para efectuar los cálculos correspondientes. En Xia, Sun y Chen (2018) es posible encontrar el detalle con respecto a las fórmulas necesarias para el cálculo de cada índice de diversidad alfa usado en este trabajo.

La riqueza es el índice más simple de calcular, este consiste únicamente en contar el número de especies distintas presentes en una muestra (en este caso ASVs). Puesto que la riqueza observada en la muestra no necesariamente refleja la real riqueza de una comunidad, se recurrió al índice Chao1 para estimar la verdadera riqueza en función de

las cantidades de *singletons* y de *doubletons* (ASVs con una y dos lecturas, respectivamente) presentes en cada muestra. Este índice asume que las ASVs raras guardan una estrecha relación con la riqueza real de la comunidad y, por tanto, es posible utilizarlas para la estimación de esta. La **Ecuación (3.1)** muestra la fórmula usada para el cálculo de este índice, mientras que la **Ecuación (3.2)** exhibe la utilizada para estimar la varianza asociada.

$$S_{Chao1} = S_{Observado} + \frac{{n_1}^2}{2n_2} \quad \textbf{(3.1)}$$

$$Var(S_{Chao1}) = n_2 \left(\frac{m^4}{4} + m^3 + \frac{m^2}{2}\right) \quad \textbf{(3.2)}$$

En las expresiones anteriores S corresponde al número de ASVs y el subíndice indica si se trata de la riqueza estimada (Chao1) o la observada; en tanto, n_1 y n_2 corresponden al número de ASVs con una y dos lecturas, respectivamente, y m es el cociente n_1/n_2.

El índice Shannon – Wiener mide la diversidad considerando simultáneamente la riqueza y homogeneidad de una muestra. Este se basa en la idea de incertidumbre y se asocia con la dificultad de predecir correctamente la especie de un individuo extraído al azar, así mientras más diversa es una comunidad más difícil será lo anterior y mayor valor tomará este índice. Teóricamente no existe un límite superior para este índice y será mayor mientras más especies existan en una comunidad o mientras más homogénea sea esta. Como observación, para un número fijo de especies, este índice alcanzará su máximo valor cuando la abundancia relativa de todas las especies sea la misma. La **Ecuación (3.3)** muestra la fórmula empleada para el cálculo del índice Shannon – Wiener (H').

$$H' = -\sum_{i=1}^{S} p_i \ln p_i \quad \textbf{(3.3)}$$

En la expresión anterior S indica el número de especies de una comunidad y p_i hace referencia a la proporción de la especie i. Debido a que este índice utiliza el logaritmo de las proporciones, las especies con menores abundancias tienden a adquirir un mayor peso en el cálculo de la diversidad.

Finalmente, el índice de Simpson se basa en que la diversidad está inversamente relacionada con la probabilidad de que dos individuos seleccionados al azar desde una misma comunidad pertenezcan a la misma especie. Al igual que Shannon – Wiener, este mide simultáneamente la riqueza y homogeneidad de una comunidad, y sigue un comportamiento similar al del otro índice: mientras más especies posea una comunidad o mientras más homogénea sea esta, mayor será el valor del índice Simpson. La **Ecuación (3.4)** exhibe la fórmula utilizada para el cálculo del índice de Simpson (D).

$$D = 1 - \sum_{i=1}^{S} {p_i}^2 \qquad \textbf{(3.4)}$$

En la fórmula anterior S indica el número de especies en la comunidad y p_i la proporción de la especie i en esta. Este índice adquiere valores entre cero y uno, siendo una comunidad más diversa mientras mayor sea el valor. A diferencia del índice Shannon – Wiener, el índice de Simpson tiende a otorgar mayor peso a las especies más abundantes en el cálculo de diversidad.

Por último, se utilizó la prueba estadística no paramétrica Wilcoxon – Mann - Whitney (función *wilcox.test* en R) para estudiar si existían diferencias significativas entre los valores de diversidad media asociados a cada subconjunto (agua y lecho), esto para cada una de las medidas de diversidad alfa.

B. Taxones y ASVs comunes

Frecuentemente los estudios de comunidades bacterianas en ambientes naturales analizan qué bacterias son comunes entre muestras que comparten algún rasgo en común, este subconjunto de bacterias es conocido como núcleo. Algunas investigaciones han analizado la existencia de este núcleo motivadas en entender la biogeografía y la dinámica de estas comunidades (Wang *et al.*, 2016; de Oliveira *et al.*, 2015, Staley *et al.*, 2015), por ejemplo, para analizar si existe un reservorio universal propio para cada ambiente, o también, para estudiar si ocurre una sucesión en la comunidad bacteriana al avanzar aguas abajo en un río. Otros estudios, en tanto, han analizado los elementos comunes entre comunidades para determinar si algunos taxones estarían relacionados con ciertas regiones espaciales, a usos de suelo específicos o con determinados factores de estrés, esto pensando en su uso como bioindicadores (Washington *et al.*, 2013).

En el contexto de esta investigación y debido a su carácter exploratorio resulta interesante analizar si existe un reservorio común entre las comunidades estudiadas, dado que se desarrollan en un mismo tipo de ambiente, pero en distintas regiones espaciales. Por otro lado, puesto que en varios estudios se ha observado que las comunidades planctónicas y las sedimentarias difieren en sus características estructurales, es coherente realizar este análisis de acuerdo con el origen de las muestras (agua o lecho).

Se decidió inspeccionar qué taxones resultaron comunes dentro de cada subconjunto de muestras (agua y lecho) y entre todas las muestras en los niveles de Filo y Clase. Para esto, primero se transformó el objeto phyloseq original mediante la función *tax_glom* (paquete phyloseq), la cual agrupa las ASVs según el taxón al cuál pertenecen, esto bajo el nivel taxonómico especificado con el argumento *taxrank*. Esta agrupación implicó modificaciones en las tablas de taxonomía y de lecturas, en el caso de la primera se eliminó la información redundante manteniendo solo una etiqueta por taxón, en tanto que para la segunda se sumaron aquellas filas asociadas a un mismo taxón. Así, tanto la tabla de

taxonomía como la de lecturas del nuevo objeto poseían tantas filas como taxones se encontraron para el nivel taxonómico especificado, existiendo una correspondencia entre las filas de ambas tablas. La tabla con las características de los PMs no sufrió cambios. Luego de generar los objetos phyloseq para los niveles señalados, se procedió con la remoción de aquellos taxones sin asignación taxonómica (identificados con la etiqueta "N/A" en la tabla de taxonomía), ya que las ASVs sin información para un determinado nivel taxonómico son agrupadas por el algoritmo valiéndose de la información del nivel inmediatamente superior, por lo que en la práctica no es posible asegurar que todas las ASVs agrupadas en un taxón "N/A" pertenecen al mismo clúster. Por último, en cada subconjunto de muestras (agua y lecho) se identificaron todos aquellos taxones cuya abundancia relativa en todas las muestras fuese mayor a cero y se recurrió a diagramas de Venn (generados con la función *VennDiagram* del paquete VennDiagram) para mostrar los taxones comunes en agua, en lecho y entre todas las muestras, esto para Filo y Clase.

Este análisis también fue efectuado al nivel de ASV (el nivel más fino), esto para determinar la composición del núcleo de agua y de lecho. Se trabajó con la matriz de lecturas original (sin aplicar filtro de abundancia), esta fue primero dividida en dos matrices (agua – lecho) y luego, en cada una de estas, se identificaron aquellas ASVs comunes entre todas las muestras del subconjunto. Una ASV fue considerada común (y, por tanto, perteneciente al núcleo) si estaba presente en todas las muestras de un conjunto, esto es, que su abundancia relativa fuese mayor a cero en todas las muestras. Así como se hizo para Filo y Clase, también se construyó un diagrama de Venn para los elementos comunes en cada subconjunto y entre todas las muestras. Además, para cada muestra dentro de un subconjunto se calculó qué proporción de sus ASVs y de sus lecturas pertenecían al núcleo. Finalmente, se utilizó la–prueba Wilcoxon – Mann – Whitney para evaluar si las diferencias entre los dos subconjuntos, en cuanto a las proporciones de ASVs y de lecturas asociadas a sus núcleos, resultaban ser significativas ($p < 0.01$).

C. Composición taxonómica

La matriz de lecturas totales fue primero transformada en una matriz de abundancias relativas y se le aplicó un filtro (abundancia/prevalencia), esto con el propósito de reducir el ruido generado por las ASVs menos abundantes, minimizar las posibilidades de incluir secuencias quimeras persistentes y, sobre todo, reducir la proporción de ceros en la matriz. Dado que no existe un método universal, se tuvo que primero inspeccionar las características de los datos y, en función de estas, definir alternativas de filtro según abundancia y/o prevalencia. Estos filtros fueron evaluados según dos aspectos: la reducción en la proporción de ceros y el porcentaje de lecturas removidas por muestra. Se probaron distintas opciones y se seleccionó aquel filtro que produjo la mayor reducción en la proporción de ceros, pero sin comprometer el volumen de información, es decir, sin producir una gran remoción de lecturas. El filtro finalmente adoptado consistió en mantener aquellas ASVs que cumpliesen con cualquiera de las siguientes condiciones: tener una abundancia relativa superior al 0.1 % en al menos una muestra o mostrar una prevalencia igual o mayor al 25 % (presente en a lo menos 4 de las 16 muestras). Este filtro también fue aplicado antes del análisis de diversidad beta.

Ya preparados los datos, se procedió a determinar la composición de las comunidades bajo los niveles taxonómicos de Filo y Clase. Dado los altos números de categorías existentes en cada uno de estos niveles, se trabajó con los 10 taxones más abundantes en el caso de Filo y con los 20 más abundantes para Clase. Los taxones excluidos se agruparon dentro de la categoría (“otros”) y aquellas ASVs sin información taxonómica para un determinado nivel, en caso de existir, se agruparon dentro de la categoría “N/A”. Posteriormente, se procedió a identificar aquellos Taxones con Abundancias Significativamente Diferentes (TASD) entre agua y lecho, lo cual se efectuó bajo un enfoque tradicional y uno composicional.

Para el primer enfoque se utilizó DESeq2, este paquete R fue desarrollado inicialmente para el estudio de la expresión diferencial de genes usando datos de secuenciación de ARN, no obstante, también se ha utilizado para determinar TASD entre muestras de microbioma, debido a la similitud que guardan este tipo de datos con los otros. DESeq2 requiere como entrada una matriz de lecturas, esta es sometida primero a un proceso de normalización (VST, del inglés *Variance Stabilizing Transformation*). Luego el algoritmo asume que las lecturas de cada taxón siguen una distribución Binomial Negativa y calcula la dispersión (un parámetro de la distribución) para cada taxón. Posteriormente, DESeq2 utiliza la prueba de Wald para comparar las dispersiones de distintos grupos de muestras y aplica el método de Benjamini – Hochberg para corregir los valores *p*. Finalmente, aquellos taxones con valores de dispersión significativamente diferentes ($p < 0.01$) entre dos o más grupos de muestras son reconocidos como diferencialmente abundantes. Además de reconocer los TASD, DESeq2 entrega una estimación del tamaño de efecto para cada taxón expresado como $\log_2$ *Fold Change* (medida de comparación entre los parámetros de dos grupos; un valor 1 indica una relación 2:1 entre parámetros).

Para implementar el enfoque composicional se trabajó con ALDEx2, este paquete R está diseñado para analizar datos generados por secuenciación de alto rendimiento (también denominada secuenciación HTS, del inglés *High-Throughput Sequencing*), considerando el carácter composicional de estos. ALDEx2 requiere como entrada una matriz de lecturas, luego, como primer paso, por cada celda de esta matriz el algoritmo genera una distribución de probabilidades posteriores en la forma de un vector con *k* valores tomados mediante el método de Monte Carlo desde una distribución Dirichlet (cada muestra tiene su propia distribución Dirichlet ajustada). Tras el proceso anterior se obtienen *k* matrices, cada una con el valor *k* de los vectores de probabilidades posteriores anteriormente generados, en seguida, ALDEx2 aplica a cada una de estas la transformación CLR (del inglés, *Centred Log-Ratio*) y procede a realizar una prueba de significancia estadística (Welch y/o Wilcoxon – Mann – Whitney) por cada taxón para evaluar diferencias entre grupos de muestras. Posteriormente, el algoritmo corrige los valores *p* (*k* valores por cada

taxón) con el método Benjamini – Hochberg y determina un único valor *p* por taxón, el cual corresponde al promedio de los valores corregidos. Finalmente, ALDEx2 reconoce como TASD aquellos que posean un valor *p* bajo un nivel de significancia predefinido (0.05 en este caso, puesto que ALDEx2 es más conservador que DESeq2). Adicionalmente, este algoritmo entrega para cada taxón una estimación del tamaño de efecto, esta corresponde a la diferencia media observada entre las *k* matrices de probabilidades posteriores para los subconjuntos previamente definidos, en este caso, de acuerdo con el origen de las muestras: subconjunto de agua y subconjunto de lecho.

D. Diversidad beta

Whittaker (1960) propuso tres enfoques para estudiar la diversidad de los ecosistemas: alfa (local o en una escala espacial reducida), gamma (regional o en una escala espacial amplia) y beta (variación espacial de la diversidad). Esta última es la conexión entre las otras dos medidas de diversidad y en la práctica se traduce en el estudio de la disimilitud, en cuanto a composición, que existe entre comunidades (Díaz *et al.*, 2021).

El estudio de la disimilitud entre comunidades bajo este enfoque es un problema multidimensional, donde cada una de estas comunidades (muestras) corresponden a un punto dentro de este espacio, el cual cuenta con tantas dimensiones como taxones (Filos, Clases o ASVs) estén presentes en el conjunto de muestras. Esto no es físicamente posible de graficar, por lo que se requiere utilizar técnicas de ordenación para generar proyecciones bidimensionales de este espacio, las cuales deben ser capaces de capturar la mayor cantidad de información posible, esto es, que las distancias observadas en la proyección se aproximen lo más posible a las reales. Para estudiar esta propiedad se utilizó tanto el enfoque tradicional como el enfoque composicional, no obstante, antes de aplicar cualquiera de estos se aplicó un filtro abundancia/prevalencia (anteriormente descrito en la **sección 3.5.3C**) a la matriz de lecturas, esto para reducir su proporción de ceros y el ruido provocado por las ASVs menos abundantes.

Bajo el enfoque tradicional se trabajó con la técnica de ordenación PCoA (del inglés, *Principal Coordinates Analysis*) utilizando la función *ordinate* del paquete Phyloseq. Esta técnica no trabaja directamente con el espacio multidimensional anteriormente señalado (una dimensión por especie), sino que usa las disimilitudes calculadas a partir de este. Como entrada requiere una matriz de disimilitud, cuyos elementos (d_{ij}) corresponden a la diferencia entre la muestra de la columna i y la de columna j. Esta diferencia (disimilitud) se determina en base a la matriz de lecturas y es posible emplear cualquier distancia o medida de disimilitud para su cálculo. Antes de determinar la matriz de disimilitud, la matriz de lecturas fue sometida a una normalización siguiendo lo señalado por McMurdie (2018), para esto se usó la función *varianceStabilizingTransformation* (VST) del paquete DESeq2, la cual asume que las lecturas siguen una distribución Binomial Negativa y aplica una transformación acorde para generar una estabilización en la varianza. Tras normalizar la matriz, se calculó la disimilitud entre muestras usando el índice Bray – Curtis, el cual, a diferencia de la distancia euclidiana, no se ve tan afectado por los "doble – ceros" típicos en matrices de lecturas pertenecientes a comunidades bacterianas. Estos "doble-ceros" (cuando dos muestras marcan cero lecturas para una misma especie) pueden provocar una falsa similitud entre dos comunidades. La **Ecuación (3.5),** extraída de Xia, Sun y Chen (2018), muestra la expresión utilizada para calcular el índice de disimilitud Bray – Curtis (BC) entre la muestra i y la muestra j.

$$BC_{ij} = \frac{\sum_{k=1}^{n} |X_{ki} - X_{kj}|}{\sum_{k=1}^{n} |X_{ki} + X_{kj}|} \qquad \textbf{(3.5)}$$

En la expresión anterior X_{ki} y X_{kj} indican las proporciones de la especie k en la muestra i y en la muestra j, respectivamente, en tanto, n hace referencia al número de especies (u otro taxón, dependiendo del nivel seleccionado) presentes en el conjunto de muestras. Este índice de disimilitud toma valores entre cero y uno, donde un valor cero implica que las muestras son iguales y un valor uno que son completamente diferentes.

El PCoA genera un set de ejes principales a partir de la matriz de disimilitud, para lo cual primero la somete a una transformación (le otorga una estructura conveniente) y luego determina los valores y vectores propios de esta nueva matriz, estos últimos son posteriormente escalados a la raíz del valor propio correspondiente. La proyección bidimensional se obtiene al graficar los dos primeros ejes principales (aquellos que mayor variabilidad explican), las coordenadas pueden ser encontradas en los vectores propios correspondientes, mientras que la variabilidad asociada a cada eje puede ser calculada mediante el cociente entre el autovalor asociado y la sumatoria de todos los autovectores. Como observación, dependiendo de qué métrica se utilice podrían obtenerse autovalores negativos, frente a lo cual sería necesario aplicar una corrección en los autovalores. Afortunadamente, no fue el caso en este estudio.

Por otro lado, para implementar el enfoque composicional primero se utilizó la función *cmultRepl* del paquete zCompositions para el tratamiento de los ceros en la matriz de lecturas, puesto que las transformaciones *log-ratio* no son compatibles con valores nulos. Esta función infiere valores de reemplazo para los ceros bajo un enfoque Bayesiano y realiza una corrección multiplicativa de los valores no nulos para mantener las proporciones originales entre las lecturas. Tras solucionar la situación de los ceros, las columnas de la matriz obtenida (una por muestra) fueron sometidas a la transformación CLR, la cual convierte el vector X en el vector X_{CLR} siguiendo la **Ecuación (3.6)**.

$$X_{CLR} = \left(\log\left(\frac{x_1}{g_x}\right), \log\left(\frac{x_2}{g_x}\right), \ldots, \log\left(\frac{x_n}{g_x}\right)\right) \quad \textbf{(3.6)}$$

En la fórmula anterior x_1, x_2, …, x_n son las componentes del vector original X y corresponden a las lecturas observadas una muestra para cada ASV. En tanto, g_x es la media geométrica del vector X, esta se calcula por medio de la **Ecuación (3.7)**.

$$g_x = \sqrt[n]{x_1 * x_2 * \ldots * x_n} \quad \textbf{(3.7)}$$

Las ecuaciones **(3.6)** y **(3.7)** pueden ser encontradas en Gloor *et al.* (2017).

Para visualizar la disimilitud entre las muestras se recurrió a la técnica PCA (del inglés, *Principal Component Analysis*), la cual fue implementada con la función *rda* (paquete vegan), con el argumento *scale = FALSE*. Como entrada se utilizó la matriz de lecturas obtenida tras el tratamiento de los ceros y la aplicación de la transformación *log-ratio*. Este algoritmo utilizó dicha matriz para determinar la matriz de dispersión, esta es una matriz de asociación entre las variables originales (muestras) cuyas celdas contienen las varianzas y covarianzas entre las distintas variables. Luego, el PCA rota los ejes de tal forma que las nuevas posiciones maximicen la variabilidad explicada, esto lo logra mediante el cálculo de los valores y vectores propios de la matriz de dispersión. Los nuevos ejes principales son ortogonales entre sí, cada uno de estos explica una fracción de la variabilidad total observada, la cual corresponde al cociente entre el autovalor asociado al eje y la sumatoria de los autovalores, y se ordenan en forma decreciente según la variabilidad que son capaces de explicar. Por último, la proyección bidimensional se construye a partir del producto entre las coordenadas antiguas de cada punto (valores ligados a cada ASV) por las componentes de los primeros dos vectores propios, esta operación entrega las nuevas coordenadas de cada punto en la proyección. No obstante, en R las coordenadas de las muestras en la proyección bidimensional se obtuvieron mediante la función *scores*, utilizando el argumento *scaling = 'sites'*.

Para evaluar si la disimilitud observada entre muestras puede ser explicada por el origen de estas se utilizó primero la prueba estadística PERMANOVA, la cual fue implementada con la función *adonis2* del paquete vegan. Esta prueba requiere como entrada una matriz de disimilitud (distancia entre muestras) y un vector con información respecto de alguna variable cualitativa o cuantitativa de interés (en este caso, el origen de las muestras).

Luego, utiliza como estadístico un pseudo cociente – F calculado como la razón entre la distancia promedio al interior de los subconjuntos y la distancia promedio entre subconjuntos. Posteriormente, recurre a un método permutacional para producir una distribución empírica de valores para el estadístico F (permuta filas y columnas de la matriz original y para cada permutación recalcula el estadístico F). Finalmente, compara el valor del estadístico F con la distribución empírica generada y determina su nivel de significancia como un valor *p*. Así, si el valor *p* se encuentra bajo cierto umbral previamente establecido (en este caso 0.01), puede rechazarse la hipótesis nula de que las disimilitudes observadas son producto del azar y aceptar la hipótesis alternativa de que las diferencias observadas se explican por la variable de interés (origen). En el caso del enfoque tradicional se utilizó el índice de disimilitud Bray-Curtis y la matriz de lecturas normalizada con VST (DESeq2), mientras que bajo el composicional se usó la matriz de lecturas derivada tras el tratamiento de ceros y la aplicación de la transformación CLR, y la distancia euclidiana como medida de disimilitud.

Adicionalmente, también se utilizó la prueba ANOSIM para evaluar lo mismo que con la prueba PERMANOVA. Se recurrió a la función *anosim* del paquete vegan para implementar esta prueba estadística. Como entrada requiere una matriz de lecturas (normalizada), una medida de disimilitud y un vector con información para la variable cualitativa de interés (origen de las muestras). ANOSIM utiliza una matriz con los rangos (*rankings*) de los valores de disimilitud observados, en base a esta calcula el estadístico R que consiste en la diferencia entre los promedios de los rangos entre los subconjuntos y al interior de los subconjuntos. Así como la prueba PERMANOVA, se vale de permutaciones para generar una distribución empírica para R y determinar el valor *p* asociado. R guarda similitud con un coeficiente de correlación, un valor de R igual a cero sugiere que las diferencias entre grupos y al interior de los grupos son iguales, mientras que un valor de uno indica que las diferencias entre grupos son mayores a las observadas al interior de estos. Un valor negativo de R es excepcional y sugiere que las disimilitudes al interior de los grupos exceden las observadas entre grupos.

En resumen, durante esta sección (**3.5.3**) se expuso lo relativo a los análisis de la estructura de cada una de las comunidades estudiadas y las diferencias que se observaron entre las comunidades de agua y las de lecho. Todo este proceso fue desarrollado en R y se utilizaron principalmente cuatro paquetes de funciones para esto: Phyloseq, Vegan, DESeq2 y ALDEx2.

Como primer paso, las ASVs obtenidas tras el procesamiento bioinformático con DADA2 fueron sometidas a un filtro de acuerdo con la información taxonómica de estas. Aquellas ASVs remanentes se utilizaron para estudiar la diversidad alfa de cada muestra, esto por medio de cuatro índices que recurrentemente aparecen en la literatura (Riqueza, Chao1, Shannon – Wiener, Simpson). Estas ASVs fueron también utilizadas para determinar aquellos elementos que resultaban ser comunes entre todas las muestras, solo entre las muestras de agua y exclusivamente entre las de lecho, el conjunto de estos elementos comunes es conocido generalmente como núcleo.

Posteriormente, las ASVs fueron sometidas a un segundo filtro (basado en abundancia y prevalencia), esto con motivo de reducir la proporción de ceros en la matriz de lecturas y el ruido generado por aquellas ASVs menos abundantes. El subconjunto de ASVs que sobrevivieron a este filtro se utilizaron para determinar la composición de cada muestra bajo los niveles de Filo y Clase, estos datos fueron posteriormente analizados bajo un enfoque tradicional y uno composicional para determinar, para cada nivel, qué taxones resultaban presentar abundancias significativamente diferentes entre muestras de agua y de lecho. Por último, estas ASVs también se usaron para analizar la variación de la diversidad entre todas las muestras, esto por medio de gráficas de ordenación siguiendo un enfoque tradicional y uno composicional, y se evaluó si las diferencias observadas en las gráficas de ordenación podían ser explicadas por el origen de las muestras.

3.6 Relación entre variables ambientales y características estructurales de CBs

Determinadas las características estructurales de las comunidades se procedió a evaluar la influencia de distintos factores externos sobre estas. Este proceso inició con la recolección de datos de múltiples variables ambientales para caracterizar los puntos de muestreo. Luego, se llevó a cabo una reducción del número variables, para lo cual fue necesario establecer ciertos criterios de exclusión y conservación. El conjunto de las variables conservadas será tratado de aquí en adelante como el conjunto de variables ambientales. Por último, se analizó la relación entre las variables ambientales y los aspectos estructurales (diversidad alfa, diversidad beta y taxones más abundantes) de las distintas comunidades, esto en forma separada para el subconjunto de agua y el de lecho. La **Figura 3.4** resume la secuencia de trabajo seguida.

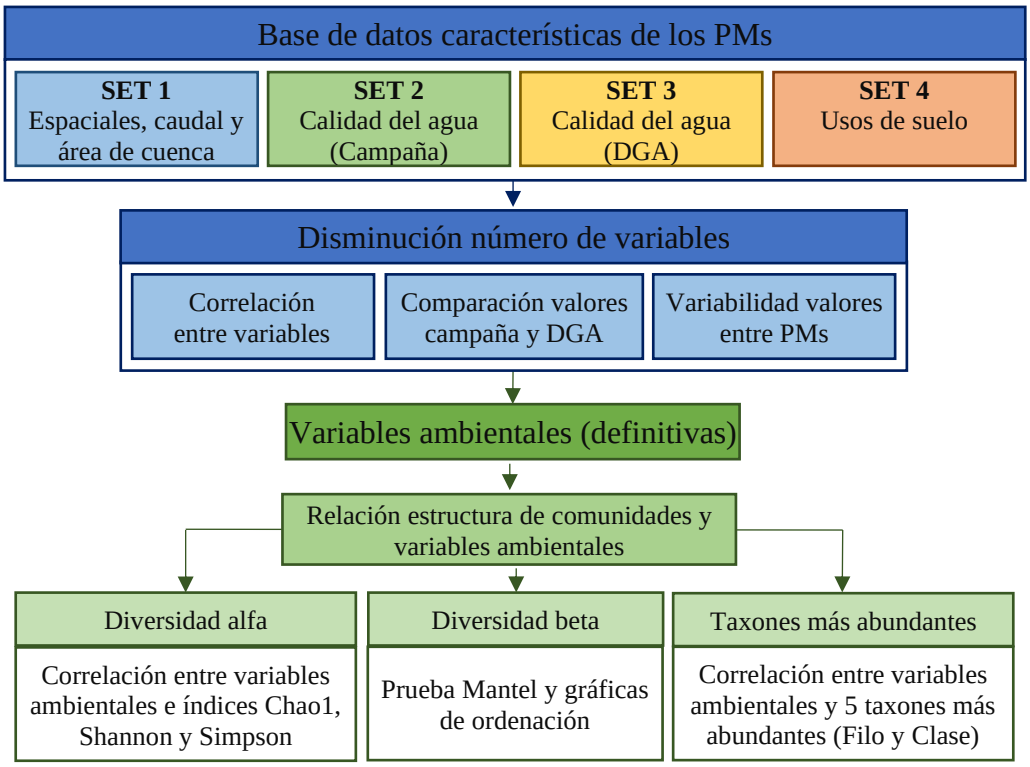

Figura 3.4 Secuencia de trabajo para el análisis de la relación entre variables ambientales y características estructurales de las comunidades.

3.6.1 Reducción del número de variables

Tras la recopilación de datos se obtuvieron valores para un total de 36 variables (ver **Tabla A 3.4.3** Puesto que un mayor número de variables implica un incremento en el número de pruebas a realizar y, por tanto, un aumento de las probabilidades de obtener falsos positivos, se decidió reducir el número de variables ambientales a evaluar.

Antes que nada, fue necesario establecer si considerar tanto las variables de calidad medidas durante la campaña como aquellas obtenidas desde los registros DGA. Para esto, primero se compararon los valores de aquellos parámetros fisicoquímicos que coincidiesen o mostrasen un alto grado de relación entre los conjuntos N°2 (DGA) y N°3 (campaña). Se encontró que las variables medidas *in-situ* durante la campaña mostraron gran similitud con su contraparte de la DGA, por el contrario, aquellas fueron obtenidos desde los análisis químicos en laboratorio mostraron diferencias importantes con los valores encontrados en los registros históricos (ver **Figura A 3.4.2**). Con respecto a esto último es necesario señalar que algunas de estas variables no coincidían plenamente con las presentes en la DGA, esto se dio en los casos del nitrógeno (mientras que en la campaña fue medido el nitrógeno de nitratos y nitritos, en los registros se encontró nitrógeno de nitratos) y del fósforo (en la campaña se midió el fósforo total, pero en los registros se encontró fósforo de ortofosfato). En vista de lo anteriormente expuesto y, principalmente, debido a que los valores promedio obtenidos desde la DGA provenían de registros con 30 años de extensión, se decidió priorizar las variables de calidad extraídas desde la DGA y excluir aquellas medidas durante la campaña, con la única excepción de la turbidez, para la cual no fue posible encontrar una contraparte en los registros históricos DGA.

Inicialmente se consideró utilizar los porcentajes de usos de suelo en las cuencas aportantes como variables a estudiar, no obstante, debido a que estos son datos composicionales (requieren de un trato especial) y, además, se contaba con un número reducido de puntos de muestreo (≤ 8), se decidió finalmente excluir estas variables, puesto

que necesariamente alzaría la tasa de falsos positivos. Otro aspecto no menos importante con respecto a los usos de suelo es que, como se verá más adelante en la **sección 4.2.2**, estos mostraron cierto nivel de asociación con la latitud, por lo que indirectamente podrían relacionarse los resultados obtenidos para dicha variable con los usos de suelo.

Las variables remanentes fueron sometidas a un análisis de correlación, esto mediante la función *cor.test* en R, la cual permite determinar el nivel de correlación (Spearman o Pearson) entre dos variables y el valor *p* asociado. Se evaluó la correlación existente entre las distintas combinaciones de variables ambientales y se utilizó la función *corrplot* (paquete corrplot) para generar representaciones gráficas de los resultados. En base al análisis anterior, se mantuvieron aquellas variables que mostraron un claro control sobre otras, no obstante, las variables aparentemente dependientes no fueron descartadas inmediatamente, sino se prefirió recurrir a gráficas de dispersión para visualizar el nivel de correlación, si esta resultaba ser clara e importante ($r > 0.7$) la variable dependiente quedaba excluida. En el caso particular de las variables DGA, además se optó por excluir aquellas cuya variabilidad entre puntos de muestreo era prácticamente nula.

Finalmente, las variables ambientales seleccionadas para los análisis futuros fueron 11: altitud, latitud, área de la cuenca aportante, turbidez (campaña), pH (DGA), conductividad eléctrica (DGA), Demanda Química de Oxígeno (DGA), concentración NO_3 – N (DGA), concentración PO_4 – P (DGA), cociente N – NO_3/ P – PO_4 (DGA) y concentración de Al (DGA). La exclusión de la longitud se explica por su alto nivel de correlación con la altitud y la latitud, en tanto que, el caudal fue excluido por mostrar un alto nivel de correlación con el área de la cuenca aportante.

3.6.2 Relación entre variables ambientales y diversidad alfa

Primero que todo, es necesario precisar que solo se consideraron los índices Chao1 (riqueza estimada), Shannon – Wiener y Simpson para este análisis y que, además, se trabajó con los conjuntos de agua y de lecho en forma separada, esto debido a que para un mismo sitio tanto la muestra de agua como la de lecho tenían asociados los mismos datos ambientales. Primero, se utilizó la función *cor.test* de R para determinar el nivel de correlación entre cada variable ambiental y cada índice de diversidad alfa (33 combinaciones), esta función también entrega el valor *p* asociado. Luego, se recurrió a la función *corrplot* (paquete corrplot) para condensar la información de las correlaciones en gráficas similares a matrices, las cuales indican mediante colores el nivel de correlación y, además, permiten marcar aquellas correlaciones que resulten ser significativas. Debido a que se efectuaron múltiples pruebas estadísticas se decidió corregir el valor *p* mediante el método Benjamini – Hochberg (Benjamini y Hochberg, 1995), para lo cual se utilizó la función *p.adjust* de R. Se aceptaron como significativas aquellas correlaciones cuyo valor *p* corregido fuese inferior a 0.5, no obstante, también se aceptaron aquellas correlaciones con valores *p* (sin corregir) menores a 0.01. Posteriormente, se generaron gráficas de dispersión para las 33 combinaciones posibles, no obstante, se centró la atención en aquellas cuyo nivel de correlación resultase ser fuerte ($r > 0.7$) y estadísticamente significativo ($p_{\text{corregido}} < 0.05$ o, en su defecto, $p < 0.01$), las combinaciones restantes fueron inspeccionadas con menor profundidad. Para aquellas gráficas de dispersión donde se observó una clara tendencia lineal en la distribución de los puntos se decidió ajustar un modelo de regresión simple como el indicado en la **Ecuación (3.8)**.

$$\hat{Y} = \beta_0 + \beta_1 * X \qquad \textbf{(3.8)}$$

En la expresión anterior $\hat{Y}$ corresponde a la variable de respuesta, en este caso un índice de diversidad alfa, en tanto X es la variable ambiental estudiada y, por último, β_0 y β_1 corresponden a los coeficientes del modelo de regresión lineal simple.

Se adoptó el mismo criterio usado en el análisis de correlación y se reportaron como estadísticamente significativas aquellas regresiones lineales con un valor *p* corregido (con el método Benjamini – Hochberg) menor a 0.05 o cuyo valor *p* fuese inferior a 0.01.

3.6.3 Relación entre variables ambientales y diversidad beta

El análisis de la relación entre la diversidad beta de las muestras y las variables ambientales se llevó a cabo mediante la prueba estadística de Mantel, la cual fue implementada con la función *mantel* (vegan). A grandes rasgos, esta prueba permite analizar el nivel de correlación entre dos matrices de disimilitud, una de estas con las diferencias entre muestras según composición y la otra con las disimilitudes entre muestras según la variable ambiental estudiada. Como ya ha sido señalado anteriormente en la **Sección 3.5.3D**, la diversidad beta puede ser determinada considerando cualquier nivel taxonómico, para efectos de este análisis se optó por utilizar los niveles Filo, Clase y ASV. La disimilitud entre muestras según composición se calculó bajo el enfoque tradicional, es decir, estas se determinaron utilizando la matriz de lecturas normalizada con VST (DESeq2) y mediante el índice Bray – Curtis.

Para cada variable ambiental se determinó una matriz de disimilitud, las celdas de estas contenían las diferencias entre muestras calculadas mediante la distancia euclidiana, para esto se recurrió a la función *dist* en R. Ya determinadas las matrices de disimilitud, el cálculo del nivel de correlación fue realizado por medio del coeficiente de correlación Spearman. Para determinar la significancia estadística de la prueba la función *mantel* utiliza el producto Hadamard (*z*) entre las dos matrices como estadístico y recurre a permutaciones, así como las pruebas PERMANOVA y ANOSIM, para generar una

distribución empírica con la cual contrastar *z*. Puesto que se efectuaron múltiples pruebas estadísticas, se decidió corregir el valor *p* con el método Benjamini – Hochberg y se utilizó como umbral para evaluar la significancia estadística el valor 0.05, no obstante, también se aceptaron como significativas aquellas correlaciones cuyo valor *p* sin corregir resultase ser inferior a 0.01.

3.6.4 Relación entre variables ambientales y taxones más abundantes

Antes que nada, fue necesario definir qué taxones serían considerados en el análisis, arbitrariamente se seleccionaron los cinco Filos y las cinco Clases con mayores abundancias relativas promedio, esto para reducir el número de relaciones a evaluar. De igual forma que para los análisis anteriores, se trabajó con las muestras de lecho y de agua en forma separada. Ya establecido qué taxones serían considerados para los análisis, se procedió a determinar el nivel de correlación que existía entre cada variable ambiental y cada uno de los taxones seleccionados (55 combinaciones en total), lo cual se efectuó utilizando la función *cor.test* en R. La secuencia de trabajo es similar a la ya explicada en la **sección 3.6.2**, por tanto, se prosiguió con la generación de las gráficas de correlaciones mediante *corrplot* y, además, con la corrección del valor *p* por medio del método Benjamini – Hochberg utilizando la función *p.adjust*. Similarmente, se generaron gráficas de dispersión para las 55 combinaciones, pero se centró la atención en aquellas que resultaron ser fuertes ($r > 0.7$) y estadísticamente significativas ($p < 0.05$ o, en su defecto, $p < 0.01$). Finalmente, para los casos donde se observó una clara tendencia lineal se decidió ajustar un modelo de regresión lineal simple como el exhibido anteriormente en la **Ecuación (3.8)**.

CAPÍTULO 4 RESULTADOS

4.1 Estructura del capítulo

En el presente capítulo se exponen y discuten los resultados obtenidos, este se subdividide en nueve secciones, considerando la actual. Comenzando en la segunda sección, esta se enfoca en la caracterización de los puntos de muestreo. En la siguiente se aborda el procesamiento bioinformático de las secuencias genéticas. La cuarta, quinta, sexta y séptima sección están totalmente centradas en las características estructurales de las comunidades bacterianas, en estas se exponen los resultados con respecto a la diversidad alfa, la composición taxonómica, el análisis de TASD y la diversidad beta de las comunidades, así como también lo obtenido del análisis de elementos comunes entre estas. En la penúltima sección se encuentra lo relativo a la relación entre las variables ambientales y las características estructurales de las comunidades. Por último, la novena sección está netamente enfocada en la discusión de los resultados.

Antes de comenzar, podría ser de utilidad para el lector revisar la **Tabla 4.1**, esta es un recordatorio del significado de los códigos usados para identificar los puntos de muestreo.

Tabla 4.1 Puntos de muestreo y sus códigos de identificación.

Punto de muestreo (PM)	Río	Sector (aproximado)
P1	Biobío	Coihue
P2	Biobío	Nacimiento
P3	Biobío	Santa Juana
P4	Itata	Puente Ruta del Itata
P5	Ñuble	Puente Ruta N-62
P6	Longaví	Puente Ruta 5 (km. 323)
P7	Maule	Puente Ruta 5 (km. 269)
P8	Teno	Puente Ruta 5 (km. 176)
P9	Tinguiririca	Puente Ruta 5 (km. 142)

4.2 Características de los puntos de muestreo

Del proceso de recolección de datos se obtuvo información para un total de 36 variables (ver **Tabla A 3.4.3**), entre las cuales se encuentran:

- 10 parámetros fisicoquímicos de calidad medidos en terreno.
- 13 parámetros fisicoquímicos de calidad extraídos desde registros DGA.
- 3 variables espaciales.
- El Área de las cuencas aportantes y los porcentajes de 8 usos de suelo.
- El caudal medio anual de los cauces.

4.2.1 Variables espaciales, caudal medio anual y área de cuenca

Entre las variables espaciales se encuentran aquellas registradas en terreno con GPS (latitud y longitud) y la altitud, cuyo valor fue estimado con Google Earth y contrastado con aquel observado en modelos de elevación digital. En la **Tabla 4.2** se exhiben los valores obtenidos para las variables espaciales señaladas y, además, los correspondientes a las áreas de las cuencas aportantes y caudal medio anual.

Tabla 4.2 Valores de variables espaciales, área de cuenca y caudal.

PM	Altitud (msnm)	Coordenadas geográficas		Área de cuenca (km^2)	Caudal medio anual (m^3/s)
		Latitud (S)	Longitud (W)		
P1	67	37.55	72.59	11 108	534.01
P2	60	37.40	72.67	16 976	694.25
P3	45	37.17	72.93	23 266	784.47
P4	40	36.71	72.45	3 825	93.77
P5	31	36.64	72.45	5 129	96.86
P6	150	36.01	71.72	834	21.44
P7	95	35.55	71.70	5 927	120.07
P8	290	34.90	71.17	1 463	28.77
P9	360	34.62	70.98	1 849	29.42

Durante la campaña de terreno se visitaron múltiples cauces avanzando preferentemente en la dirección sur – norte, por tanto, la variable central de este estudio fue la latitud. En la tabla anterior es posible notar que la altitud y longitud muestran tendencias en la dirección sur – norte, la primera tiende a aumentar sus valores mientras que para la segunda estos van reduciéndose, el análisis de correlación arrojó que estas variables están correlacionadas significativamente con la latitud (ver **Figura A 4.2.1**) la primera en forma negativa y la segunda positivamente. Este comportamiento es consecuencia de haberse desplazado por la Ruta 5, esta vía tiene una inclinación sureste – noroeste en el área de estudio, aproximándose hacia la cordillera de los Andes cerca de Santiago. El área de la cuenca aportante y el caudal medio anual también están significativamente correlacionados con la latitud (ver **Figura A 4.2.1**), motivo por el que se incluyen en esta tabla, ambas variables se reducen hacia el norte. En el caso del área de la cuenca aportante esto es consecuencia de la geografía del país y de la posición de la Ruta 5: la franja terrestre disminuye su ancho a medida que se aproxima a la capital; en tanto, la Ruta 5 se aproxima hacia la cordillera en dicho sentido, por lo que no permite mayor desarrollo de las cuencas. En el caso del caudal esto se debe a la reducción de la superficie de la cuenca aportante y a la disminución de las precipitaciones medias anuales en el sentido sur – norte.

4.2.2 Usos de suelo en cuencas aportantes

La **Figura 4.1** exhibe la proporción de los usos de suelo en las cuencas aportantes mediante un gráfico de barras, estas se ordenan según el resultado del análisis clúster jerárquico (mostrado a la izquierda). La figura incluye, además, una gráfica de ordenación PCA para visualizar las similitudes y posibles agrupaciones entre los puntos de muestreo en función de los usos de suelo, esta se construyó considerando el carácter composicional de estos datos, vale decir, tras verificar que no hubiese valores cero se aplicó la transformación CLR y luego se realizó un PCA.

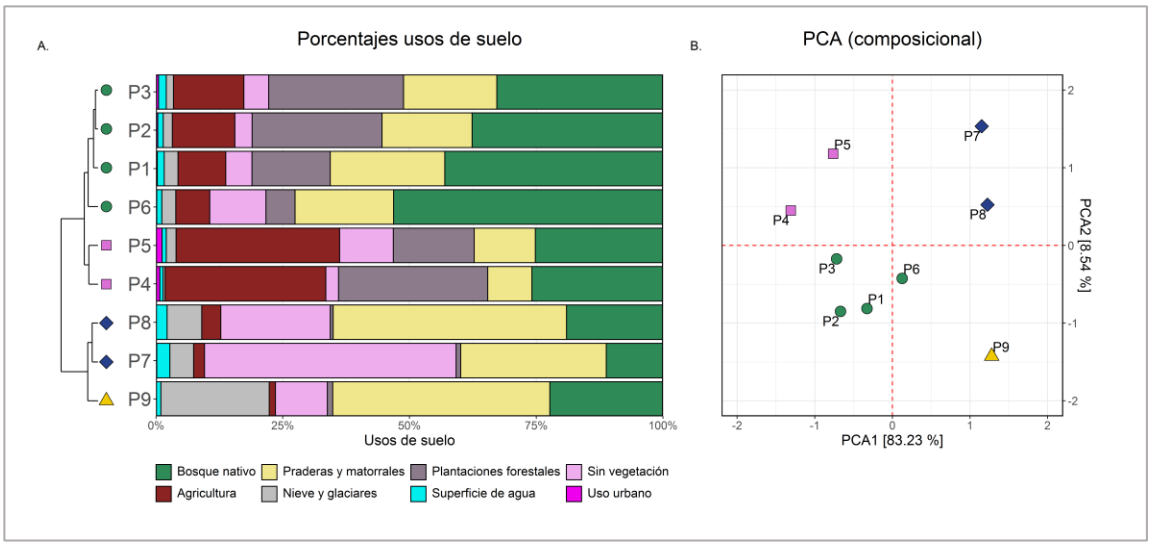

Figura 4.1 (A) Proporciones de los usos de suelo en cuencas aportantes. (B) Gráfica de ordenación PCA de los puntos de muestreo según usos de suelo.

De acuerdo con el análisis clúster jerárquico, los puntos pueden ser agrupados en cuatro conjuntos según sus usos de suelo: el primer grupo (verde – círculo) considera los puntos de los ríos Biobío (P1, P2 y P3) y Longaví (P6), las cuencas de este clúster se caracterizan por un alto porcentaje de bosque nativo (41.6% promedio), proporciones no menores de praderas y/o matorrales (19.6 %) y plantaciones forestales (18.3 %), la agricultura, en tanto, ocupa un porcentaje levemente superior al 10 %; el segundo clúster (lila – cuadrado) agrupa los puntos de los ríos Itata y Ñuble (P4 y P5), sus cuencas destacan por un alto uso agrícola (32 %), pero también altos porcentajes de bosque nativo (25.5 %) y plantaciones forestales (22.7 %), además, son las que presentan los mayores porcentajes de uso urbano (0.96 %). El tercero (azul – rombo) engloba los puntos de los ríos Maule (P7) y Teno (P8), en estas cuencas predominan los usos de praderas y/o matorrales (37.4 %) y sin vegetación (35.6 %), también se observa presencia de bosque nativo, pero su porcentaje resulta menor al 16 %; por último, el cuarto clúster (dorado – triángulo) está conformado solo por P9 (Tinguiririca), su cuenca destaca por un alto porcentaje de praderas y/o matorrales (42.8 %) y por presentar el mayor porcentaje de nieve y/o glaciares (21.4 %), también muestra una superficie no menor de bosque nativo (22.3 %).

El resultado del PCA concuerda con el del análisis clúster jerárquico, en ambos es posible reconocer un patrón espacial en la agrupación de los puntos. En general, se observa que aquellos puntos geográficamente más próximos entre sí muestran mayor similitud en los usos de suelo de sus cuencas, siendo agrupados en un mismo clúster (ver Figura A 4.2.2). La única excepción fue P6 (Longaví), cuya cuenca mostró mayor similitud con las del río Biobío que con las de los ríos Itata o Ñuble. Un último aspecto por mencionar es que el análisis de correlación detectó que las proporciones de tres de los ocho usos de suelo están correlacionadas con la latitud (ver **Figura A 4.2.1**) tanto nieve y/o glaciares como praderas y/o matorrales aumentan sus porcentajes al avanzar hacia el norte, mientras que el porcentaje de plantaciones forestales lo hace hacia el sur.

4.2.3 Variables de calidad medidas en la campaña de terreno

De las diez variables medidas en la campaña de terreno, cinco fueron registradas durante el trabajo de campo y cinco se obtuvieron de los análisis de las muestras de agua recolectadas. La **Tabla 4.3** expone los valores obtenidos en cada uno de los nueve puntos de muestreo para las diez variables de calidad medidas en la campaña.

Tabla 4.3 Valores de parámetros fisicoquímicos medidos en la campaña de terreno.

PM	Mediciones en terreno					Análisis en laboratorio				
	pH	Temp. (°C)	O.D. (mg/l)	Cond. (S/cm)	Turbidez (NTU)	P total (mg/l)	N total (mg/l)	$NO_3 + NO_2$ (mg/l)	TKN (mg/l)	DQO (mg/l)
P1	5.62	8.80	11.82	72	5.9	0.293	0.779	0.161	0.618	1
P2	6.15	10.53	10.93	65	4.4	0.130	0.392	0.222	0.170	2
P3	6.59	11.76	11.44	80	4.9	0.133	1.070	0.171	0.900	4
P4	7.41	12.27	10.91	72	3.9	0.204	1.370	0.312	1.060	7
P5	6.94	13.88	10.82	73	5.0	0.157	1.040	0.337	0.698	3
P6	7.12	12.95	10.42	87	0	0.145	1.370	0.531	0.842	1
P7	8.65	15.91	11.44	200	2.0	0.163	0.724	0.239	0.485	ND*
P8	7.60	18.09	9.20	547	8.3	0.080	0.669	0.147	0.522	2
P9	7.47	19.44	9.24	295	24.8	0.135	0.420	0.074	0.347	6

* No detectado.

En la tabla se observa que el pH, la temperatura y la conductividad tienden a aumentar hacia el norte. Este patrón es respaldado por el análisis de correlación (ver **Figura A 4.2.1**), según el cual las tres variables señaladas tienen una correlación negativa y significativa con la latitud, siendo esta más marcada en el caso de la temperatura. Sin embargo, debido a que los puntos desde el río Longaví hacia el norte fueron visitados consecutivamente dentro del mismo día, es probable que la tendencia observada para la temperatura se haya visto acentuada artificialmente. El pH se mantuvo generalmente en un rango de 6 a 8, con valores levemente ácidos en la zona sur y levemente básicos en la norte, además, destacan los valores de los puntos P1 (Coihue) y P8 (Teno), esto por ser el más ácido y el más básico, respectivamente. Para la conductividad se observa un quiebre en el río Maule, con valores menores hacia el sur y mayores hacia el norte, presentándose el mayor valor en el río Teno (P8). El oxígeno disuelto estuvo generalmente cercano al límite de saturación (en equilibrio con la atmósfera), aunque mostró una correlación positiva con la altitud, es probablemente se deba a la relación inversa que existe entre esta variable y la temperatura. En cuanto a la turbidez, los ríos visitados mostraron aguas claras, con baja o nula turbidez, la excepción fue Tinguiririca (P9), donde se observó una gran carga de sedimentos finos en el río (fondo no visible) y planicie de inundación durante la campaña.

Los parámetros medidos en el laboratorio no mostraron correlaciones significativas con variables espaciales o usos de suelo (ver **Figura A 4.2.1**), es posible que estos estén controlados por factores locales que no fueron considerados durante este estudio (por ejemplo, descargas de aguas residuales), o bien, se vean influenciados por múltiples factores simultáneamente. El fósforo se mantuvo en el rango de 0.1 mg/l a 0.2 mg/l, observándose el mayor y menor valor en los puntos P1 (Coihue) y P8 (Teno), respectivamente. En el caso del nitrógeno, su forma orgánica (TKN) predomina sobre sus formas inorgánicas en casi todos los puntos, con la excepción de P2 (Nacimiento). Por último, la demanda química de oxígeno (DQO) tuvo valores muy bajos en todos los puntos e incluso, en uno de estos la concentración de DQO fue tan baja que no logró ser detectada.

4.2.4 Variables de calidad extraídas desde los registros históricos DGA

Se encontraron más de 60 parámetros fisicoquímicos en los registros DGA, no obstante, debido a la falta de continuidad en algunos de estos, la ausencia de información por largos períodos en otros y la poca variabilidad existente en otros tantos, solo se extrajeron los datos de 13 parámetros de calidad, de los cuales solo siete coindicen o son comparables con aquellos medidos en terreno. Los valores promedio para estos siete parámetros de calidad se muestran a continuación en la **Tabla 4.4**.

Tabla 4.4 Valores promedio para los parámetros fisicoquímicos de calidad extraídos desde los registros históricos de la DGA (1/2).

PM	pH	Temp. (°C)	O.D. (mg/l)	Cond. (S/cm)	DQO (mg/l)	$N\text{-}NO_3$ (mg/l)	$P\text{-}PO_4$ (mg/l)
P1	7.67	13.62	10.21	82.07	13.33	0.33	0.13
P2	-	-	-	-	-	-	-
P3	7.35	16.55	9.40	95.16	16.79	0.30	0.24
P4	7.57	16.37	9.00	101.61	19.12	0.17	0.36
P5	7.93	16.13	9.42	98.45	17.05	0.21	0.20
P6	7.60	16.25	10.39	101.15	15.54	0.33	0.30
P7	8.07	15.27	10.72	134.01	16.93	0.19	0.16
P8	8.50	18.17	11.30	376.43	18.33	0.55	0.17
P9	7.76	16.80	10.17	318.31	14.16	0.91	0.33

Al comparar lo de la tabla anterior con lo encontrado en la campaña se detectan diferencias porcentuales importantes en más de una variable (ver **Figura A 3.4.2**). Estas diferencias podrían ser consecuencia de que los protocolos de la DGA y los de la campaña difieran entre sí, pero también producto de que los valores DGA corresponden a promedios históricos (sin discriminar entre estaciones), mientras que los de campaña provienen de una única medición hecha en primavera. Estas diferencias son mayores para las variables medidas en laboratorio, no obstante, debe considerarse que en el caso del nitrógeno y del fósforo las variables medidas en la campaña no coinciden con las disponibles en los registros DGA. Para el nitrógeno, en la DGA se encontró la concentración de nitrógeno

de nitratos, mientras que en la campaña se midió el nitrógeno total y sus formas orgánicas e inorgánicas (esta última forma se usó en la comparación); respecto al fósforo, en la DGA se encontró el fósforo de ortofosfato, mientras que en la campaña fue medido el fósforo total. A lo anterior se debe agregar que los registros DGA para $N\text{-}NO_3$ y $P\text{-}PO_4$ presentaron importantes discontinuidades, entre estas, un vacío de información para los siete años anteriores al 2019. Para la DQO se observó gran variabilidad en los registros DGA, tomando valores tan bajos como los vistos en la campaña y tan altos como 100 mg/l.

Aunque las diferencias porcentuales de las variables medidas *in situ* fueron menores al 50 % (ver **Figura A 3.4.2**), existen también diferencias relevantes en cuanto a las tendencias y los rangos de estas variables que no pueden ser ignoradas. El análisis de correlación indicó que, a diferencia de lo observado en la campaña, solo la conductividad exhibe una correlación significativa (y positiva) con la latitud (ver **Figura A 4.2.3**), mientras que el pH y la temperatura carecen de patrones espaciales claros. En cuanto a los valores de pH DGA, todos resultaron mayores a siete, lo que resulta discordante con lo observado en la campaña, donde algunos puntos mostraron valores levemente ácidos. No obstante, en ambos casos (DGA y campaña) los mayores valores de pH están en el norte, específicamente en los ríos Maule (P7) y Teno (P8). Para la temperatura, el rango de los valores DGA es más acotado y no existe una diferencia marcada entre norte y sur. Esta mayor uniformidad en los valores puede ser consecuencia de que los puntos del norte están más próximos a la cordillera de los Andes y a una mayor elevación, lo que podría derivar en la ocurrencia de bajas temperaturas en el año, reduciéndose el promedio anual para estos puntos. La conductividad muestra una tendencia similar a la observada en la campaña, con mayores valores en el norte, sin embargo, la magnitud estos es menor, además, el quiebre entre valores bajos y altos ya no ocurre en el río Maule, sino que en el Teno. Por último, los valores de oxígeno disuelto fueron muy similares a los observados en la campaña y, dado los valores de temperatura que observados en los registros DGA, es probable que las concentraciones de oxígeno estén cercanas al límite de saturación.

En general, los valores de $P - PO_4$ en los registros DGA fueron levemente mayores a los de fósforo total observados en la campaña (sería esperable que fuesen menores), la única excepción fue P1 (Coihue), donde el valor DGA resultó considerablemente menor al de campaña. Los valores de $P - PO_4$ se mantuvieron entre 0.13 mg/l y 0.36 mg/l, y no se detectaron correlaciones significativas con otras variables (ver **Figura A 4.2.3**). En el caso del nitrógeno, la mitad de los puntos tuvieron concentraciones de nitratos y nitritos (como conjunto) mayores a las concentraciones de $N - NO_3$ en los registros históricos, mientras que la otra mitad mostró valores inferiores. Las concentraciones de $N - NO_3$ (DGA) estuvieron entre 0.17 mg/l y 0.91 mg/l y se observó una correlación significativa (y positiva) entre esta variable y la altitud, con los mayores valores en los ríos Teno y Tinguiririca. Por último, los valores de DQO (DGA) se mantuvieron entre 10 mg/l y 20 mg/l (concentraciones mayores a las de campaña) y no se detectaron correlaciones significativas entre esta variable con otras.

Los valores promedio para los otros seis parámetros fisicoquímicos de calidad obtenidos desde los registros DGA se muestran en la **Tabla 4.5**, cabe destacar que la primera de estas variables (Cociente $N\text{-}NO_3/ P\text{-}PO_4$) no se encontró directamente en dichos registros, sino que fue determinada en función de estos.

Tabla 4.5 Valores promedio para los parámetros fisicoquímicos de calidad extraídos desde los registros históricos de la DGA (2/2).

PM	Cociente $N\text{-}NO_3/ P\text{-}PO_4$	Cloruro (mg/l)	Sulfato (mg/l)	Mn total (mg/l)	Al total (mg/l)	Fe total (mg/l)
P1	15.48	5.16	7.39	0.02	0.37	0.40
P2	-	-	-	-	-	-
P3	7.00	5.64	9.16	0.03	0.54	0.51
P4	6.42	4.08	4.82	0.05	0.48	0.44
P5	7.52	4.85	5.99	0.03	0.67	0.46
P6	11.36	3.77	7.66	0.02	0.96	0.31
P7	13.02	8.64	16.31	0.02	1.10	0.43
P8	18.88	30.67	73.40	0.05	0.77	0.94
P9	24.55	15.97	62.43	0.11	3.62	2.14

En la tabla anterior se aprecia que casi todas las variables presentan sus menores valores en la zona sur y los más altos en el norte, no obstante, según el análisis de correlación, este patrón se relaciona con la altitud y no directamente con la latitud (todas se correlacionan significativamente con la altitud, pero solo cuatro con la latitud). El cociente $N\text{-}NO_3$/ $P\text{-}PO_4$ tomó valores entre 6.4 y 24.6 y fue la variable con un más claro patrón asociado a la latitud, aunque P1 (Coihue) se separó de la tendencia general. Por otro lado, tanto el cloruro como el sulfato se correlacionaron significativamente con la conductividad, con valores considerablemente más altos en la zona norte, específicamente en los ríos Tinguiririca (P9) y Teno (P8). En el caso del cloruro, las concentraciones fueron generalmente inferiores a 10 mg/l, con la excepción P8 y P9, donde fueron de 16 y 31 mg/l, respectivamente. En cuanto al sulfato, la mayor parte de los puntos tuvo concentraciones menores a 17 mg/l, mientras que en P8 y P9 se observaron valores entre 60 y 75 mg/l (considerablemente mayores a los de otros puntos). Con respecto a los metales, para los tres se observaron mayores concentraciones en el punto P9 (Tinguiririca), con concentraciones de 0.11, 3.62 y 2.14 mg/l para manganeso, aluminio y hierro, respectivamente. En los puntos restantes las concentraciones de manganeso fueron iguales o inferiores a 0.05 mg/l, mientras que para el aluminio inferiores a 1.2 mg/l y para el hierro fueron menores a 1 mg/l.

4.3 Secuenciación y procesamiento bioinformático

Pese a que se recolectaron 9 muestras de lecho (L) y 9 muestras de agua (A) para ser analizadas solo fue posible secuenciar 16 de las 18 muestras: en las muestras de lecho del río Maule (P7L) y de agua del río Tinguiririca (P9A) no se encontró suficiente material genético para ser secuenciado. Tras los procesos de extracción, amplificación, purificación y secuenciación del material genético se obtuvieron un total de 1 517 269 secuencias, estas fueron posteriormente procesadas con el paquete DADA2, luego de lo cual el número de secuencias bajó a 1 178 685 (agrupadas en 7212 ASVs), lo que equivale

al 77.7 % de la cantidad inicial. La **Tabla A 4.3.1** muestra, a modo de resumen, los resultados del procesamiento bioinformático con DADA2 expresado en función del número de lecturas por muestra tras cada etapa (filtro de calidad, remoción de ruido y eliminación de quimeras). Las secuencias que se mantuvieron tras todos los procesos anteriores fueron sometidas a un filtro según taxonomía, tras este quedaron un total de 1 139 210 secuencias válidas (agrupadas en 6 762 ASVs), con el máximo y mínimo número de lecturas encontrados en las muestras de lecho pertenecientes al río Biobío, específicamente en las tomadas en Nacimiento (P2L) y Coihue (P1L), con 40 211 y 118 820 secuencias totales, respectivamente.

Como paso previo a los análisis de composición taxonómica y diversidad beta se aplicó un filtro a la matriz de lecturas basado en la abundancia relativa y prevalencia de las ASVs, los efectos de este en los números de lecturas y de ASVs se muestran en la **Tabla A 4.3.2**. A grandes rasgos, este filtro redujo el número de ASVs a 4 372 (30.3 % menor) y la cantidad de secuencias válidas a 1 094 974 (3.9 % menor), en tanto que el valor máximo decayó a 116 288 y el mínimo a 31 675, manteniéndose estos en las muestras P2L y P1L, respectivamente. Además, se observó una caída desde 65.9 % a 53.8 % en la proporción de ceros presentes en la matriz de lecturas.

Por otro lado, este filtro de abundancia/prevalencia también produjo cambios en la cantidad de taxones por nivel taxonómico (ver **Tabla A 4.3.3**). No obstante, debe señalarse que dicha tabla solo considera taxones con taxonomía conocida hasta el nivel de interés, en consecuencia, no es posible confiar en la información provista para los niveles inferiores (Género/Especie), puesto que estos presentan grandes proporciones de ASVs sin información taxonómica (ver **Tabla A 4.3.4**). Pese a lo anterior, sí es posible afirmar con seguridad que existe un mayor número de taxones en el conjunto de muestras de lecho que en el de agua, esto para todos los niveles taxonómicos.

4.4 Diversidad alfa de las comunidades

4.4.1 Riqueza observada y riqueza estimada

En la **Figura 4.2** se muestra la variación de la riqueza observada (A) y de la riqueza estimada (B) entre los distintos puntos de muestreo, tanto para las muestras de agua (azul) como para las de lecho (rojo), adicionalmente, se incluyen gráficos de cajas y bigotes para visualizar las diferencias entre estos subconjuntos.

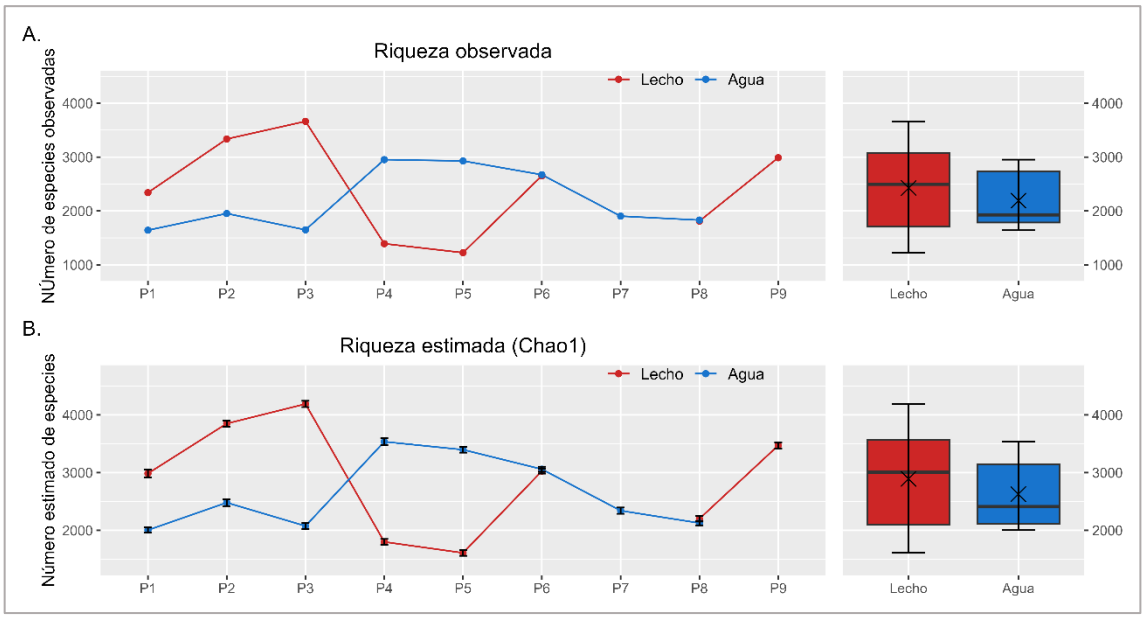

Figura 4.2 Valores de riqueza observada (A) y riqueza estimada (B) para las muestras de agua y de lecho en cada punto de muestreo.

Se aprecia que la riqueza observada y la estimada siguen la misma tendencia, existiendo un desfase prácticamente constante entre ambas curvas tanto para agua como lecho. De acuerdo con la **Tabla A 4.4.1,** el valor de riqueza observado es, en promedio, un 82.9% del valor de riqueza real (estimado usando Chao1 como referencia), lo cual indica que se logró capturar un porcentaje importante de los individuos presentes en cada ambiente (en algunos casos, como P6L y P3L, cercano a un 90 %).

En las muestras de agua la riqueza observada se mantuvo entre 1 642 y 2 952 ASVs, mientras que la estimada entre 2 004 y 3 535. Los mayores valores se encontraron en P4 (Itata) y P5 (Ñuble), cabe señalar que estos dos puntos están próximos entre sí y sus cuencas poseen una distribución de usos de suelo similar (ver **Figura 4.1**). A partir de Ñuble la riqueza decrece de sur a norte, alcanzando una riqueza estimada de 2 126 en P8 (Teno), el tercer valor más bajo en agua. Los menores valores se encontraron en el río Biobío, con riquezas estimadas de 2 004 y 2 075 en los puntos P1 (Coihue) y P3 (Santa Juana), respectivamente. No se halló un patrón en el sentido este – oeste en el río Biobío, dado que P2 (Nacimiento) mostró mayor riqueza que P1 y P3. Lo anterior podría deberse a la confluencia con el río Vergara aguas arriba de P2, cauce que se ve influido por la cordillera de Nahuelbuta, la cual presenta gran superficie de bósques nativos.

Centrándose ahora en lecho, la riqueza observada estuvo entre 1 125 y 3 663 ASVs y la estimada entre 1 609 y 4 189, lo que denota mayor variabilidad en lecho que en agua. Las mayores riquezas se encontraron en el río Biobío, específicamente en P2 (Nacimiento) y P3 (Santa Juana), mientras que las menores se hallaron en P4 (Itata) y P5 (Ñuble), contrario a lo visto para agua. Enfocando ahora la atención en los gráficos de cajas y bigotes, se oberva que el subconjunto de agua presenta valores medios de riqueza (observada y estimada) levemente menores que el subconjunto de lecho, no obstante, como se aprecia en la **Tabla 4.6**, las diferencias entre las riquezas en agua y en lecho (observada y estimada) no son significativas.

Tabla 4.6 Resultados de la prueba Wilcoxon – Mann – Whitney para comparar la riqueza de las muestras según origen.

Riqueza	Origen: lecho		Origen: agua		Estadístico (W)	Valor *p*	Conclusión*
	Promedio	D.E.	Promedio	D.E			
Observada	2 426	896	2 191	563	36	0.71	No se puede rechazar Ho
Estimada	2 891	947	2627	616	36	0.71	No se puede rechazar Ho

* Ho: las medias de los subconjuntos de agua y de lecho son iguales.

4.4.2 Índices Shannon – Wiener y Simpson

La **Figura 4.3** exhibe los resultados de los índices Shannon – Wiener (A) y Simpson (B) para las muestras de agua y lecho en cada punto de muestreo. Nuevamente, se incluyen gráficos de cajas y bigotes para comparar las muestras según origen.

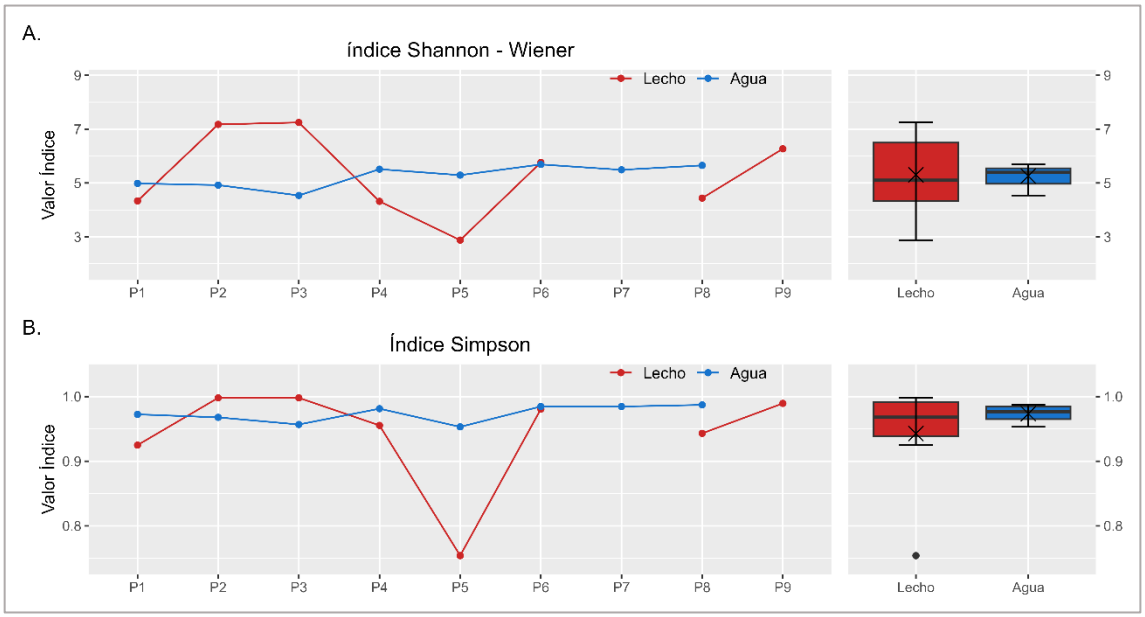

Figura 4.3 Valores de índices Shannon – Wiener (A) y Simpson (B) para las muestras de agua y de lecho en cada punto de muestreo.

Se observa que el índice Shannon – Wiener (S-W) varió más en lecho, con valores entre 2.88 (P5 – Ñuble) y 7.25 (P2 – Santa Juana). La forma de esta curva se asemeja a las de riqueza y nuevamente los puntos P2 y P3 exhibieron los mayores valores de diversidad, lo que no es de extrañar, dado que este índice se ve fuertemente influido por la riqueza (aumenta con la riqueza). No obstante, existen ciertas diferencias en las relaciones entre puntos, por ejemplo, P1 (Coihue) mostró mayor riqueza que P4 (Itata) y P8 (Teno), pero su nivel de diversidad S-W no difiere mayormente con las de estos dos puntos, lo que indicaría una distribución más homogénea entre ASVs en estos últimos puntos que en P1. En tanto, P5 (Ñuble) nuevamente fue el punto con menor diversidad de lecho.

Centrándose ahora en el conjunto de agua, el índice Shannon- Wiener estuvo entre 5.26 (P3 – Santa Juana) y 5.69 (P6 – Longaví). A diferencia de lecho, en agua existen aparentes patrones espaciales. Se observa un incremento de S-W en el sentido sur – norte, con los menores valores ubicados en el río Biobío y los mayores en los ríos Longaví (P6) y Teno (P8). En este conjunto se aprecia también una tendencia en el sentido este – oeste, específicamente en el río Biobío (P1 - Coihue, P2 - Nacimiento y P3 – Santa Juana), el índice S-W incrementa su valor al avanzar hacia el oeste. Los dos subconjuntos de muestras presentaron valores medios de diversidad Shannon – Wiener muy similares (5.30 en lecho v/s 5.26 en agua), no obstante, la desviación estándar en el subconjunto de lecho fue mucho más alta que la del subconjunto de agua (1.56 v/s 0.41). Un aspecto interesante es que prácticamente todas las muestras de lecho mostraron valores más altos o bajos que los observados en el subconjunto de agua, la única excepción fue el punto P6 (Longaví) donde las diversidades de lecho y de agua fueron prácticamente iguales. Finalmente, la prueba Wilcoxon – Mann – Whitney indicó que ninguno de los conjuntos exhibe diferencias significativas para este índice (ver **Tabla 4.7**).

El índice Simpson varió en lecho, tomando valores entre 0.754 (P1 – Coihue) y 0.998 (P2 – Santa Juana), pero muy poco dentro del subconjunto de agua, donde adquirió valores entre 0.953 (P5 – Ñuble) y 0.987 (P8 – Teno). En el caso de las muestras de lecho, nuevamente se observa que P2 (Nacimiento) y P3 (Santa Juana) muestran los mayores valores de diversidad dentro de lecho e incluso entre todas las muestras, mientras que el menor valor, tanto para este conjunto como entre todas las muestras, se encontró en la de muestra de lecho de P5 (Ñuble). En el caso particular de las muestras de agua, se observan los mismos patrones espaciales detectados en los valores del índice S-W, tanto en el sentido este – oeste como también en el sentido sur – norte, aunque para este último el punto P5 (Ñuble) no se apega a la tendencia general, con un valor mucho menor al de sus pares. Para este índice se observa un valor medio de diversidad más alto en el conjunto de agua que en el de lecho (0.973 v/s 0.943) y otra vez se encontró una menor desviación estándar en el conjunto de agua (0.013 v/s 0.081 en lecho). Según la prueba

Wilcoxon – Mann – Whitney no existen diferencias significativas entre agua y lecho (ver **Tabla 4.7**).

Tabla 4.7 Resultados de la prueba Wilcoxon – Mann – Whitney para comparar índices Shannon – Wiener y Simpson entre muestras según origen.

Índice de diversidad	Origen: lecho		Origen: agua		Estadístico (W)	Valor *p*	Conclusión*
	Promedio	D.E.	Promedio	D.E			
Shannon	5.30	1.56	5.26	0.41	32	1.00	No se puede rechazar Ho
Simpson	0.943	0.081	0.973	0.013	29	0.79	No se puede rechazar Ho

* Ho: las medias de los subconjuntos de agua y de lecho son iguales.

Es necesario señalar la prueba Wilcoxon – Mann – Whitney requiere que exista homocedasticidad entre los grupos comparados. De acuerdo con los resultados de la prueba Fligner – Killeen (ver **Tabla A 4.4.2)**, es posible asumir esta condición para los subconjuntos tanto en la riqueza observada como para la estimada, pero no así para los índices Shannon – Wiener y Simpson. Pese a lo anterior, se efectuó la prueba Wilcoxon – Mann –Whitney para estos últimos casos, aunque solo con fines informativos.

4.5 Taxones y ASVs comunes entre comunidades

La **Figura 4.4** muestra diagramas de Venn para los niveles Filo, Clase y ASV, en cada uno de estos exhibe la cantidad de elementos comunes entre todas las muestras (morado), los exclusivamente comunes entre muestras de agua (azul) y los exclusivamente comunes en las de lecho (rosado). La suma entre los elementos comunes a todas las muestras (morado) y aquellos exclusivamente comunes en un determinado subconjunto (rosa o azul) entregan como resultado el total de elementos comunes de dicho subconjunto. Por último, el número entre paréntesis que acompaña la etiqueta de cada nivel indica el total

de elementos detectados en este, en tanto que los números entre paréntesis que acompañan las etiquetas de agua y de lecho indican el total de elementos en cada subconjunto.

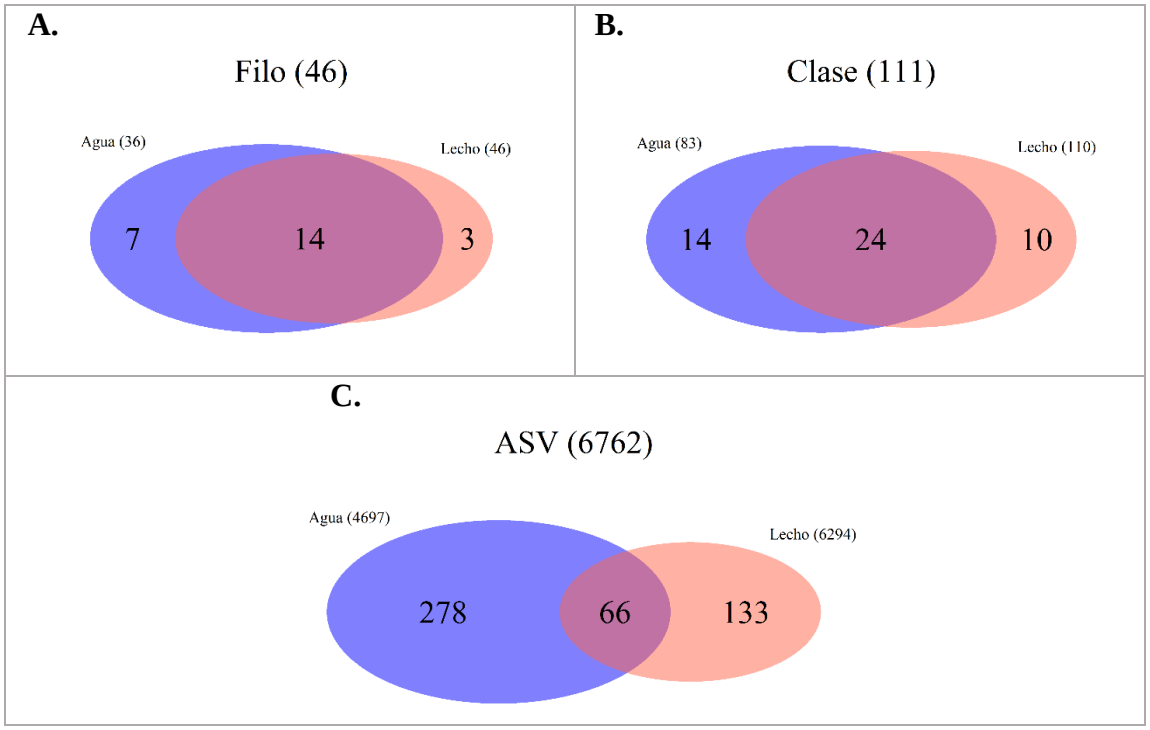

Figura 4.4 Diagramas de Venn para los taxones comunes bajo los niveles de Filo (A), Clase (B) y ASV (C).

Lo primero que se observa en los diagramas Venn es que el subconjunto de agua siempre mostró un mayor número de elementos comunes que lecho, de igual forma, se aprecia que el área compartida entre ambos conjuntos (morado) reduce su tamaño relativo con respecto a las áreas totales de cada conjunto al disminuir el nivel taxonómico. En la **Tabla** 4.8 se muestran los cocientes (%) entre elementos comunes y totales para el total de muestras, para agua y para lecho, estos valores se dedujeron de la **Figura 4.4**.

Tabla 4.8 Cociente entre elementos comunes y elementos totales en conjunto total de muestras y en subconjuntos de agua y de lecho.

Nivel	Todas las muestras	Subconjunto agua	Subconjunto lecho
Filo	30.4 %	58.3 %	37.0 %
Clase	21.6 %	45.8 %	30.9 %
ASV	1.0 %	7.3 %	3.2 %

Según se aprecia en la tabla anterior, el subconjunto de agua no solo muestra un mayor número de elementos comunes que el subconjunto de lecho, sino que estos constituyen una mayor proporción del total en este subconjunto que en el de lecho, esto en los tres niveles. También se observa que la razón entre elementos comunes a todas las muestras y el total de elementos encontrados entre todas estas (primera columna) se reduce al pasar a niveles inferiores, alcanzando el mínimo valor en el nivel ASV (1 %). En base a estos resultados se esperaría que en los próximos análisis las muestras de agua tiendan a mostrar mayor similitud entre ellas, mientras que las de lecho difieran más entre sí.

En el caso particular del nivel ASV se efectuó un análisis más detallado, los resultados se exhiben en la **Figura 4.5**. En esta se muestran las proporciones (como porcentajes) de ASVs (A) y de lecturas (B) pertenecientes al núcleo, esto para las muestras de agua (azul) y de lecho (rojo). Se incluyen, además, gráficos de cajas y bigotes para facilitar la comparación entre los subconjuntos.

En la figura se observa que las muestras de agua, en promedio, presentan mayores proporciones de ASVs y lecturas comunes que las de lecho, siendo estas diferencias estadísticamente significativas (ASVs: W = 6, $p < 0.01$, efecto = 0.68; Lecturas: W = 0, $p < 0.001$, efecto = 0.84). Tanto en agua como en lecho se observa que las ASVs pertenecientes al núcleo representan una pequeña fracción del total de una muestra (Agua: 17%; Lecho:10%), pero el porcentaje de lecturas asociadas resulta importante (Agua: 60; Lecho: 27%).

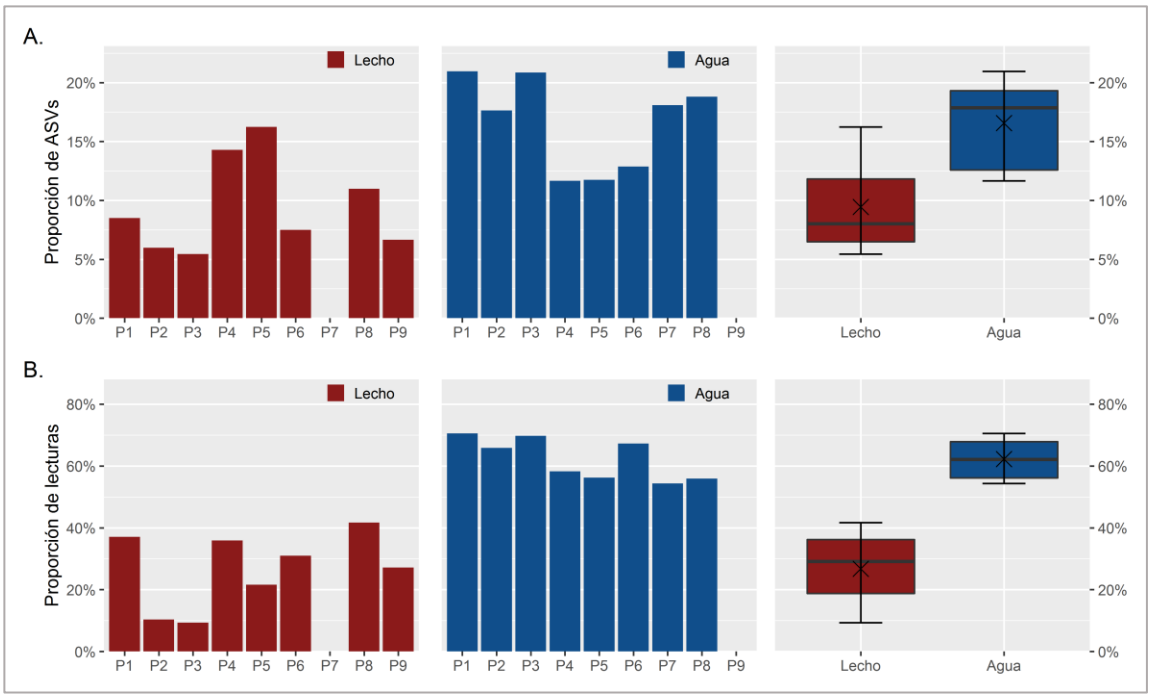

Figura 4.5 Porcentajes de ASVs comunes (A) y de lecturas comunes (B) en los subconjuntos de muestras de agua y de lecho.

4.6 Composición taxonómica de las comunidades

Tras aplicar el filtro de abundancia/prevalencia y agrupar ASVs según taxonomía, se lograron identificar un total de 40 Filos y 109 Clases entre todas las muestras (ver **Tabla A 4.3.3**). En el caso particular del subconjunto de lecho se encontraron 40 Filos y 107 Clases, en tanto que en el subconjunto de agua se hallaron 27 Filos y 82 Clases.

4.6.1 Nivel taxonómico de Filo

La **Figura 4.6** muestra la composición de las comunidades bacterianas para el nivel de Filo, esto con respecto a los 10 taxones más abundantes entre todas las muestras, los que se ordenan según sus abundancias relativas promedio. Los valores numéricos, en tanto, pueden ser revisados en la **Tabla A 4.6.1**.

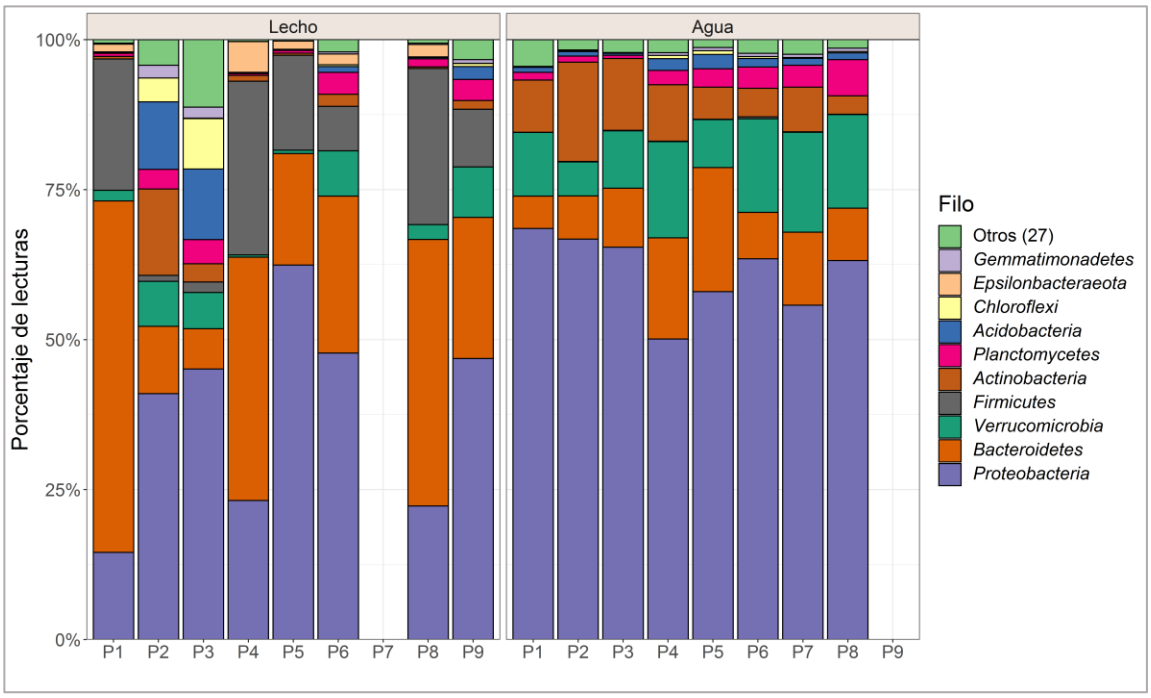

Figura 4.6 Composición de las comunidades bacterianas a nivel de Filo.

Como se aprecia en la figura, los Filos más abundantes fueron *Proteobacteria* (abundancia promedio de 49.7 %), *Bacteroidetes* (19.9 %), *Verrucomicrobia* (8.27 %), *Firmicutes* (7.1 %) y *Actinobacteria* (5.6 %). Otros Filos con proporciones importantes fueron: *Planctomycetes*, *Acidobacteria*, *Chloroflexi*, *Epsilonbacteraeota* y *Gemmatimonadetes*, con abundancias relativas promedio entre 0.5 % y 2.4 %. Las muestras de agua presentaron composiciones bastante parecidas entre sí y se observa que más del 93 % de las lecturas presentes en estas estuvieron asociadas a los Filos *Proteobacteria*, *Bacteroidetes*, *Verrucomicrobia* y *Actinobacteria*, siendo estos taxones, sin considerar *Bacteroidetes*, más abundantes en este subconjunto que en el de lecho (*Proteobacteria:* 61.4 v/s 37.9 %; *Verrucomicrobia*: 12.2 v/s 4.3 %; *Actinobacteria*: 8.4 v/s 2.9 %). En el caso del subconjunto de lecho se observa gran similitud entre casi todas las muestras excepto por P2L y P3L (Nacimiento y Santa Juana), las cuales resultan similares entre sí, pero notoriamente diferentes a las restantes. P2L y P3L presentan particularmente gran abundancia de *Acidobacteria*, *Actinobacteria*, *Chloroflexi* y *Gemmatimonadetes*. Los Filos *Bacteroidetes*, *Firmicutes* y *Acidobacteria* resultaron ser más abundantes en el

subconjunto de lecho que en el de agua, con abundancias promedio de 28.7 %, 14.1 % y 3.4 %, respectivamente. El Filo *Firmicutes* resultó considerablemente escaso entre las muestras de agua, con una abundancia relativa promedio inferior al 0.2 %.

Como se indicó en la **sección 3.5.3C**, se recurrió a dos paquetes R para determinar aquellos taxones con abundancias significativamente diferentes (TASD) entre las muestras de agua y de lecho. En la **Figura 4.7** se exponen, por un lado, las abundancias relativas promedio junto con la desviación estándar de los diez Filos más abundantes entre todas las muestras (izquierda) y, por otro lado, aquellos Filos identificados como TASD entre muestras de agua y de lecho (derecha). Esto último mediante una tabla que exhibe el tamaño del efecto (DESeq2: log_2FC; ALDEx2: efecto medio) e indica mediante colores si un taxón es significativamente más abundante en lecho (valor en rojo), en agua (valor en azul) o en ninguno (valor en gris), además, la columna "sig" (de significativo) indica con un círculo verde los TASD según DESeq2 y ALDEX2 (simultáneamente).

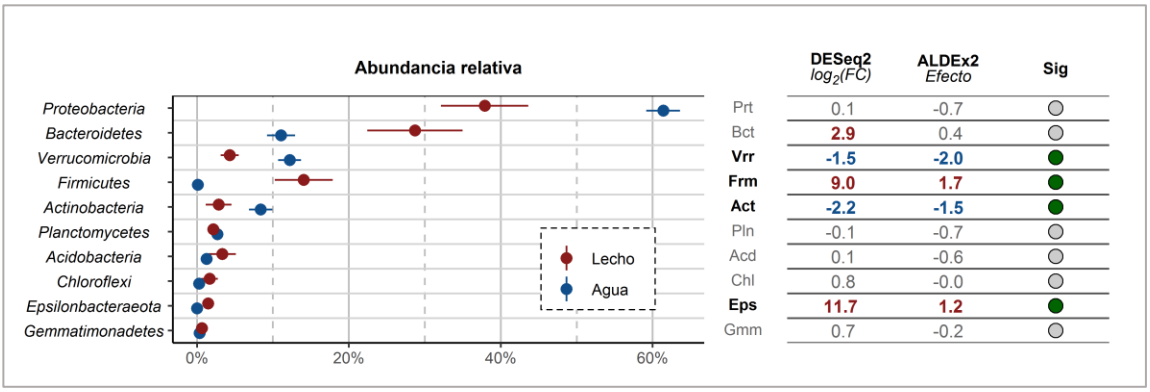

Figura 4.7 Filos identificados como TASD entre muestras de agua y de lecho.

DESeq2 determinó que cinco de los diez Filos más abundantes eran TASD, de estos, tres fueron más abundantes en las muestras de lecho (*Bacteroidetes*, *Firmicutes y Epsilonbacteraeota*) y dos en de agua (*Verrucomicrobia* y *Actinobacteria*). El algoritmo de ALDEx2, por su parte, identificó cuatro Filos como TASD, entre estos aquellos reconocidos por DESeq2 como más abundantes en agua y dos de los tres anteriormente

reconocidos como más abundantes en lecho, específicamente aquellos que mostraron un mayor tamaño de efecto según DESeq2 (*Epsilonbacteraeota* y *Firmicutes*).

4.6.2 Nivel taxonómico de Clase

Con respecto al nivel Clase, la **Figura 4.8** exhibe la composición de cada comunidad bajo según los 20 taxones más abundantes (ordenados por abundancia relativa promedio). Los valores numéricos pueden ser encontrados en la **Tabla A 4.6.2**.

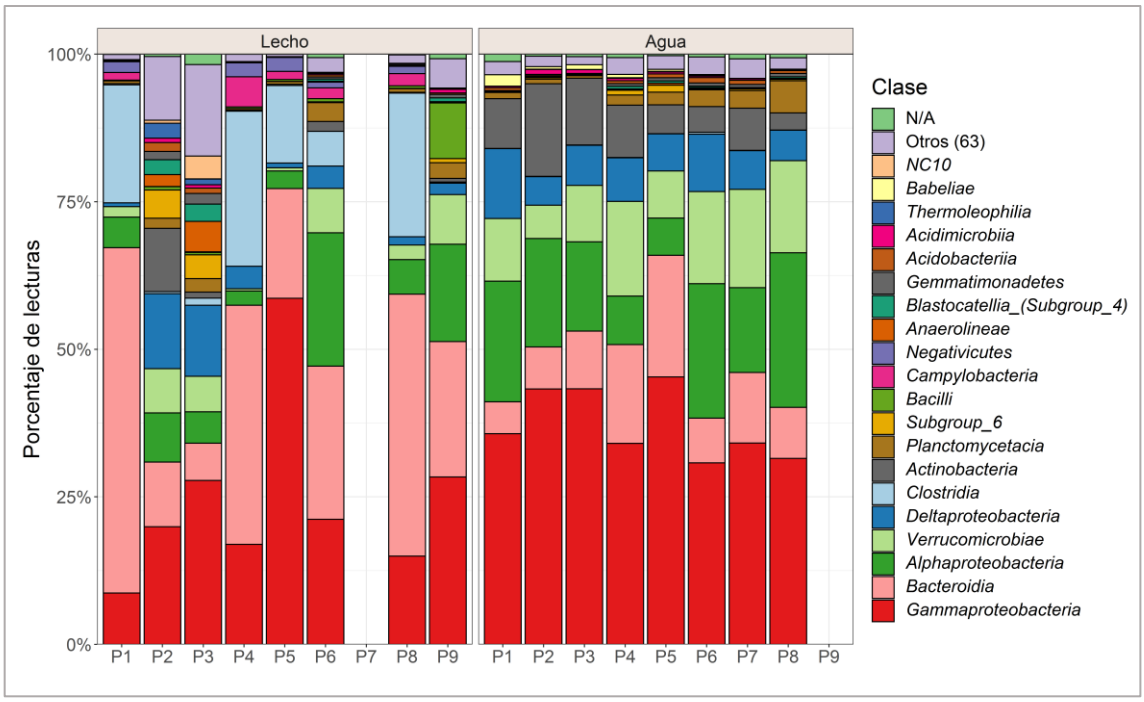

Figura 4.8 Composición de las comunidades bacterianas a nivel de Clase.

Es posible apreciar que las cinco Clases más abundantes, entre todas las muestras, corresponden a *Gammaproteobacteria* (30.9 %), *Bacteroidia* (19.7 %), *Alphaproteobacteria* (12.6 %), *Verrucomicrobiae* (8.3 %) y *Deltaproteobacteria* (6.0 %), de las cuales tres pertenecen al Filo *Proteobacteria*. Otras Clases con abundancias

relativas promedio (entre todas las muestras) superiores al 1.0 % fueron: *Clostridia* (5.8 %), *Actinobacteria* (4.9 %) y *Planctomycetacia* (1.8 %).

En este nivel también se observa gran homogeneidad entre las muestras de agua, estas se encuentran dominadas principalmente por seis taxones: *Gammaproteobacteria*, *Bacteroidia*, *Alphaproteobacteria*, *Verrucomicrobiae*, *Deltaproteobacteria* y *Actinobacteria*. Las clases anteriores constituyen, en promedio, un poco más del 92 % de las lecturas totales de las muestras de este conjunto. Las Clases *Gammaproteobacteria* y *Alphaproteobacteria* resultan más abundantes en este subconjunto, en tanto que las Clases *Verrucomicrobiae* y *Actinobacteria* son notoriamente más abundantes que en este conjunto que en de lecho, especialmente *Actinobacteria* (7.9 % v/s 1.8 %). Un último aspecto que destaca es que, a diferencia de lo observado en lecho, la categoría "otros" ocupa una proporción bastante uniforme entre las distintas muestras de agua.

Con respecto al conjunto de lecho, la disimilitud que ya se observaba en Filo es ahora bastante más evidente, siendo posible reconocer a lo menos tres grupos bastante definidos: uno conformado por las muestras de los puntos P1L, P4L, P5L, P6L y P8L; otro constituido por las muestras de los puntos P2L, P3L; y, por último, la muestra del punto P9L en solitario. En cuanto al primero, sus muestras están dominadas primordialmente por cuatro Clases: *Gammaproteobacteria*, *Bacteroidia*, *Alphaproteobacteria* y *Clostridia*, constituyendo estas, en promedio, un poco más del 73 % de las lecturas totales de estas muestras. Pese a estar dominadas por cuatro taxones, las muestras del primer grupo exhiben gran variabilidad en cuanto a la abundancia de estas Clases: P1L, P4L y P8L muestran una gran proporción de *Bacteroidia* y *Clostridia*, y particularmente P4L presenta gran abundancia de *Campylobacteria* (5.1 %); P5L destaca por su abundancia de *Gammaproteobacteria*, la cual resulta incluso mayor a las encontradas en cualquiera de las muestras de agua; por último, P6L exhibe gran proporción de *Alphaproteobacteria* y *Verrucomicrobiae*, en relación a las otras muestras de este grupo. El segundo subconjunto (P2L y P3L) posee un alto grado de similitud y presenta diferencias importantes con los

otros, por ejemplo, mientras que P1, P4, P5, P6 y P8 estaban constituidas principalmente por cuatro taxones, en este subconjunto dichos taxones representan un poco más del 40 % de una muestra. Las muestras P2L y P3L destacan por estar compuestas por un mayor número de taxones, aunque con abundancias relativas bajas. Esto queda en evidencia en el tamaño de las barras asociadas a las 20 Clases más abundantes, pero también en la proporción que ocupa la categoría "otros" dentro de estas muestras (13.1 %). Interesantemente, las Clases *Deltaproteobacteria*, *Subgroup_6*, *Anaerolinae y Blastocatellia (Subgroup_4)* resultan notoriamente más abundantes en estas muestras que en las otras de lecho, especialmente la primera de estas (12.4 %). Particularmente, la muestra P3L presenta, a diferencia de todas las otras muestras, abundancias importantes de *anaerolinae* (5.2 %) y de *NC10* (3.9 %). Un último aspecto destacable de ese grupo es la abundancia de *Actinobacteria* en la muestra P2L (10.7 %), la cual resulta comparable con las observadas en las muestras de agua. El tercer subconjunto está conformado únicamente por P9L, esta muestra resulta bastante diferente al resto de las de lecho. A diferencia de P2L y P3L, P9L tiene más del 76 % de sus lecturas asociadas a las Clases *Gammaproteobacteria*, *Bacteroidia*, *Alphaproteobacteria* y *Verrucomicrobiae*, y, en contraste con P1L, P4L, P5L, P6L y P8L, la Clase *Clostridia* constituye menos del 1 % de esta muestra. La Clase *Bacilli* resulta ser considerablemente abundante en P9L, esta tiene asociada una abundancia relativa de 9.4 %, muy superior a las observadas en cualquiera de las otras muestras de lecho.

Por último, considerando ahora el subconjunto de lecho globalmente, pareciera ser que las Clases *Clostridia*, *Bacilli*, *Campylobacteria* y *Negativicutes* son exclusivas de este subconjunto. La primera de estas tiene una abundancia de 13 % en lecho, mientras que en agua *Clostridia* solo alcanza un 0.1 %. Para las Clases restantes, en lecho se observan porcentajes superiores al 1.3 %, mientras que en agua estos resultan ser inferiores a 0.02 % en el caso de *Bacilli* e inferiores a 0.002 % en cuanto a *Campylobacteria* y *Negativicutes*.

Bajo este nivel también se identificaron los TASD entre agua y lecho, los resultados de este análisis son mostrados a continuación en la **Figura 4.9** cuya lectura sigue la misma lógica de la imagen mostrada en la sección de Filos, con la única diferencia de que ahora se incluye también información con respecto a qué Filo pertenece cada taxón.

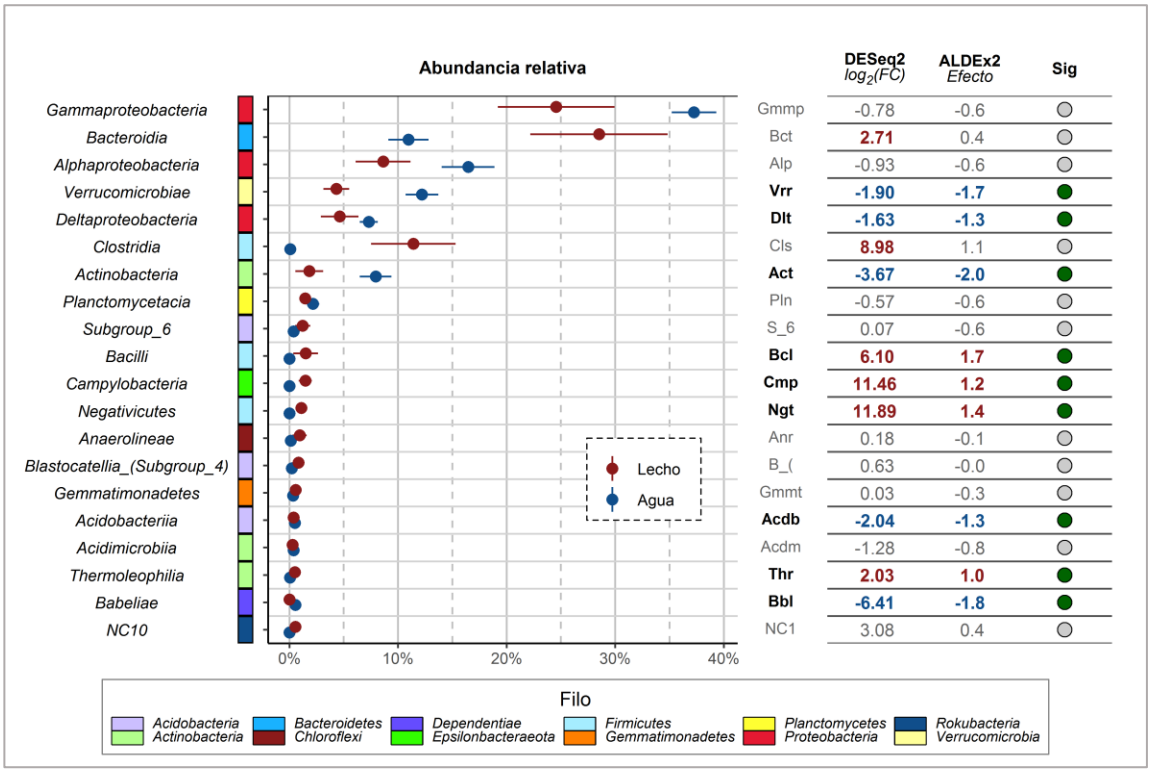

Figura 4.9 Clases identificadas como TASD entre muestras de agua y de lecho.

De acuerdo con la figura anterior, DESeq2 marcó 11 de las veinte Clases mostradas como TASD, mientras que ALDEx2 solo reconoció nueve de estas como TASD. Ambos protocolos coinciden en sus resultados para el subconjunto de agua, estos identificaron las Clases *Verrucomicrobiae*, *Deltaproteobacteria*, *Actinobacteria*, *Acidobacteriia* y *Babeliae* como más abundantes dentro de este subconjunto. Con respecto del subconjunto de lecho, ALDEx2 marcó los taxones *Bacilli*, *Campylobacteria*, *Negativicutes* y *Thermoleophilia* como más abundantes en este grupo que en el de agua, por su parte DESeq2 identificó estos mismos taxones más las Clases *Bacteroidia* y *Clostridia*.

4.7 Diversidad beta de las comunidades

4.7.1 Nivel Filo

La **Figura 4.10** expone los resultados del análisis de diversidad beta mediante gráficas de ordenación para el nivel de Filo, esto bajo los enfoques tradicional (izquierda) y composicional (derecha). Cada flecha verde está asociada a un taxón que mostró una correlación fuerte y significativa con alguno de los ejes principales, las proyecciones de estas flechas indican el nivel de correlación del taxón asociado con cada eje, no obstante, debe señalarse que este nivel de correlación (Spearman) está escalado por un factor (K_e).

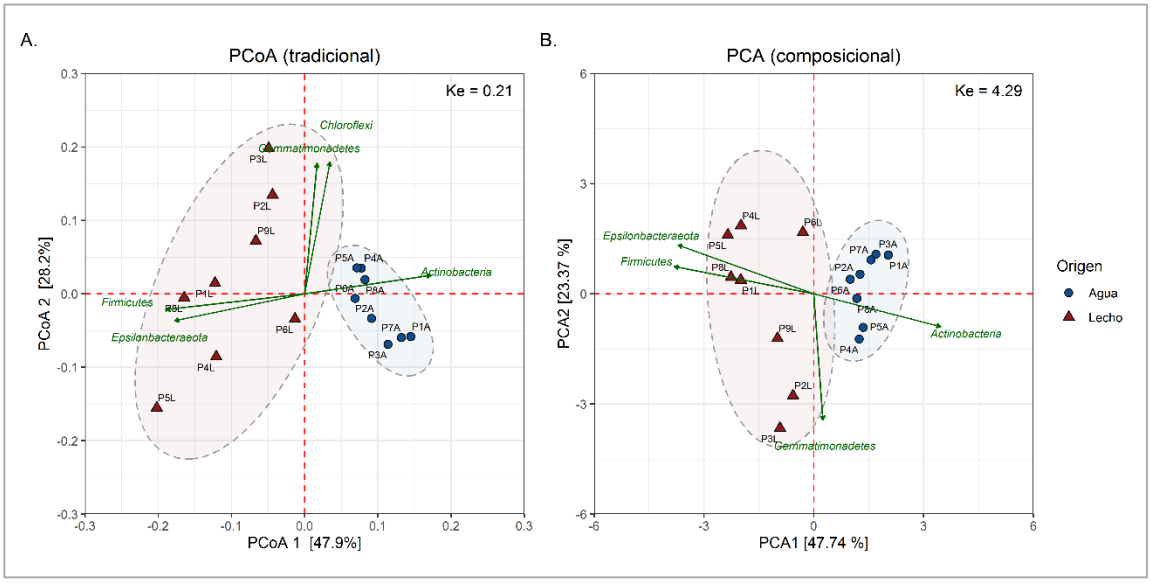

Figura 4.10 Gráficas de ordenación con disimilitudes en el nivel Filo, bajo enfoque tradicional (A) y enfoque composicional (B).

Se observa que en la gráfica PCoA los dos primeros ejes explican un 76.1 % de la variabilidad presente en los datos, mientras que para la PCA sus ejes explican un 71.11 %. A grandes rasgos estas dos gráficas muestran gran similitud entre sí, aunque las muestras de lecho en la PCoA presentan levemente mayor dispersión. Bajo este nivel ya se aprecia una clara separación entre las muestras de agua y las de lecho, observación que es

respaldada tanto por los resultados de la prueba PERMANOVA (ver **Tabla A 4.7.1**), como por los de la prueba ANOSIM (ver **Tabla A 4.7.2**), estas indican que existe una diferencia significativa entre las muestras y que esta es atribuible al origen de estas. Otro aspecto relevante es que las muestras de agua están más cercanas entre sí, lo cual indica mayor similitud entre ellas, resultado que es consistente con la **Figura 4.6**, donde las muestras de agua mostraron gran homogeneidad en sus composiciones. Por el contrario, las muestras de lecho exhiben gran dispersión, lo cual es concordante con las disimilitudes observadas anteriormente para este subconjunto en la **Figura 4.6**.

Centrando ahora la atención en las flechas verdes, ambos enfoques coinciden en que las abundancias de los Filos *Actinobacteria, Firmicutes y Epsilonbacteraeota* están fuertemente correlacionadas con el eje horizontal, asimismo, ambos concuerdan con que la abundancia de *Gemmatimonadetes* estaría correlacionada con el eje vertical, la única diferencia entre ambos enfoques se encuentra en el Filo *Chloroflexi*, el cual solo según el enfoque tradicional estaría fuertemente correlacionado con el eje vertical. Puesto que las muestras de lecho están posicionadas hacia la izquierda y las de agua hacia la derecha, aquellos Filos fuertemente correlacionados con el eje horizontal deberían coincidir con los identificados como TASD por ALDEx2 y DESeq2 (ver **Figura 4.7**). De acuerdo con el sentido de las flechas verdes, *Firmicutes* y *Epsilonbacteraeota* serían más abundantes en las muestras de lecho, mientras que *Actinobacteria* lo sería en las de agua, esto resulta consistente con lo obtenido del análisis de TASD, la única diferencia está en el taxón *Verrucomicrobia*, el cual no mostró una correlación significativa con el eje horizontal, pero que anteriormente fue reconocido como un TASD con mayor abundancia en las muestras de agua (ver **Figura 4.7**).

4.7.2 Nivel Clase

En la **Figura 4.11** se presentan las gráficas de ordenación obtenidas para la diversidad beta bajo el nivel de Clase, nuevamente la gráfica de la izquierda se asocia al enfoque tradicional (A) y la de la derecha al composicional (B). Las flechas verdes presentes en cada gráfica siguen la misma lógica usada en la **Figura 4.10**.

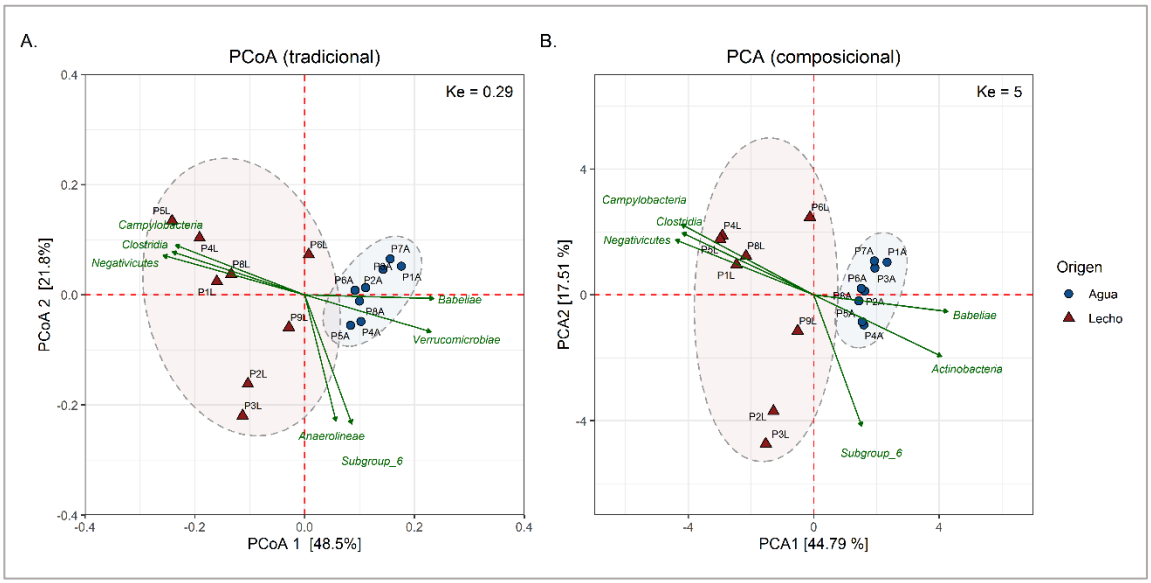

Figura 4.11 Gráficas de ordenación con disimilitudes en el nivel Clase, bajo enfoque tradicional (A) y enfoque composicional (B).

En el caso de la gráfica PCoA sus ejes lograron capturar el 70.3 % de la variabilidad presente en los datos, mientras que para la PCA se tiene que sus ejes solo fueron capaces de explicar un 62.3 %. Los valores anteriores resultan ser menores a los observados para el nivel de Filo, cabe señalar que mientras menor sea la variabilidad capturada por los ejes, mayor es la discordancia entre la distancia observada en la gráfica y la real. Así como en el nivel Filo, en Clase también se observa que en la gráfica PCoA las muestras de lecho están más dispersas, no obstante, al revisarlo globalmente, los patrones observados en cada gráfica resultan consistentes entre sí. Bajo este nivel las muestras de lecho y las de agua también exhiben una clara separación, cuestión que es respaldada por los resultados

de las pruebas estadísticas PERMANOVA y ANOSIM (ver **Tabla A 4.7.1** y **Tabla A 4.7.2**, respectivamente), las cuales detectaron diferencias significativas en la distribución de las muestras y que se explicaría por el origen de estas. Nuevamente se aprecia que las muestras de agua se posicionan muy cercanas entre sí, lo cual da cuenta de la gran similitud existente dentro de este subconjunto, lo que es consistente con lo visualizado en las gráficas de composición taxonómica (ver **Figura 4.8**). En cuanto a las de lecho, la distribución de estas muestras en el espacio sigue un patrón muy similar al detectado visualmente en la **Figura 4.8**, donde se encontró que, de acuerdo con sus composiciones, era posible separar las muestras de lecho en tres clústeres: uno conformado por P1L, P4L, P5L, P6L y P8L; otro agrupando P2L y P3L; y un tercero conformado únicamente por P9L. En ambas gráficas de ordenación se observa que P6L es la única muestra que no sigue la tendencia anteriormente expuesta, este punto se muestra muy diferente a cualquiera de los otros pertenecientes al subconjunto de lecho.

Pasando ahora a la interpretación de las flechas, bajo ambos enfoques las abundancias de las Clases *Campylobacteria, Clostridia, Negativicutes* y *Babeliae* están fuertemente correlacionadas con el eje horizontal. Dado que en este nivel las muestras de lecho también están ubicadas hacia la izquierda y las de agua hacia la derecha, se tiene que los tres primeros taxones señalados anteriormente serían más abundantes en el subconjunto de lecho, mientras que *Babeliae* lo sería en el de agua. A lo anterior hay que agregar que las Clases *Verrucomicrobiae* y *Actinobacteria* también mostraron correlaciones con el eje horizontal, la primera solo bajo el enfoque tradicional y la segunda únicamente bajo el composicional, por lo demás, ambas serían más abundantes en las muestras de agua. Omitiendo el caso de la Clase *Clostridia*, los taxones anteriores y sus patrones (en qué muestras fueron más abundantes) coinciden con lo observado en el análisis de TASD con DESeq2 y ALDEx2. Sin embargo, se aprecia que las Clases *Deltaproteobacteria, Bacilli*, *Acidobacteria* y *Thermoleophilia*, pese a ser reconocidos como TASD, no presentaron correlaciones significativas con el eje horizontal. Una posible explicación de lo anterior es que en el análisis de TASD se trabajó con los subconjuntos de agua y de lecho en forma

separada, mientras que en el análisis de correlación se trabajó con todas las muestras simultáneamente, por tanto, si existen patrones a nivel interno en las muestras de un subconjunto, un TASD no necesariamente mostrará correlación significativa con el eje. Por ejemplo, si existe un aumento de *Deltaproteobacteria* (TASD más abundante en agua) en las muestras de agua al avanzar a la derecha, entonces no será posible notar que es más abundante en las muestras de agua bajo el enfoque de correlaciones. Con respecto al eje vertical, ambos enfoques coinciden en que la abundancia de la Clase Subgroup_6 se incrementa en el sentido inverso al del eje, el enfoque tradicional adicionalmente reconoció que la Clase *Anaerolineae* está fuertemente correlacionada con el eje vertical.

4.7.3 Nivel ASV

Por último, la **Figura 4.12** entrega los resultados obtenidos para el análisis de diversidad beta bajo el nivel de ASV, nuevamente el tradicional se muestra a la izquierda (A) y el composicional en la derecha (B).

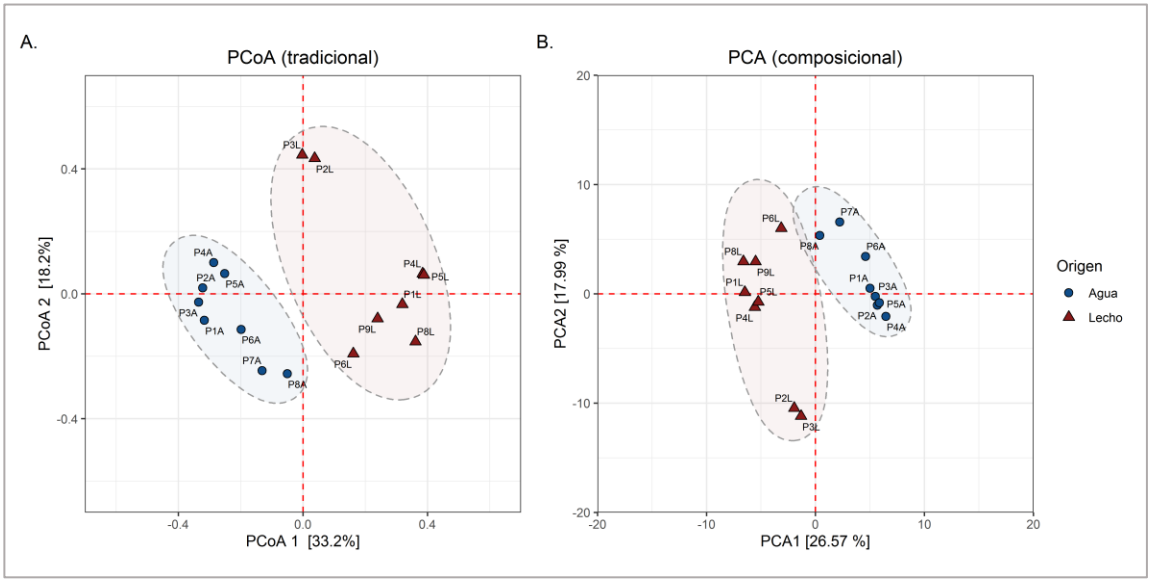

Figura 4.12 Gráficas de ordenación con disimilitudes en el nivel ASV, bajo enfoque tradicional (A) y enfoque composicional (B).

El nivel de variabilidad explicada los ejes de la gráfica PCoA es ahora de 51.4 %, mientras que los ejes de la gráfica PCA logran explicar un 44.56 %. Existe una clara reducción en la cantidad de información que logran capturar los ejes a medida que se pasa a niveles inferiores, esto se debe a que al agrupar ASVs bajo un mismo taxón se genera una reducción en la disimilitud observada entre muestras, en el caso del PCoA esto se explica por la desigualdad triangular, mientras que para el PCA se debe a la disminución en la variabilidad observada en los datos (al conglomerar variables mediante sumas se reduce la covarianza entre muestras). Lo anterior se evidencia también en los gráficos de composición, a medida que se pasa a niveles inferiores las muestras comienzan a mostrar mayor disimilitud entre ellas (ver **Figura A 4.7.1**).

En este nivel (ASV) también se observa que las muestras de agua y de lecho están separadas y que las muestras de agua exhiben gran cercanía entre ellas, no obstante, existen algunas diferencias que vale la pena destacar, ambas asociadas a las muestras de lecho. Primero, en las gráficas para los niveles de Filo y Clase las muestras de lecho se mostraron relativamente centradas con respecto del eje horizontal, pero ahora estas muestran cierta asimetría en relación con este eje. Segundo, es posible notar que en el subconjunto de lecho las muestras tienden a agruparse en dos clústeres bastante definidos: uno conformado por P2L y P3L; y otro constituido por las muestras restantes, las cuales, cabe destacar, están muy próximas entre sí. Resulta interesante y destacable que el clúster conformado por P2L y P3L no solo se posiciona alejado del resto de muestras de lecho, sino que también se encuentra separado de las muestras de agua. Evidentemente, estas dos muestras son muy diferentes a las otras 14 en cuanto a sus composiciones.

4.8 Influencia de variables ambientales

4.8.1 En la diversidad alfa

En la **Figura 4.13** se presenta un resumen de los análisis de correlación realizados, en esta figura el nivel de correlación se indica mediante colores (rojizos para correlaciones negativas y azulados para positivas), en tanto, las correlaciones no significativas se encuentran tachadas con una equis blanca y solo en los casos donde estas fueron estadísticamente significativas se proporciona el valor del coeficiente de correlación Spearman. Cabe señalar que la letra L entre paréntesis hace referencia al subconjunto de lecho, mientras que la letra A se ha reservado para el subconjunto de agua. El detalle de los coeficientes de correlación y los valores *p* calculados (con y sin corrección), se muestran en la **Tabla A 4.8.1** (lecho) y la **Tabla A 4.8.2** (agua).

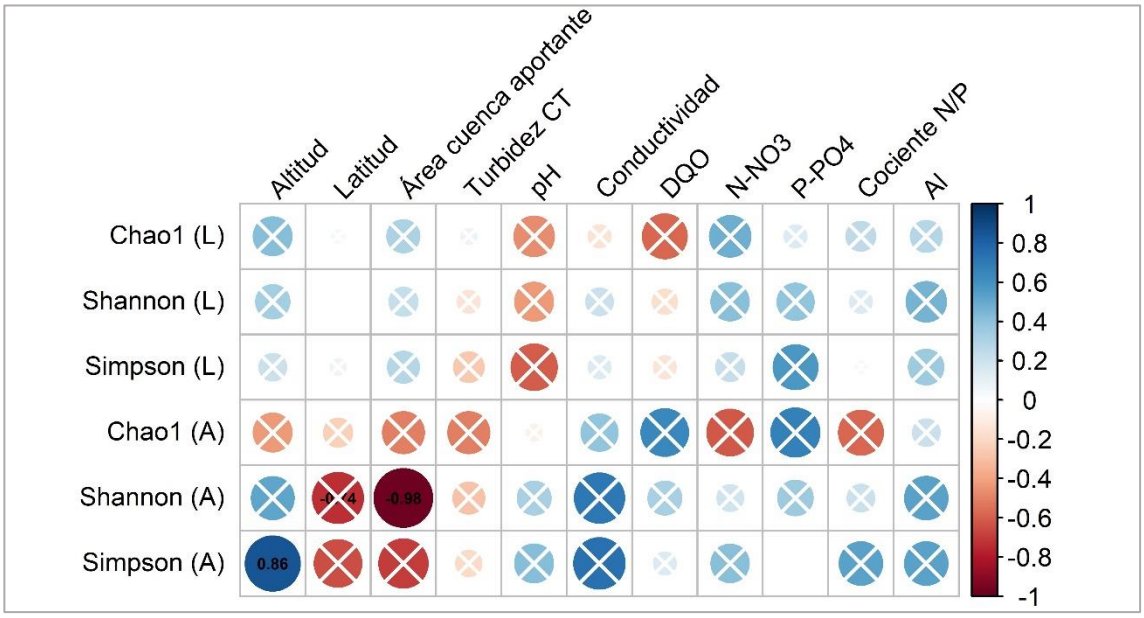

Figura 4.13 Correlaciones entre variables ambientales e índices de diversidad alfa de los subconjuntos de lecho y de agua.

Solo en dos de las 33 combinaciones se encontró un nivel de correlación significativo y fuerte ($r > 0.7$), siendo estas observadas únicamente dentro del subconjunto de agua y específicamente para los índices Shannon – Wiener y Simpson. El índice S-W se correlaciona negativa y fuertemente con el área de la cuenca aportante ($r = -0.98$). Por otro lado, el índice Simpson posee una correlación fuerte y positiva con la altitud ($r = 0.86$). Es posible notar que estos índices también parecen exhibir correlaciones fuertes con la conductividad (pero no significativas), no obstante, al revisar las gráficas de dispersión asociadas no fue posible encontrar patrones claros que respalden esto (no se incluyen).

Para los tres casos donde se observaron correlaciones fuertes y significativas (todos asociados al subconjunto de agua) se construyeron gráficas de dispersión, estas se muestran en la **Figura 4.14**. Para cada una de estas se ajustó un modelo de regresión lineal simple a los puntos graficados (recta de color negro), la ecuación y el valor de R^2 asociados a este ajuste son exhibidos en la esquina superior derecha de las gráficas. En la figura señalada se observa que solo una de las tres combinaciones índice de diversidad alfa – variable ambiental presenta un ajuste excelente ($R > 0.9$), específicamente la relación entre el índice Shannon – Wiener y área de la cuenca aportante. Este ajuste lineal no solo presenta un buen valor de R^2, sino que visualmente puede verificarse que la recta se ajusta bien a los puntos sin existir influencia de valores extremos (pueden elevar artificalmente el valor de R^2), por tanto, no parece que la relación entre estas dos variables sea consecuencia del azar. El índice Shannon – Wiener también mostró una correlación fuerte y significativa con la latitud, pero al examinar el ajuste lineal se encontró que este no fue muy satisfactorio ($R^2 < 0.7$). No obstante, aunque los puntos parecen un tanto distantes con respecto de la recta, si estos son analizados en conjuntos (agrupando aquellos puntos latitudinalmente más próximos entre sí), es evidente que sí existe una tendencia decreciente de este índice con la latitud, con un fuerte descenso en los puntos del río Biobío, por tanto, es posible que sí exista una relación entre este índice con la latitud, pero que la baja cantidad de datos no hayan permitido detectarla apropiadamente. Por último, la tercera combinación con una correlación fuerte y significativa fue la del índice Simpson

con la altitud, sin embargo, la gráfica de dispersión asociada no tuvo un R^2 aceptable, por el contrario, este fue muy bajo, es probable que el nivel de correlación entre el índice Simpson y la altitud haya sido afectado por el alto valor de altitud en P8 (Teno).

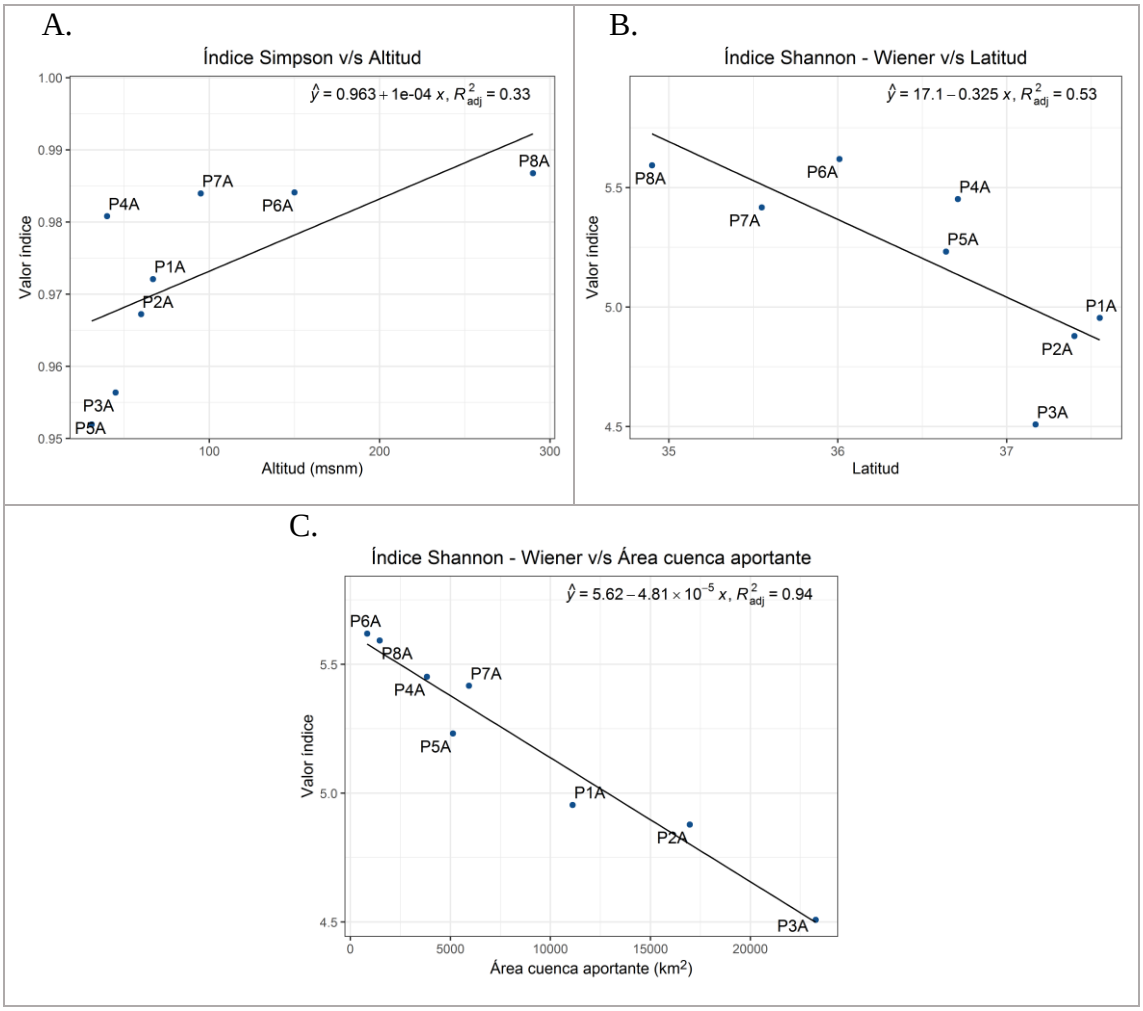

Figura 4.14 Gráficas de dispersión de aquellas combinaciones con niveles de correlación fuertes y significativos.

4.8.2 En la diversidad beta

Se muestran a continuación los resultados obtenidos para las pruebas Mantel, estas se efectuaron bajo los niveles de Filo, Clase y ASV, es decir, las disimilitudes observadas de acuerdo con cada variable fueron comparadas con la diversidad beta observada bajo cada uno de estos niveles. La **Tabla 4.9** exhibe los resultados de esta prueba para el subconjunto de lecho, cabe señalar que "R" da cuenta del nivel de correlación observado (Spearman).

Tabla 4.9 Resultados prueba Mantel entre variables ambientales y diversidad beta en muestras de lecho bajo los niveles Filo, Clase y ASV.

Variable	Filo			Clase			ASV		
	R	***p***	***p* corregido**	**R**	***p***	***p* corregido**	**R**	***p***	***p* corregido**
Altitud	-0.13	0.707	0.939	-0.06	0.573	0.919	0.00	0.376	0.627
Latitud	-0.15	0.772	0.939	-0.05	0.521	0.919	0.11	0.194	0.627
Área cuenca aportante	0.42	0.035	0.385	0.55	0.017	0.187	0.69	0.014	0.154
Turbidez CT	-0.18	0.796	0.939	-0.12	0.668	0.919	-0.12	0.666	0.814
pH DGA	-0.02	0.461	0.939	-0.10	0.599	0.919	0.02	0.381	0.627
Conductividad DGA	-0.35	0.939	0.939	-0.29	0.847	0.940	-0.11	0.526	0.723
DQO DGA	-0.10	0.683	0.939	-0.25	0.877	0.940	-0.36	0.981	0.987
N-NO3 DGA	-0.11	0.561	0.939	-0.05	0.488	0.919	0.04	0.338	0.627
P-PO4 DGA	-0.23	0.868	0.939	-0.29	0.940	0.940	-0.35	0.987	0.987
Cociente N/P DGA	-0.11	0.606	0.939	-0.08	0.544	0.919	0.12	0.222	0.627
Al DGA	-0.07	0.538	0.939	0.05	0.355	0.919	-0.02	0.399	0.627

Según lo mostrado en la tabla anterior, ninguna de las variables ambientales presentó una correlación significativa ni tampoco fuerte con la diversidad beta de las muestras de lecho en ninguno de los tres niveles. La variable ambiental que mostró la mayor correlación con la diversidad beta fue el área de la cuenca aportante (Filo: r = 0.42; Clase: r = 0.55; ASV: r = 0.69). Si se flexibilizase el límite hasta valor $p < 0.05$, entonces la correlación entre el área de la cuenca aportante y la diversidad beta sería significativa en los tres niveles, no obstante, el valor *p* corregido es muy alto.

Siguiendo con las pruebas Mantel, en la **Tabla 4.10** se presentan los resultados obtenidos para el subconjunto de agua, en esta tabla se han destacado aquellos valores *p* corregidos inferiores a 0.05 y también los valores *p* menores a 0.01.

Tabla 4.10 Resultados prueba Mantel entre variables ambientales y diversidad beta en muestras de agua bajo los niveles Filo, Clase y ASV.

Variable	Filo			Clase			ASV		
	R	***p***	***p* corregido**	**R**	***p***	***p* corregido**	**R**	***p***	***p* corregido**
Altitud	-0.02	0.545	0.833	0.12	0.271	0.647	0.68	0.017	**0.047**
Latitud	0.27	0.079	0.435	0.38	0.031	0.341	0.93	**0.001**	**0.011**
Área cuenca aportante	0.29	0.066	0.435	0.21	0.136	0.647	0.06	0.329	0.402
Turbidez CT	-0.03	0.553	0.833	-0.05	0.614	0.750	0.48	0.062	0.097
pH DGA	-0.14	0.761	0.898	0.02	0.401	0.735	0.61	0.012	**0.044**
Conductividad DGA	-0.06	0.606	0.833	0.13	0.217	0.647	0.81	**0.002**	**0.011**
DQO DGA	0.15	0.210	0.770	-0.01	0.513	0.741	-0.19	0.731	0.778
N-NO3 DGA	-0.19	0.834	0.898	-0.10	0.781	0.781	0.29	0.164	0.226
P-PO4 DGA	-0.04	0.507	0.833	-0.13	0.710	0.781	-0.18	0.778	0.778
Cociente N/P DGA	-0.26	0.898	0.898	-0.06	0.539	0.741	0.40	0.054	0.097
Al DGA	0.08	0.285	0.784	0.08	0.294	0.647	0.43	0.058	0.097

Se observa que para este subconjunto existen cuatro variables ambientales cuyo nivel de correlación con la diversidad beta resultó ser significativo bajo los criterios establecidos en la **sección 3.6.3**, estas fueron: altitud, latitud, pH (DGA) y conductividad (DGA), aunque solo en el nivel ASV. Sin embargo, los niveles de correlación del pH (DGA) y de la altitud con la diversidad beta no lograron superar el valor 0.7, el cual se estableció como límite para considerar que una correlación era fuerte, aunque la altitud estuvo bastante próxima (r = 0.68). A diferencia de lo que se observó con el área de la cuenca aportante para el subconjunto de lecho, la cual mostró siempre el mayor nivel de correlación sin importar el nivel considerado, en el caso del subconjunto de agua ninguna de las cuatro variables ambientales presentó un patrón que persistiese en todos los niveles.

4.8.3 En los taxones más abundantes

En la **Figura 4.15** se presentan los resultados de los análisis de correlación de las variables ambientales con los cinco Filos más abundantes en lecho (L) y los cinco Filos más abundantes en agua (A), esta figura sigue la misma lógica que la **Figura 4.13** para su interpretación. En tanto que, el detalle con respecto de los valores *p* (con y sin corrección) y los valores de los coeficientes de correlación obtenidos se muestran en la **Tabla A 4.8.3**, para lecho, y en la **Tabla A 4.8.4**, para agua.

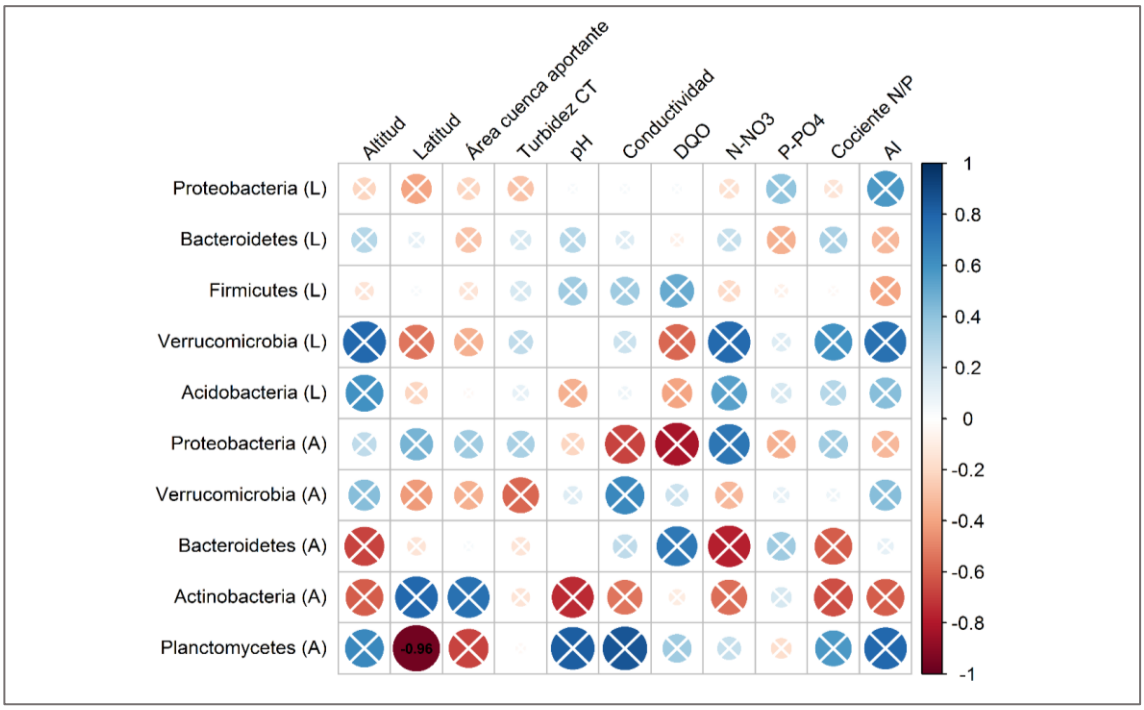

Figura 4.15 Correlaciones entre variables ambientales y Filos más abundantes.

El análisis de correlación arrojó que ninguno de los cinco Filos más abundantes en lecho mostró una correlación significativa con alguna de las variables ambientales estudiadas. En cuanto a la intensidad de las correlaciones, se observa que el Filo *Verrucomicrobia* exhibe correlaciones fuertes con la altitud, con la concentración de nitrógeno de nitratos y con la concentración de aluminio, estas no resultan ser significativas bajo los criterios

establecidos en la **sección 3.6.4**, pero si se libera el límite para el valor *p* hasta 0.05, entonces las dos primeras pasarían a ser significativas (ver **Tabla A 4.8.3**). Cabe señalar que la altitud y la concentración de nitrógeno de nitratos son variables que presentan una correlación fuerte y significativa entre sí (ver **Figura A 4.2.3**). Se decidió efectuar gráficas de dispersión para examinar con mayor profundidad la relación de este Filo con la altitud y con la concentración de nitrógeno de nitrato, estas se muestran en la **Figura A 4.8.1**. Como es posible apreciar en la figura señalada, no se detecta una tendencia en este Filo con respecto a ninguna de las dos variables ambientales.

Solo en el subconjunto de agua fue posible encontrar una correlación significativa y fuerte bajo los criterios adoptados, específicamente entre la abundancia de *Planctomycetes* y la latitud (r = -0.96). Para examinar con mayor detalle esta relación se construyó la gráfica de dispersión correspondiente, esta se muestra a continuación en la **Figura 4.16**. En esta figura se observa una clara tendencia decreciente para la abundancia de este Filo en las muestras de agua con la latitud, adicionalmente, se observa que la recta se ajusta muy bien a los puntos, con un valor R^2 superior a 0.85.

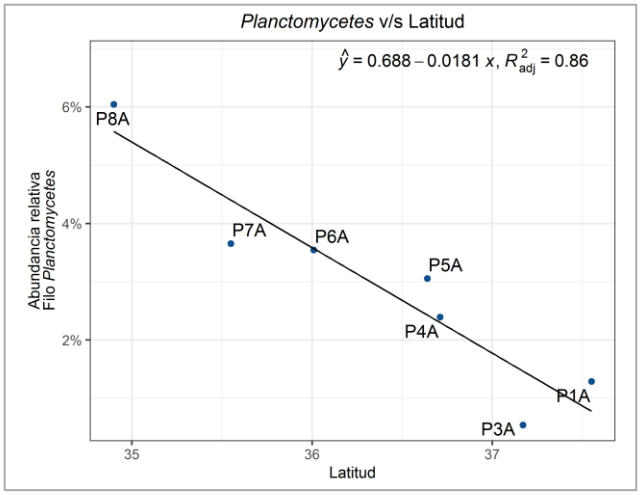

Figura 4.16 Relación del Filo *Planctomycetes* (agua) y la latitud.

Retornando al análisis de la **Figura 4.15**, visualmente se detectan otras combinaciones con correlaciones fuertes dentro esta, al revisarlas en la **Tabla A 4.8.4** es posible notar que algunas de estas correlaciones podrían ser significativas si se libera la restricción del valor *p* (sin corregir) hasta un valor de 0.05, entre estas las correlaciones de: la abundancia de *Proteobacteria* con la DQO, la de *Bacteroidetes* con la concentración $N - NO_3$, la de *Actinobacteria* con la latitud y la de *Planctomycetes* simultáneamente con la altitud, el pH y la conductividad. Para estas relaciones también se construyeron gráficas de dispersión, estas se exhiben en anexos (ver **Figura A 4.8.2**, **Figura A 4.8.3** y **Figura A 4.8.4**). En las figuras señaladas se observa que solo la relación del Filo *Planctomycetes* (agua) y el pH parece tener potencial, la recta del modelo lineal simple presenta un buen ajuste ($R^2 > 0.7$) y no se observa distorsión producto de valores extremos.

Continuando, en la **Figura 4.17** se aprecian los resultados de los análisis de correlación de las variables ambientales con las cinco Clases más abundantes en lecho (L) y con las cinco más abundantes en agua (A). El detalle con respecto de los coeficientes de correlación y los valores *p* (con y sin corrección) se muestran en la **Tabla A 4.8.5**, para lecho, y en la **Tabla A 4.8.6**, para el subconjunto de agua. Nuevamente en el subconjunto de lecho no se observa ningún taxón con una correlación significativa con alguna de las variables ambientales, no obstante, dos Clases presentan correlaciones fuertes con algunas de estas, específicamente *Alphaproteobacteria* y *Deltaproteobacteria*. Al revisar los valores *p* (con y sin corrección) asociados a estas Clases, esto en la **Tabla A 4.8.5**, se observa que si se permiten valores *p* hasta un máximo de 0.05, entonces la Clase *Alphaproteobacteria* presenta correlaciones significativas con la altitud y con la concentración de nitrógeno de nitratos. Se optó por construir gráficas de dispersión para examinar con mayor detalle estas dos correlaciones (ver Figura A 4.8.5), con lo cual se encontró que únicamente la correlación entre la abundancia de *Alphaproteobacteria* con la concentración de nitrógeno de nitratos presenta potencial. Aunque el modelo de regresión lineal ajustado a esta gráfica no dispone de un buen R^2 (< 0.1), es muy probable que el punto P6L sea un valor anómalo.

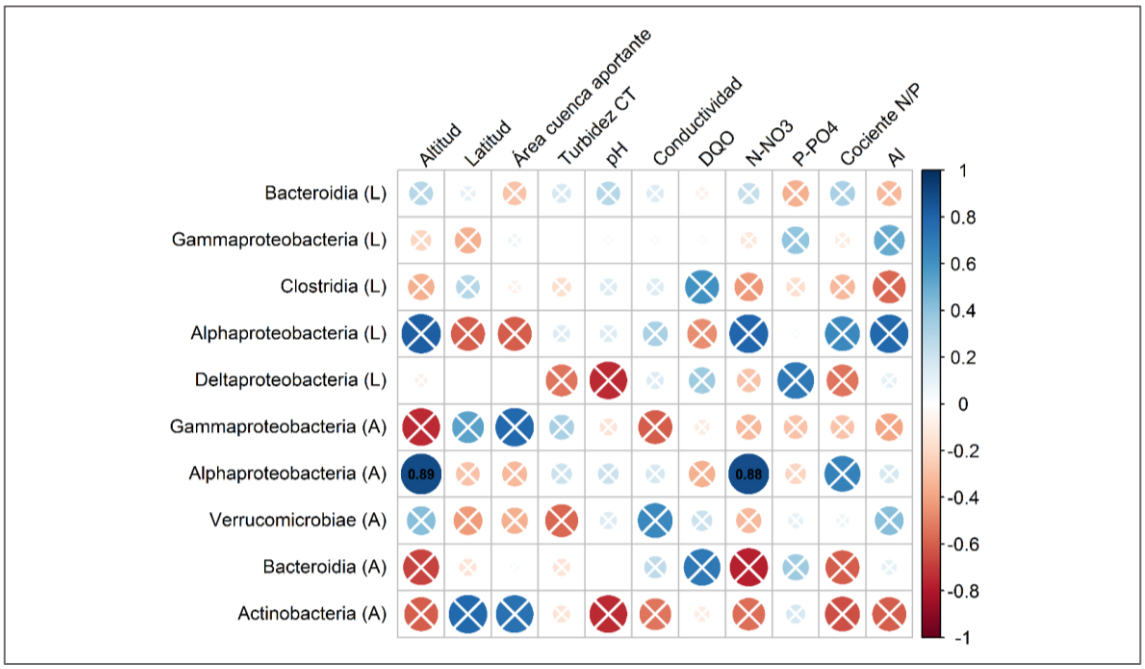

Figura 4.17 Correlaciones entre variables ambientales y Clases más abundantes.

En el caso del subconjunto de agua se detectaron correlaciones fuertes y significativas entre la abundancia de *Alphaproteobacteria* con la altitud y con la concentración de nitrógeno de nitratos. Al revisar las gráficas de dispersión de estas dos combinaciones (ver **Figura A 4.8.6**) se encontró que la gráfica de la abundancia de *Alphaproteobacteria* v/s concentración de nitrógeno de nitratos mostró un mejor ajuste lineal que aquella asociada a la altitud ($R^2_{altitud} = 0.59 < R^2_{N\text{-}NO3} = 0.69$) y, además, exhibió una menor distorsión producto de los valores extremos. En la **Figura 4.17** se observan también otras correlaciones fuertes (pero no significativas), al revisar la **Tabla A 4.8.6** se encontró que si se aumenta el límite para el valor *p* hasta un valor de 0.05, entonces la correlación entre *Gammaproteobacteria* y el área de la cuenca aportante, la de *Bacteroidia* y la concentración de nitrógeno de nitratos y la de *Actinobacteria* con la latitud, pasan a ser significativas. Se optó por generar gráficas de dispersión para estas tres correlaciones (ver **Figura A 4.8.7**), de acuerdo con estas solo para la correlación entre la *Gammaproteobacteria* y el área de la cuenca aportante se observa una clara tendencia en los puntos, con la única excepción de P5A, el cual claramente es un valor anómalo.

4.9 Discusión de resultados

En esta investigación se estudiaron conjuntamente comunidades bacterianas planctónicas (agua) y sedimentarias (lecho) presentes en ríos de la zona centro sur de Chile, región que es conocida a nivel mundial por su gran biodiversidad y albergar especies con un alto grado de endemismo (Figueroa *et al.*, 2013; Fierro *et al.*, 2017). Aún existen grandes vacíos de conocimiento en torno a esta biodiversidad en la zona, siendo esta carencia aún más profunda para la diversidad microbiana (Habit *et al.*, 2019), por tanto, la presente investigación corresponde a uno de los primeros acercamientos al estudio de las comunidades bacterianas de ríos dentro de esta región geográfica y los factores que las modulan, lo cual resulta una etapa necesaria para avanzar en el establecimiento de una línea base de información que permita el uso futuro de las comunidades bacterianas como agentes bioindicadores de la calidad del agua en sistemas acuáticos naturales.

4.9.1 Variabilidad en las condiciones ambientales

Chile presenta gran diversidad de condiciones climáticas y fisiográficas (Andreoli *et al.*, 2012), lo que conlleva a que sus ríos muestren una alta diversidad en hidrología, morfología y también en sus características fisicoquímicas. Parte de esta variabilidad fue capturada durante este estudio, en el cual se observaron grandes diferencias entre los cauces ubicados en los extremos latitudinales de la zona de estudio. El caudal de los ríos y el caudal normalizado por unidad de área de la cuenca decrecieron hacia el norte, patrón que es coherentemente explicado por los gradientes de precipitaciones y temperaturas que se desarrollan en la macrozona centro de Chile (región que contiene a la zona de estudio). Mientras las precipitaciones se reducen hacia el norte, las temperaturas se incrementan (Castro y Gironás, 2021), ocasionando una reducción del volumen de los recursos hídricos en dicho sentido. Además de presentar menores caudales, los ríos en el norte de la zona de estudio se caracterizaron por mostrar mayores conductividades, valores de pH levemente básicos, mayores concentraciones de iones ($SO_4{}^{-2}, Cl^-$), nitratos ($NO_3{}^-$) y de

metales (Mn, Fe, Al). Dos factores que podrían explicar en parte lo anterior son: por un lado, en el norte se encontraron grandes proporciones de suelo sin vegetación, estos podrían ser más propensos a la erosión del agua, generándose un mayor ingreso de metales, sales y nutrientes vía escorrentía superficial. Por otro lado, en terreno se detectaron extracciones de áridos en las planicies de inundación, dependiendo de la intensidad de esta actividad podría estar generándose la re-suspensión de sustancias desde el lecho hacia la columna de agua. Otra explicación se relaciona con las características geológicas de las cuencas. Peña-Guerrero *et al.* (2020) atribuyeron las altas conductividades y concentraciones de iones observadas en ríos de la cuenca del Maipo a la interacción de estos con las evaporitas, andesitas y granodioritas que abundan en la zona centro de Chile. Por consiguiente, dada la cercanía espacial entre dicha cuenca y las ubicadas en el norte de la zona de estudio, es posible que estas compartan características geológicas y, por ende, sus ríos se asemejen en características fisicoquímicas.

Tanto el alcance geográfico como muestral de un estudio sobre biodiversidad microbiológica dependerán de los objetivos propuestos, no obstante, en vista de la variabilidad observada en caudales y características fisicoquímicas del agua, así como también en los usos de suelo, es recomendable que en futuros estudios se opte por definir escalas espaciales más acotadas, o bien, considerar un mayor número de puntos de muestreo, esto con el fin de capturar adecuadamente la variabilidad. También se sugiere utilizar algún indicador del nivel de intervención antropogénica, puesto que esto permite definir sitios de referencia. Por ejemplo, se propone utilizar un índice como el IDI (del inglés, *Integrated Disturbance Index*) empleado por Fierro *et al.*, 2018, el cual considera simultáneamente alteraciones locales y a nivel de cuenca. Esta estrategia posibilita identificar factores externos que influyan de manera dinámica en las comunidades microbiológicas que habitan la columna de agua y en el lecho de los ríos.

4.9.2 Diferencias entre comunidades bacterianas de agua y de lecho

Aunque la variabilidad geográfica e hidrológica resultan ser componentes centrales para definir estrategias de análisis y evaluación de las características de las comunidades bacterianas, esta no es la única. El origen de las muestras (agua o lecho) determina el hábitat natural de las comunidades, este hábitat es un factor clave y de acuerdo con los resultados de este estudio, en conjunto con lo encontrado en la literatura, debe ser considerado al estudiar la biogeografía de las comunidades bacterianas.

Al centrarse en los Filos más abundantes, los hábitats agua y lecho mostraron gran similitud con respecto a los taxones detectados: *Proteobacteria* fue el Filo dominante entre todas las muestras seguido en abundancia por *Bacteroidetes*, *Verrucomicrobia*, *Firmicutes* y *Actinobacteria*. La presencia de estos Filos, todos pertenecientes al grupo TFB (del inglés, *Typical Freshwater Bacteria*) (Newton *et al.*, 2011), y la dominancia de *Proteobacteria*, tanto en agua como en lecho, son aspectos que han sido observados anteriormente en ambientes de agua dulce (Gibbons *et al.*, 2014; Read *et al.*, 2015; Staley *et al.*, 2015; Wang *et al.*, 2018; Adhikari *et al.*, 2019; Wang *et al.*, 2019; Ouyang *et al.*, 2020). Sin embargo, al revisar las abundancias relativas, se detectó que varios de estos taxones tuvieron diferencias significativas entre agua y lecho, motivo que explica por qué en las gráficas de ordenación las muestras se agruparon según origen (agua o lecho). Así como en otros estudios que han comparado comunidades de agua y de lecho (Ouyang *et al.*2020; Zárate *et al.*, 2020), *Proteobacteria* fue más abundante en agua (61.4 %) que en lecho (37.9 %), aunque esta diferencia no fue estadísticamente significativa. *Firmicutes* y *Actinobacteria*, por otro lado, si tuvieron diferencias significativas, el primero mostró gran abundancia en lecho y casi nula presencia en agua, mientras que el segundo fue más abundante en agua. Las disimilitudes entre comunidades se acentuaron aún más al descender al nivel Clase, la separación entre muestras de agua y de lecho se tornó más evidente en las gráficas de ordenación y se incrementó el número de taxones cuyas abundancias resultaron ser significativamente diferentes entre agua y lecho. Bajo el nivel

de Clase se observó que *Bacilli*, *Campylobacteria* y *Negativicutes* exhibieron grandes abundancias en lecho, pero casi nula presencia en agua, en tanto que las Clases *Verrucomicrobia*, *Deltaproteobacteria* y *Actinobacteria* fueron significativamente más abundantes en agua. Dado el carácter exploratorio de este estudio solo se trabajó con niveles taxonómicos superiores, no obstante, se sugiere que en futuras investigaciones se opte por niveles más finos, dado que incluso para Familia y Género es posible observar diferencias en cuanto los individuos dentro de un mismo taxón.

En lo que respecta a la diversidad alfa, las comunidades de agua y de lecho mostraron patrones diferentes dentro de la zona de estudio: en el río Biobío se encontró mayor diversidad en lecho, en los ríos Itata y Ñuble se observó mayor diversidad en agua y en los cauces restantes los niveles de diversidad entre ambas comunidades fueron prácticamente iguales ente sí. En los pocos estudios que han considerado conjuntamente comunidades de agua y de lecho se ha observado un comportamiento como el del río Biobío, esto es, mayor diversidad en lecho que en agua (Staley *et al.*, 2015; Mao *et al.*, 2019; Ouyang *et al.*, 2020) y se ha atribuido esto a una mayor complejidad en dichas comunidades. Esta explicación resulta consistente con el hecho de que las comunidades en lecho forman biopelículas (Flemming *et al.*, 2016), lo que favorece la formación de gradientes de concentraciones, pero también con que estas comunidades disponen de más recursos que su contraparte planctónica, ya que el lecho de los ríos tiende a acumular nutrientes y materia orgánica, las cuales propician el desarrollo de las bacterias. Un factor importante que puede influir en la diversidad son los tiempos de respuesta de las comunidades a cambios en sus entornos: mientras las que habitan suspendidas en el agua responden a cambios en el corto plazo, las del lecho lo hacen frente a alteraciones a largo plazo (Ouyang *et al.*, 2020). Por este motivo, no es posible descartar que lo observado durante la campaña, particularmente en los ríos Itata y Ñuble (mayor diversidad en el agua), se deba a una observación puntual y no a una tendencia.

En vista de lo encontrado sería pertinente que, además de considerar simultáneamente comunidades bacterianas de agua y de lecho, se opte por estudiar ríos en forma independiente, reconociendo los procesos propios de cada uno y cómo estos influyen en el desarrollo de sus respectivas comunidades bacterianas. Dentro de los procesos característicos de cada río, se encuentran relevantes variables hidrológicas, geológicas y el nivel de intervención antrópica.

4.9.3 Núcleo de las comunidades bacterianas

Gibbons *et al.* (2013) plantearon que en comunidades bacterianas marinas existiría un *núcleo* compuesto por bacterias cuyas proporciones podrían variar en el tiempo y espacio, pero siempre estarían presentes. En línea con lo anterior, estudios han sugerido que podría existir un núcleo compartido por comunidades bacterianas que habitan dentro de un mismo cauce (Gibbons *et al.*, 2014; de Oliveira *et al.* 2015; Staley *et al.* 2015; Wang *et al.*, 2016; Wang *et al.*, 2019), para analizarlo estos han determinado el conjunto de bacterias comunes entre dichas comunidades. Aquellas investigaciones que se han centrado en comunidades planctónicas han reportado que entre 67 y 95 % de las lecturas totales de una muestra han estado asociadas al núcleo (Staley *et al.*, 2015; Wang *et al.*, 2016; Wang *et al.*, 2019), en tanto que para comunidades en lecho un estudio encontró que cerca del 50 % de las lecturas en una muestra formaban parte del núcleo (Gibbons *et al.*, 2014). En la presente investigación el núcleo (conjunto común) encontrado en agua también mostró una mayor proporción de lecturas asociadas que aquel identificado en lecho (60 % v/s 25 %), siendo esta diferencia estadísticamente significativa. Ambos porcentajes son inferiores a los reportados anteriormente para comunidades bacterianas en ríos, no obstante, debe considerarse que estos se obtuvieron con muestras recolectadas en cauces diferentes, mientras que los informados en estudios previos se determinaron con muestras procedentes de un mismo río, por tanto, conectadas entre sí mediante dispersión (Ruiz-González *et al.*, 2015).

Destaca notablemente la similitud observada entre las comunidades planctónicas en esta investigación, esto considerando que se trabajó con múltiples cuencas, algunas de ellas distantes entre sí. Una fracción importante de las bacterias presentes en estas comunidades pueden ser arrastradas desde los suelos (Ruiz-González *et al.*, 2015), por lo tanto, la similitud observada podría indicar que los microbiomas bacterianos de las distintas cuencas se asemejan entre sí. En el estudio de Niño-García *et al.* (2016), centrado en las comunidades planctónicas de Quebec (Canadá), es posible advertir que, pese a provenir de distintas regiones geográficas, las comunidades que habitaban ambientes con Tiempos de Retención Hidráulica (TRH) altos (ríos mayores y lagos) se asemejaron en sus composiciones, más aún, la similitud fue mayor entre aquellas comunidades cuyos ambientes de origen tenían valores de pH cercanos. En base a lo anterior, la similitud entre las comunidades planctónicas en la zona de estudio no se explicaría solo por una posible semejanza entre los microbiomas de las distintas cuencas, también debe considerarse el efecto del transporte dentro de la red fluvial, el cual conlleva a un incremento del TRH y de la influencia que tienen las condiciones fisicoquímicas en las comunidades (Niño-García *et al.*, 2016). De esta forma, las bacterias comienzan a ser seleccionadas según su tolerancia a dichas condiciones y de acuerdo con qué tan competitivas son en ambientes con mayores TRH (Read *et al.*, 2015). Aun cuando se observó variabilidad en las características fisicoquímicas de los ríos dentro de la zona de estudio, no se encontraron diferencias extremas entre los distintos puntos de muestreo, esto podría explicar por qué, pese a provenir de distintas cuencas, las comunidades planctónicas mostraron una importante proporción de bacterias en común. Las comunidades de lecho, por el contrario, fueron más disímiles entre ellas, el núcleo de estas se conformó por un menor número de bacterias y tuvo un bajo porcentaje de lecturas asociadas. Esta situación podría relacionarse con el carácter sésil de estas comunidades, que se traduce en una mayor conexión de éstas con su entorno, lo que conlleva a que las condiciones locales influyan fuertemente en cómo se estructuran y evolucionan.

Identificar núcleos entre comunidades desarrolladas en ambientes con condiciones similares permitiría reconocer qué bacterias podrían estar vinculadas con actividades humanas específicas, cuáles serían típicas de ambientes no perturbados y si existen individuos que serían propios de ciertas regiones geográficas (Washington *et al.*, 2013), todo lo anterior resulta primordial para el desarrollo y potencial implementación de las comunidades bacterianas como bioindicadores. Asimismo, los núcleos y por consiguiente las comunidades que no forman parte de ellos, conforman grupos que deben ser reconocidos ya que la evolución de su estructura podría indicar cambios importantes generados debido a procesos antrópicos o ambientales de origen natural.

4.9.4 Influencia de factores ambientales

Las comunidades bacterianas han mostrado ser altamente sensibles con su entorno y ser capaces de responder relativamente rápido frente a alteraciones en éste, por esto, se ha planteado que éstas podrían ser potenciales bioindicadores (Niu *et al.*, 2018; Astudillo-García *et al.*, 2019), especialmente en zonas cuyos niveles de contaminación han provocado la desaparición de otros indicadores de tipo biológico (Washington *et al.*, 2013; Li *et al.*, 2017, Niu *et al.*, 2018). No obstante, una cualidad central que deben tener los indicadores bióticos es que los cambios en sus organismos y/o comunidades puedan vincularse inequívocamente con alteraciones específicas, siendo deseable además que estos cambios sean cuantificables (Astudillo-García *et al.*, 2019). Pese a que los estudios en torno a las comunidades bacterianas en ríos se han incrementado en los últimos años (Li *et al.*, 2021), éstos son aún insuficientes para alcanzar un nivel de comprensión que permita la vinculación entre cambios en las comunidades con alteraciones en sus entornos. Entre los motivos que dificultan esta tarea están la redundancia funcional que exhiben estas comunidades (Peter *et al.*, 2011), las interacciones sinérgicas o antagónicas que pueden existir entre factores de estrés (Osorio *et al.*, 2014; Corcoll *et al.*, 2015; Romero, Sabater, Timoner y Acuña, 2018) y la influencia del TRH de los cuerpos de agua, el cual no siempre es considerado al estudiar estas comunidades (Niño-García *et al.*, 2016). En la

práctica, un río puede entenderse como un sistema con un flujo de pistón con entradas y salidas puntuales y dispersas de agua y sustancias, el que además está asociado con superficies en las cuales se desarrollan procesos de reacción y transporte. Debido a esto, las concentraciones de sustancias y el desarrollo de organismos a lo largo de su extensión (cauce) están estrechamente relacionadas con la configuración hidráulica del sistema.

En este estudio la diversidad alfa (Shannon – Wiener) de las comunidades en agua se correlacionó negativamente con el área de la cuenca aportante. Anteriormente, estudios que han abarcado extensiones espaciales amplias han detectado una reducción en la diversidad alfa de las comunidades planctónicas con el avance dentro de la red fluvial (Savio *et al.*, 2015; Niño-García *et al.*, 2016; Ruiz-González *et al.*, 2017). Esto se debería a que inicialmente, en las cabeceras, la mayor conexión que existe entre la red fluvial y las áreas terrestres permite el ingreso constante de una gran cantidad de bacterias procedentes de los suelos, siendo estas capaces de mantener sus números, sin embargo, al avanzar dentro de la red la influencia de lo terrestre se reduce, el TRH aumenta y las características fisicoquímicas del agua se tornan relevantes, con lo cual solo las bacterias más competitivas y tolerantes en estas nuevas condiciones logran prosperar, disminuyendo así la diversidad (Savio *et al.*, 2015; Niño-García *et al.*, 2016; Ruiz-González *et al.*, 2017). Puesto que los puntos de muestreo de los ríos localizados más al sur estuvieron más cercanos a la Cordillera de la Costa y los del norte más próximos a la Cordillera de los Andes, y que a su vez la superficie de la cuenca aportante decreció en el sentido sur – norte, el área de la cuenca aportante estaría relacionada indirectamente con la posición dentro de la red fluvial, lo que explicaría el patrón observado en la diversidad alfa de las comunidades planctónicas en este estudio.

Se encontraron correlaciones significativas entre la diversidad beta de las comunidades y condiciones ambientales, pero solo en el nivel ASV. De acuerdo con la prueba Mantel la disimilitud entre muestras de agua se correlacionó con el pH y la conductividad de los ríos. Con respecto a esto, Niño-García *et al.* (2016) encontraron que el pH fue el principal

factor ambiental detrás de las disimilitudes observadas entre las comunidades planctónicas de Quebec, en tanto que Zárate *et al.* (2020) identificaron que la salinidad, en conjunto con el ORP y la concentración de clorofila, explicaban las disimilitudes entre las comunidades bacterianas que estudiaron. La disimilitud entre las muestras de agua también se correlacionó con variables espaciales, las que a su vez se correlacionan con el pH y conductividad. Al seleccionar la variable ambiental de interés que explica la disimilitud es necesario conocer la relación existente entre variables asociadas a parámetros fisicoquímicos y variables que indican características espaciales o geográficas. Del mismo modo, se espera que la variabilidad fisicoquímica de un río sea explicada en parte por factores geográficos los que a su vez determinan la geología e hidrología del sistema acuático. Se espera que la ejecución de más estudios que conecten variables ambientales y estructuras de comunidades microbiológicas con alcances apropiados temporales y espaciales, logre develar la compleja evolución de estos organismos para su potencial seguimiento con fines de monitoreo de la calidad del agua de los ríos.

CAPÍTULO 5 CONCLUSIONES

En las últimas décadas la biodiversidad ha experimentado un fuerte declive a nivel mundial producto de la acción humana, con los ecosistemas de agua dulce, vitales para la supervivencia y desarrollo de la humanidad, entre los más afectados. Se teme que de continuar esta tendencia estos ambientes sean incapaces de mantener sus funciones ecológicas, siendo afectados a su vez los servicios ecosistémicos que estos proveen. En este escenario, resulta crucial contar con herramientas que permitan detectar tempranamente alteraciones perjudiciales para la integridad ecológica de estos ecosistemas, antes de que desemboquen en situaciones más complejas de remediar o, incluso, irreversibles. Actualmente, además de mediciones de parámetros fisicoquímicos, las campañas de biomonitoreo utilizan elementos bióticos (bioindicadores) para evaluar el estado ecológico de los sistemas naturales, esto por medio de índices calculados en función de las poblaciones de determinados organismos, como peces, macroinvertebrados y algas bentónicas. Sin embargo, la recolección e identificación de estos especímenes es muy costosa, en tiempo y recursos, y requiere de expertos para concretarse, además, estos índices no pueden ser usados en ambientes cuyo nivel de deterioro ha provocado la desaparición de los organismos bioindicadores. Las bacterias son un componente central de los ecosistemas acuáticos, estas participan activamente en los ciclos biogeoquímicos, son responsables de la degradación de la materia orgánica y recirculación de nutrientes, forman parte de la base de la cadena trófica y ejercen fuerte influencia en la calidad del agua, adicionalmente, estos microorganismos muestran una alta sensibilidad y rápida respuesta frente a las alteraciones (físicas, químicas y biológicas) que sufre su entorno. Por todo lo anterior, sumado al carácter ubicuo de estos microorganismos, se ha sugerido que las bacterias podrían ser integradas como bioindicadores en el monitoreo ecológico, mejorando la capacidad de detectar amenazas anticipadamente y expandiendo el nivel de comprensión de los ecosistemas. Con el fuerte desarrollo que se ha dado en las tecnologías de secuenciación en los últimos años y la continua disminución de sus costos (operativos y de inversión), esta idea se ha tornado factible, no obstante, aún existe gran

desconocimiento en torno a la biogeografía de las bacterias en sistemas de agua dulce y con respecto a cómo responden estas comunidades frente a distintos factores de estrés.

Aquí se estudiaron las comunidades bacterianas planctónicas y sedimentarias presentes en siete ríos de la zona centro sur de Chile, región con escasa información en torno a la diversidad microbiana de sus ríos, con el fin de sentar bases para futuras investigaciones de esta naturaleza en el país. Este trabajo entrega herramientas para el procesamiento de los datos genéticos (Código R en **Anexo 5.1**) y el análisis de la estructura de las comunidades bacterianas (Código R en **Anexo 5.2**), además, aporta observaciones que debiesen considerarse en el diseño de futuras investigaciones. Para el estudio de estas comunidades se usó secuenciación de alto desempeño (Illumina MiSeq) y se trabajó en el entorno de R, un lenguaje de programación libre, tanto para el procesamiento y análisis de los datos genéticos generados, así como también para la estadística posterior. En tanto, la caracterización de los puntos de muestreo se realizó con datos recolectados en la campaña y también datos históricos de caudales, de calidad del agua y de usos de suelo.

Las comunidades bacterianas de agua y de lecho mostraron composiciones taxonómicas, a nivel de Filo y de Clase, típicas de ambientes de agua dulce, pese a ello, comunidades plantónicas y sedimentarias sí exhibieron composiciones diferentes, reconociéndose múltiples taxones con abundancias significativamente diferentes entre agua y lecho. En cuanto a diversidad, globalmente no se encontraron diferencias significativas entre agua y lecho, pero localmente se detectaron patrones distintos entre los ríos estudiados: en algunos se halló mayor diversidad en lecho, en otros se observó mayor en agua y en los restantes, ambos hábitats tuvieron niveles equiparables. Pese a la importante extensión espacial considerada y la variabilidad detectada en las condiciones ambientales, las comunidades planctónicas mostraron gran similitud entre sí: En estas una fracción importante de las lecturas detectadas en una muestra formaban parte de un núcleo común (sobre 60 %). Por el contrario, las comunidades de lecho tuvieron mayores diferencias entre ellas y bajos porcentajes de taxones en común, lo que podría indicar una mayor

influencia de los factores locales en la composición de estas comunidades. El grado de similitud en comunidades de agua y de lecho es una variable que podría explicarse en base a diversos factores ambientales de los sitios de estudio, como el nivel de intervención antrópica. Por esto, la investigación integrada de los parámetros y relaciones que puedan develar esta interacción se considera como un área de sumo interés para poder predecir el comportamiento de las comunidades frente a factores externos de diverso origen.

Se encontró gran variabilidad en la zona de estudio con respecto a caudales, usos de suelo y algunos parámetros fisicoquímicos (pH, conductividad, Cl^-, SO_4^{-2}, Al, Fe, Mn), con marcadas diferencias entre los puntos (ríos) del norte y los del sur. Caudales y usos de suelo fueron los aspectos en los que mayores diferencias se observaron, mientras que las características fisicoquímicas, aun cuando mostraron variabilidad entre los puntos de muestreo, no tuvieron diferencias significativas entre estos. En este estudio no se consideró un indicador del nivel de intervención humana en los puntos de muestreo, esto debiese ser incorporado a futuro para reconocer ecosistemas sanos que puedan usarse como referencia. En las comunidades de agua se encontró que la diversidad de estas se correlacionó significativamente con el área de la cuenca aportante, considerando que la superficie de la cuenca estuvo indirectamente relacionada con la posición de los puntos en sus respectivas redes fluviales, lo observado concuerda con lo hallado en otros estudios: esto es, que la diversidad de las comunidades tiende a disminuir al avanzar aguas abajo en la red fluvial. También se encontró que las disimilitudes entre comunidades planctónicas, medidas a nivel de ASV, se correlacionaron fuertemente con variables espaciales (latitud, altitud) y de calidad (pH, conductividad), no obstante, puesto que estas variables se correlacionaron a su vez entre sí, no es posible determinar con certeza estadística si las disimilitudes entre comunidades se explican principalmente por diferencias en las condiciones ambientales, o bien, son consecuencia de diferencias en la distribución espacial de estos microorganismos dentro de la zona de estudio.

La biogeografía de las comunidades bacterianas es un tema complejo de estudiar y del cual aún existen vacíos importantes de conocimiento. Aun cuando el uso de estas comunidades como bioindicadores no exige una comprensión absoluta con respecto de estas, al menos debe poder establecerse una línea base de referencia y determinar relaciones inequívocas entre alteraciones en los ecosistemas y cambios en las comunidades. Con el desarrollo de la presente investigación exploratoria se logró establecer una base bibliográfica para sustentar futuros estudios, se encontró un entorno de trabajo adecuado para el procesamiento y análisis de datos genéticos de comunidades bacterianas, y se obtuvieron resultados y conclusiones que sería pertinente tener presentes en nuevos estudios, entre las más relevantes están:

- Dadas las diferencias observadas entre comunidades planctónicas y sedimentarias, y considerando lo indicado en la literatura, se recomienda que las comunidades de agua y de lecho sean estudiadas simultáneamente, y separar las comunidades planctónicas según dos grupos de bacterias: las que flotan libremente y aquellas que lo hacen adheridas a partículas.
- Es claro que, para la extensión geográfica considerada en este estudio, el número de puntos de muestreo fue reducido. Futuros estudios debiesen abarcar extensiones espaciales más acotadas y un mayor número de puntos, con tal de capturar mejor la variabilidad en condiciones ambientales y de las comunidades, pero también con motivo de obtener resultados estadísticamente significativos.
- Como ya se ha señalado anteriormente, debe incorporarse un indicador que dé cuenta del nivel de impacto antropogénico con el fin de reconocer ecosistemas que puedan usarse como referencia (ecosistemas sanos).
- Al definir los puntos de muestreo debiesen preferirse zonas que en sus cercanías dispongan de estaciones con registros históricos de caudales y de calidad relativamente completos y extensos, para poder comparar con lo medido durante terreno y, además, caracterizar mejor los puntos de muestreo.

- Dada la gran complejidad que presentan las comunidades bacterianas sería adecuado estudiarlas considerándolas como un todo, por ejemplo, utilizando redes de co-ocurrencia, método que permite encontrar relaciones positivas o negativas entre los integrantes de estas comunidades, así como también qué bacterias serían los pivotes de la comunidad.
- Un aspecto que no fue posible tratar en este estudio, dadas sus limitaciones, y que debiese abordarse en futuras investigaciones es el análisis del perfil funcional de las comunidades, el cual podría entregar información complementaria a la composición taxonómica de estas.

REFERENCIAS

Adhikari, N. P., Liu, Y., Liu, K., Zhang, F., Adhikari, S., Chen, Y., y Liu, X. (2019). Bacterial community composition and diversity in Koshi River, the largest river of Nepal. *Ecological indicators*, *104*, 501-511. https://doi.org/10.1016/j.ecolind.2019.05.009

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 139-160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

Alonso, Á., Figueroa, R., y Castro-Díez, P. (2017). Pollution assessment of the Biobío River (Chile): Prioritization of substances of concern under an ecotoxicological approach. *Environmental management*, *59*(5), 856-869. https://doi.org/10.1007/s00267-017-0824-5

Andreoli, A., Mao, L., Iroume, A., Arumi, J. L., Nardini, A., Pizarro, R., … y Link, O. (2012). The need for a hydromorphological approach to Chilean river management. *Revista Chilena de Historia Natural*, *85*(3), 339-343. https://doi.org/10.4067/S0716-078X2012000300008

Aphalo, P. J. (2022). Ggpmisc: Miscellaneous Extensions ggplot2. R package versión 0.4.5. https://CRAN.R-project.org/package=ggpmisc

Arriagada, L., Rojas, O., Arumí, J. L., Munizaga, J., Rojas, C., Farias, L., y Vega, C. (2019). A new method to evaluate the vulnerability of watersheds facing several stressors: A case study mediterraneanean Chile. *Science of the Total Environment, 651*, 1517-1533. https://doi.org/10.1016/j.scitotenv.2018.09.237

Arriagada, P., Karelovic, B., y Link, O. (2021). Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine learning algorithm. *Journal of Hydrology*, *598*, 126454. https://doi.org/10.1016/j.jhydrol.2021.126454

Astudillo-García, C., Hermans, S. M., Stevenson, B., Buckley, H. L., y Lear, G. (2019). Microbial assemblages and bioindicators as proxies for ecosystem health status: potential and limitations. *Applied microbiology and biotechnology*, *103*(16), 6407-6421. https://doi.org/10.1007/s00253-019-09963-0

Aufdenkampe, A. K., Mayorga, E., Raymond, P. A., Melack, J. M., Doney, S. C., Alin, S. R., … y Yoo, K. (2011). Riverine coupling of biogeochemical cycles between land, oceans, and atmosphere. *Frontiers in Ecology and the Environment*, *9*(1), 53-60. https://doi.org/10.1890/100014

Balian, E. V., Segers, H., Martens, K., y Lévéque, C. (2007). The freshwater animal diversity assessment: an overview of the results. *Freshwater animal diversity assessment*, 627-637. https://doi.org/10.1007/978-1-4020-8259-7_61

Barbosa, O., Colson, D., Durán, A. P., Godoy, K., Jones, A., Jones, G., … y Trippier, B. (2019). JNCC Report No: 634.

Battin T. J., Besemer, K., Bengtsson, M. M., Romani, A. M., y Packmann, A. I. (2016) The ecology and biogeochemistry of stream biofilms. *Nature Reviews Microbiology*, *14*(4), 251-263. https://doi.org/10.1038/nrmicro.2016.15

Beiko, R. G., Hsiao, W., y Parkinson, J. (Eds.). (2018). *Microbiome analysis: methods and protocols*. Humana Press, New York. https://doi.org/10.1007/978-1-4939-8728-3

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**, 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Borcard, D., Gillet, F., y Legendre, P. (2011). *Numerical ecology with R* (Vol. 2, p. 688). Springer, New York. https://doi.org/10.1007/978-1-4419-7976-6

Borruso, L., Zerbe, S., y Brusetti, L. (2015). Bacterial community structures as a diagnostic tool for watershed quality assessment. *Research in microbiology*, *166*(1), 38-44. https://doi.org/10.1016/j.resmic.2014.11.004

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., y Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature methods*, *13*(7), 581-583. https://doi.org/10.1038/nmeth.3869

Callahan, Benjamin. (2018). Silva taxonomic training data formatted for DADA2 (Silva version 132) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.1172783

Caracciolo, A. B., Topp, E., y Grenni, P. (2015). Pharmaceuticals in the environment: biodegradation and effects on natural microbial communities. A review. *Journal of pharmaceutical and biomedical analysis*, *106*, 25-36. https://doi.org/10.1016/j.jpba.2014.11.040

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., ... y Naeem, S. (2012). Biodiversity loss and its impact on humanity. *Nature*, *486*(7401), 59-67. https://doi.org/10.1038/nature11148

Castro, L., y Gironás, J. (2021). Precipitation, Temperature and Evaporation. En B. Fernández y J. Gironás (Eds.), *Water Resources of Chile* (Vol 8, pp. 31-60). Springer, Cham. https://doi.org/10.1007/978-3-030-56901-3

Centro Nacional del Medio Ambiente (CENMA) (2017). Monitoreo para la vigilancia ambiental de la cuenca del río Biobío. Unidad de Biodiversidad, Centro Nacional del Medio Ambiente, Fundación de la Universidad de Chile.

Chen, H. (2022). VennDiagram: Generate High-Resolution Venn and Euler Plots. R package version 1.7.1. https://CRAN.R-project.org/package=VennDiagram

Clark, D. R., Ferguson, R. M., Harris, D. N., Matthews Nicholass, K. J., Prentice, H. J., Randall, K. C., ... y Dumbrell, A. J. (2018). Streams of data from drops of water: 21st century molecular microbial ecology. *Wiley Interdisciplinary Reviews: Water*, *5*(4), e1280. https://doi.org/10.1002/wat2.1280

Corcoll, N., Casellas, M., Huerta, B., Guasch, H., Acuña, V., Rodríguez-Mozaz, S., … y Sabater, S. (2015). Effects of flow intermittency and pharmaceutical exposure on the structure and metabolism of stream biofilms. *Science of the total environment*, *503*, 159-170. https://doi.org/10.1016/j.scitotenv.2014.06.093

Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., ... y Lanzén, A. (2020). Ecosystems monitoring powered by environmental genomics: a review of current strategies with an implementation roadmap. *Molecular ecology*. https://doi.org/10.1111/mec.15472

de Oliveira, L. F. V., y Margis, R. (2015). The source of the river as a nursery for microbial diversity. *PLoS One*, *10*(3), e0120608. https://doi.org/10.1371/journal.pone.0120608

de Vries, A., y Ripley, B. D. (2022). ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'. R package version 0.1.23. https://CRAN.R-project.org/package=ggdendro

Díaz, G., Górski, K., Heino, J., Arriagada, P., Link, O., y Habit, E. (2021). The longest fragment drives fish beta diversity in fragmented river networks: Implications for river management and conservation. *Science of The Total Environment*, *766*, 144323. https://doi.org/10.1016/j.scitotenv.2020.144323

Dirección General de Aguas (DGA) (2016). Atlas del Agua Chile 2016.

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z. I., Knowler, D. J., Lévêque, C., ... y Sullivan, C. A. (2006). Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological reviews*, *81*(2), 163-182. https://doi.org/10.1017/S1464793105006950

Feio, M. J., Hughes, R. M., Callisto, M., Nichols, S. J., Odume, O. N., Quintella, B. R., ... y Alonso-EguíaLis, P. (2021). The Biological Assessment and Rehabilitation of the World's Rivers: An Overview. *Water 2021*, *13*, 371. https://doi.org/10.3390/w13030371

Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G., y Gloor, G. B. (2013). ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. *PloS one*, *8*(7), e67019. https://doi.org/10.1371/journal.pone.0067019

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., y Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, *2*(1), 1-13. https://doi.org/10.1186/2049-2618-2-15

Fierro, P., Bertrán, C., Tapia, J., Hauenstein, E., Peña-Cortés, F., Vergara, C., ... y Vargas-Chacoff, L. (2017). Effects of local land-use on riparian vegetation, water quality, and the functional organization of macroinvertebrate assemblages. *Science of the Total Environment*, *609*, 724-734. https://doi.org/10.1016/j.scitotenv.2017.07.197

Fierro, P., Arismendi, I., Hughes, R. M., Valdovinos, C., y Jara-Flores, A. (2018). A benthic macroinvertebrate multimetric index for Chilean Mediterranean streams. *Ecological Indicators*, *91*, 13-23. https://doi.org/10.1016/j.ecolind.2018.03.074

Fierro, P., Valdovinos, C., Arismendi, I., Díaz, G., Jara-Flores, A., Habit, E., y Vargas-Chacoff, L. (2019). Examining the influence of human stressors on benthic algae, macroinvertebrate, and fish assemblages in Mediterranean streams of Chile. *Science of the total environment*, *686*, 26-37. https://doi.org/10.1016/j.scitotenv.2019.05.277

Figueroa, R., Bonada, N., Guevara, M., Pedreros, P., Correa-Araneda, F., Díaz, M. E., y Ruiz, V. H. (2013). Freshwater biodiversity and conservation mediterraneanean climate streams of Chile. *Hydrobiologia*, *719*(1), 269-289. https://doi.org/10.1007/s10750-013-1685-4

Flemming, H. C., Wingender, J., Szewzyk, U., Steinberg, P., Rice, S. A., y Kjelleberg, S. (2016). Biofilms: an emergent form of bacterial life. *Nature Reviews Microbiology*, *14*(9), 563. https://doi.org/10.1038/nrmicro.2016.94

Gibbons, S. M., Caporaso, J. G., Pirrung, M., Field, D., Knight, R., y Gilbert, J. A. (2013). Evidence for a persistent microbial seed bank throughout the global ocean. *Proceedings of the National Academy of Sciences*, *110*(12), 4651-4655. https://doi.org/10.1073/pnas.1217767110

Gibbons, S. M., Jones, E., Bearquiver, A., Blackwolf, F., Roundstone, W., Scott, N., ... y Gilbert, J. A. (2014). Human and environmental impacts on river sediment microbial communities. *PloS one*, *9*(5), e97435. https://doi.org/10.1371/journal.pone.0097435

Gloor, G. B., Macklaim, J. M., y Fernandes, A. D. (2016). Displaying variation in large datasets: plotting a visual summary of effect sizes. *Journal of Computational and Graphical Statistics*, *25*(3), 971-979. https://doi.org/10.1080/10618600.2015.1131161

Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V., y Egozcue, J. J. (2016). It's all relative: analyzing microbiome data as compositions. *Annals of epidemiology*, *26*(5), 322-329. https://doi.org/10.1016/j.annepidem.2016.03.003

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., y Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, *8*, 2224. https://doi.org/10.3389/fmicb.2017.02224

Habit, E., Górski, K., Alò, D., Ascencio, E., Astorga, A., N. Colin, … y Woelfl, S. (2019). Biodiversidad de ecosistemas de agua dulce. En P. A. Marquet *et al.* (Eds.), *Biodiversidad y cambio climático en Chile: Evidencia científica para la toma de decisiones*. Informe de la mesa de Biodiversidad. Santiago: Comité Científico COP25; Ministerio de Ciencia, Tecnología, Conocimiento e Innovación

Herman, M. R., y Nejadhashemi, A. P. (2015). A review of macroinvertebrate-and fish-based stream health indices. *Ecohydrology & Hydrobiology*, *15*(2), 53-67. https://doi.org/10.1016/j.ecohyd.2015.04.001

Hvitfeldt, E. (2021). paletteer: Comprehensive Collection of Color Palettes. version 1.3.0. https://github.com/EmilHvitfeldt/paletteer

Ibekwe, A. M., Ma, J., y Murinda, S. E. (2016). Bacterial community composition and structure in an Urban River impacted by different pollutant sources. *Science of the Total Environment*, *566*, 1176-1185. https://doi.org/10.1016/j.scitotenv.2016.05.168

Karrasch, B., Parra, O., Cid, H., Mehrens, M., Pacheco, P., Urrutia, R., ... y Zaror, C. (2006). Effects of pulp and paper mill effluents on the microplankton and microbial self-purification capabilities of the Biobio River, Chile. *Science of the Total Environment*, *359*(1-3), 194-208. https://doi.org/10.1016/j.scitotenv.2005.03.029

Laperriere, S. M., Hilderbrand, R. H., Keller, S. R., Trott, R., y Santoro, A. E. (2020). Headwater stream microbial diversity and function across agricultural and urban land use gradients. *Applied and Environmental Microbiology*, *86*(11), e00018-20. https://doi.org/10.1128/AEM.00018-20

Lau, K. E., Washington, V. J., Fan, V., Neale, M. W., Lear, G., Curran, J., y Lewis, G. D. (2015). A novel bacterial community index to assess stream ecological health. *Freshwater Biology*, *60*(10), 1988-2002. https://doi.org/10.1111/fwb.12625

Legendre, P., y Legendre, L. (1998). *Numerical ecology* (2 ed.). Elsevier Science, Amsterdam, Netherlands.

Li, J., Li, Y., Qian, B., Niu, L., Zhang, W., Cai, W., ... y Wang, C. (2017). Development and validation of bacteria-based index of biotic integrity for assessing the ecological status of urban rivers: a case study of Qinhuai River basin in Nanjing, China. *Journal of environmental management*, *196*, 161-167. https://doi.org/10.1016/j.jenvman.2017.03.003

Li, Y., Yang, N., Qian, B., Yang, Z., Liu, D., Niu, L., y Zhang, W. (2018). Development of a bacteria-based index of biotic integrity (Ba-IBI) for assessing ecological health of the Three Gorges Reservoir in different operation periods. *Science of the Total Environment*, *640*, 255-263. https://doi.org/10.1016/j.scitotenv.2018.05.291

Li, K., Hu, J., Li, T., Liu, F., Tao, J., Liu, J., ... y Che, R. (2021). Microbial abundance and diversity investigations along rivers: current knowledge and future directions. *Wiley Interdisciplinary Reviews: Water*, *8*(5), e1547. https://doi.org/10.1002/wat2.1547

Love, M. I., Huber, W., y Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 1-21. https://doi.org/10.1186/s13059-014-0550-8

Mao, Y., Liu, Y., Li, H., He, Q., Ai, H., Gu, W., y Yang, G. (2019). Distinct responses of planktonic and sedimentary bacterial communities to anthropogenic activities: case study of a tributary of the Three Gorges Reservoir, China. *Science of the Total Environment*, *682*, 324-332. https://doi.org/10.1016/j.scitotenv.2019.05.172

Masotti, I., Aparicio-Rizzo, P., Yevenes, M. A., Garreaud, R., Belmar, L., y Farías, L. (2018). The influence of river discharge on nutrient export and phytoplankton biomass off the central chile coast (33–37° S): seasonal cycle and interannual variability. *Frontiers in Marine Science*, *5*, 423. https://doi.org/10.3389/fmars.2018.00423

McMurdie, P. J. (2018). Normalization of microbiome profiling data. In *Microbiome Analysis* (pp. 143-168). Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-8728-3_10

McMurdie, P. J., y Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, *8*(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Mora-Gómez J., A. Freixa, N. Perujo y L. Barral-Fraga (2016) Limits of the biofilm concept and types of aquatic biofilms. En A. M. Romaní, H. Guasch y M. D. Balaguer (eds.). *Aquatic biofilms: Ecology, Water Quality and Wastewater Treatment*. Caister Academic Press. https://doi.org/10.21775/9781910190173.01

Neuwirth, E. (2022). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., y Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiology and molecular biology reviews*, *75*(1), 14-49. https://doi.org/10.1128/mmbr.00028-10

Niño-García, J. P., Ruiz-González, C., y Del Giorgio, P. A. (2016). Interactions between hydrology and water chemistry shape bacterioplankton biogeography across boreal freshwater networks. *The ISME journal*, *10*(7), 1755-1766. https://doi.org/10.1038/ismej.2015.226

Niu, L., Li, Y., Wang, P., Zhang, W., Wang, C., Li, J., y Wu, H. (2018). Development of a microbial community-based index of biotic integrity (MC-IBI) for the assessment of ecological status of rivers in the Taihu Basin, China. *Ecological Indicators*, *85*, 204-213. https://doi.org/10.1016/j.ecolind.2017.10.051

O'Brien, A., Townsend, K., Hale, R., Sharley, D., y Pettigrove, V. (2016). How is ecosystem health defined and measured? A critical review of freshwater and estuarine studies. *Ecological Indicators*, *69*, 722-729. https://doi.org/10.1016/j.ecolind.2016.05.004

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, … y Wagner, H. (2022). vegan: Community Ecology Package. R package version 2.5-7. https://CRAN.R-project.org/package=vegan

Osorio, V., Proia, L., Ricart, M., Pérez, S., Ginebreda, A., Cortina, L. J., … y Barceló, D. (2014). Hydrological variation modulates pharmaceutical levels and biofilm responses in a Mediterranean river. Science of the total environment, 472, 1052-1061. https://doi.org/10.1016/j.scitotenv.2013.11.069

Ouyang, L., Chen, H., Liu, X., Wong, M. H., Xu, F., Yang, X., ... y Li, S. (2020). Characteristics of spatial and seasonal bacterial community structures in a river under anthropogenic disturbances. *Environmental Pollution*, *264*, 114818. https://doi.org/10.1016/j.envpol.2020.114818

Palarea-Albaladejo, J., y Martín-Fernández, J. A. (2015). zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, *143*, 85-96. https://doi.org/10.1016/j.chemolab.2015.02.019

Pander, J., y Geist, J. (2013). Ecological indicators for stream restoration success. *Ecological indicators*, *30*, 106-118. https://doi.org/10.1016/j.ecolind.2013.01.039

Parada, A. E., Needham, D. M., y Fuhrman, J. A. (2016). Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental microbiology*, *18*(5), 1403-1414. https://doi.org/10.1111/1462-2920.13023

Pauchard, A., Aguayo, M., Peña, E., y Urrutia, R. (2006). Multiple effects of urbanization on the biodiversity of developing countries: the case of a fast-growing metropolitan area (Concepción, Chile). *Biological conservation*, *127*(3), 272-281. https://doi.org/10.1016/j.biocon.2005.05.015

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., ... y Kahlert, M. (2018). The future of biotic indices in the ecogenomic era: Integrating (e) DNA metabarcoding in biological assessment of aquatic ecosystems. *Science of the Total Environment*, *637*, 1295-1310. https://doi.org/10.1016/j.scitotenv.2018.05.002

Pedersen, T. L. (2021). ggforce: Accelerati‘g 'ggpl’t2'. R package version 0.3.3. https://CRAN.R-project.org/package=ggforce

Pedersen, T. L. (2020). patchwork: The Composer of Plots. https://github.com/thomasp85/patchwork.

Peña-Guerrero, M. D., Nauditt, A., Muñoz-Robles, C., Ribbe, L., y Meza, F. (2020). Drought impacts on water quality and potential implications for agricultural production in the Maipo River Basin, Central Chile. *Hydrological Sciences Journal*, *65*(6), 1005-1021. https://doi.org/10.1080/02626667.2020.1711911

Peter, H., Ylla, I., Gudasz, C., Romani, A. M., Sabater, S., y Tranvik, L. J. (2011). Multifunctionality and diversity in bacterial biofilms. *PloS one*, *6*(8), e23225. https://doi.org/10.1371/journal.pone.0023225

QGIS.org, 2022. QGIS Geographic Information System. QGIS Association. http://www.qgis.org

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... y Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*, *41*(D1), D590-D596. https://doi.org/10.1093/nar/gks1219

Quinn, T. P., Erb, I., Richardson, M. F., y Crowley, T. M. (2018). Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, *34*(16), 2870-2878. https://doi.org/10.1093/bioinformatics/bty175

Quinn, T. P., Erb, I., Gloor, G., Notredame, C., Richardson, M. F., y Crowley, T. M. (2019). A field guide for the compositional analysis of any-omics data. *GigaScience*, *8*(9), giz107. https://doi.org/10.1093/gigascience/giz107

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Read, D. S., Gweon, H. S., Bowes, M. J., Newbold, L. K., Field, D., Bailey, M. J., y Griffiths, R. I. (2015). Catchment-scale biogeography of riverine bacterioplankton. *The ISME journal*, *9*(2), 516-526. https://doi.org/10.1038/ismej.2014.166

Reid, A. J., Carlson, A. K., Creed, I. F., Eliason, E. J., Gell, P. A., Johnson, P. T., ... y Cooke, S. J. (2019). Emerging threats and persistent conservation challenges for freshwater biodiversity. *Biological Reviews*, *94*(3), 849-873. https://doi.org/10.1111/brv.12480

Roberto, A. A., Van Gray, J. B., y Leff, L. G. (2018). Sediment bacteria in an urban stream: spatiotemporal patterns in community composition. *Water research*, *134*, 353-369. https://doi.org/10.1016/j.watres.2018.01.045

Romaní, A. M., Amalfitano, S., Artigas, J., Fazi, S., Sabater, S., Timoner, X., Ylla, I. y Zoppini, A. (2013). Microbial biofilm structure and organic matter use Mediterraneanean streams. *Hydrobiologia*, *719*(1), 43-58. https://doi.org/10.1007/s10750-012-1302-y

Romero, F., Sabater, S., Timoner, X., y Acuña, V. (2018). Multistressor effects on river biofilms under global change conditions. Science of The Total Environment, 627, 1-10. https://doi.org/10.1016/j.scitotenv.2018.01.161

Ruiz-González, C., Niño-García, J. P., y Del Giorgio, P. A. (2015). Terrestrial origin of bacterial communities in complex boreal freshwater networks. *Ecology letters*, *18*(11), 1198-1206. https://doi.org/10.1111/ele.12499

Ruiz-González, C., Niño-García, J. P., Berggren, M., y Del Giorgio, P. A. (2017). Contrasting dynamics and environmental controls of dispersed bacteria along a hydrologic gradient. https://doi.org/10.4081/aiol.2017.7232

Sabater, S., Guasch, H., Romaní, A., y Muñoz, I. (2002). The effect of biological factors on the efficiency of river biofilms in improving water quality. *Hydrobiologia*, *469*(1-3), 149-156. https://doi.org/10.1023/A:1015549404082

Sagova-Mareckova, M., Boenigk, J., Bouchez, A., Cermakova, K., Chonova, T., Cordier, T., ... y Stoeck, T. (2020). Expanding ecological assessment by integrating microorganisms into routine freshwater biomonitoring. *Water Research*, 116767. https://doi.org/10.1016/j.watres.2020.116767

Savio, D., Sinclair, L., Ijaz, U. Z., Parajka, J., Reischer, G. H., Stadler, P., ... y Eiler, A. (2015). Bacterial diversity along a 2600 km river continuum. *Environmental microbiology*, *17*(12), 4994-5007. https://doi.org/10.1111/1462-2920.12886

Scheffers, B. R., De Meester, L., Bridge, T. C., Hoffmann, A. A., Pandolfi, J. M., Corlett, R. T., ... y Watson, J. E. (2016). The broad footprint of climate change from genes to biomes to people. *Science*, *354*(6313). https://doi.org/10.1126/science.aaf7671

Sigee D. (2005). *Freshwater microbiology: biodiversity and dynamic interactions of microorganisms in the aquatic environment*. John Wiley & Sons. https://doi.org/10.1002/0470011254

Slowikowski, K. (2021). ggrepel: Automatically Position N–n - Overlapping Text Labels with ggplot2. R package version 0.9.1. https://CRAN.R-project.org/package=ggrepel

Staley, C., Gould, T. J., Wang, P., Phillips, J., Cotner, J. B., y Sadowsky, M. J. (2015). Species sorting and seasonal dynamics primarily shape bacterial communities in the Upper Mississippi River. *Science of the Total Environment*, *505*, 435-445. https://doi.org/10.1016/j.scitotenv.2014.10.012

Tickner, D., Opperman, J. J., Abell, R., Acreman, M., Arthington, A. H., Bunn, S. E., ... y Young, L. (2020). Bending the curve of global freshwater biodiversity loss: an emergency recovery plan. *BioScience*, *70*(4), 330-342. https://doi.org/10.1093/biosci/biaa002

Tsilimigras, M. C., y Fodor, A. A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of epidemiology*, *26*(5), 330-335. https://doi.org/10.1016/j.annepidem.2016.03.002

Uribe, H. (2020). Recursos hídricos y riego en la Región de Ñuble. *Colección Libros INIA-Instituto de Investigaciones Agropecuarias*.

Valdovinos, C., y Parra, O. (2006). La cuenca del río Biobío: historia natural de un ecosistema de uso múltiple. *Publicaciones Centro EULA*, 1-25.

Vicuña Díaz, S., y Meza, F. J. (2012). Los nuevos desafíos para la gestión de los recursos hídricos en Chile en el marco del cambio global. Centro de Políticas Públicas UC.

Wang, P., Chen, B., Yuan, R., Li, C., y Li, Y. (2016). Characteristics of aquatic bacterial community and the influencing factors in an urban river. *Science of the Total Environment*, *569*, 382-389. https://doi.org/10.1016/j.scitotenv.2016.06.130

Wang, L., Zhang, J., Li, H., Yang, H., Peng, C., Peng, Z., y Lu, L. (2018). Shift in the microbial community composition of surface water and sediment along an urban river. *Science of the Total Environment*, *627*, 600-612. https://doi.org/10.1016/j.scitotenv.2018.01.203

Wang, P., Zhao, J., Xiao, H., Yang, W., y Yu, X. (2019). Bacterial community composition shaped by water chemistry and geographic distance in an anthropogenically disturbed river. *Science of the Total Environment*, *655*, 61-69. https://doi.org/10.1016/j.scitotenv.2018.11.234

Washington, V. J., Lear, G., Neale, M. W., y Lewis, G. D. (2013). Environmental effects on biofilm bacterial communities: a comparison of natural and anthropogenic factors in New Zealand streams. *Freshwater Biology*, *58*(11), 2277-2286. https://doi.org/10.1111/fwb.12208

Wei, T., y Simko, V. (2021). R package corrplot: Visualization of a Correlation Matrix (Version 0.92). https://github.com/taiyun/corrplot

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., ... y Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1), 1-18. https://doi.org/10.1186/s40168-017-0237-y

Whittaker, R. H. (1960). Vegetation of the Siskiyou mountains, Oregon and California. *Ecological monographs*, *30*(3), 279-338. https://doi.org/10.2307/1943563

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... y Yutani, H. (2019). Welcome to the Tidyverse. *Journal of open-source software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilke, C. O. (2020). cowplot: Streamlined Plot Theme and Plot Annotations f‘r 'ggpl’t2'. R package version 1.1.1. https://CRAN.R-project.org/package=cowplot

Wilke, C. O. (2021). ggtext: Improved Text Rendering Support for ggplot2. R package version 0.1.1. https://CRAN.R-project.org/package=ggtext

Xia, Y., Sun, J., y Chen, D. G. (2018). *Statistical analysis of microbiome data with R* (Vol. 847). Singapore: Springer. https://doi.org/10.1007/978-981-13-1534-3

Zárate, A., Dorador, C., Araya, R., Guajardo, M., Florez, J. Z., Icaza, G., ... y Valdés, J. (2020). Connectivity of bacterial assemblages along the Loa River in the Atacama Desert, Chile. *PeerJ*, *8*, e9927. https://doi.org/10.7717/peerj.9927

Zhang, X., Gu, Q., Long, X. E., Li, Z. L., Liu, D. X., Ye, D. H., ... y Chen, X. P. (2016). Anthropogenic activities drive the microbial community and its function in urban river sediment. *Journal of soils and sediments*, *16*, 716-725. https://doi.org/10.1007/s11368-015-1246-8

ANEXOS

Anexo 3.4 Ubicación de estaciones DGA, variables extraídas desde registros DGA y contraste entre valores de campaña y DGA

Figura A 3.4.1 Mapa con la ubicación geográfica de las estaciones DGA utilizadas.

Tabla A 3.4.1 Coordenadas e información de las estaciones DGA utilizadas.

Estación	Tipo	PM asociado	Código mapa	Coordenadas geográficas		Coordenadas UTM		
				Latitud S	Longitud W	Huso	Norte	Este
Río Biobío en Coihue	F, C	P1	E1	37° 33’ 35’’	72° 35’ 20’’	18H	5840299	712955
Río Biobío en Santa Juana	C	P3	E2	37° 10’ 14’’	72° 56’ 09’’	18H	5884207	683264
Río Biobío en Desembocadura	F	P3	E3	36° 50’ 16’’	73° 03’ 41’’	18H	5921362	672852
Río Itata en Balsa Nueva Aldea	F, C	P4	E4	36° 39’ 20’’	72° 27’ 00’’	18H	5940282	727913
Río Ñuble en Confluencia	C	P5	E5	36° 38’ 14’’	72° 27’ 08’’	18H	5942344	727773
Río Longaví en Longitudinal	C	P6	E6	36° 00’ 14’’	71° 43’ 30’’	19H	6012181	254382
Río Maule en Longitudinal	F, C	P7	E7	35° 33’ 34’’	71° 42’ 39’’	19H	6061507	254294
Río Teno antes Junta Mataquito	C	P8	E8	34° 58’ 14’’	71° 22’ 08’’	19H	6127661	283737
Río Tinguiririca en Los Olmos	C	P9	E9	34° 29’ 51’’	71° 22’ 30’’	19H	6180119	281938

La columna Tipo indica si la estación es fluviométrica (F), de monitoreo de cal©(C) o ambas (F, C).

Tabla A 3.4.2 Detalle de variables ambientales extraídas desde los registros DGA.

Parámetro	Método de medición	Unidades
Aluminio Total	Espectrofotometría de absorción atómica	mg/L Al
Cloruro	Potenciométrico – Argentométrico – Titulación	mg/l Cl
Conductividad Específica	Conductivímetro	mhos/cm
Demanda Química de Oxígeno	Reflujo Dicromato de Potasio – Colorimetría	mg/l O_2
Fierro Total	Espectrofotometría de absorción atómica	mg/l Fe
Fósforo de Ortofosfato	Kjeldahl – Colorimetría	mg/l PO4
Manganeso Total	Colorimetría – Persulfato	mg/l Mn
Nitrógeno de Nitrito	Ned-Dicloruro-Colorimetría	mg/l
Oxígeno Disuelto		mg/l
Ph	Potenciométrico	unidades. Ph
Sulfato	Turbidimétrico	mg/l
Temperatura	Termómetro	Grados Celsius

Tabla A 3.4.3 Conjuntos de variables ambientales recopilados.

Conjunto N° 1 (5)	Conjunto N° 2 (10)	Conjunto N° 3 (13) *	Conjunto N° 4 (8)
Variables espaciales, caudal y área de cuenca	Parámetros fisicoquímicos (CT)	Parámetros fisicoquímicos (DGA)	Usos de suelo
Altitud Latitud Longitud Caudal medio anual Área de cuenca	pH Temperatura Conductividad Oxígeno disuelto Turbidez Fósforo total Nitrógeno total NO_3+NO_2 TKN DQO	pH Temperatura Conductividad Oxígeno disuelto DQO N-NO3 P-PO4 Cociente N/P Cloruro Sulfato Mn total Al total Fe Total	Agrícola Urbano Bosque nativo Sup. de agua Nieve y glaciares Plantación forestal Praderas y matorrales Sin vegetación

* P2 no dispone de información para las variables ambientales de este conjunto.

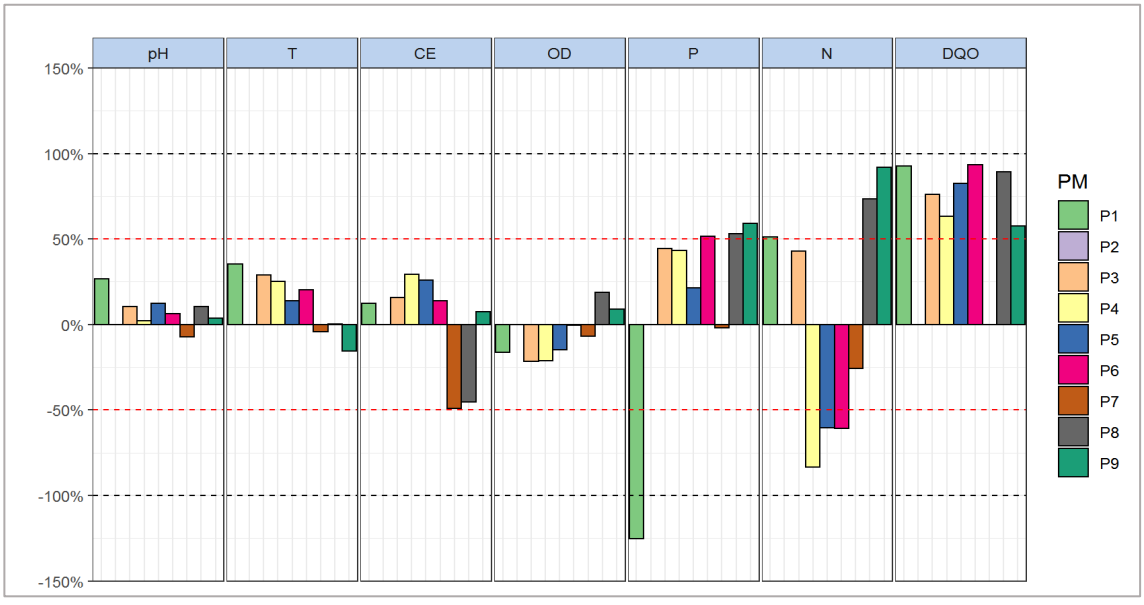

Figura A 3.4.2 Comparación entre valores de campaña y DGA.

Anexo 4.2 Mapa con clasificación de PMs según usos de suelo y correlaciones entre variables ambientales

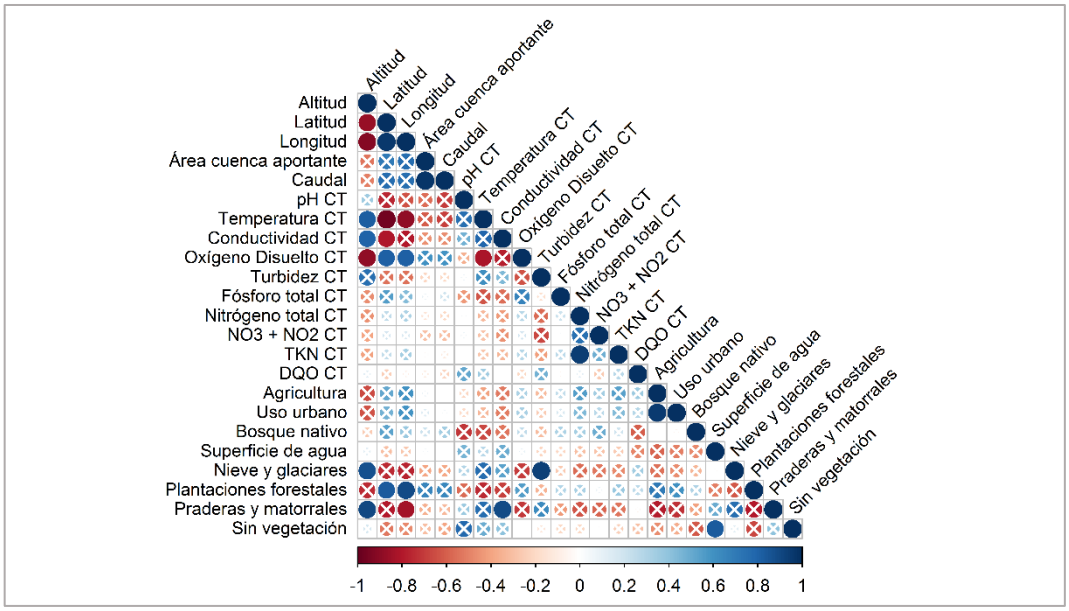

Figura A 4.2.1 Correlación entre variables de los conjuntos N° 1, N° 2 y N° 4.

Figura A 4.2.2 Mapa clasificación de los puntos de muestreo según usos de suelo.

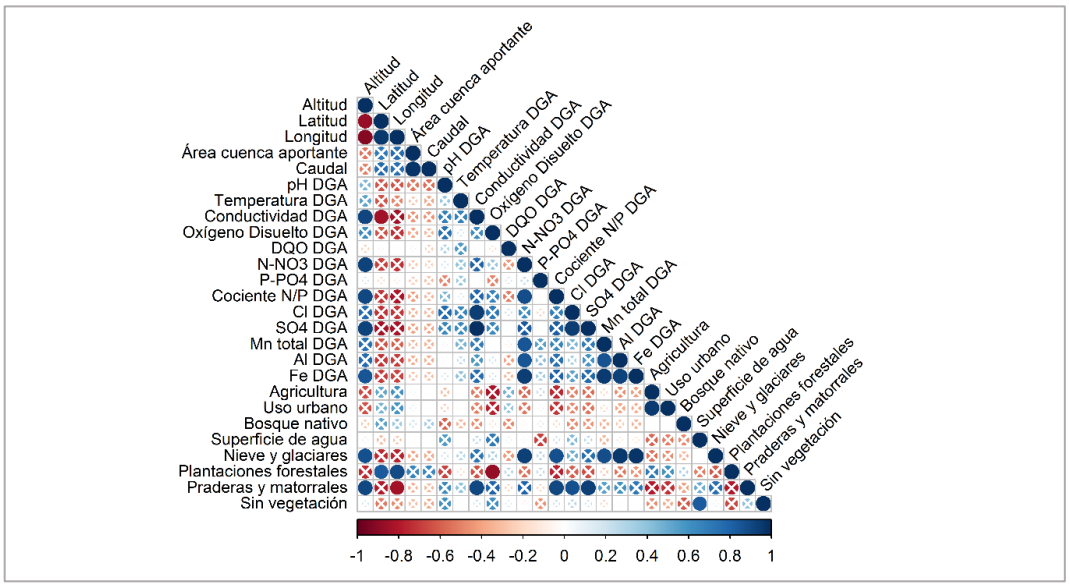

Figura A 4.2.3 Correlación entre las variables de los conjuntos N° 1, N° 3 y N° 4.

Anexo 4.3 Resumen del procesamiento bioinformático de las secuencias genéticas y efectos del filtro abundancia/prevalencia (ab/prev) en cada muestra

Tabla A 4.3.1 Resultados parciales del procesamiento bioinformático con DADA2.

Muestra	Secuencias iniciales	Número de secuencias conservadas tras					Variación (%)
		Filtro de calidad	Remoción de ruido		Unión F-R*	Remoción quimeras	
			En F*	En R*			
P1L	147 901	125 854	125 272	125 287	120 668	118 918	80.4
P2L	83 557	71 135	68 054	69 236	52 690	51 559	61.7
P3L	97 973	83 377	80 032	81 153	64 883	63 438	64.8
P4L	79 724	67 857	67 536	67 600	65 138	64 628	81.1
P5L	95 017	79 896	79 655	79 559	76 696	76 457	80.5
P6L	89 351	76 193	75 430	75 614	70 836	69 991	78.3
P8L	89 951	76 882	76 480	76 587	74 107	73 742	82.0
P9L	102 047	87 166	85 906	86 392	80 077	78 514	76.9
P1A	65 194	55 218	54 931	54 990	53 246	52 270	80.2
P2A	86 749	73 910	73 573	73 623	70 977	69 565	80.2
P3A	86 097	73 355	73 170	73 143	71 275	70 507	81.9
P4A	104 837	89 159	88 358	88 403	83 384	82 319	78.5
P5A	117 641	100 370	99 681	99 807	95 265	94 482	80.3
P6A	108 362	92 753	91 852	92 038	87 757	86 251	79.6
P7A	99 954	83 967	83 583	83 458	79 765	78 555	78.6
P8A	62 914	52 605	52 127	52 223	48 479	47 489	75.5
Global	1 517 269	1 289 697	1 275 640	1 279 113	1 195 243	1 178 685	77.7

*F: secuencias *Forward*; R: secuencias *Reverse*.

Tabla A 4.3.2 Efectos del filtro ab/prev en el N° de lecturas y de ASVs por muestra.

Muestra	Antes del filtro		Después del filtro		Variación	
	N° lecturas	N° ASVs	N° lecturas	N° ASVs	Lecturas	ASVs
P1L	118 720	2 340	116 288	2 080	-2.05	-11.11
P2L	40 211	3 335	31 675	2 393	-21.23	-28.25
P3L	58 319	3 663	41 457	2 482	-28.91	-32.24
P4L	64 151	1 392	63 413	1 264	-1.15	-9.20
P5L	75 973	1 225	75 511	1 134	-0.61	-7.43
P6L	67 879	2 652	66 501	2 432	-2.03	-8.30
P8L	73 537	1 811	72 158	1 618	-1.88	-10.66
P9L	77 540	2 988	72 328	2 507	-6.72	-16.10
P1A	50 479	1 642	50 142	1 573	-0.67	-4.20
P2A	68 505	1 952	67 843	1 851	-0.97	-5.17
P3A	69 454	1 649	69 084	1 564	-0.53	-5.15
P4A	81 062	2 952	80 027	2 708	-1.28	-8.27
P5A	88 574	2 928	87 428	2 727	-1.29	-6.86
P6A	81 027	2 672	79 622	2 475	-1.73	-7.37
P7A	76 983	1 903	75 428	1 777	-2.02	-6.62
P8A	46 796	1 830	46 069	1 705	-1.55	-6.83
Global	1 139 210	6 762	1 094 974	4 372	-3.88	-35.34

Tabla A 4.3.3 Efectos del filtro de ab/prev en el número de taxones por nivel.

Nivel	Muestras de lecho (L)		Muestras de agua (A)		Todas las muestras	
	Antes	Después	Antes	Después	Antes	Después
Filo	46	37	36	36	46	37
Clase	110	82	83	77	111	83
Orden	229	182	188	172	235	183
Familia	304	251	252	241	307	252
Género	558	446	447	407	566	448
Especie	122	109	106	101	126	112

Solo se consideraron aquellos taxones cuya taxonomía era conocida hasta el nivel en cuestión.

Tabla A 4.3.4 Efectos del filtro ab/prev en la fracción de ASVs sin identificación para el nivel taxonómico indicado.

Nivel taxonómico	Muestras de lecho		Muestras de agua		Todas las muestras	
	Antes	Después	Antes	Después	Antes	Después
Filo	0	0	0	0	0	0
Clase	3.3	1.8	2.7	2.1	3.5	2
Orden	14.5	10.7	13	11.1	15.1	11.2
Familia	26.7	20.4	23.3	20.6	27.5	20.9
Género	57.5	51.7	55.3	52.7	58.2	52
Especie	97.7	96.9	97.4	97.0	97.8	97

Anexo 4.4 Valores de los índices de diversidad alfa en cada muestra y resultados de prueba estadística Fligner – Killen para evaluar homocedasticidad

Tabla A 4.4.1 Valores de los índices de diversidad alfa en cada muestra.

Muestra	Riqueza			Shannon – Wiener	Simpson
	Observada	Estimada	Razón obs/est (%)		
P1L	2340	2985	78.4	4.3	0.925
P2L	3335	3850	86.6	7.2	0.998
P3L	3663	4189	87.4	7.2	0.998
P4L	1392	1799	77.4	4.3	0.955
P5L	1225	1609	76.1	2.9	0.754
P6L	2652	3026	87.6	5.8	0.981
P8L	1811	2200	82.3	4.4	0.943
P9L	2988	3470	86.1	6.3	0.989
Promedio lecho	2426	2891	82.8	5.3	0.943
P1A	1642	2004	82.0	5.0	0.972
P2A	1952	2480	78.7	4.9	0.968
P3A	1649	2075	79.5	4.5	0.957
P4A	2952	3535	83.5	5.5	0.981
P5A	2928	3397	86.2	5.3	0.953
P6A	2672	3059	87.3	5.7	0.985
P7A	1903	2341	81.3	5.5	0.985
P8A	1830	2126	86.1	5.7	0.987
Promedio agua	2191	2627	83.1	5.3	0.973
Promedio global	2308	2759	82.9	5.3	0.958

Tabla A 4.4.2 Resultados prueba Fligner – Killeen en índices de diversidad alfa.

Índice de diversidad alfa	Prueba Fligner – Killeen		
	Estadístico (χ^2)	Valor *p*	Conclusión*
Riqueza observada	2.36	0.124	No se puede rechazar Ho
Riqueza estimada (Chao1)	1.45	0.229	No se puede rechazar Ho
Shannon - Wiener	7.80	< 0.01	Es posible aceptar Ha
Simpson	7.48	< 0.01	Es posible aceptar Ha

* Ho: las varianzas de los subconjuntos de agua y de lecho son iguales; Ha: las varianzas de estos subconjuntos no son iguales.

Anexo 4.6 Abundancias relativas de taxones más abundantes bajo Filo y Clase

Tabla A 4.6.1 Abundancias relativas observadas en cada muestra para los 10 Filos más abundantes.

Filo	Muestras de lecho (L)								Muestras de agua (A)								Promedios		
	P1L	P2L	P3L	P4L	P5L	P6L	P8L	P9L	P1A	P2A	P3A	P4A	P5A	P6A	P7A	P8A	L	A	T*
Proteobacteria	14.5	41.0	45.1	23.2	62.4	47.8	22.3	46.9	68.5	66.8	65.4	50.1	58.0	63.5	55.7	63.2	37.9	61.4	49.6
Bacteroidetes	58.6	11.2	6.7	40.6	18.6	26.2	44.4	23.5	5.4	7.2	9.8	16.9	20.7	7.7	12.2	8.7	28.7	11.1	19.9
Verrucomicrobia	1.8	7.5	6.0	0.4	0.5	7.5	2.5	8.4	10.6	5.7	9.6	16.0	8.0	15.6	16.7	15.6	4.3	12.2	8.3
Firmicutes	21.9	1.0	1.7	29.0	15.8	7.4	26.0	9.6	0.0	0.1	0.1	0.1	0.1	0.3	0.1	0.0	14.1	0.1	7.1
Actinobacteria	0.4	14.4	3.0	1.0	0.3	2.0	0.3	1.4	8.7	16.6	12.0	9.4	5.3	4.7	7.4	3.1	2.9	8.4	5.6
Planctomycetes	0.5	3.3	4.0	0.3	0.4	3.7	1.3	3.5	1.3	1.0	0.5	2.4	3.1	3.5	3.7	6.0	2.1	2.7	2.4
Acidobacteria	0.2	11.2	11.8	0.2	0.1	1.0	0.2	2.1	0.8	0.8	0.3	2.0	2.4	1.4	1.2	1.1	3.4	1.3	2.3
Chloroflexi	0.1	4.0	8.4	0.1	0.1	0.3	0.1	0.5	0.1	0.2	0.1	0.5	0.6	0.3	0.1	0.2	1.7	0.3	1.0
Epsilonbacteraeota	1.3	0.0	0.1	5.1	1.4	1.8	2.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.5	0.0	0.7
Gemmatimonadetes	0.1	2.1	1.8	0.0	0.0	0.3	0.2	0.6	0.0	0.1	0.0	0.4	0.5	0.5	0.5	0.6	0.7	0.3	0.5
Otros (27)	0.6	4.3	11.2	0.3	0.2	2.0	0.6	3.3	4.4	1.7	2.1	2.2	1.3	2.3	2.4	1.4	2.8	2.2	2.5

* Entre todas las muestras

Tabla A 4.6.2 Abundancias relativas observadas en cada muestra para las 20 Clases más abundantes.

Clase	Muestras de lecho (L)								Muestras de agua (A)								Promedio		
	P1	P2	P3	P4	P5	P6	P8	P9	P1	P2	P3	P4	P5	P6	P7	P8	L	A	T*
Gammaproteobacteria	8.7	19.9	27.8	17.0	58.7	21.2	15.0	28.4	35.7	43.3	43.3	34.1	45.3	30.8	34.1	31.5	24.6	37.3	30.9
Bacteroidia	58.5	10.9	6.3	40.5	18.6	25.9	44.4	23.0	5.4	7.1	9.8	16.7	20.6	7.6	12.0	8.6	28.5	11.0	19.7
Alphaproteobacteria	5.2	8.3	5.3	2.4	3.0	22.6	5.9	16.5	20.5	18.4	15.1	8.2	6.3	22.8	14.4	26.2	8.6	16.5	12.6
Verrucomicrobiae	1.8	7.5	6.0	0.4	0.5	7.5	2.5	8.4	10.6	5.7	9.6	16.0	8.0	15.6	16.7	15.6	4.3	12.2	8.3
Deltaproteobacteria	0.6	12.7	12.0	3.8	0.8	3.8	1.4	2.0	11.9	4.8	6.8	7.4	6.3	9.7	6.6	5.2	4.6	7.3	6.0
Clostridia	20.0	0.4	1.2	26.3	13.1	5.9	24.3	0.2	0.0	0.1	0.1	0.1	0.0	0.3	0.1	0.0	11.4	0.1	5.8
Actinobacteria	0.2	10.7	1.0	0.1	0.2	1.7	0.1	0.6	8.4	15.7	11.3	8.9	4.9	4.4	7.1	2.9	1.8	7.9	4.9
Planctomycetacia	0.5	1.7	2.3	0.3	0.4	3.2	0.6	2.7	1.1	0.8	0.4	1.7	2.2	2.9	3.0	5.4	1.5	2.2	1.8
Subgroup_6	0.1	4.8	4.0	0.0	0.0	0.1	0.0	0.7	0.1	0.2	0.1	0.8	1.2	0.2	0.3	0.2	1.2	0.4	0.8
Bacilli	0.1	0.6	0.4	0.3	0.3	0.5	0.4	9.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.5	0.0	0.8
Campylobacteria	1.3	0.0	0.1	5.1	1.4	1.8	2.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.5	0.0	0.7
Negativicutes	1.8	0.0	0.0	2.4	2.4	1.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	0.0	0.6
Anaerolineae	0.0	2.0	5.2	0.0	0.0	0.2	0.0	0.2	0.1	0.1	0.0	0.2	0.3	0.2	0.0	0.1	1.0	0.1	0.5
Blastocatellia (Subgroup_4)	0.0	2.5	2.9	0.1	0.0	0.4	0.1	0.7	0.1	0.2	0.1	0.5	0.4	0.2	0.2	0.3	0.8	0.2	0.5
Gemmatimonadetes	0.1	1.4	1.8	0.0	0.0	0.3	0.2	0.6	0.0	0.1	0.0	0.4	0.5	0.5	0.5	0.6	0.6	0.3	0.5
Acidobacteriia	0.0	1.5	0.9	0.0	0.0	0.3	0.0	0.2	0.5	0.2	0.1	0.5	0.6	0.9	0.7	0.5	0.4	0.5	0.4
Acidimicrobiia	0.1	0.7	0.6	0.1	0.1	0.2	0.1	0.6	0.2	0.8	0.7	0.4	0.3	0.2	0.3	0.1	0.3	0.4	0.3
Thermoleophilia	0.1	2.5	1.0	0.1	0.0	0.1	0.0	0.2	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.5	0.1	0.3
Babeliae	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	1.9	0.5	0.8	0.6	0.3	0.1	0.0	0.1	0.0	0.5	0.3
NC10	0.0	0.5	3.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.3
Otros (63)	0.9	10.8	15.5	1.2	0.4	2.5	1.4	5.0	2.2	1.7	1.4	2.8	2.3	3.0	3.3	1.9	4.7	2.3	3.5
N/A	0.0	0.4	1.8	0.0	0.0	0.6	0.1	0.8	1.3	0.3	0.4	0.6	0.3	0.5	0.8	0.6	0.5	0.6	0.5

* Entre todas las muestras

Anexo 4.7 Resultados pruebas estadísticas PERMANOVA y ANOSIM, y composición taxonómica bajo nivel de Familia

Tabla A 4.7.1 Resultados prueba PERMANOVA para los niveles Filo, Clase y ASV.

Nivel	Enfoque tradicional			Enfoque composicional		
	Estadístico F	R^2	Valor *p*	Estadístico F	R^2	Valor *p*
Filo	9.44	0.40	0.001	9.98	0.42	0.001
Clase	9.34	0.40	0.001	8.89	0.39	0.001
ASV	5.63	0.29	0.001	4.49	0.24	0.001

R^2 indica el porcentaje de variabilidad explicada.

Tabla A 4.7.2 Resultados prueba ANOSIM para los niveles Filo, Clase y ASV.

Nivel	Enfoque tradicional		Enfoque composicional	
	R	Valor *p*	R	Valor *p*
Filo	0.70	0.002	0.76	0.002
Clase	0.68	0.002	0.77	0.002
ASV	0.73	0.002	0.69	0.002

R indica el nivel de correlación, un valor 0 se traduce en que no existe correlación.

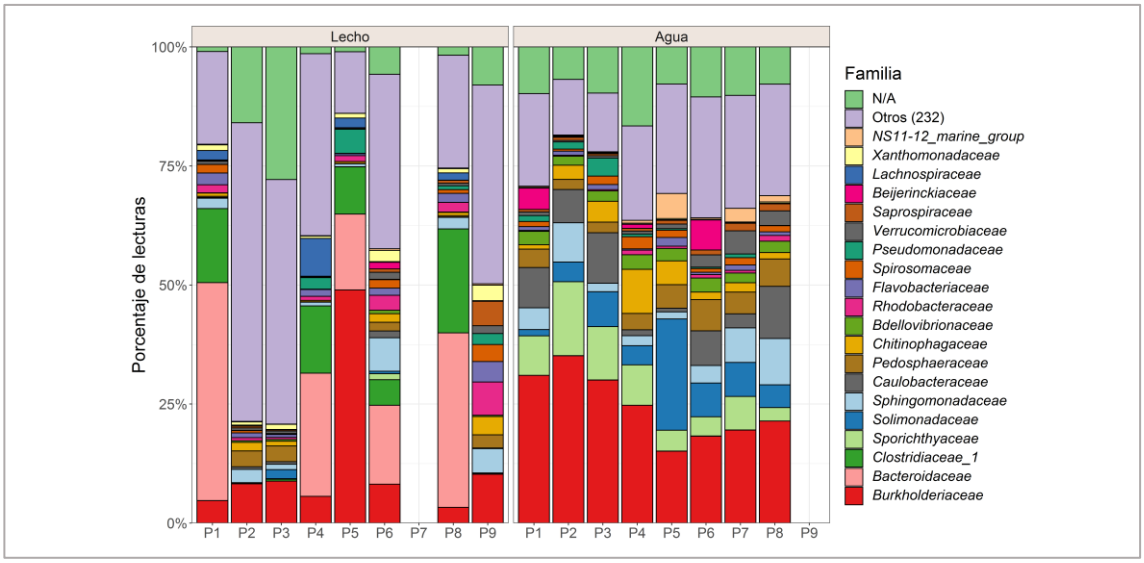

Figura A 4.7.1 Composición de las comunidades bacterianas a nivel de Familia.

Anexo 4.8 Relación entre características estructurales de las comunidades bacterianas y variables ambientales

Tabla A 4.8.1 Detalle del análisis de correlación entre variables ambientales e índices de diversidad alfa para las muestras de lecho.

Variable	Riqueza estimada (Chao1)			Shannon - Wiener			Simpson		
	r	*p*	*p* corregido	r	*p*	*p* corregido	r	*p*	*p* corregido
Altitud	0.43	0.289	0.896	0.33	0.420	0.896	0.21	0.610	0.896
Latitud	0.05	0.911	0.969	0.00	1.000	1.000	0.07	0.867	0.953
Área cuenca aportante	0.31	0.456	0.896	0.24	0.570	0.896	0.29	0.493	0.896
Turbidez campaña	0.07	0.867	0.953	-0.14	0.736	0.896	-0.26	0.531	0.896
pH DGA	-0.46	0.294	0.896	-0.43	0.337	0.896	-0.61	0.148	0.896
Conductividad DGA	-0.14	0.760	0.896	0.21	0.645	0.896	0.14	0.760	0.896
DQO DGA	-0.57	0.180	0.896	-0.18	0.702	0.896	-0.14	0.760	0.896
N-NO3 DGA	0.49	0.268	0.896	0.41	0.355	0.896	0.23	0.613	0.896
P-PO4 DGA	0.14	0.760	0.896	0.39	0.383	0.896	0.57	0.180	0.896
Cociente N/P DGA	0.25	0.589	0.896	0.14	0.760	0.896	-0.04	0.939	0.969
Al DGA	0.29	0.535	0.896	0.46	0.294	0.896	0.36	0.432	0.896

Aquellos valores $p < 0.01$ y valores *p* corregidos < 0.05 se encuentran destacados con negrita.

Tabla A 4.8.2 Detalle del análisis de correlación entre variables ambientales e índices de diversidad alfa para las muestras de agua.

Variable	Riqueza estimada (Chao1)			Sh–nnon - Wiener			Simpson		
	r	*p*	*p* corregido	r	*p*	*p* corregido	r	*p*	*p* corregido
Altitud	-0.43	0.289	0.531	0.52	0.183	0.418	0.86	**0.007**	0.108
Latitud	-0.24	0.570	0.724	-0.74	**0.037**	0.383	-0.64	0.086	0.387
Área cuenca aportante	-0.50	0.207	0.418	-0.98	**<0.001**	**0.001**	-0.69	0.058	0.383
Turbidez CT	-0.50	0.207	0.418	-0.29	0.493	0.650	-0.19	0.651	0.737
pH DGA	-0.07	0.879	0.907	0.32	0.482	0.650	0.43	0.337	0.586
Conductividad DGA	0.39	0.383	0.602	0.71	0.071	0.387	0.75	0.052	0.383
DQO DGA	0.64	0.119	0.418	0.32	0.482	0.650	0.14	0.760	0.809
N-NO3 DGA	-0.61	0.144	0.418	0.20	0.670	0.737	0.41	0.355	0.586
P-PO4 DGA	0.68	0.094	0.387	0.36	0.432	0.647	0.00	1.000	1.000
Cociente N/P DGA	-0.57	0.180	0.418	0.21	0.645	0.737	0.54	0.215	0.418
Al DGA	0.21	0.645	0.737	0.54	0.215	0.418	0.54	0.215	0.418

Aquellos valores $p < 0.01$ y valores *p* corregidos < 0.05 se encuentran destacados con negrita.

Tabla A 4.8.3 Detalle del análisis de correlación entre cinco Filos más abundantes en lecho y variables ambientales

Variable	*Proteobacteria*			*Bacteroidetes*			*Verrucomicrobia*			*Firmicutes*			*Acidobacteria*		
	r	***p***	***p* cor.**	**r**	***p***	***p* cor.**	**r**	***p***	***p* cor.**	**r**	***p***	***p* cor.**	**r**	***p***	***p* cor.**
Altitud	-0.21	0.645	0.957	0.29	0.535	0.957	0.79	0.036	0.957	-0.14	0.760	0.957	0.61	0.148	0.957
Latitud	-0.39	0.383	0.957	0.11	0.819	0.957	-0.54	0.215	0.957	0.04	0.939	0.957	-0.21	0.645	0.957
Área cuenca aportante	-0.21	0.645	0.957	-0.29	0.535	0.957	-0.36	0.432	0.957	-0.14	0.760	0.957	-0.04	0.939	0.957
Turbidez CT	-0.29	0.535	0.957	0.18	0.702	0.957	0.25	0.589	0.957	0.18	0.702	0.957	0.11	0.819	0.957
pH DGA	0.04	0.939	0.957	0.29	0.535	0.957	0.00	1.000	1.000	0.36	0.432	0.957	-0.36	0.432	0.957
Conductividad DGA	0.04	0.939	0.957	0.14	0.760	0.957	0.21	0.645	0.957	0.36	0.432	0.957	0.07	0.879	0.957
DQO DGA	0.04	0.939	0.957	-0.07	0.879	0.957	-0.57	0.180	0.957	0.50	0.253	0.957	-0.39	0.383	0.957
N-NO3 DGA	-0.16	0.728	0.957	0.23	0.613	0.957	0.77	0.041	0.957	-0.20	0.670	0.957	0.54	0.210	0.957
P-PO4 DGA	0.39	0.383	0.957	-0.36	0.432	0.957	0.14	0.760	0.957	-0.07	0.879	0.957	0.18	0.702	0.957
Cociente N/P DGA	-0.14	0.760	0.957	0.32	0.482	0.957	0.61	0.148	0.957	-0.04	0.939	0.957	0.29	0.535	0.957
Al DGA	0.57	0.180	0.957	-0.32	0.482	0.957	0.75	0.052	0.957	-0.39	0.383	0.957	0.43	0.337	0.957

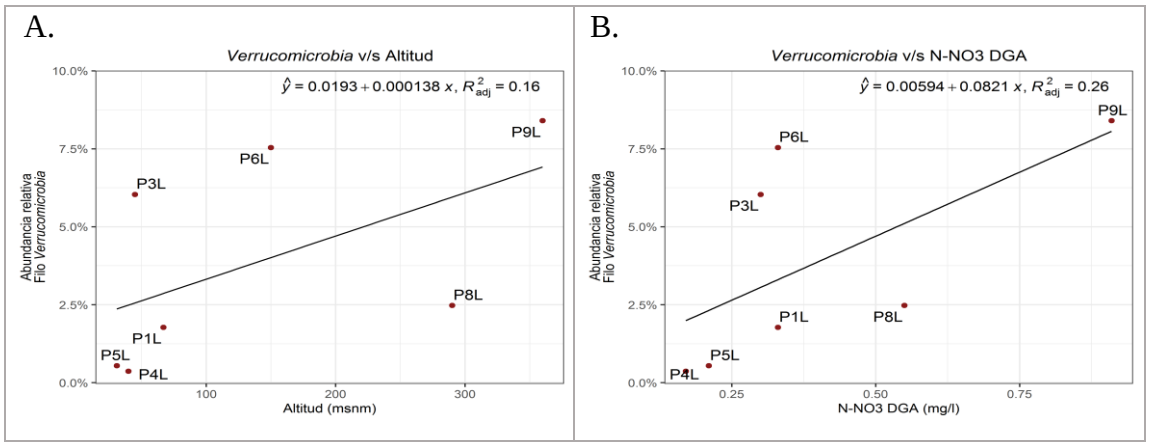

Figura A 4.8.1 Relación del Filo *Verrucomicrobia* (lecho) con la altitud (A) y con N-NO_3 (B).

Tabla A 4.8.4 Detalle del análisis de correlación entre cinco Filos más abundantes en agua y variables ambientales.

Variable	*Proteobacteria*			*Verrucomicrobia*			*Bacteroidetes*			*Actinobacteria*			*Planctomycetes*		
	r	***p***	***p* cor.**	**R**	***p***	***p* cor.**	**R**	***p***	***p* cor.**	**R**	***p***	***p* cor.**	**R**	***p***	***p* cor.**
Altitud	0.25	0.589	0.830	0.43	0.337	0.663	-0.68	0.094	0.368	-0.61	0.148	0.408	0.64	0.119	0.386
Latitud	0.46	0.294	0.647	-0.43	0.337	0.663	-0.14	0.760	0.871	0.79	0.036	0.319	-0.96	**<0.01**	**0.025**
Área cuenca aportante	0.36	0.432	0.698	-0.36	0.432	0.698	0.04	0.939	0.957	0.75	0.052	0.319	-0.68	0.094	0.368
Turbidez CT	0.32	0.482	0.717	-0.57	0.180	0.451	-0.14	0.760	0.871	-0.14	0.760	0.871	-0.04	0.939	0.957
pH DGA	-0.21	0.645	0.844	0.14	0.760	0.871	0.00	1.000	1.000	-0.75	0.052	0.319	0.82	0.023	0.319
Conductividad DGA	-0.68	0.094	0.368	0.64	0.119	0.386	0.25	0.589	0.830	-0.54	0.215	0.493	0.86	0.014	0.319
DQO DGA	-0.82	0.023	0.319	0.21	0.645	0.844	0.71	0.071	0.357	-0.11	0.819	0.883	0.36	0.432	0.698
N-NO3 DGA	0.72	0.068	0.357	-0.32	0.478	0.717	-0.77	0.041	0.319	-0.56	0.192	0.460	0.23	0.613	0.843
P-PO4 DGA	-0.36	0.432	0.698	0.11	0.819	0.883	0.36	0.432	0.698	0.18	0.702	0.871	-0.18	0.702	0.871
Cociente N/P DGA	0.36	0.432	0.698	0.07	0.879	0.930	-0.61	0.148	0.408	-0.64	0.119	0.386	0.57	0.180	0.451
Al DGA	-0.32	0.482	0.717	0.43	0.337	0.663	0.11	0.819	0.883	-0.61	0.148	0.408	0.79	0.036	0.319

Aquellos valores $p < 0.01$ y valores p corregidos < 0.05 se encuentran destacados con negrita.

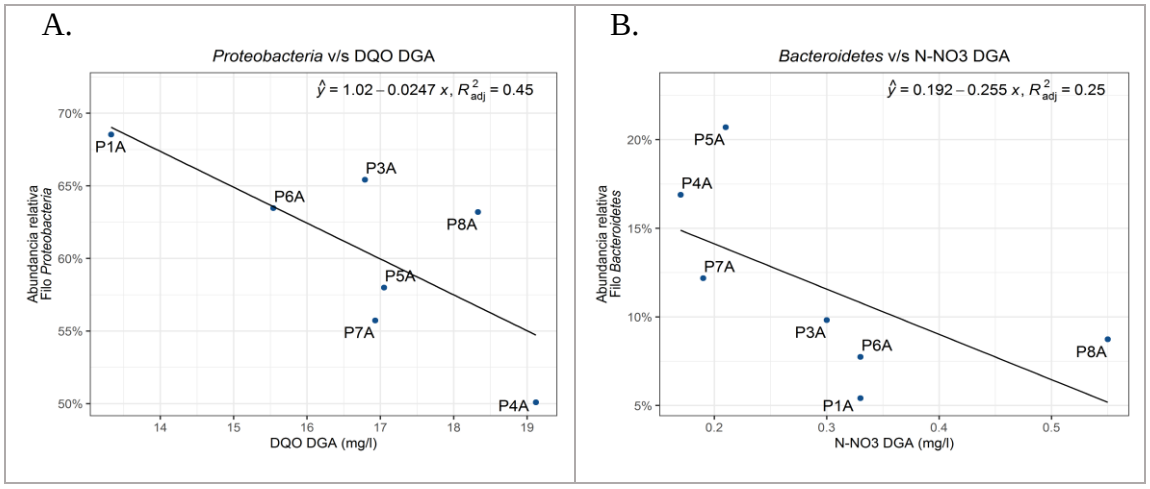

Figura A 4.8.2 Relación del Filo *Proteobacteria* (agua) con DQO (A) y del Filo *Bacteroidetes* (agua) con $N - NO_3$ (B).

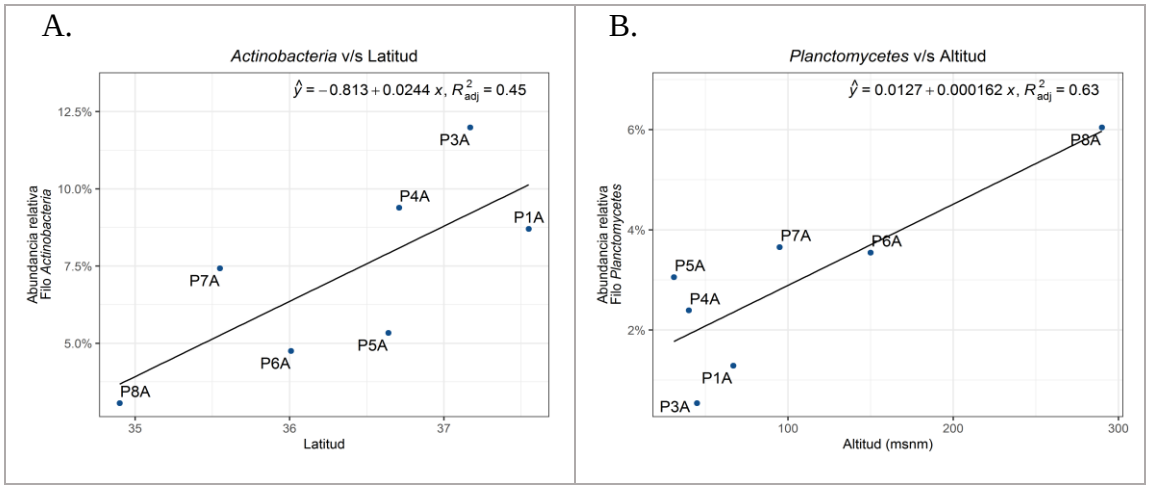

Figura A 4.8.3 Relación del Filo *Actinobacteria* (agua) con la latitud (A) y del Filo *Planctomycetes* (agua) con la altitud (B)

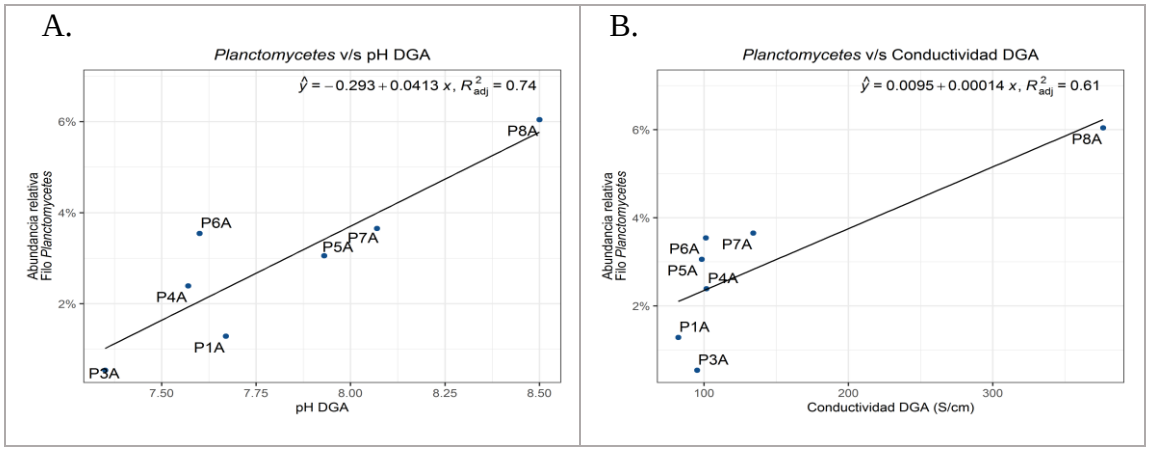

Figura A 4.8.4 Relación del Filo *Planctomycetes* (agua) con el pH (A) y con la conductividad (B).

Tabla A 4.8.5 Detalle del análisis de correlación entre las cinco Clases más abundantes en lecho y variables ambientales.

Variable	*Bacteroidia*			*Gammaproteobacteria*			*Clostridia*			*Alphaproteobacteria*			*Deltaproteobacteria*		
	r	***p***	***p* cor.**	**r**	***p***	***p* cor.**	**r**	***p***	***p* cor.**	**r**	***p***	***p* cor.**	**r**	***p***	***p* cor.**
Altitud	0.29	0.535	0.994	-0.21	0.645	0.994	-0.36	0.432	0.994	0.82	0.023	0.664	-0.07	0.879	0.994
Latitud	0.11	0.819	0.994	-0.36	0.432	0.994	0.29	0.535	0.994	-0.61	0.148	0.906	0.00	1.000	1.000
Área cuenca aportante	-0.29	0.535	0.994	0.07	0.879	0.994	-0.07	0.879	0.994	-0.61	0.148	0.906	0.00	1.000	1.000
Turbidez CT	0.18	0.702	0.994	0.00	1.000	1.000	-0.18	0.702	0.994	0.14	0.760	0.994	-0.54	0.215	0.986
pH DGA	0.29	0.535	0.994	0.04	0.939	0.994	0.14	0.760	0.994	0.14	0.760	0.994	-0.75	0.052	0.717
Conductividad DGA	0.14	0.760	0.994	0.04	0.939	0.994	0.14	0.760	0.994	0.32	0.482	0.994	0.14	0.760	0.994
DQO DGA	-0.07	0.879	0.994	0.04	0.939	0.994	0.61	0.148	0.906	-0.46	0.294	0.994	0.36	0.432	0.994
N-NO3 DGA	0.23	0.613	0.994	-0.13	0.788	0.994	-0.43	0.333	0.994	0.79	0.033	0.664	-0.29	0.531	0.994
P-PO4 DGA	-0.36	0.432	0.994	0.39	0.383	0.994	-0.18	0.702	0.994	0.04	0.939	0.994	0.71	0.071	0.785
Cociente N/P DGA	0.32	0.482	0.994	-0.11	0.819	0.994	-0.32	0.482	0.994	0.64	0.119	0.906	-0.54	0.215	0.986
Al DGA	-0.32	0.482	0.994	0.50	0.253	0.994	-0.57	0.180	0.986	0.79	0.036	0.664	0.11	0.819	0.994

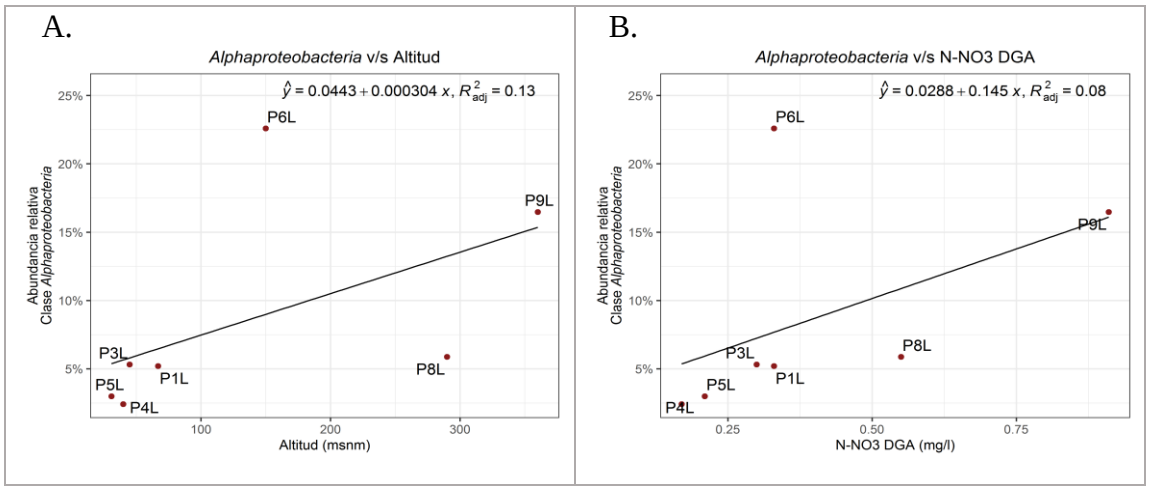

Figura A 4.8.5 Relación de la Clase *Alphaproteobacteria* (lecho) con la altitud (A) y con $N - NO_3$ (B).

Tabla A 4.8.6 Detalle del análisis de correlación entre las cinco Clases más abundantes en agua y variables ambientales.

Variable	*Gammaproteobacteria*			*Alphaproteobacteria*			*Verrucomicrobiae*			*Bacteroidia*			*Actinobacteria*		
	r	***p***	***p* cor.**	**r**	***p***	***p* cor.**	**r**	***p***	***p* cor.**	**r**	***p***	***p* cor.**	**r**	***p***	***p* cor.**
Altitud	-0.75	0.052	0.359	0.89	**0.007**	0.232	0.43	0.337	0.773	-0.68	0.094	0.469	-0.61	0.148	0.480
Latitud	0.54	0.215	0.564	-0.29	0.535	0.840	-0.43	0.337	0.773	-0.14	0.760	0.866	0.79	0.036	0.359
Área cuenca aportante	0.79	0.036	0.359	-0.32	0.482	0.829	-0.36	0.432	0.829	0.04	0.939	0.957	0.75	0.052	0.359
Turbidez CT	0.32	0.482	0.829	0.21	0.645	0.866	-0.57	0.180	0.551	-0.14	0.760	0.866	-0.14	0.760	0.866
pH DGA	-0.14	0.760	0.866	0.21	0.645	0.866	0.14	0.760	0.866	0.00	1.000	1.000	-0.75	0.052	0.359
Conductividad DGA	-0.61	0.148	0.480	0.18	0.702	0.866	0.64	0.119	0.480	0.25	0.589	0.866	-0.54	0.215	0.564
DQO DGA	-0.11	0.819	0.866	-0.36	0.432	0.829	0.21	0.645	0.866	0.71	0.071	0.436	-0.11	0.819	0.866
N-NO3 DGA	-0.32	0.478	0.829	0.88	**0.008**	0.232	-0.32	0.478	0.829	-0.77	0.041	0.359	-0.56	0.192	0.557
P-PO4 DGA	-0.29	0.535	0.840	-0.21	0.645	0.866	0.11	0.819	0.866	0.36	0.432	0.829	0.18	0.702	0.866
Cociente N/P DGA	-0.29	0.535	0.840	0.68	0.094	0.469	0.07	0.879	0.912	-0.61	0.148	0.480	-0.64	0.119	0.480
Al DGA	-0.39	0.383	0.829	0.18	0.702	0.866	0.43	0.337	0.773	0.11	0.819	0.866	-0.61	0.148	0.480

Aquellos valores $p < 0.01$ y valores p corregidos < 0.05 se encuentran destacados con negrita.

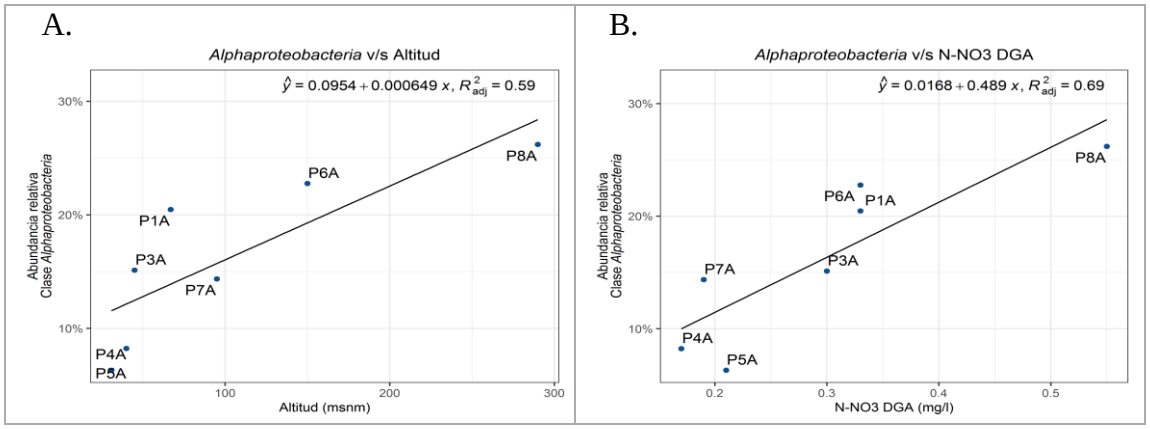

Figura A 4.8.6 Relación de la Clase *Alphaproteobacteria* (agua) con la altitud (A) y con $N - NO_3$ (B).

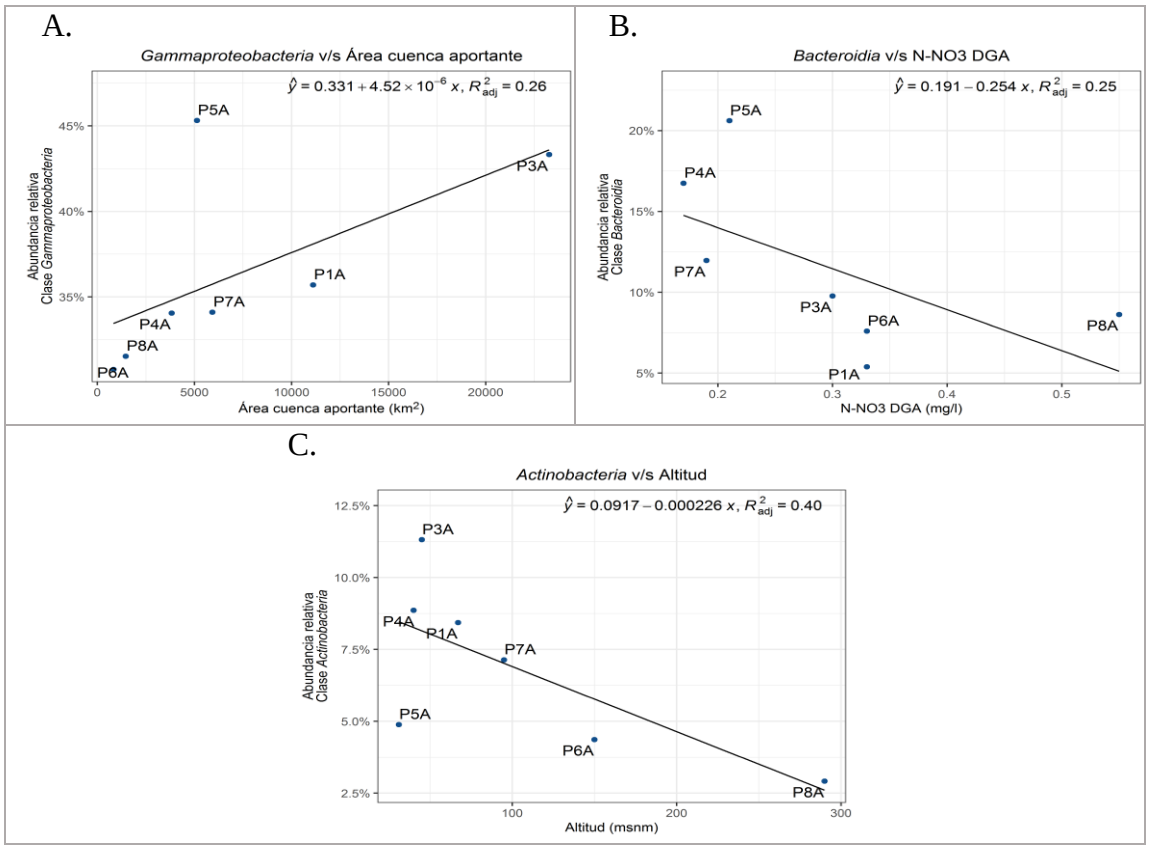

Figura A 4.8.7 Relación de los Filos *Gammaproteobacteria*, *Bacteroidia* y *Actinobacteria* con área de cuenca (A), N - NO_3 (B) y altitud (C), respectivamente.

Anexo 5.1 Código R: procesamiento de secuencias y asignación taxonómica

Información de Sesión:

```
R version 4.3.2 (2023-10-31 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 11 x64 (build 22631)
Matrix products: default
locale:
[1] LC_COLLATE=Spanish_Chile.utf8  LC_CTYPE=Spanish_Chile.utf8  LC_MONETARY=Spanish_Chile.utf8 LC_NUMERIC=C
[5] LC_TIME=Spanish_Chile.utf8
time zone: America/Santiago
tzcode source: internal
attached base packages:
 [1] parallel  stats4  grid  stats  graphics  grDevices utils  datasets  methods  base
other attached packages:
 [1] ALDEx2_1.34.0  latticeExtra_0.6-30  vegan_2.6-4  lattice_0.21-9
 [5] permute_0.9-7  zCompositions_1.5.0-1  truncnorm_1.0-9  NADA_1.6-1.1
 [9] survival_3.5-7  MASS_7.3-60  ggdendro_0.2.0  corrplot_0.92
[13] VennDiagram_1.7.3  futile.logger_1.4.3  cowplot_1.1.3  ggforce_0.4.2
[17] patchwork_1.2.0  ggpmisc_0.5.5  ggpp_0.5.6  ggtext_0.1.2
[21] ggrepel_0.9.5  broom_1.0.5  scales_1.3.0  paletteer_1.6.0
[25] RColorBrewer_1.1-3  GGally_2.2.1  sna_2.7-2  network_1.18.2
[29] statnet.common_4.9.0  qgraph_1.9.8  igraph_2.0.2  philr_1.28.0
[33] microbiome_1.24.0  DESeq2_1.42.0  SummarizedExperiment_1.32.0  Biobase_2.62.0
[37] MatrixGenerics_1.14.0  matrixStats_1.2.0  GenomicRanges_1.54.1  BiocManager_1.30.22
[41] phangorn_2.11.1  ape_5.7-1  DECIPHER_2.30.0  RSQLite_2.3.5
[45] Biostrings_2.70.2  GenomeInfoDb_1.38.6  XVector_0.42.0  IRanges_2.36.0
[49] S4Vectors_0.40.2  BiocGenerics_0.48.1  dada2_1.30.0  Rcpp_1.0.12
[53] phyloseq_1.46.0  ggpubr_0.6.0  xtable_1.8-4  kableExtra_1.4.0
[57] gridExtra_2.3  plyr_1.8.9  lubridate_1.9.3  forcats_1.0.0
[61] stringr_1.5.1  dplyr_1.1.4  purrr_1.0.2  readr_2.1.5
[65] tidyr_1.3.1  tibble_3.2.1  ggplot2_3.5.0  tidyverse_2.0.0
[69] knitr_1.45
```

Código:

```
# Ingreso datos necesarios ------------------------------------------------

# Directorio central (dirección carpeta PROCESAMIENTO SECUENCIAS)
dir.core <- 'C:/Users/joaqu/Desktop/PROCESAMIENTO SECUENCIAS'
# Carpeta con secuencias raw
rawsec <- 'SECUENCIAS RAW'
# Nombre carpeta de resultados y carpeta de respaldos
car.resu <- 'RESULTADOS_132pseudo'
car.resp <- 'RESPALDO_132pseudo'
# Nombre carpeta de base de datos genéticos
bd.gen <- 'SILVA132'
# Nombre de archivos de datos genéticos para asignación taxonómica
asigtax1 <- 'silva_nr_v132_train_set.fa'        # Taxonomía hasta género
asigtax2 <- 'silva_species_assignment_v132.fa'    # Taxonomía especie

# Paquetes ----------------------------------------------------------------
# Paquetes CRAN
cran.paq <- c("knitr", "tidyverse", "plyr", "grid",
              "gridExtra", "kableExtra", "xtable", "ggpubr")
# Paquetes Bioconductor
bioc.paq <- c("phyloseq", "dada2", "DECIPHER", "phangorn", "ggpubr",
              "BiocManager","DESeq2", "microbiome", "philr")
# Instalar paquetes CRAN
.inst <- cran.paq %in% installed.packages()
if(any(!.inst)) {
  install.packages(cran.paq[!.inst])
}
# Instalar paquetes BioConductor
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
.inst <- bioc.paq %in% installed.packages()
if(any(!.inst)) {
  BiocManager::install(bioc.paq[!.inst])
}
# Cargar paquetes
sapply(c(cran.paq, bioc.paq), require, character.only = TRUE)
```

```
# Procesamiento de secuencias ----------------------------------------------
set.seed(1996)

# Directorios
dir.sraw <- file.path(dir.core,rawsec)  # Archivos FASTQ
dir.meta <- file.path(dir.core,'METADATA')        # Metadata
dir.dtbs <- file.path(dir.core,'BASES DE DATOS')  # Bases de datos genéticos
dir.silv <- file.path(dir.dtbs,bd.gen)

# Creación directorios "RESPALDO" y "RESULTADOS"
dir.resp <- file.path(dir.core, car.resp)           # Carpeta respaldos
if (file.exists(dir.resp)==F) {dir.create(dir.resp)}
dir.resu <- file.path(dir.core, car.resu)          # Carpeta resultados
if (file.exists(dir.resu)==F) {dir.create(dir.resu)}

# Archivos Forward y Reverse
Fs <- sort(list.files(dir.sraw, pattern = '_R1.fastq'))
Rs <- sort(list.files(dir.sraw, pattern = '_R2.fastq'))

# subset de muestras
IDMuestra <- read.csv(file.path(dir.meta, 'IDMuestra.csv'))
Fs <- Fs[order(match(sapply(strsplit(Fs, '_'), `[`, 1), IDMuestra[,1]))]
Rs <- Rs[order(match(sapply(strsplit(Rs, '_'), `[`, 1), IDMuestra[,1]))]

# Remoción de "primers" (no necesario para estos datos)
# Directorio filtrado
dir.fltr <- file.path(dir.sraw, 'FILTRADO')
if(!file_test('-d', dir.fltr)) dir.create(dir.fltr)
Fs.fltr <- file.path(dir.fltr, paste0(IDMuestra[,1], '_F_fltr.fastq.gz'))
Rs.fltr <- file.path(dir.fltr, paste0(IDMuestra[,1], '_R_fltr.fastq.gz'))
# Directorios completos para Forward y Reverse
Fs <- file.path(dir.sraw, Fs)
Rs <- file.path(dir.sraw, Rs)

# Calidad de secuencias RAW
dir.cali <- file.path(dir.resu,'CALIDAD SECUENCIAS')
if(file.exists(dir.cali)==F){dir.create(dir.cali)}
print(plotQualityProfile(Fs)) + ggtitle('FORWARD')
ggsave('FORWARD.png', plot = last_plot(), path = dir.cali, scale = 2,
       width = 15, height = 12, units = 'cm', dpi = 500, limitsize = T)
print(plotQualityProfile(Rs)) + ggtitle('REVERSE')
ggsave('REVERSE.png', plot = last_plot(), path = dir.cali, scale = 2,
       width = 15, height = 12, units = 'cm', dpi = 500, limitsize = T)

# Filtro y truncado
fltr.out <- filterAndTrim(Fs, Fs.fltr, Rs, Rs.fltr, truncLen = c(145,140),
                          maxN = 0, maxEE = c(2,2), truncQ = 2, rm.phix = T,
                          matchIDs = T,compress = T, multithread = F)
saveRDS(fltr.out,file.path(dir.resp,'fltr_out.rds'))

# Dereplicado
Fs.derep <- derepFastq(Fs.fltr, verbose = T)
Rs.derep <- derepFastq(Rs.fltr, verbose = T)
names(Fs.fltr) <- IDMuestra[,1]
names(Rs.fltr) <- IDMuestra[,1]
saveRDS(Fs.derep,file.path(dir.resp,'Fs_derep.rds'))
saveRDS(Rs.derep,file.path(dir.resp,'Rs_derep.rds'))

# Errores y gráficas de errores
Fs.err <- learnErrors(Fs.fltr, multithread = F)
Rs.err <- learnErrors(Rs.fltr, multithread = F)

dir.err <- file.path(dir.resu,'ERRORES')
if(file.exists(dir.err) == F) {dir.create(dir.err)}
print(plotErrors(Fs.err)) + ggtitle('FORWARD')
ggsave('FORWARD.png', plot = last_plot(), path = dir.err, scale = 2,
       width = 15, height = 12, units = 'cm', dpi = 500, limitsize = T)
print(plotErrors(Rs.err)) + ggtitle('REVERSE')
ggsave('REVERSE.png', plot = last_plot(), path = dir.err, scale = 2,
       width = 15, height = 12, units = 'cm', dpi = 500, limitsize = T)

saveRDS(Fs.err,file.path(dir.resp,'Fs_err.rds'))
saveRDS(Rs.err,file.path(dir.resp,'Rs_err.rds'))

# Objetos dada (inferencia de ASVs)
Fs.dada <- dada(Fs.derep, err = Fs.err, multithread = F, pool = 'pseudo')
Rs.dada <- dada(Rs.derep, err = Rs.err, multithread = F, pool = 'pseudo')
saveRDS(Fs.dada,file.path(dir.resp,'Fs_dada.rds'))
saveRDS(Rs.dada,file.path(dir.resp,'Rs_dada.rds'))

# unión de secuencias
merged <- mergePairs(Fs.dada, Fs.derep, Rs.dada, Rs.derep)

# Tabla con todas las secuencias
sectab <- makeSequenceTable(merged)
table(nchar(getSequences(sectab)))

# Tabla sin secuencias quimeras
sectab.noquim <- removeBimeraDenovo(sectab)
saveRDS(sectab.noquim,file.path(dir.resp,'sectab_noquim.rds'))

# Resumen de "pipeline"
getN <- function(x) sum(getUniques(x))
track <- cbind(fltr.out, sapply(Fs.dada, getN), sapply(Rs.dada, getN),
               sapply(merged, getN), rowSums(sectab.noquim))
colnames(track) <- c("input","filtered","denoisedF",
                     "denoisedR","merged","nonchim")
rownames(track) <- IDMuestra[,1]
write.csv(track,file.path(dir.resu,'resumen.csv'))

# Asignación taxonómica (hasta Género)
fastaRef <- file.path(dir.silv,asigtax1)
taxtab <- assignTaxonomy(sectab.noquim, refFasta = fastaRef,
multithread = T)
saveRDS(taxtab,file.path(dir.resp,'taxtab.rds'))
gc()

# Asignación taxonómica (Especie)
sp.ref <- file.path(dir.silv,asigtax2)
taxtab.sp <- addSpecies(taxtab, sp.ref, n = 2, verbose=TRUE)
saveRDS(taxtab.sp,file.path(dir.resp,'taxtab_sp.rds'))
gc()

# Objeto phyloseq (sin árbol)
lectab <- t(sectab.noquim)
colnames(lectab) <- IDMuestra[,2]
meta <- read.csv(file.path(dir.meta,'Metadata.csv'),
                 header=TRUE)
rownames(meta) <- meta$Muestra
ps.notree <- phyloseq(otu_table(lectab, taxa_are_rows = T),
                      tax_table(taxtab),
                      sample_data(meta))
taxa_names(ps.notree) <- paste0('ASV',seq(ntaxa(ps.notree)))
saveRDS(ps.notree,file.path(dir.resu,'ps_notree.rds'))

# Árbol filogenético
secs <- getSequences(sectab.noquim)
names(secs) <- secs
alignment <- AlignSeqs(DNAStringSet(secs), anchor = NA,verbose = T)
phangAlign <- phyDat(as(alignment, "matrix"), type="DNA")
dm <- dist.ml(phangAlign)
treeNJ <- NJ(dm)
fit = pml(treeNJ, data = phangAlign)
fitGTR <- update(fit, k=4, inv=0.2)
fitGTR <- optim.pml(fitGTR, model="GTR", optInv=TRUE,
optGamma=TRUE,
                    rearrangement = "stochastic",
                    control = pml.control(trace = 0))
saveRDS(fitGTR,file.path(dir.resp,'fitGTR.rds'))

# Creación de objeto phyloseq (con Árbol)
ps.base <- phyloseq(otu_table(lectab, taxa_are_rows = T),
                    sample_data(meta),
                    tax_table(taxtab.sp),
                    phy_tree(fitGTR$tree))
taxa_names(ps.base) <- paste0('ASV',seq(ntaxa(ps.base)))
saveRDS(ps.base,file.path(dir.resu,'ps_base.rds'))
```

Anexo 5.2 Código R: Análisis estructurales de comunidades y estadística

Código:

```
# Activación de paquetes ----------------------------------------------

# Listado de paquetes
cran.paq <- c('igraph','qgraph','sna','network','GGally',
              'RColorBrewer','paletteer','tidyverse','scales','broom',
              'ggrepel','ggtext','ggpmisc','patchwork','ggforce','cowplot',
              'VennDiagram','corrplot','ggdendro','zCompositions')
bioc.paq <- c('vegan', 'phyloseq', 'ALDEx2', 'DESeq2')


# Instalación de paquetes faltantes
.inst <- cran.paq %in% installed.packages()
if(any(!.inst)) {
  install.packages(cran.paq[!.inst])
}

if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

.inst <- bioc.paq %in% installed.packages()
if(any(!.inst)) {
  BiocManager::install(bioc.paq[!.inst])
}


# Activar paquetes
sapply(c(cran.paq, bioc.paq), require, character.only = TRUE)

# Directorios de trabajo ----------------------------------------------

# Elementos a proporcionar por usuario:
# Definir directorio central
dir.core <- 'C:/Users/joaqu/Desktop/ANÁLISIS COMUNIDAD'

# Definir directorio objeto phyloseq
dir.phy <- paste0(dir.core,'/INFORMACIÓN BASE/ASV_S132')

# Definir directorio Metadata (IDMuestra, Auxivar)
dir.meta <- paste0(dir.core,'/METADATA')

# Definir nombre de directorio Resultados
RES <- 'ASV132pool_DGA_prom_anual' # Nombre carpeta

dir.resu <- paste0(dir.core,'/RESULTADOS/',RES)
if (file.exists(dir.resu)==F){dir.create(dir.resu)}

# Creación de directorios (automático)
# Directorios principales (Resultados)
dir.fltr <- paste0(dir.resu,'/00 - EFECTOS DE FILTRO PREVALENCIA')
dir.abun <- paste0(dir.resu,'/01 - ABUNDANCIAS RELATIVAS')
dir.comu <- paste0(dir.resu,'/02 - TAXONES COMUNES')
dir.alfa <- paste0(dir.resu,'/03 - DIVERSIDAD ALFA')
dir.beta <- paste0(dir.resu,'/04 - DIVERSIDAD BETA')
dir.vamb <- paste0(dir.resu,'/05 - VARIABLES AMBIENTALES')
dir.alfv <- paste0(dir.resu,'/06 - ALFA_VAR (REGRESIONES)')
dir.betv <- paste0(dir.resu,'/07 - BETA_VAR (ORDENACIÓN)')
dir.taxv <- paste0(dir.resu,'/08 - TAX_VAR (REGRESIONES)')

# creación de carpetas
for (i in c(dir.core,dir.fltr,dir.abun,dir.comu,dir.alfa,dir.beta,
            dir.vamb,dir.alfv,dir.betv,dir.taxv)){
  if (file.exists(i)==F){
    dir.create(i)
  }
}

# Ingreso de datos, conf. previas y elementos auxiliares ----------------
-

# Objeto phyloseq original (proporcionar nombre)
PS.base <- readRDS(file.path(dir.phy,'ps_pool_tree.rds'))

# Archivo auxiliar variables ambientales (proporcionar nombre)
auxivar <- read.csv(file.path(dir.meta,'Auxivar.csv'),
                    row.names = 1)

# Guardar gráficos y tablas (definir si guardar o no resultados)
GUARDAR <- T

# COMIENZA TRABAJO DEL PROGRAMA (no se requiere más del usuario)

# Arreglos en objeto phyloseq
tax_table(PS.base)[is.na(tax_table(PS.base))] <- 'N/A'

# Filtro taxonomía inapropiada, sin asignación o AR = 0% en todas las muestras
PS.base <- subset_taxa(PS.base, Dominio == 'Bacteria' &
                       !(Filo == 'N/A') &
                       !(Orden == 'Chloroplast') &
                       !(Familia == 'Mitochondria'))
PS.base <- filter_taxa(PS.base,function(x) sum(x)>0,prune = T)

# Filtro prevalencia/abundancia
# (prevalencia 25% de las muestras o Ab.rel. > 0.1% en mínimo 1 muestra)
PSr.base <- transform_sample_counts(PS.base,function(x) x/sum(x))
fltr <- filter_taxa(PSr.base, function(x) sum(x>0.001) > 0 | sum(x>0) > 3)
PS <- prune_taxa(fltr,PS.base)


# Efectos de filtro prev. en ASVs totales y lecturas totales
Nfltr <- data.frame(Lecturas.pre = sample_sums(PS.base),
                    ASVs.pre = apply(otu_table(PS.base),2,
                                     function(x){sum(x>0)}),
                    Lecturas.post = sample_sums(PS),
                    ASVs.post = apply(otu_table(PS),2,
                                      function(x) {sum(x>0)}))
Nfltr <- rbind(Nfltr,global = c(sum(Nfltr[,1]),ntaxa(PS.base),
                                sum(Nfltr[,3]),ntaxa(PS)))
Nfltr$Lec.var <- round((Nfltr[,3] - Nfltr[,1])/Nfltr[,1] * 100,digits=2)
Nfltr$ASV.var <- round((Nfltr[,4] - Nfltr[,2])/Nfltr[,2] * 100,digits=2)

if (GUARDAR==T){
  write.csv(Nfltr,paste0(dir.fltr,'/dif_ASVs_LECs.csv'))
}

# Efectos de filtro prev. en número de taxones por nivel taxonómico
ps.list <- list(pre = PS.base, pos = PS)
ite1 <- T
for (i in 2:7){
  for (f in c('pre','pos')){
    # Construcción tabla OTUs
    PS.ori <- ps.list[[f]]
    for (o in c('Lecho','Agua')) {
      cont <- subset_samples(PS.ori, Origen == o) %>%
        otu_table() %>%
        data.frame() %>%
        cbind(.,apply(.,1,function(x) {as.numeric(sum(x) > 0)}))

      colnames(cont)[ncol(cont)] <- paste('T',o,sep = '.')
```

```
    if (o=='Lecho') {
      cont.tab <- cont
    } else {
      cont.tab <- cbind(cont.tab,cont)
    }
  }

  # Construcción tabla taxonomía - conteos (tc)
  cont.tab <- cbind(cont.tab,
                    Total = apply(otu_table(PS.ori),1,
                                  function(x) as.numeric(sum(x) > 0)))
  tc <- data.frame(tax_table(PS.ori)[,i], cont.tab)
  tc[nrow(tc) + 1,] <- data.frame('N/A',t(rep(0,ncol(tc)-1)))
  tc <- aggregate(tc[-1], list(tc[,1]), function(x) sum(x > 0))
  tc <- rbind(apply(tc[-1][!(tc[,1]=='N/A'),], 2, function(x) sum(x > 0)),
              tc[-1][tc[,1]=='N/A',])
  tc <- data.frame(t(tc))
  colnames(tc) <- c('A','B')
  tc$B <- round(tc$B/apply(cont.tab,2,function(x) sum(x > 0)),digits = 3)*100
  colnames(tc) <- rep(paste(rank_names(PS.ori)[i],f,sep='.'),2)

  if (ite1==T){
    tc.tab <- tc[1]
    na.tab <- tc[2]
    ite1 <- F
  } else {
    tc.tab <- cbind(tc.tab,tc[1])
    na.tab <- cbind(na.tab,tc[2])
  }

 }

}

if (GUARDAR == T){
 write.csv(tc.tab,paste0(dir.fltr,'/Numero_taxones.csv'))  # N taxones
 write.csv(na.tab,paste0(dir.fltr,'/Numero_NAs.csv'))    # N N/As
}


# Proporción de ceros en tablas de lecturas
Pceros <- data.frame(ori = sum(otu_table(PS.base)==0)/(16*ntaxa(PS.base)),
                     fltr = sum(otu_table(PS)==0)/(16*ntaxa(PS)))

# Agrupación por taxonomía ("original")
Taxa.list.base <- list(Filo = tax_glom(PS.base,taxrank='Filo'),
                       Clase = tax_glom(PS.base,taxrank='Clase'),
                       Familia = tax_glom(PS.base,taxrank='Familia'),
                       Género = tax_glom(PS.base,taxrank='Género'),
                       ASV = PS.base)

# Agrupación por taxonomía (filtrado)
Taxa.list <- list(Filo = tax_glom(PS,taxrank='Filo'),
                  Clase = tax_glom(PS,taxrank='Clase'),
                  Familia = tax_glom(PS,taxrank='Familia'),
                  Género = tax_glom(PS,taxrank='Género'),
                  ASV = PS)

TAXA <- names(Taxa.list)


# AUXILIAR: Función para establecer límites de gráficos
xylim <- function(inf,sup){
 dif <- sup-inf
 esc <- 10^floor(log(dif,10))
 i <- floor(inf/esc)*esc
 s <- ceiling(sup/esc)*esc
 if (max(s-sup,inf-i)>dif/5){
   esc <- 10^floor(log(dif/3,10))
   i <- floor(inf/esc)*esc
   s <- ceiling(sup/esc)*esc
 }
 return(c(i,s))
}


# Limpieza
rm(ps.list, Nfltr, PS.ori, cont, cont.tab, tc, tc.tab, na.tab)


# Diversidad alfa (gráficos/tablas) ----------------------------------------

# Directorios
dir.auxi <- paste(dir.alfa,'01 - GRÁFICOS', sep='/')
if (file.exists(dir.auxi )==F){dir.create(dir.auxi)}


# Dataframe diversidad
div.alfa <- data.frame(Shannon = c('índice Shannon - Wiener','Valor Índice'),
                       Simpson = c('Índice Simpson','Valor Índice'),
                       Chao1 = c('Riqueza estimada (Chao1)',
                                 'Número estimado de especies'),
                       Observed = c('Riqueza observada',
                                    'NÚmero de especies observadas'),
                       row.names = c('titulo','etiqueta'))

DIV <- data.frame(sample_data(PS.base)[,c('Origen','Sitio','Muestra')])
DIV$Origen <- factor(DIV$Origen, levels = c('Lecho','Agua'))
DIV <- data.frame(complete(DIV,Origen,Sitio))
DIV2 <- data.frame(estimate_richness(PS.base,measures = colnames(div.alfa)))
DIV2 <- cbind(Muestra=rownames(DIV2),DIV2)
DIV <- left_join(DIV,DIV2,by='Muestra')
tag <- data.frame(Shannon = 'A.', Simpson = 'B.',Chao1 = 'B.', Observed = 'A.')
DIV$Origen <- factor(DIV$Origen, c('Lecho','Agua'))

pdiv.list <- list()

for (i in colnames(div.alfa)){
 if(i == 'Chao1'){
   ylim <- xylim(min(DIV[,i] - max(DIV[,'se.chao1'],na.rm =T), na.rm=T),
                 max(DIV[,i] + max(DIV[,'se.chao1'],na.rm =T),na.rm=T))
   } else{
     ylim <- xylim(min(DIV[,i],na.rm=T),max(DIV[,i],na.rm=T))
   }

 p1 <- ggplot(DIV, aes(x = .data[['Sitio']], y = .data[[i]],
                       group = Origen, color = Origen)) +
   geom_line(na.rm = T) +
   geom_point(size = 1.5, na.rm=T) +
   scale_color_manual(values=c('firebrick3','dodgerblue3')) +
   scale_y_continuous(position = 'left' , limits = ylim,
                      expand = expansion(mult = c(.1,.2), add = 0)) +
   xlab(NULL) + ylab(div.alfa['etiqueta',i]) + labs(tag = tag[i]) +
   theme(legend.position = c(0.8,0.96),
         legend.title = element_blank(),
         legend.direction = 'horizontal',
         legend.background = element_blank(),
         legend.key = element_blank(),
         legend.text = element_text(size=10))
 if (i=='Chao1'){
   p1 <- p1 + geom_errorbar(aes(x=Sitio,ymin=Chao1-se.chao1,
                                ymax=Chao1+se.chao1),
                            width=0.1,color='black')
 }

 p2 <- ggplot(DIV,aes(x = .data[['Origen']],y = .data[[i]], fill = Origen)) +
   geom_boxplot(na.rm=T) +
   scale_fill_manual(values=c('firebrick3','dodgerblue3')) +
   stat_summary(fun = mean, geom = 'point' , size=4, shape = 4, na.rm = T) +
```

```
   scale_y_continuous(position = 'right' , limits = ylim ,
                      expand = expansion(mult = c(.1,.2),add=0)) +
   stat_boxplot(geom='errorbar',width=0.3,na.rm=T) +
   theme(legend.position = 'none') +
   xlab(NULL) + ylab(NULL)

 p1 + p2 + plot_layout(width = c(3,1)) +
   plot_annotation(title=div.alfa['titulo',i]) &
   theme(plot.title = element_text(hjust=0.5))

 if (GUARDAR==T){
   ggsave(paste(i,'.png', sep = ''), path = dir.auxi, plot = last_plot(),
          scale = 1.5, width = 15,  height = 5,units = 'cm',  dpi = 500,
          limitsize = TRUE)
 }

 pdiv.list[[i]] <- (p1 + ggtitle(div.alfa['titulo',i]) +
                      theme(plot.title = element_text(hjust = 0.5))) +
   p2 + plot_layout(width = c(3,1))

}

if (GUARDAR==T){
 p.riq <- pdiv.list$Observed/pdiv.list$Chao1
 ggsave('riqueza_Chao1.png', path = dir.auxi, plot = last_plot(),
        scale = 1.7, width = 15,  height = 8,units = 'cm',  dpi = 500,
        limitsize = TRUE)

 p.div <- pdiv.list$Shannon/pdiv.list$Simpson
 ggsave('Shannon_Simpson.png', path = dir.auxi, plot = last_plot(),
        scale = 1.7, width = 15,  height = 8,units = 'cm',  dpi = 500,
        limitsize = TRUE)
}


# Limpieza
rm(dir.auxi,DIV2,p1,p2)


# Diversidad alfa (Estadística/tablas) ----------------------------------

# Directorio
dir.auxi <- paste0(dir.alfa,'/02 - TABLAS')
if (file.exists(dir.auxi )==F){dir.create(dir.auxi)}

DIV2 <- DIV[is.na(DIV$Muestra)==F,]
DIV2$se.chao1 <- NULL

ite1 <- T
for (i in colnames(div.alfa)){
 L <- as.numeric(DIV2[DIV2$Origen=='Lecho',i])
 A <- as.numeric(DIV2[DIV2$Origen=='Agua',i])
 Wt <- wilcox.test(L,A,alternative='two.sided',exact = F)
 flig <- fligner.test(x = list(L,A))

 aux <- data.frame(indice = i, mn.lecho = mean(L), sd.lecho = sd(L),
                   mn.agua = mean(A), sd.agua = sd(A),
                   FK.est = flig$statistic, FK.pval = flig$p.value,
                   sig.FKt = ifelse(flig$p.value < 0.05,'*',''),
                   W.est = Wt$statistic, pval = Wt$p.value,
                   sig.Wt = ifelse(Wt$p.value < 0.05,'*',''))
 rownames(aux) <- NULL

 if(ite1 == T){
   est.div <- aux
   ite1 <- F
 } else {
   est.div <- rbind(est.div,aux)
 }

}

if (GUARDAR==T){
 write.csv(est.div,paste(dir.auxi,'/','Wilcoxon_test_alfa.csv',sep=''))
}

DIV2 <- DIV2[,-c(1,2)] ; rownames(DIV2) <- DIV2[,1] ; DIV2 <- DIV2[-1]
colnames(DIV2) <- c('Riqueza','Chao1','Shannon','Simpson')

if (GUARDAR==T){
 write.csv(DIV2,paste(dir.auxi,'Diversidad_alfa.csv',sep='/'))
}

# Limpieza
rm(dir.auxi,DIV2,est.div,Wt,flig,L,A)


# Taxones Comunes ----------------------------------------------------------

# Directorio
dir.aux1 <- paste(dir.comu,'01 - GRÁFICOS',sep='/')
if (file.exists(dir.aux1)==F){dir.create(dir.aux1)}
dir.aux2 <- paste(dir.comu,'02 - TABLAS',sep='/')
if (file.exists(dir.aux2)==F){dir.create(dir.aux2)}

OTU.COM <- as.data.frame(matrix(NA,nrow=nsamples(PS.base),ncol=8))
OTU.COM[1:2] <- sample_data(PS.base)[,c('Muestra','Origen')]
colnames(OTU.COM) <- c('Muestra','Origen','asc.tot','lec.tot','asv.com',
                       'lec.com','prop.asv.com','prop.lec.com')
OTU.COM[3] <- apply(otu_table(PS.base),2,function(x) sum(x>0))
OTU.COM[4] <- apply(otu_table(PS.base),2,sum)

count.l <- otu_table(subset_samples(PS.base,Origen=='Lecho'))
count.a <- otu_table(subset_samples(PS.base,Origen=='Agua'))

OTU.COM[5] <- c(rep(sum(apply(count.l,1,min)>0),ncol(count.l)),
                rep(sum(apply(count.a,1,min)>0),ncol(count.a)))
OTU.COM[6] <- c(apply(count.l[apply(count.l,1,min)>0,],2,sum),
                apply(count.a[apply(count.a,1,min)>0,],2,sum))
OTU.COM[7] <- OTU.COM[5]/OTU.COM[3]
OTU.COM[8] <- OTU.COM[6]/OTU.COM[4]


# Gráficos
OTU.CM <- data.frame(sample_data(PS.base)[,c('Origen','Sitio','Muestra')])
OTU.CM$Origen <- factor(OTU.CM$Origen,levels=c('Lecho','Agua'))
OTU.CM <- data.frame(complete(OTU.CM,Origen, Sitio,
fill=list(value=NA)))
OTU.CM <- left_join(OTU.CM,OTU.COM[,c('Muestra','prop.asv.com','prop.lec.com')],
                    by='Muestra')
colnames(OT.CM)[4:5] <- c('ASVs_Comunes','Lecturas_Comunes')
ylab.com <- data.frame(ASVs_Comunes = 'Proporción de ASVs',
                       Lecturas_Comunes = 'Proporción de lecturas')
po.list <- list()
pl.list <- list()

for (i in c('ASVs_Comunes','Lecturas_Comunes')){
 ylim <- xylim(0,max(OT.CM[,i],na.rm=T))
 p1 <- ggplot(OT.CM[OT.CM$Origen=='Lecho',],
              aes(x=.data[['Sitio']], y = .data[[i]], fill=Origen)) +
   xlab(NULL) + ylab(ylab.com[i]) +
   geom_bar(stat='identity',position='dodge',na.rm=T) +
   scale_fill_manual(values='firebrick4') +
   scale_y_continuous(labels=percent_format(accuracy = 1),limits = ylim,
                      expand = expansion(mult = c(0,.1),add = 0)) +
   theme(legend.position = c(0.8,0.95),
         legend.title = element_blank(),
         legend.direction = 'horizontal',
         legend.background = element_blank(),
         legend.key = element_blank(),
         legend.key.size = unit(0.4,'cm'),
         panel.grid.major.x = element_blank())
```

```
  p2 <- ggplot(OT.CM[OT.CM$Origen=='Agua',],
               aes(x =.data[['Sitio']], y = .data[[i]], fill = Origen)) +
    xlab(NULL) + ylab(NULL) +
    geom_bar(stat='identity',position='dodge',na.rm=T) +
    scale_fill_manual(values='dodgerblue4') +
    scale_y_continuous(labels=NULL, limits = ylim,
                       expand = expansion(mult = c(0,.1),add = 0)) +
    theme(legend.position = c(0.8,0.95),
          legend.title = element_blank(),
          legend.direction = 'horizontal',
          legend.background = element_blank(),
          legend.key = element_blank(),
          legend.key.size = unit(0.4,'cm'),
          axis.ticks.y = element_blank(),
          panel.grid.major.x = element_blank())

  p3 <- ggplot(OT.CM,aes(x = .data[['Origen']], y = .data[[i]],
fill=Origen)) +
    geom_boxplot(na.rm=T) +
    scale_fill_manual(values = c('firebrick4','dodgerblue4')) +
    stat_summary(fun = mean , geom = 'point', size = 4,shape = 4,na.rm
= T) +
    scale_y_continuous(position                                    =
'right',labels=percent_format(accuracy = 1),
                       limits = ylim,
                       expand = expansion(mult = c(0,.1),add = 0)) +
    stat_boxplot(geom='errorbar',width=0.3,na.rm=T) +
    xlab(NULL) + ylab(NULL) +
    theme(legend.position = 'none')

  p1 + p2 + p3 + plot_layout(width=c(1,1,1)) +
    plot_annotation(gsub('_',' ',i)) &
    theme(plot.title = element_text(hjust=0.5))

  if (GUARDAR==T){
    ggsave(paste(i,'.png'), path = dir.aux1, plot = last_plot(),scale = 1,
           width = 15,  height = 8, units = 'cm',dpi = 500,  limitsize =
TRUE)
  }

  if (i=='ASVs_Comunes'){
    po.list[[1]] <- p1 + labs(tag = 'A.')
    po.list[[2]] <- p2
    po.list[[3]] <- p3
  } else {
    pl.list[[1]] <- p1 + labs(tag = 'B.')
    pl.list[[2]] <- p2
    pl.list[[3]] <- p3
  }

}

(po.list[[1]] + po.list[[2]] + po.list[[3]])/
  (pl.list[[1]] + pl.list[[2]] + pl.list[[3]])


if (GUARDAR==T){
  ggsave(paste('comunes.png'), path = dir.aux1, plot = last_plot(),scale
= 1.6,
         width = 15,  height = 9, units = 'cm',dpi = 500,  limitsize = TRUE)
}


# Diagramas de Venn (Filo, Clase, Familia, OTU)
flog.threshold(futile.logger::ERROR, name = "VennDiagramLogger")
venn.list <- list()
taxi.aux <- c('Filo','Clase','Familia')

for (i in 1:5){
  ps.t <- Taxa.list.base[[i]]
  ps.l <- subset_samples(ps.t,Origen=='Lecho')
  ps.a <- subset_samples(ps.t,Origen=='Agua')

  # Eliminación de N/As (solo Filo, Clase y Familia)
  if (i<4){
    otu_tab1                                                    <-
otu_table(ps.l)[which(!(tax_table(ps.l)[,taxi.aux[i]]=='N/A')),]
    otu_tab2                                                    <-
otu_table(ps.a)[which(!(tax_table(ps.a)[,taxi.aux[i]]=='N/A')),]
    otu_tab3                                                    <-
otu_table(ps.t)[which(!(tax_table(ps.t)[,taxi.aux[i]]=='N/A')),]
    ntax.l <- sum(apply(otu_tab1,1,sum) > 0)
    ntax.a <- sum(apply(otu_tab2,1,sum) > 0)
    ntaxo <- nrow(otu_tab3)

  } else {
    otu_tab1 <- otu_table(ps.l)
    otu_tab2 <- otu_table(ps.a)
    otu_tab3 <- otu_table(ps.t)
    ntax.l <- sum(apply(otu_tab1,1,sum) > 0)
    ntax.a <- sum(apply(otu_tab2,1,sum) > 0)
    ntaxo <- nrow(otu_tab3)
  }

  venn.list[[1]] <- names(which(apply(otu_tab1,1,min)> 0))
  venn.list[[2]] <- names(which(apply(otu_tab2,1,min)> 0))
  if (length(venn.list[[1]])>=length(venn.list[[2]])){
    cat.name <- c(paste0('Lecho (',ntax.l,')'),
                  paste0('Agua (',ntax.a,')'))
    cat.posi <- c(50,-50)
    cat.fill <- c('tomato','blue')
    cat.cols <- c('brown','lightsteelblue')
    } else {
    cat.name <- c(paste0('Lecho (',ntax.l,')'),
                  paste0('Agua (',ntax.a,')'))
    cat.posi <- c(-140,140)
    cat.fill <- c('tomato','blue')
    cat.cols <- c('brown','lightsteelblue')
    }
  if (GUARDAR==T){
    venn <- venn.diagram(x=venn.list,print.mode = c('raw'),
                         main = paste(TAXA[i],' (',ntaxo,')',sep = ''),
                         main.pos = c(0.5,0.63),
                         lty = 'blank',
                         width = 7.5,
                         height = 4,
                         units = 'cm',
                         resolution = 1000,
                         category.names=cat.name,
                         cat.cex = 0.5,
                         cat.pos = cat.posi, cat.dist=c(0.07,0.07),
                         cex = 1,
                         filename=paste(dir.aux1,'/Venn_2',
                                        TAXA[i],'.png',sep=''),
                         fill = cat.fill,
                         colors = cat.cols,
                         margin = 1.2,
                         disable.logging = T)
  }
}


# OTUs/Lecturas únicas
UNI <- data.frame(matrix(NA , nrow = nrow(OTU.COM), ncol=2))
rownames(UNI) <- OTU.COM[,'Muestra']; colnames(UNI) <-
c('otu.uni','lec.uni')
UNI.L <- apply(count.l,1,function(x) sum(sum(x>0)==1))
UNI.A <- apply(count.a,1,function(x) sum(sum(x>0)==1))

for (i in 1:nrow(UNI)){
  if (i<=9){
    UNI[i,1] <- sum((otu_table(PS.base)[,i]>0)*UNI.L)
    UNI[i,2]                                                    <-
sum((otu_table(PS.base)[,i]>0)*otu_table(PS.base)[,i]*UNI.L)
  }
  UNI[i,1] <- sum((otu_table(PS.base)[,i]>0)*UNI.A)
  UNI[i,2]                                                      <-
sum((otu_table(PS.base)[,i]>0)*otu_table(PS.base)[,i]*UNI.A)
}
```

```
OTU.COM.UNI <- cbind(OTU.COM,UNI)
OTU.COM.UNI$prop.otu.uni <-
OTU.COM.UNI[,9]/OTU.COM.UNI[,3]
OTU.COM.UNI$prop.lec.uni <-
OTU.COM.UNI[,10]/OTU.COM.UNI[,4]

if (GUARDAR==T){

write.csv(OTU.COM.UNI,paste(dir.aux2,'OTU_COM_UNI.csv',sep=
'/'))
}

# Limpieza
rm(OTU.COM, count.l, count.a, OT.CM,venn.list, ps.a, ps.l, cat.name,
   cat.posi, cat.fill, cat.cols, UNI, UNI.L, UNI.A, OTU.COM.UNI)


# Abundancias relativas ----------------------------------------------

# Directorio
dir.aux1 <- paste(dir.abun,'01 - TABLAS',sep = '/')
if (file.exists(dir.aux1)==F){dir.create(dir.aux1)}
dir.aux2 <- paste(dir.abun,'02 - GRÁFICOS',sep = '/')
if (file.exists(dir.aux2)==F){dir.create(dir.aux2)}

# Colores
qual_col_pals = brewer.pal.info[brewer.pal.info$category == 'qual',]
col_vector = unlist(mapply(brewer.pal, qual_col_pals$maxcolors,
                           rownames(qual_col_pals)))
# Otros
TT <- c(10,20,20,20) # Top Taxones

# Abundancias relativas (PS filtro)
top.tax.list <- list()
for (i in 1:4) {
  AUX <- cbind(tax_table(Taxa.list[[i]])[,TAXA[i]],
               data.frame(otu_table(Taxa.list[[i]])))
  colnames(AUX)[1] <- 'TAXA'
  AUX <- aggregate(.~ TAXA,AUX,sum)
  rownames(AUX) <- AUX[,1]; AUX <- AUX[-1]
  AUX[] <- apply(AUX,2,function(x) {x/sum(x)})
  AUX <- AUX[order(apply(AUX,1,mean),decreasing=T),]
  top.tax.list[[i]] <-
rownames(AUX[!(rownames(AUX)=='N/A'),])[1:TT[i]]
  NT <- nrow(AUX[!(rownames(AUX)=='N/A'),])-TT[i]
  AUX <- rbind(AUX[!(rownames(AUX)=='N/A'),][1:TT[i],],

apply(AUX[!(rownames(AUX)=='N/A'),][(TT[i]+1):(nrow(AUX)-
1),],
               2,sum),
             AUX[rownames(AUX)=='N/A',])
  rownames(AUX)[(TT[i]+1)]<-paste('Otros
(',as.character(NT),')',sep='')

  # Archivo CSV
  if (GUARDAR==T){
    write.csv(100*AUX,paste(dir.aux1,'/',TAXA[i],'.csv',sep = ''))
  }

  rownames(AUX)[1:TT[i]] <-
paste('*',rownames(AUX)[1:TT[i]],'*',sep='')
  AUX <- t(AUX)
  orden <- colnames(AUX)
  OSM <-
data.frame(sample_data(PS.base)[,c('Origen','Sitio','Muestra')])
  AUX <- cbind(OSM, AUX)
  AUX$Origen <- factor(AUX$Origen,levels = c('Lecho','Agua'))
  AUX <-
data.frame(complete(AUX,Origen,Sitio,fill=list(value=NA)),
                    check.names = F)
  AUX<- AUX %>%
    gather(Taxa,Abundancia,colnames(AUX)[-c(1,2,3)])
  AUX[,'Taxa'] <- factor(AUX[,'Taxa'],rev(orden))

  # Gráfico
  ggplot(AUX,aes(x=Sitio,y=Abundancia,fill=Taxa,order(orden))) +
    geom_bar(stat='identity',position='stack',color = 'black',na.rm = T)
+
    facet_grid(~Origen,scale="free") +
    scale_fill_manual(values=col_vector) +
    scale_y_continuous(labels=percent, expand = expansion(mult = 0,
add = 0)) +
    xlab('')+ylab('Porcentaje de lecturas') + theme_bw() +
    labs(fill=TAXA[i]) + guides(fill=guide_legend(ncol=1)) +
    theme(text=element_text(size=20),
          legend.text = element_markdown(),
          legend.key = element_rect(color = NA, fill =NA),
          legend.key.size = unit(0.75,'cm'),
          strip.background = element_rect(fill = 'seashell2'))

  if (GUARDAR==T){
    ggsave(paste(TAXA[i],'.png',sep = ''), path = dir.aux2,
           plot = last_plot(), scale = 1.8, width = 20,  height = 12,
           units = 'cm',  dpi = 500,  limitsize = TRUE)
  }

}


# Limpieza
# rm(dir.aux1,dir.aux2,AUX,orden,i)


# Diferencias en abundancias -----------------------------------------

# Directorios
dir.aux1 <- paste(dir.abun,'03 - DIFERENCIAS
ABUNDANCIAS',sep = '/')
if (file.exists(dir.aux1)==F){dir.create(dir.aux1)}
dir.aux2 <- paste(dir.abun,'04 - TABLAS AD',sep = '/')
if (file.exists(dir.aux2)==F){dir.create(dir.aux2)}

# DESeq2/ALDEx2
dsq.list <- list()
ald.list <- list()
for (i in c(1,2,4)){
  # DESeq2
  dsq <- phyloseq_to_deseq2(Taxa.list[[i]],~Origen)
  set.seed(1996)
  dsq <- DESeq(dsq)
  res <- results(dsq, alpha = 0.05)
  res <- data.frame(res)
  res <-
cbind(tax=as.character(tax_table(Taxa.list[[i]])[,TAXA[i]]),res)
  res <- res[!(res$tax=='N/A'),]
  res$log.padj <- -log(res$padj,10)
  dsq.list[[i]] <- res

  # ALDEx2
  tax.cont <- data.frame(otu_table(Taxa.list[[i]]))
  cond <- sample_data(Taxa.list[[i]])$Origen
  set.seed(1996)
  ald <- tax.cont %>%
    aldex.clr(cond,mc.samples=128,verbose=F,useMC = T)
  ald.t <- aldex.ttest(ald, paired.test = F)
  ald.e <- aldex.effect(ald,include.sample.summary = F,verbose=F)

  ald <- data.frame(tax =
as.character(tax_table(Taxa.list[[i]])[,TAXA[i]]),
                    ald.e,ald.t,log.wiep=-log(ald.t$wi.ep,10),
                    log.wieBH=-log(ald.t$wi.eBH,10))
  ald <- ald[!(ald$tax=='N/A'),]
  ald.list[[i]] <- ald
}


# Gráfico top k taxones más abundantes
for (i in c(1,2,4)){
  # DESeq2
  dsq <- dsq.list[[i]]
```

```
  dsq.ab <- dsq[dsq$tax %in% top.tax.list[[i]],]
  dsq.ab <- dsq.ab[order(dsq.ab$log2FoldChange,decreasing=T),]
  dsq.ab$tax <- paste('*',dsq.ab$tax,'*',sep='')
  dsq.ab$tax <- ifelse(dsq.ab$padj<0.01,paste(dsq.ab$tax,' __*__
',sep=''),
                paste(dsq.ab$tax,' ',sep=''))
  dsq.ab$tax <- factor(dsq.ab$tax,levels=dsq.ab$tax)

  xlim <- xylim(-max(abs(dsq.ab$log2FoldChange),na.rm=T),
            max(abs(dsq.ab$log2FoldChange),na.rm=T))

  p1 <- ggplot(dsq.ab,aes(x=log2FoldChange,y=tax,color =
log2FoldChange > 0)) +
    geom_point(size=2) + theme_classic() +
    scale_color_manual(name=NULL, labels = c('más
abundante<br>en agua',
                                'más abundante<br>en lecho'),
                values = c('dodgerblue4','firebrick4')) +
    scale_x_continuous(limits = xlim) +
    geom_vline(xintercept = 0, linetype = 'dashed' , color='black') +
    ylab(NULL) + xlab('Log<sub>2</sub>(FC)') +
    ggtitle(paste('DESeq2 (',TAXA[i],')',sep='')) +
    theme(plot.title = element_text(hjust=0.5),
          axis.title.x = element_markdown(),
          axis.text.y = element_markdown(colour = 'black'),
          panel.grid.major.y =
element_line(color='gray',linetype='dashed'),
          legend.text = element_markdown(),
          legend.position = 'right',
          legend.key = element_blank(),
          legend.background = element_blank(),
          legend.key.height = unit(0.8,'cm'),
          legend.box.background = element_rect(linetype = 'dashed',
                                    color = 'black'))

  # ALDEx2
  ald <- ald.list[[i]]
  ald.ab <- ald[ald$tax %in% top.tax.list[[i]],]
  ald.ab <- ald.ab[order(ald.ab$effect,decreasing=T),]
  ald.ab$tax <- paste('*',ald.ab$tax,'*',sep='')
  ald.ab$tax <- ifelse(ald.ab$wi.eBH<0.05,paste(ald.ab$tax,' __*__
',sep=''),
                paste(ald.ab$tax,' ',sep=''))
  ald.ab$tax <- factor(ald.ab$tax,levels=ald.ab$tax)

  xlim <- xylim(-max(abs(ald.ab$effect),na.rm=T),
            max(abs(ald.ab$effect),na.rm=T))

  p2 <- ggplot(ald.ab,aes(x=effect,y=tax,color = effect > 0)) +
    geom_point(size=2) + theme_classic() + scale_x_continuous(limits
= xlim) +
    scale_color_manual(name=NULL, labels = c('más
abundante<br>en agua',
                                'más abundante<br>en lecho'),
                values = c('dodgerblue4','firebrick4')) +
    geom_vline(xintercept = 0, linetype = 'dashed' , color='black') +
    xlab('Efecto') + ylab(NULL) +
    ggtitle(paste('ALDEx2 (',TAXA[i],')',sep='')) +
    theme(plot.title = element_text(hjust=0.5),
          axis.text.y = element_markdown(colour = 'black'),
          panel.grid.major.y =
element_line(color='gray',linetype='dashed'),
          legend.text = element_markdown(),
          legend.position = 'right',
          legend.key = element_blank(),
          legend.background = element_blank(),
          legend.key.height = unit(0.8,'cm'),
          legend.box.background = element_rect(linetype = 'dashed',
                                    color = 'black'))

  # Gráfica compuesta
  p3 <- p1 + p2 + plot_layout(guides = 'collect') &
    theme(legend.text = element_markdown(),
          legend.position = 'right',
          legend.key = element_blank(),
          legend.background = element_blank(),
          legend.key.height = unit(0.8,'cm'),
          legend.box.background = element_rect(linetype = 'dashed',
                                    color = 'black'))

  if (GUARDAR==T){
    ggsave(paste('toptax_',TAXA[i],'_DESeq2.png',sep=''), path =
dir.aux1,
           plot = p1, scale = 1.7,  width = 15, height = 0.3*TT[i] + 2,
           units = 'cm',  dpi = 500,  limitsize = TRUE)
    ggsave(paste('toptax_',TAXA[i],'_ALDEx2.png',sep=''), path =
dir.aux1 ,
           plot = p2, scale = 1.7,  width = 15, height = 0.3*TT[i] + 2,
           units = 'cm',  dpi = 500,  limitsize = TRUE)
    ggsave(paste('toptax_',TAXA[i],'.png',sep=''), path = dir.aux1,
           plot = p3, scale = 1.7,  width = 15, height = 0.3*TT[i] + 2,
           units = 'cm',  dpi = 500,  limitsize = TRUE)
  }

}

# Tablas
for (i in c(1,2,4)){
  ald.tab <- ald.list[[i]][,c("tax","effect",'wi.ep','wi.eBH')]
  dsq.tab <- dsq.list[[i]][,c('tax','log2FoldChange','pvalue','padj')]
  dif.ab.tab <- inner_join(ald.tab, dsq.tab, by = 'tax')

  if (GUARDAR == T){
    write.csv(dif.ab.tab,paste(dir.aux2,'/',TAXA[i],'_diftab.csv',sep =
''))

  }

}


# Gráficos abundancia/tabla taxones diferentes ----------------------------
-

# Directorio
dir.auxi <- paste(dir.abun,'05 - GRÁFICOS_TABLAS', sep = '/')
if (file.exists(dir.auxi)==F){dir.create(dir.auxi)}

TT <- c(10,20,20,20)

for (i in c(1,2,4)){
  # Objeto phyloseq
  ps <- Taxa.list[[i]]

  # Data frame abundancias relativas por taxón
  tax.cont <- cbind(tax_table(ps)[,TAXA[i]],
data.frame(otu_table(ps)))
  colnames(tax.cont)[1] <- 'TAXA'
  tax.cont <- aggregate(.~ TAXA,tax.cont,sum)
  tax.cont[-1] <- apply(tax.cont[-1],2,function(x) {x/sum(x)})
  tax.cont <- tax.cont[order(apply(tax.cont[-
1],1,mean),decreasing=T),]
  tax.cont <- tax.cont[!(tax.cont$TAXA=='N/A'),]
  rownames(tax.cont) <- tax.cont[,1]
  tax.cont <- data.frame(tax.cont[-1])
  orden <- rownames(tax.cont)
  tax.cont <- t(tax.cont)
  colnames(tax.cont) <- orden
  tax.cont <- tax.cont[,1:TT[i]]
  orden <- rev(orden[1:TT[i]]) %>%
    factor(.,levels = .)
  OSM <- data.frame(sample_data(ps)[,c('Origen','Sitio','Muestra')])
  tax.cont <- data.frame(OSM, tax.cont, check.names = F)
  tax.cont <- tax.cont %>%
    gather(Taxa,Abundancia,colnames(tax.cont)[-c(1,2,3)])
  tax.cont[,'Taxa'] <- factor(tax.cont[,'Taxa'],orden)


  # Gráfico abundancias relativas: promedio/desv. est.
```

```
  p1 <- ggplot(tax.cont,aes(x = Abundancia, y = Taxa, color = Origen))
+
    geom_hline(yintercept = seq(1,TT[i]-1) + 0.5, color = 'lightgray') +
    geom_hline(yintercept = TT[i] + 0.5, color = 'black', linewidth =
0.5) +
    scale_y_discrete(expand = expansion(mult = 0,add = 0.5)) +
    scale_x_continuous(label = percent) +
    stat_summary(fun.data = mean_se, geom = 'pointrange',
                 position = position_dodge(width = 0.4)) +
    scale_color_manual(name = NULL,
                       breaks = c('Lecho','Agua'),
                       labels = c('Lecho','Agua'),
                       values = c('Firebrick4','Dodgerblue4')) +
    labs(y = NULL, x = NULL) + ggtitle('Abundancia relativa') +
    theme_classic() +
    theme(plot.title =  element_text(hjust = 0.5, vjust = -8,
                                     size = 10, face = 'bold'),
          axis.text.y = element_text(hjust = 0.5,
                                     color = 'black',
                                     face = 'italic'),
          legend.position = c(0.8,0.3),
          legend.margin = margin(.1,.4,.1,.2, unit = 'cm'),
          legend.background = element_blank(),
          legend.box.background = element_rect(linetype = 'dashed',
                                               color = 'black'),
          panel.grid.major.x = element_line(color = 'gray',
                                            linetype = 1),
          panel.grid.minor.x = element_line(color = 'gray',
                                            linetype = 2))

  if (i > 1){
    # Gráfico auxiliar para leyenda inferior (colores)
    if (i==2){
      df.tax <- distinct(data.frame(tax_table(ps)[,c('Filo',TAXA[i])]))
    } else {
      df.tax <- distinct(data.frame(tax_table(ps)[,c('Clase',TAXA[i])]))
    }

    colnames(df.tax) <- c('A','B')
    df.tax <- df.tax[match(rev(orden),df.tax$B),]
    df.tax$B <- factor(df.tax$B,orden)
    df.tax$C <- as.numeric(df.tax$B)

    set.seed(1996)
    pal                                                              <-
sample(as.character(c(paletteer_d("colorBlindness::PairedColor12Ste
ps"),
                          'firebrick4','dodgerblue4')),
                nrow(distinct(df.tax['A'])))

    p0 <- ggplot(df.tax, aes(x=0,  color = A)) +
      geom_tile(aes(x = 0, y = C, width = 0.1, fill = A),
                color = 'black', size = 0.7) +
      scale_fill_manual(values = pal) +
      scale_y_discrete(limits = rev(df.tax$B),
                       expand = expansion(mult = 0,add = 0.5)) +
      scale_x_continuous(limits = c(-0.05,0.05),
                         expand = expansion(mult = 0, add = 0))
    if (i == 2){
      p0 <- p0 + guides(fill = guide_legend(nrow = 2, title = 'Filo',
                                            title.position = 'top'))
    } else {
      p0 <- p0 + guides(fill = guide_legend(nrow = 2, title = 'Clase',
                                            title.position = 'top'))
    }

    p0 <- p0 + theme(panel.background = element_rect(fill = 'white'),
                     axis.title.x = element_blank(),
                     axis.text.x = element_blank(),
                     axis.ticks.x = element_blank(),
                     axis.text.y = element_text(hjust = 0.5,
                                                color = 'black',
                                                face = 'italic',
                                                size = 9),
                     axis.title.y = element_blank(),
                     axis.ticks.y = element_blank(),
                     legend.position = 'bottom',
                     legend.text = element_markdown(face = 'italic',
                                                    size = 8),
                     legend.title = element_text(hjust = 0.5),
                     legend.key = element_rect(color = 'black'),
                     legend.key.height = unit(0.01,'cm'),
                     legend.background = element_blank(),
                     legend.box.background = element_rect(linetype = 1,
                                                          color = 'gray50'),
                     plot.margin = unit(c(0,0,0,0),'cm'))

    lgnd <- get_plot_component(p0,'guide-box-bottom')
    p0 <- p0 + theme(legend.position = 'none')

  }


  # Tabla DESeq2/ ALDEx2 (para graficar)
  dsq.tab <- dsq.list[[i]][dsq.list[[i]]$tax %in% orden,]
  dsq.tab <- data.frame(tax = dsq.tab$tax,
                        dsq = format(dsq.tab$log2FoldChange,
                                     digits = 2, nsmall = 1),
                        dsq.p = ifelse(dsq.tab$padj < 0.01,'*',' '))
  ald.tab <- ald.list[[i]][ald.list[[i]]$tax %in% orden,]
  ald.tab <- data.frame(tax =ald.tab$tax,
                        ald = format(ald.tab$effect,
                                     digits = 1, nsmall = 1),
                        ald.p = ifelse(ald.tab$wi.eBH < 0.05,'*',' '))
  tab <- left_join(data.frame(tax = rev(orden)), dsq.tab, by = 'tax')
  tab <- left_join(tab, ald.tab, by = 'tax')
  tab$sig <- ifelse(tab$dsq.p == '*' & tab$ald.p == '*','*','')
  tab <- data.frame(tab[1], abv = abbreviate(tab$tax, minlength = 3),
                    y = rev(seq(1,TT[i])), tab[-1])
  tab$fon <- ifelse(tab$sig == '*', 'bold','plain')
  tab$col <- ifelse(tab$sig == '*', 'black','gray40')
  tab$f1 <- ifelse(tab$dsq.p == '*','bold','plain')
  tab$c1 <- ifelse(tab$dsq.p == ' ','gray40',
                   ifelse(as.numeric(tab$dsq) < 0,'dodgerblue4', 'firebrick4'))
  tab$f2 <- ifelse(tab$ald.p == '*','bold','plain')
  tab$c2 <- ifelse(tab$ald.p == ' ','gray40',
                   ifelse(as.numeric(tab$ald) < 0,'dodgerblue4', 'firebrick4'))
  rownames(tab) <- NULL

  # Gráfica de tabla
  p2 <- ggplot(tab) +
    scale_y_discrete(limits = rev(tab$abv),
                     expand = expansion(mult = 0,add = 0.5)) +
    scale_x_continuous(limits = c(1,4), position = 'top',
                       breaks = c(1.5,2.5,3.5),
                       labels = c('**DESeq2**<br>*log<sub>2</sub>(FC)*',
                                  '**ALDEx2**<br>*Efecto*',
                                  '**Sig**')) +
    geom_text(data = tab, aes(x = 1.5, y = y, label = dsq,
                              colour = c1, fontface = f1),
              size = 3.5) +
    geom_text(data = tab, aes(x = 2.5, y = y, label = ald,
                              colour = c2, fontface = f2),
              size = 3.5) +
    geom_point(aes(x = 3.5, y = y, fill = sig),
               size = 3.5, shape = 21, color = 'black') +
    scale_color_manual(values = c('dodgerblue4','firebrick4','gray40',
                                  'dodgerblue4','firebrick4','gray40')) +
    scale_fill_manual(values = c('gray80','darkgreen'))

  # Líneas horizontales
  for (k in 0:TT[i]){
    linea <- data.frame(x=c(1,4), y = k + 0.5)
    p2 <- p2 + geom_line(data = linea, aes(x = x, y = y),
                         inherit.aes = F, color = 'gray30')
  }

  p2 <- p2 + guides(color = 'none', fill = 'none') +
    theme(panel.background = element_rect(fill = 'white'),
          axis.title.x = element_blank(),
```

```
      axis.text.x.top = element_markdown(size = 9,
                                         color = 'black',
                                         hjust = 0.5,
                                         vjust = 0.5),
      axis.ticks.x = element_blank(),
      axis.text.y = element_markdown(hjust = 0.5,
                                     face = rev(tab$fon),
                                     color = rev(tab$col)),
      axis.title.y = element_blank(),
      axis.ticks.y = element_blank(),
      legend.position = 'none')

 # Gráfica compuesta
 if (i == 1){
  p3 <- p1 + p2 + plot_layout(width = c(3,2))
 } else {
  p1 <- p1 + theme(axis.text.y = element_blank())
  p3 <- (p0 + p1 + p2 + plot_layout(width = c(0.1,3,2)))/lgnd

  if (i==2){
   p3 <- p3 + plot_layout(heights = c(9,1))
  } else {
   p3 <- p3 + plot_layout(heights = c(7,1))
  }
 }


 if (GUARDAR==T){
  ggsave(paste(TAXA[i],'.png',sep = ''), path = dir.auxi,
         plot = p3, scale = 1.5, width = 15,  height = TT[i]/2,
         units = 'cm',  dpi = 500,  limitsize = TRUE)
 }
}


# Diversidad Beta (PCoA, PCA, PERMANOVA) ------------------------
----------

# Directorios
dir.aux1 <- paste(dir.beta,'01 - TABLAS',sep = '/')
if (file.exists(dir.aux1)==F){dir.create(dir.aux1)}
dir.aux2 <- paste(dir.beta,'02 - PCoA (fltr)',sep = '/')
if (file.exists(dir.aux2)==F){dir.create(dir.aux2)}
dir.aux3 <- paste(dir.beta,'03 - PCoA (VST)',sep = '/')
if (file.exists(dir.aux3)==F){dir.create(dir.aux3)}
dir.aux4 <- paste(dir.beta,'04 - PCA (COM)',sep = '/')
if (file.exists(dir.aux4)==F){dir.create(dir.aux4)}
dir.aux5 <- paste(dir.beta,'05 - VST+COM',sep = '/')
if (file.exists(dir.aux5)==F){dir.create(dir.aux5)}


# taxonomía - abundancias (auxiliar)
tax.ab.list <- list()
for (i in 1:4){
 ps <- Taxa.list[[i]]
 tax.ab <- data.frame(tax =
data.frame(tax_table(ps)[,TAXA[i]])[,,drop = T],
                      data.frame(otu_table(ps)) %>%
                       apply(.,2, function(x){x/sum(x)}))
 tax.ab <- tax.ab[!(tax.ab$tax=='N/A'),]
 tax.ab <- tax.ab[order(apply(tax.ab[-1],1,mean),decreasing =
T),][1:TT[i],]
 rownames(tax.ab) <- tax.ab[,1]
 tax.ab <- t(tax.ab[-1])
 tax.ab <- data.frame(Muestra = rownames(tax.ab), tax.ab,
check.names = F) %>%
   pivot_longer(-Muestra, names_to = 'tax',  values_to = 'rel_ab')
 tax.ab.list[[i]] <- tax.ab
}

# PCoA (solo filtrado)
for (i in c(1,2,5)){
 # Ordenación
 PSr <- transform_sample_counts(Taxa.list[[i]],function(x)
{x/sum(x)})
 ord <- ordinate(PSr,'PCoA','bray')
 p.ord <- plot_ordination(PSr,ord)
 p.ord$data$Origen <- as.factor(p.ord$data$Origen)

 xy.lim <- max(abs(p.ord$data$Axis.1),abs(p.ord$data$Axis.2),na.rm
= T)
 xy.lim <- xylim(-xy.lim,xy.lim)

 # Gráfica
 p <- ggplot(p.ord$data,aes(x=Axis.1,y=Axis.2)) +
  theme_bw() + coord_fixed() +
  geom_hline(yintercept = 0, color = 'red', linetype = 'dashed') +
  geom_vline(xintercept = 0, color = 'red', linetype = 'dashed') +
  geom_point(aes(fill = Origen, shape = Origen),size = 3.5,color =
'black') +
  scale_color_manual(values = c('dodgerblue4','firebrick4')) +
  scale_shape_manual(values = c(21,24)) +
  geom_text_repel(aes(label=p.ord$data$Muestra), size = 3,
                  min.segment.length = 0.1, force = 10, force_pull = 4) +
  scale_x_continuous(limits = xy.lim) +
  scale_y_continuous(limits = xy.lim) +
  geom_mark_ellipse(aes(fill = Origen), linetype = 2, color = 'gray30',
                    expand = 0.05,alpha = 0.1, show.legend = F) +
  scale_fill_manual(values=c('dodgerblue4','firebrick4')) +
  ggtitle(paste('PCoA ','(', TAXA[i],')',sep='')) +
  xlab(gsub('Axis.1 ','PCoA 1',p.ord$labels$x)) +
  ylab(gsub('Axis.2 ','PCoA 2',p.ord$labels$y)) +
  theme(plot.title=element_text(hjust=0.5),
        legend.key = element_blank())

 if (GUARDAR==T){
  ggsave(paste('PCoA_fltr_',TAXA[i],'.png',sep=''), plot = last_plot(),
         path = dir.aux2, scale = 1.2,  width = 20,  height = 16,
         units = 'cm',  dpi = 500,  limitsize = TRUE)
 }

}

# PCoA (fltr + Normalización VST - DESeq2)
ps.vst.list <- list()
pcoa.vst.list <- list()
for (i in c(1,2,5)){
 ps <- Taxa.list[[i]]
 dsq <- phyloseq_to_deseq2(ps,~Origen)
 set.seed(1996)
 dsq.vst <- t(assay(varianceStabilizingTransformation(dsq)))
 dsq.vst[dsq.vst < 0] <- 0
 otu_table(ps) <- t(otu_table(dsq.vst,F))
 ps.vst.list[[i]] <- ps

 pcoa <- plot_ordination(ps,ordinate(ps,'PCoA','bray'))
 pcoa$data$Origen <- as.factor(pcoa$data$Origen)

 xy.lim <- max(abs(pcoa$data$Axis.1),abs(pcoa$data$Axis.2),na.rm
= T)
 xy.lim <- xylim(-2.2*xy.lim, 2.2*xy.lim)

 p <- ggplot(pcoa$data,aes(x=Axis.1,y=Axis.2)) + theme_bw() +
coord_fixed() +
  geom_hline(yintercept = 0, color = 'red', linetype = 'dashed') +
  geom_vline(xintercept = 0, color = 'red', linetype = 'dashed') +
  geom_mark_ellipse(aes(fill = Origen), color = 'gray60', linetype =
2,
                    alpha = 0.05, show.legend = F, inherit.aes = T) +
  scale_fill_manual(values=c('dodgerblue4','firebrick4')) +
  geom_point(aes(fill = Origen, shape = Origen),
             color = 'black', size = 2.5) +
  scale_shape_manual(values = c(21, 24)) +
  geom_text_repel(aes(label = pcoa$data$Muestra), size = 2.5, color
= 'black',
                  force = 0.1, force_pull = 0.2, min.segment.length = 1) +
  scale_x_continuous(limits = xy.lim,
                     expand = expansion(mult = 0, add = 0)) +
  scale_y_continuous(limits = xy.lim,
                     expand = expansion(mult = 0, add = 0)) +
```

```
  scale_color_manual(values=c('dodgerblue4','firebrick4')) +
  ggtitle(paste0('PCoA ','(', TAXA[i],')')) +
  xlab(gsub('Axis.1 ','PCoA 1',pcoa$labels$x)) +
  ylab(gsub('Axis.2 ','PCoA 2',pcoa$labels$y)) +
  theme(plot.title=element_text(hjust=0.5))

 if (GUARDAR==T){
  ggsave(paste0('PCoA_vst_',TAXA[i],'.png'),
        plot = last_plot(), path = dir.aux3, scale = 1.2,  width = 20,
        height = 16,  units = 'cm',  dpi = 500,  limitsize = TRUE)
 }


 # Correlación taxones-ejes x e y
 if (!(TAXA[i] == 'ASV')){
  K.esc <- abs(xy.lim[2]/1.5)
  cor_x<- inner_join(pcoa$data[,c('Muestra','Axis.1','Axis.2')],
                tax.ab.list[[i]], by = 'Muestra') %>%
    group_by(tax) %>%
    nest(data = -tax) %>%
    mutate(cor.x = map(data,~cor.test(.x$rel_ab, .x$Axis.1,
                                method = 'spearman',
                                alternative = 'two.sided',
                                exact = F) %>%
                     tidy())) %>%
    unnest(cor.x) %>%
    mutate(estimate = estimate * K.esc) %>%
    dplyr::select(tax, estimate, p.value) %>%
    rename(c('estimate' = 'rho', 'p.value'='pval'))

  cor_y<- inner_join(pcoa$data[,c('Muestra','Axis.1','Axis.2')],
                tax.ab.list[[i]], by = 'Muestra') %>%
    group_by(tax) %>%
    nest(data = -tax) %>%
    mutate(cor.y = map(data,~cor.test(.x$rel_ab, .x$Axis.2,
                                method = 'spearman',
                                alternative = 'two.sided',
                                exact = F) %>%
                     tidy())) %>%
    unnest(cor.y) %>%
    mutate(estimate = estimate * K.esc) %>%
    dplyr::select(tax, estimate, p.value) %>%
    rename(c('estimate' = 'rho', 'p.value'='pval'))

  cor.tab <- inner_join(cor_x, cor_y, by = 'tax') %>%
    mutate(abv = abbreviate(tax, minlength = 3)) %>%
    filter(abs(rho.x) > 0.8*K.esc & pval.x < 0.01 |
           abs(rho.y) > 0.8*K.esc & pval.y < 0.01)

  p <- p + geom_segment(data = cor.tab,
                    aes(x = 0, xend = rho.x, y = 0, yend = rho.y),
                    color = 'darkgreen', linejoin = 'round',
                    arrow = arrow(type = 'closed',
                                  length = unit(0.01,'npc')),
                    inherit.aes = F) +
    geom_text_repel(data = cor.tab, aes(x = 1.3*rho.x, y = 1.3*rho.y,
                                  label = tax, fontface = 'italic'),
              color = 'darkgreen', force_pull = 1, min.segment.length = Inf,
              force = .001, size = 2.8, max.overlaps = 30, max.time = 1) +
    annotate('text',x = 0.95*xy.lim[2], y = 0.95*xy.lim[2],
             label = paste0('Ke = ',round(K.esc, digits = 2)),
             hjust = 1, vjust = 1)

  if (GUARDAR==T){
    ggsave(paste0('PCoA_vst_',TAXA[i],'_vec.png'),
          plot = last_plot(), path = dir.aux3, scale = 1.2,  width = 20,
          height = 16,  units = 'cm',  dpi = 500,  limitsize = TRUE)
  }

 }
 pcoa.vst.list[[i]] <- p + ggtitle('PCoA (tradicional)') +
   theme(legend.position = 'none')
}

# PCA (fltr + Composicional)
ps.com.list <- list()
pca.com.list <- list()
for (i in c(1,2,5)){
  # Objeto phyloseq
  ps <- Taxa.list[[i]]

  # Reemplazo de ceros (zCompositions - cmultRepl() requiere
muestras en filas)
  otu.comp <- data.frame(otu_table(ps))
  set.seed(1996)
  otu.comp  <-  t(cmultRepl(t(otu.comp),method='CZM',  output='p-
counts'))
  otu.comp <- apply(otu.comp,2,function(x) {log(x) - mean(log(x))})

  # Nuevo objeto phyloseq
  otu_table(ps) <- otu_table(otu.comp, taxa_are_rows = T)
  ps.com.list[[i]] <- ps

  # PCA
  pca <- rda(t(otu.comp), scale = F) # rda() requiere muestras en filas
  pca.data <- data.frame(Origen = sample_data(ps)$Origen,
                scores(pca, scaling = 'sites')$sites)

  var.pca                                                    <-
as.character(round(pca$CA$eig[1:2]/sum(pca$CA$eig)*100,digits  =
1))
  var.pca <- paste0('[',var.pca,' %',']')

  xy.lim <- max(abs(pca.data$PC1),abs(pca.data$PC2),na.rm = T)
  xy.lim <- xylim(-2.2*xy.lim, 2.2*xy.lim)

  p <- ggplot(pca.data,aes(x = PC1,y = PC2, color = Origen)) +
    theme_bw() + coord_fixed() +
    geom_hline(yintercept = 0, color = 'red' , linetype = 'dashed') +
    geom_vline(xintercept = 0, color = 'red' , linetype = 'dashed') +
    geom_mark_ellipse(aes(fill = Origen), color = 'gray60', linetype =
2,
                      alpha = 0.05, show.legend = F) +
    scale_fill_manual(values = c('dodgerblue4','firebrick4')) +
    geom_point(aes(shape = Origen, fill = Origen),
               color = 'black', size = 2.5) +
    scale_shape_manual(values = c(21, 24)) +
    scale_color_manual(values=c('dodgerblue4','firebrick4')) +
    scale_x_continuous(limits = xy.lim,
                       expand = expansion(mult = 0, add = 0)) +
    scale_y_continuous(limits = xy.lim,
                       expand = expansion(mult = 0, add = 0)) +
    geom_text_repel(aes(label = rownames(pca.data)), size = 2.5,
                    color = 'black', force_pull = 0.2, force = 0.1,
                    min.segment.length = 1) +
    ggtitle(paste0('PCA (',TAXA[i],')')) +
    xlab(paste0('PCA1 ',var.pca[1])) +
    ylab(paste0('PCA2 ',var.pca[2])) +
    theme(plot.title=element_text(hjust=0.5))

  if (GUARDAR==T) {
    ggsave(paste0('PCA_',TAXA[i],'.png'), path = dir.aux4,
          plot = last_plot(), scale = 1, width = 20,  height = 16,
          units = 'cm', dpi = 500,  limitsize = TRUE)
  }


  # Correlación taxones-ejes x e y
  if (!(TAXA[i] == 'ASV')){
    K.esc <- abs(xy.lim[2]/1.5)
    pca.data  <-  data.frame(Muestra  =  rownames(pca.data),
pca.data[,2:3])
    cor_x<- inner_join(pca.data[,c('Muestra','PC1','PC2')],
                  tax.ab.list[[i]], by = 'Muestra') %>%
      group_by(tax) %>%
      nest(data = -tax) %>%
      mutate(cor.x = map(data,~cor.test(.x$rel_ab, .x$PC1,
                                  method = 'spearman',
                                  alternative = 'two.sided',
                                  exact = F) %>%
```

```
                    tidy())) %>%
    unnest(cor.x) %>%
    mutate(estimate = estimate * K.esc) %>%
    dplyr::select(tax, estimate, p.value) %>%
    rename(c('estimate' = 'rho', 'p.value'='pval'))

  cor_y<- inner_join(pca.data[,c('Muestra','PC1','PC2')],
                     tax.ab.list[[i]], by = 'Muestra') %>%
    group_by(tax) %>%
    nest(data = -tax) %>%
    mutate(cor.y = map(data,~cor.test(.x$rel_ab, .x$PC2,
                                      method = 'spearman',
                                      alternative = 'two.sided',
                                      exact = F) %>%
                       tidy())) %>%
    unnest(cor.y) %>%
    mutate(estimate = estimate * K.esc) %>%
    dplyr::select(tax, estimate, p.value) %>%
    rename(c('estimate' = 'rho', 'p.value'='pval'))

  cor.tab <- inner_join(cor_x, cor_y, by = 'tax') %>%
    mutate(abv = abbreviate(tax, minlength = 3)) %>%
    filter(abs(rho.x) > 0.8*K.esc & pval.x < 0.01 |
           abs(rho.y) > 0.8*K.esc & pval.y < 0.01)

  p <- p + geom_segment(data = cor.tab,
                        aes(x = 0, xend = rho.x, y = 0, yend = rho.y),
                        color = 'darkgreen', linejoin = 'round',
                        arrow = arrow(type = 'closed',
                                      length = unit(0.01,'npc')),
                        inherit.aes = F) +
    geom_text_repel(data = cor.tab, aes(x = 1.3*rho.x, y = 1.3*rho.y,
                                        label = tax, fontface = 'italic'),
              color = 'darkgreen', min.segment.length = Inf, force_pull =
1,
              force = 0.001, size = 2.8, max.overlaps = 30, max.time = 1)
+
    annotate('text',x = 0.95*xy.lim[2], y = 0.95*xy.lim[2],
             label = paste0('Ke = ',round(K.esc, digits = 2)),
             hjust = 1, vjust = 1)

  if (GUARDAR==T) {
    ggsave(paste0('PCA_',TAXA[i],'_vec.png'), path = dir.aux4,
           plot = last_plot(), scale = 1, width = 20,  height = 16,
           units = 'cm', dpi = 500,  limitsize = TRUE)
  }

 }

 pca.com.list[[i]] <- p + ggtitle('PCA (composicional)')

}

# Gráficas compuestas (VST/COMP)
for (i in c(1,2,5)){
 pcoa.vst.list[[i]] + pca.com.list[[i]] +
   plot_annotation(tag_levels = 'A', tag_suffix = '.') +
   plot_layout(guides = 'collect')

 if (GUARDAR==T){
   ggsave(paste0('vstcomp_', TAXA[i], '.png'), path = dir.aux5,
          plot = last_plot(), scale = 1.9, width = 15, height = 7.5,
          units = 'cm', dpi = 500, limitsize = T)
 }

}

# PERMANOVA y ANOSIM (VST y COMP)
ite1 <- T
for (i in c(1,2,5)){
 # Vector origen
 cond <- data.frame(sample_data(PS)[,'Origen',drop = F])
 cond$Origen <- factor(cond$Origen, levels = c('Lecho','Agua'))

 # Gráfica ordenación tradicional (VST + PCoA-BC)
 otu.trad <- t(otu_table(ps.vst.list[[i]]))
 set.seed(1996)
 pnova.trad <- adonis2(otu.trad ~ Origen,
                       cond,permutations = 999, method = 'bray')
 ansm.trad <- anosim(otu.trad, cond$Origen,
                     permutations = 999, distance = 'bray')

 # Gráfica ordenación composicional (zcomp + PCA-Aitchison)
 otu.comp <- t(otu_table(ps.com.list[[i]]))
 set.seed(1996)
 pnova.comp <- adonis2(otu.comp ~ Origen, cond,
                       permutations = 999, method = 'euclidian')
 ansm.comp <- anosim(otu.comp, cond$Origen,
                     permutations = 999, distance = 'euclidian')

 # Tabla PERMANOVA
 pnova <- data.frame(tax = TAXA[i],
                     F.trad = pnova.trad$F[1],
                     R2.trad = pnova.trad$R2[1],
                     p.trad = pnova.trad$`Pr(>F)`[1],
                     F.comp = pnova.comp$F[1],
                     R2.comp = pnova.comp$R2[1],
                     p.comp = pnova.comp$`Pr(>F)`[1])

 # Tabla ANOSIM
 ansm <- data.frame(tax = TAXA[i],
                    R.trad = ansm.trad$statistic,
                    p.trad = ansm.trad$signif,
                    R.comp = ansm.comp$statistic,
                    p.comp = ansm.comp$signif)

 if (ite1 == T){
  pnova.tab <- pnova
  ansm.tab <- ansm
  ite1 <- F
 } else {
  pnova.tab <- rbind(pnova.tab,pnova)
  ansm.tab <- rbind(ansm.tab,ansm)
 }

}

if (GUARDAR==T){
 write.csv(pnova.tab,paste0(dir.aux1,'/PERMANOVA.csv'))
 write.csv(ansm.tab,paste0(dir.aux1,'/ANOSIM.csv'))
}


# Diversidad beta subconjuntos lecho y agua (VST y com)
ord.vst.list <- list()
eig.vst.list <- list()
otu.vst.list <- list()
ord.com.list <- list()
eig.com.list <- list()
otu.com.list <- list()
for (i in c(1,2,5)){
 ttls <- data.frame(vst = c('PCoA1','PCoA2','PCoA'),
                    com = c('PCA1','PCA2','PCA'))
 col.ori <- data.frame(Lecho = 'firebrick4', Agua = 'dodgerblue4')
 for (q in c('vst','com')){
  # Directorio
  if (q=='vst'){
   dir.aux6 <- paste(dir.beta,'06 - VST agua - lecho',sep = '/')
  } else {
   dir.aux6 <- paste(dir.beta,'07 - COM agua - lecho',sep = '/')
  }

  if (file.exists(dir.aux6)==F){dir.create(dir.aux6)}

  ord.list <- list()
  eig.list <- list()
  otu.list <- list()
  p.list <- list()
  for (k in c('Lecho','Agua')) {
   ps <- subset_samples(Taxa.list[[i]], Origen == k) %>%
```

```
      filter_taxa(function(x) sum(x) > 0, prune = T)
    if (q=='vst') {
     # Aplicación de VST
     dsq <- phyloseq_to_deseq2(ps,~1)
     set.seed(1996)
     dsq.vst <- t(assay(varianceStabilizingTransformation(dsq)))
     dsq.vst[dsq.vst < 0] <- 0
     otu.list[[k]] <- dsq.vst
     ord <- dsq.vst %>%
      vegdist() %>%
      wcmdscale(eig = T, k = 2)
     eig <- ord$eig
     ord.data <- data.frame(Muestra = rownames(ord$points),
                            XX = ord$points[,1], YY = ord$points[,2])
     eig.list[[k]] <- eig
     ord.list[[k]] <- ord.data
    } else {
     # Tratamiento de ceros
     otu.tab <- data.frame(otu_table(ps))
     set.seed(1996)
     otu.tab <- t(cmultRepl(t(otu.tab),method='CZM', output='p-
counts'))
     otu.tab <- apply(otu.tab,2,function(x) {log(x) - mean(log(x))})
     otu.list[[k]] <- t(otu.tab)
     ord <- t(otu.tab) %>%
      rda(scale = F)
     eig <- ord$CA$eig
     ord.data <- data.frame(Muestra = rownames(ord$CA$u),
                            XX = ord$CA$u[,1], YY = ord$CA$u[,2])
     eig.list[[k]] <- eig
     ord.list[[k]] <- ord.data
    }

    # Límites Gráfica de ordenación
    ord.lim <- max(abs(ord.data[,2]), abs(ord.data[,3]))
    xy.lim <- xylim(-2*ord.lim, 2*ord.lim)

    if (!(TAXA[i]=='ASV')) {
     # Tabla abundancias relativas de taxones más abundantes
     tax.ab <- data.frame(tax = data.frame(tax_table(ps)[,
                                           TAXA[i]])[,,drop = T],
                          data.frame(otu_table(ps)) %>%
                           apply(.,2, function(x){x/sum(x)}))
     tax.ab <- tax.ab[!(tax.ab$tax=='N/A'),]
     tax.ab <- tax.ab[order(apply(tax.ab[-1],1,mean),
                            decreasing = T),][1:TT[i],]
     rownames(tax.ab) <- tax.ab[,1]
     tax.ab <- t(tax.ab[-1])
     tax.ab <- data.frame(Muestra = rownames(tax.ab),
                          tax.ab, check.names = F) %>%
      pivot_longer(-Muestra, names_to = 'tax', values_to = 'rel_ab')

     # Correlación taxones - ejes
     K.esc <- xy.lim[2]*0.8
     cor_x<- inner_join(ord.data, tax.ab, by = 'Muestra') %>%
      group_by(tax) %>%
      nest(data = -tax) %>%
      mutate(cor.x = map(data,~cor.test(.x$rel_ab, .x$XX,
                                        method = 'spearman',
                                        alternative = 'two.sided',
                                        exact = F) %>%
                           tidy())) %>%
      unnest(cor.x) %>%
      mutate(estimate = estimate * K.esc) %>%
      dplyr::select(tax, estimate, p.value) %>%
      rename(c('estimate' = 'rho', 'p.value'='pval'))

     cor_y<- inner_join(ord.data, tax.ab, by = 'Muestra') %>%
      group_by(tax) %>%
      nest(data = -tax) %>%
      mutate(cor.y = map(data,~cor.test(.x$rel_ab, .x$YY,
                                        method = 'spearman',
                                        alternative = 'two.sided',
                                        exact = F) %>%
                           tidy())) %>%
      unnest(cor.y) %>%
      mutate(estimate = estimate * K.esc) %>%
      dplyr::select(tax, estimate, p.value) %>%
      rename(c('estimate' = 'rho', 'p.value'='pval'))

     cor.tab <- inner_join(cor_x, cor_y, by = 'tax') %>%
      mutate(abv = abbreviate(tax, minlength = 3)) %>%
      filter(abs(rho.x) > 0.8*K.esc & pval.x < 0.01 |
             abs(rho.y) > 0.8*K.esc & pval.y < 0.01)
    }

    var.eje <- as.character(round(eig[1:2]/sum(eig)*100, digits = 2))
    var.eje <- paste0('[',var.eje,' %]')

    p <- ggplot(ord.data,aes(x = XX, y = YY)) +
     theme_bw() + coord_fixed() +
     geom_hline(yintercept = 0, color = 'red', linetype = 2) +
     geom_vline(xintercept = 0, color = 'red', linetype = 2) +
     geom_point(size = 3.5, shape = 21, fill = col.ori[1,k],
                color = 'black') +
     geom_text_repel(aes(label = Muestra), size = 3, color = 'black') +
     scale_x_continuous(limits = xy.lim,
                        expand = expansion(mult = 0 , add = 0)) +
     scale_y_continuous(limits = xy.lim,
                        expand = expansion(mult = 0 , add = 0)) +
     xlab(paste(ttls[q][1,],var.eje[1])) +
     ylab(paste(ttls[q][2,],var.eje[2])) +
     ggtitle(paste(ttls[q][3,],k)) +
     theme(plot.title = element_text(hjust = 0.5, vjust = 0))

    if (nrow(cor.tab) > 0 & !(TAXA[i]=='ASV')){
     p <- p + geom_segment(data = cor.tab,
                           aes(x = 0, xend = rho.x, y = 0, yend = rho.y),
                           color = 'darkgreen', linejoin = 'round',
                           arrow = arrow(type = 'closed',
                                         length = unit(0.01,'npc')),
                           inherit.aes = F) +
      geom_text_repel(data = cor.tab, aes(x = 1.2*rho.x, y =
1.2*rho.y,
                                          label = tax, fontface = 'italic'),
                      color = 'darkgreen', min.segment.length = Inf,
                      force_pull = 1, force = 0.001, size = 2.5,
                      max.overlaps = 50, max.time = 2) +
      annotate('text',x = 0.95*xy.lim[2], y = 0.95*xy.lim[2],
               label = paste0('Ke = ',round(K.esc, digits = 2)),
               hjust = 1, vjust = 1)
    }

    p.list[[k]] <- p

   }

   if(q == 'vst'){" }
   p.list$Lecho + p.list$Agua + plot_layout(width = c(1,1)) +
    plot_annotation(title = ifelse(q=='vst',
                                   paste0('Enfoque tradicional (',
                                          TAXA[i],')'),
                                   paste0('Enfoque composicional (',
                                          TAXA[i],')')),
                    tag_levels = 'A', tag_suffix = '.') &
    theme(plot.title = element_text(hjust = 0.5),
          plot.tag.position = c(0,0.98))

   if (GUARDAR == T){
    fn <- paste0(ttls[3,q],'_',TAXA[i],'_',q,'.png')
    ggsave(fn, plot = last_plot(), path = dir.aux6, scale = 2, width =
15,
           height = 8, units = 'cm', dpi = 500, limitsize = TRUE)
   }

   if (q == 'vst') {
    ord.vst.list[[i]] <- ord.list
    eig.vst.list[[i]] <- eig.list
    otu.vst.list[[i]] <- otu.list
   } else {
```

```
    ord.com.list[[i]] <- ord.list
    eig.com.list[[i]] <- eig.list
    otu.com.list[[i]] <- otu.list
   }

 }

}


# Variables ambientales ----------------------------------------------------

# Directorio
dir.aux1 <- paste(dir.vamb,'01 - TABLAS',sep = '/')
if (file.exists(dir.aux1)==F){dir.create(dir.aux1)}
dir.aux2 <- paste(dir.vamb,'02 - GRÁFICOS',sep = '/')
if (file.exists(dir.aux2)==F){dir.create(dir.aux2)}
dir.aux2.1 <- paste(dir.aux2,'01 - ALTITUD VS',sep = '/')
if (file.exists(dir.aux2.1)==F){dir.create(dir.aux2.1)}
dir.aux2.2 <- paste(dir.aux2,'02 - LATITUD VS',sep = '/')
if (file.exists(dir.aux2.2)==F){dir.create(dir.aux2.2)}

# Colores
qual_col_pals = brewer.pal.info[brewer.pal.info$category == 'qual',]
col_vector = unlist(mapply(brewer.pal, qual_col_pals$maxcolors,
                           rownames(qual_col_pals)))

# Previo (creación df variables ambientales)
var.data <- sample_data(PS) %>%
  data.frame() %>%
  dplyr::select(Sitio,starts_with('V')) %>%
  distinct()
var.data <- var.data[order(var.data$Sitio),]
rownames(var.data) <- var.data$Sitio
var.data$Sitio <- NULL

var.cam <- paste0('V',c(seq(15),seq(29,36)))
var.dga <- paste0('V',c(seq(5),seq(16,36)))

# Correlogramas
for (k in c('var.cam','var.dga')){
  # Nombre variables
  var <- var.data[,get(k)]

  # Matriz base
  CC.v <- matrix(NA, nrow = length(var), ncol = length(var))
  rownames(CC.v) <- auxivar[get(k),'Variable']
  colnames(CC.v) <- auxivar[get(k),'Variable']
  CC.p <- CC.v

  for (i in 1:length(get(k))){
    for (j in 1:length(get(k))){

      ct <- cor.test(var[,i],var[,j],'pearson',
                     alternative = 'two.sided', na.rm = T)
      CC.v[i,j] <- ct$estimate
      CC.p[i,j] <- ct$p.value
    }
  }

  # Gráfica
  png(paste0(dir.aux2,'/cor_',gsub('var.','',k),'.png'), width = 15,
      height = 15, units = 'cm', res = 800)
  corrplot(CC.v, p.mat = CC.p, sig.level = 0.01, insig = 'blank',
           type = 'lower', tl.col='black', tl.cex = 0.8, tl.srt = 45)
  dev.off()

}


# Altitud vs otras variables
alt.data <- cbind(Sitio = rownames(var.data), var.data)
for (i in colnames(var.data)){
  if (!(i=='V1')){
    ggplot(alt.data, aes(x = .data[['V2']], y = .data[[i]])) + theme_bw() +
      geom_point(color = 'red', na.rm = T) +
      geom_text_repel(aes(label = Sitio), na.rm = T) +
      xlab(auxivar['V1','var.uni']) +
      ylab(auxivar[i,'var.uni']) +
      scale_y_continuous(expand = expansion(mult = c(0.05,0.2))) +
      geom_smooth(method = 'lm', formula = y~x, se = F, na.rm = T,
                  linetype = 'solid', color = 'black')+
      ggtitle(paste(auxivar[i,'Variable'],
                    auxivar['V1','Variable'],sep = ' v/s ')) +
      stat_poly_eq(formula =y~x, eq.with.lhs = "italic(hat(y))~`=`~",
                   aes(label = paste(..eq.label.., ..adj.rr.label..,
                                     sep = "*plain(\",\")~")),
                   parse = TRUE, label.x=0.95,label.y = 0.98, na.rm = T) +
      theme(plot.title = element_text(hjust = 0.5),
            axis.title.x = element_markdown(),
            axis.title.y = element_markdown())

    if (GUARDAR == T) {
      ggsave(paste('Altitud_',gsub('/','div',auxivar[i,'Variable'])
                   ,'.png',sep = ''),
             plot = last_plot(), path = dir.aux2.1, scale = 1.5, width = 15,
             height = 10, units = 'cm', dpi = 500, limitsize = T)
    }

  }

}


# Latitud vs otras variables
lat.data <- cbind(Sitio=rownames(var.data),var.data)
for (i in colnames(var.data)){
  if (!(i=='V2')){
    ggplot(lat.data,aes( x = .data[['V2']], y = .data[[i]])) + theme_bw() +
      geom_point(color = 'red',na.rm = T) +
      geom_text_repel(aes(label = Sitio), na.rm = T) +
      xlab(auxivar['V2','var.uni']) +
      ylab(auxivar[i,'var.uni']) +
      scale_y_continuous(expand = expansion(mult = c(0.05,0.2))) +
      geom_smooth(method = 'lm', formula = y~x, se = F, na.rm = T,
                  linetype = 'solid', color = 'black')+
      ggtitle(paste(auxivar[i,'Variable'],
                    auxivar['V2','Variable'],sep = ' v/s ')) +
      stat_poly_eq(formula =y~x, eq.with.lhs = "italic(hat(y))~`=`~",
                   aes(label = paste(..eq.label.., ..adj.rr.label..,
                                     sep = "*plain(\",\")~")),
                   parse = TRUE, label.x=0.95,label.y = 0.98, na.rm = T) +
      theme(plot.title = element_text(hjust = 0.5),
            axis.title.x = element_markdown(),
            axis.title.y = element_markdown())

    if (GUARDAR == T) {
      ggsave(paste('Latitud_',gsub('/','div',auxivar[i,'Variable'])
                   ,'.png',sep = ''),
             plot = last_plot(), path = dir.aux2.2, scale = 1.5, width = 15,
             height = 10, units = 'cm', dpi = 500, limitsize = T)
    }

  }

}


# Comparación DGA-Terreno
var.nom <- c('pH','T','CE','OD','P','N','DQO')
var.sit <- rownames(var.data)
var.cam <- var.data[,c('V6','V7','V8','V9','V11','V13','V15')]
var.dga <- var.data[,c('V16','V17','V18','V19','V22','V21','V20')]

ite1=0
for (i in 1:7){
  auxtab <- data.frame(PM = var.sit, Variable = rep(var.nom[i],9),
                       cam = var.cam[,i], DGA = var.dga[,i])
  ct.p <- cor.test(var.cam[,i], var.dga[,i], 'pearson',
                   alternative = 'two.sided', na.rm = T)
```

```
  ct.s <- cor.test(var.cam[,i], var.dga[,i],'spearman',
                   alternative = 'two.sided', exact = F, na.rm = T)
  auxrow <- data.frame(Variable = var.nom[i], pearson = ct.p$estimate,
             p.p.val = ct.p$p.value, spearman = ct.s$estimate,
             s.p.val = ct.s$p.value)

  if (ite1<1){
    cam.dga.dif <- auxtab
    cam.dga.cor <- auxrow
    ite1 <- 1
  } else{
    cam.dga.dif <- rbind(cam.dga.dif,auxtab)
    cam.dga.cor <- rbind(cam.dga.cor,auxrow)
  }

}

rownames(cam.dga.cor) <- NULL
cam.dga.dif$dif <- (cam.dga.dif$DGA -
cam.dga.dif$cam)/cam.dga.dif$DGA
cam.dga.dif$Variable <- factor(cam.dga.dif$Variable,var.nom)

ggplot(cam.dga.dif, aes(x=PM,y=dif, fill=PM)) + theme_bw() +
  geom_bar(stat = 'identity', position = 'dodge', width = 1, color =
'black',
           na.rm = T) +
  geom_hline(yintercept = c(-0.5,0.5), color = 'red', linetype = 'dashed')
+
  geom_hline(yintercept = c(-1,1), color = 'black', linetype = 'dashed')
+
  geom_hline(yintercept = 0, color = 'black', linetype = 'solid') +
  facet_grid(~Variable) + xlab(NULL) + ylab(NULL) +
  scale_y_continuous(labels = percent, limits = c(-1.5,1.5), n.breaks =
6,
                     expand = expansion(mult=0,add=0)) +
  scale_fill_manual(values = col_vector) +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank(),
        strip.background = element_rect(fill='lightsteelblue2'),
        panel.spacing.x = unit(0.1,'lines'))

ggsave('diferencia_cam_dga.png', last_plot(), path = dir.aux2, scale =
1.4,
       width = 15, height = 8, units = 'cm', dpi = 200, limitsize = T)


# PCA usos de suelo

# Dataframe Usos de suelos
df.usos <- var.data[paste0('V',seq(29,36))]

# Dataframe Usos de suelo (con transformación CLR)
df.usos.clr <-  df.usos %>%
  apply(.,2,function(x) {x/sum(x)}) %>%
  apply(.,2,function(x) {log(x) - mean(log(x))})

# Dendrograma
dd.usos <- df.usos.clr %>%
  dist(.) %>%
  hclust(.,method = 'ward.D2') %>%
  dendro_data(.)

clus.cat <- as.character(cutree(hclust(dist(df.usos.clr),
                                method = 'ward.D2'),k=4))
df.usos.clr <- data.frame(clus = clus.cat, df.usos.clr)

dd.usos$labels$clus <- df.usos.clr[,1,drop = F][dd.usos$labels$label,]
dd.usos$labels$y <- -2

p1 <- ggplot(segment(dd.usos)) +
  geom_segment(aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_point(data = dd.usos$labels,aes(x=x,y=y-0.1,fill = clus, shape
= clus),
             color = 'black', size = 5) +
  scale_fill_manual(values = c('seagreen','orchid',
                               'royalblue4','gold2')) +
  scale_shape_manual(values = c(21,22,23,24)) +
  coord_flip() +
  scale_y_reverse(expand = c(0.2, 0)) +
  theme_dendro() +
  labs(tag = 'A.') +
  theme(plot.margin = unit(c(0,0,0,1),'cm'),
        legend.position = 'none')


# Gráfico de barras
bar.usos <- cbind(Sitio = rownames(df.usos),df.usos/100)
bar.usos$Sitio <- factor(bar.usos$Sitio,levels = dd.usos$labels$label)
colnames(bar.usos)[-1] <- auxivar[colnames(df.usos),'Variable']
orden <- colnames(bar.usos[-1])[order(apply(bar.usos[-1],2,mean),
                                      decreasing = T)]
bar.usos <- bar.usos %>%
  gather(Uso,Porcentaje,colnames(bar.usos)[-1])
bar.usos[,'Uso'] <- factor(bar.usos[,'Uso'],orden)

c.vec <- c('seagreen','khaki','thistle4','plum2',
           'brown4','gray','cyan2','magenta2')

p2 <-
ggplot(bar.usos,aes(x=Sitio,y=Porcentaje,fill=Uso,order(orden))) +
  geom_bar(stat='identity',position='stack',color = 'black',na.rm = T) +
  scale_fill_manual(values = c.vec) +
  scale_x_discrete(expand = expansion(mult = 0, add = 0)) +
  scale_y_continuous(labels=percent, expand = expansion(mult = 0,
add = 0),
                     position = 'left') +
  xlab('')+ylab('Usos de suelo') +
  ggtitle('Porcentajes usos de suelo') +
  theme_bw() +
  guides(fill=guide_legend(nrow=2, byrow = T)) +
  coord_flip() +
  theme(plot.title = element_text(hjust = 0.5, vjust = 6, size = 20),
        legend.text = element_markdown(size = 13),
        legend.title = element_blank(),
        legend.position = 'bottom',
        axis.title.x = element_text(size = 15),
        axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 20),
        panel.border = element_rect(color = 'black', fill = NA, linewidth
= 1),
        plot.margin = unit(c(1,1,1,0),'cm'))


# Gráfico de ordenación (PCA)
pca <- rda(df.usos.clr[,-1],scale=F)
SC <- data.frame(clus = df.usos.clr[,1],scores(pca, scaling =
'sites')$sites)

var.pca <-
as.character(round(pca$CA$eig[1:2]/sum(pca$CA$eig)*100,digits=2
))
var.pca <- paste('[',var.pca,' %',']',sep='')

xy.lim <- max(abs(SC$PC1),abs(SC$PC2),na.rm = T)
xy.lim <- xylim(-xy.lim,xy.lim)

p3 <- ggplot(SC,aes(x=PC1,y=PC2,fill = clus, shape = clus)) +
theme_bw() +
  geom_hline(yintercept = 0, color = 'red' , linetype = 'dashed') +
  geom_vline(xintercept = 0, color = 'red' , linetype = 'dashed') +
  geom_point(size = 5, color = 'black') +
  scale_fill_manual(values = c('seagreen','orchid',
                               'royalblue4','gold2')) +
  scale_shape_manual(values = c(21,22,23,24)) +
  geom_text_repel(aes(label=rownames(SC)),size=5,force_pull =
2,force=0.5) +
  xlim(xy.lim) +
  scale_y_continuous(limits = xy.lim, position = 'right') +
  xlab(paste('PCA1 ',var.pca[1],sep='')) +
  ylab(paste('PCA2 ',var.pca[2],sep='')) +
  ggtitle('PCA (composicional)') +
```

```
  coord_fixed() + labs(tag = 'B.') +
  theme(legend.position = 'none',
        plot.title=element_text(hjust = 0.5, vjust = 6, size = 20),
        axis.title.x = element_text(size = 15),
        axis.text.x = element_text(size = 12),
        axis.title.y = element_text(size = 15),
        axis.text.y = element_text(size = 12),
        panel.border = element_rect(color = 'black', fill = NA, linewidth
= 1),
        plot.margin = unit(c(0,1,0,0),'cm'))

p1 + p2 + p3 + plot_layout(width = c(1,8,6))

if (GUARDAR == T){
  ggsave('dendro_bar.png', last_plot(), path = dir.aux2, scale = 2.8,
         width = 15, height = 7, units = 'cm', dpi = 500, limitsize = T)
}


# Variables ambientales y diversidad alfa (regresión lineal) ----------
--

# Directorio
dir.aux1 <- paste0(dir.alfv,'/01 - TABLAS')
if (file.exists(dir.aux1)==F){dir.create(dir.aux1)}

# Auxiliar
sc.var <- paste0('V',c(1,2,4,10,16,18,20,21,22,23,27))

col.ori <- data.frame(Lecho = 'firebrick4', Agua = 'dodgerblue4')

div.alfa <- data.frame(Shannon = c('Indice Shannon - Wiener','Valor
Indice'),
                       Simpson = c('Indice Simpson','Valor Indice'),
                       Chao1 = c('Riqueza estimada (Chao1)',
                                 'Número estimado de especies'),
                       row.names = c('titulo','etiqueta'))

for (o in c('Lecho','Agua')){
  # sub directorio
  if (o == 'Lecho'){
    dir.aux2 <- paste0(dir.alfv,'/02 - GRÁFICOS LECHO')
  } else {
    dir.aux2 <- paste0(dir.alfv,'/03 - GRÁFICOS AGUA')
  }
  if (file.exists(dir.aux2)==F){dir.create(dir.aux2)}

  ps <- subset_samples(PS, Origen == o)

  df.div <- estimate_richness(ps,measures =
c('Chao1','Shannon','Simpson')) %>%
    mutate(Muestra = rownames(.)) %>%
    dplyr::select(Muestra, Chao1, Shannon, Simpson)

  df.var <- sample_data(ps) %>%
    data.frame() %>%
    dplyr::select(Muestra, all_of(sc.var))

  df.divar <- inner_join(df.div, df.var, by = 'Muestra')

  ite1 <- T
  for(v in colnames(df.var)[-1]) {
    for(d in colnames(df.div)[-1]) {
      # Test Correlación Spearman
      ct.s <- cor.test(df.divar[,v], df.divar[,d], method = 'spearman',
                       alternative ='two.sided',exact=F,na.rm = T)

      # Test Correlación Pearson
      ct.p <- cor.test(df.divar[,v], df.divar[,d], method = 'pearson',
                       alternative ='two.sided',na.rm = T)

      # regresión Lineal
      rl <- lm(as.formula(paste(d,v,sep=' ~ ')), df.divar)
      s.rl <- summary(rl)
      a.rl <- anova(rl)

      # Fila f,d
      fila <- data.frame(Origen = o, var = auxivar[v,'Variable'], div = d,
                         sp.r = ct.s$estimate, sp.p = ct.s$p.value,
                         pe.r = ct.p$estimate, pe.p = ct.p$p.value,
                         B0 = s.rl$coefficients[1,1],
                         B0.p = s.rl$coefficients[1,4],
                         B1 = s.rl$coefficients[2,1],
                         B1.p = s.rl$coefficients[2,4],
                         r2 = s.rl$r.squared,
                         r2.adj = max(s.rl$adj.r.squared,0),
                         rl.F = a.rl$`F value`[1],
                         rl.p = a.rl$`Pr(>F)`[1])

      # Gráfico
      p <- ggplot(df.divar, aes(x = .data[[v]], y = .data[[d]])) +
theme_bw() +
        geom_point(color = col.ori[,o], na.rm = T) +
        geom_text_repel(aes(label = Muestra), size = 4, na.rm = T) +
        geom_smooth(method='lm', formula = y ~ x, se=F, na.rm = T,
                    linetype='solid', color='black', linewidth = 0.5) +
        scale_y_continuous(expand = expansion(mult = c(0.05,0.2))) +
        xlab(auxivar[v,'var.uni']) + ylab(div.alfa['etiqueta',d]) +
        ggtitle(paste(div.alfa['titulo',d],
                      auxivar[v,'Variable'], sep = ' v/s ')) +
        theme(plot.title = element_text(hjust = 0.5),
              axis.title.x = element_markdown()) +
        stat_poly_eq(formula =y~x, eq.with.lhs = "italic(hat(y))~`=`~",
                     aes(label = paste(..eq.label.., ..adj.rr.label..,
                                       sep = "*plain(\",\")~")),
                     parse = TRUE, label.x=0.95,label.y = 0.98, na.rm = T)

      if (GUARDAR==T){
        if(abs(fila$pe.r) > 0.7 & fila$pe.p < 0.05 |
           abs(fila$sp.r) > 0.7 & fila$sp.p < 0.05 ){
          ggsave(paste0('ÚTIL_',d,'~',
                        gsub('/','_',auxivar[v,'Variable']),'.png'),
                 plot = p, path = dir.aux2, scale = 1.8,
                 width = 7.5, height = 6, units = 'cm', dpi = 500,
                 limitsize = TRUE)
        } else {
          ggsave(paste0('DESC_',d,'~',
                        gsub('/','_',auxivar[v,'Variable']),'.png'),
                 plot = p, path = dir.aux2, scale = 1.8,
                 width = 7.5, height = 6, units = 'cm', dpi = 500,
                 limitsize = TRUE)
        }
      }

      if (ite1 == T){
        vardiv <- fila
        ite1 <- F
      } else {
        vardiv <- rbind(vardiv, fila)
      }

    }

  }

  vardiv <- data.frame(vardiv[1:5],
                       sp.padj = p.adjust(vardiv$sp.p,'BH'), vardiv[6:7],
                       pe.padj = p.adjust(vardiv$pe.p,'BH'), vardiv[8:15])

  if (o == 'Lecho'){
    vardiv.tab <- vardiv
  } else {
    vardiv.tab <- rbind(vardiv.tab, vardiv)
  }

}

if (GUARDAR==T){
  write.csv(vardiv.tab, paste0(dir.aux1,'/vardiv.csv'))
}
```

```
vardiv.tab$var <- gsub(' DGA','',vardiv.tab$var)
vardiv.tab$div <- paste0(vardiv.tab$div,' (',
                         substr(vardiv.tab$Origen,1,1),')')

sp.rho <- vardiv.tab %>%
  dplyr::select(var, div, sp.r) %>%
  pivot_wider(names_from = var, values_from = sp.r) %>%
  column_to_rownames(var = 'div') %>%
  as.matrix()

sp.pval <- vardiv.tab %>%
  dplyr::select(var, div, sp.p) %>%
  pivot_wider(names_from = var, values_from = sp.p) %>%
  column_to_rownames(var = 'div') %>%
  as.matrix()

sp.col <- ifelse(sp.pval < 0.05, 'black', '#1C00ff00')

pe.rho <- vardiv.tab %>%
  dplyr::select(var, div, pe.r) %>%
  pivot_wider(names_from = var, values_from = pe.r) %>%
  column_to_rownames(var = 'div') %>%
  as.matrix()

pe.pval <- vardiv.tab %>%
  dplyr::select(var, div,pe.p) %>%
  pivot_wider(names_from = var, values_from = pe.p) %>%
  column_to_rownames(var = 'div') %>%
  as.matrix()

pe.col <- ifelse(pe.pval < 0.05, 'black', '#1C00ff00')

if (GUARDAR == T){
  # Gráfica Spearman
  png(paste(dir.alfv,'/spearman_div_alfa.png',sep=''),
      width = 15, height = 8, units = 'cm', res = 800)
  corrplot(sp.rho, p.mat = sp.pval, sig.level = 0.01, insig = 'pch',
           pch.cex = 4, pch.col = 'white', addCoef.col = sp.col,
           number.cex = 0.5, tl.col='black', tl.cex=0.8, tl.srt = 45)
  dev.off()


  # Gráfica Pearson
  png(paste(dir.alfv,'/pearson_div_alfa.png',sep=''),
      width = 15, height = 8, units = 'cm', res = 800)
  corrplot(pe.rho,  p.mat = pe.pval, sig.level = 0.01, insig = 'pch',
           pch.cex = 4, pch.col = 'white', addCoef.col = pe.col,
           number.cex = 0.5, tl.col='black', tl.cex=0.8, tl.srt = 45)
  dev.off()
}


# Variables ambientales y diversidad beta (Gráficos ordenación) ------
-----

sc.var <- paste0('V',c(1,2,4,10,16,18,20,21,22,23,27))

for (i in c(1,2,5)){
  # Directorios
  dir.auxi <- paste0(dir.betv,'/0',i,' - ',TAXA[i])
  if (file.exists(dir.auxi)==F){dir.create(dir.auxi)}
  dir.aux1 <- paste0(dir.auxi,'/01 - TABLAS')
  if (file.exists(dir.aux1)==F){dir.create(dir.aux1)}

  ttls <- data.frame(vst = c('PCoA1','PCoA2','PCoA'),
                     com = c('PCA1','PCA2','PCA'))
  col.ori <- data.frame(Lecho = 'firebrick4', Agua = 'dodgerblue4')

  for(q in c('vst','com')){
    var.data.list <- list()
    # Directorio
    if(q == 'vst'){
      dir.aux2 <- paste0(dir.auxi,'/02 - GRÁFICOS TRADICIONAL')
    } else {
      dir.aux2 <- paste0(dir.auxi,'/03 - GRÁFICOS
COMPOSICIONAL')
    }
    if (file.exists(dir.aux2)==F){dir.create(dir.aux2)}

    # Sets de variables
    for (k in c('Lecho','Agua')){
      var.data.list[[k]] <- Taxa.list[[i]] %>%
        subset_samples(Origen == k) %>%
        sample_data() %>%
        data.frame() %>%
        dplyr::select(Muestra, all_of(sc.var))
    }

    if (q == 'vst') {
      ord.list <- ord.vst.list[[i]]
      eig.list <- eig.vst.list[[i]]
    } else{
      ord.list <- ord.com.list[[i]]
      eig.list <- eig.com.list[[i]]
    }

    # Correlación variables - ejes x e y
    p.vec.list <- list()
    cor.tab.list <- list()
    for (k in c('Lecho', 'Agua')){
      # dataframe coordenadas PCoA + variables
      ord.var <- inner_join(var.data.list[[k]],
                            ord.list[[k]],
                            by = 'Muestra') %>%
        pivot_longer(-c('Muestra','XX','YY'),
                     names_to = 'var',
                     values_to = 'val') %>%
        group_by('var') %>%
        nest(data = -var)

      # Correlación eje x
      cor_x <- ord.var %>%
        mutate(cor.x = map(data,~cor.test(.x$val, .x$XX,
                                          method = 'spearman',
                                          alternative = 'two.sided',
                                          na.rm = T,
                                          exact = F) %>%
                             tidy())) %>%
        unnest(cor.x) %>%
        dplyr::select(var, estimate, p.value) %>%
        rename('estimate' = 'rho', 'p.value' = 'pval')

      # Correlación eje y
      cor_y <- ord.var %>%
        mutate(cor.y = map(data,~cor.test(.x$val, .x$YY,
                                          method = 'spearman',
                                          alternative = 'two.sided',
                                          na.rm = T,
                                          exact = F) %>%
                             tidy())) %>%
        unnest(cor.y) %>%
        dplyr::select(var, estimate, p.value) %>%
        rename('estimate' = 'rho', 'p.value' = 'pval')

      # DF correlaciones con ejes
      cor.tab <- inner_join(cor_x,cor_y, by = 'var') %>%
        filter(abs(rho.x) > 0.7 & pval.x < 0.01 |
                 abs(rho.y) > 0.7 & pval.y < 0.01)

      # Límites Gráfica de ordenación
      ord.lim <- max(abs(ord.list[[k]][,2]), abs(ord.list[[k]][,3]))
      xy.lim <- xylim(-1.2*ord.lim, 1.2*ord.lim)

      # Escalamiento de vectores Correlación (según límites Gráfica
ordenación)
      K.esc <- xy.lim[2]*0.8
      cor.tab <- cor.tab %>%
        mutate(rho.x = K.esc*rho.x, rho.y = K.esc*rho.y) %>%
        mutate(comp = gsub('<br>','\n',auxivar[var,'coraxis'])) %>%
```

```
      mutate(var = auxivar[var,'Variable'])

    # PCoA
    ord.data <- ord.list[[k]]

    # Variabilidad explicada por cada eje
    eig <- eig.list[[k]]
    var.eje <- as.character(round(eig[1:2]/sum(eig)*100, digits = 2))
    var.eje <- paste0('[',var.eje,' %]')

    # Gráfica de ordenación + vectores-Correlación variables ambientales
    p <- ggplot(ord.data,aes(x = XX,y = YY)) +
     theme_bw() + coord_fixed() +
     geom_hline(yintercept = 0, color = 'red', linetype = 2) +
     geom_vline(xintercept = 0, color = 'red', linetype = 2) +
     geom_point(size = 3.5, shape = 21, fill = col.ori[1,k], color = 'black') +
     geom_text_repel(aes(label = Muestra), size = 3, color = 'black') +
     scale_x_continuous(limits = xy.lim,
                        expand = expansion(mult = 0 , add = 0)) +
     scale_y_continuous(limits = xy.lim,
                        expand = expansion(mult = 0 , add = 0)) +
     xlab(paste(ttls[q][1,],var.eje[1])) +
     ylab(paste(ttls[q][2,],var.eje[2])) +
     annotate('text',x = 0.95*xy.lim[2], y = 0.95*xy.lim[2],
              label = paste0('Ke = ',round(K.esc, digits = 2)),
              hjust = 1, vjust = 1) +
     ggtitle(paste(ttls[q][3,],k)) +
     theme(plot.title = element_text(hjust = 0.5, vjust = 0))

    if (nrow(cor.tab) > 0){
     p <- p + geom_segment(data = cor.tab,
                           aes(x = 0, xend = rho.x, y = 0, yend = rho.y),
                           color = 'darkgreen', linejoin = 'round',
                           arrow = arrow(type = 'closed',
                                         length = unit(0.01,'npc')),
                           inherit.aes = F) +
       geom_text_repel(data = cor.tab, aes(x = 1.2*rho.x, y = 1.2*rho.y,
                                           label = comp, fontface = 'bold'),
                       color = 'darkgreen', min.segment.length = Inf,
                       force_pull = 1, force = 0.01, size = 2.5,
                       max.overlaps = 50, max.time = 2)

    }

    cor.tab.list[[k]] <- cor.tab
    p.vec.list[[k]] <- p

  }

  p.vec.list$Lecho + p.vec.list$Agua + plot_layout(width = c(1,1)) +
   plot_annotation(tag_levels = 'A', tag_suffix = '.') &
   theme(plot.title = element_text(hjust = 0.5),
         plot.tag.position = c(0,0.98))

  if (GUARDAR==T){
   ggsave('cor_var_axis.png', plot = last_plot(), path = dir.aux2,
          scale = 1.5, width = 20,  height = 10, units = 'cm',
          dpi = 500, limitsize = TRUE)
  }


  # Gráficos gradiente variables
  col.grad <- data.frame(Lecho = c('cyan','red2'),
                         Agua = c('gold','blue'))
  for (t in sc.var){
   p.list <- list()
   vlim.l <- xylim(min(var.data.list$Lecho[,t], na.rm = T),
                   max(var.data.list$Lecho[,t],na.rm = T))
   vlim.a <- xylim(min(var.data.list$Agua[,t], na.rm = T),
                   max(var.data.list$Agua[,t],na.rm = T))

    vlim <- data.frame(Lecho = vlim.l, Agua = vlim.a)

  for (k in c('Lecho','Agua')){
   set.seed(1996)
   ord.data <- data.frame(ord.list[[k]],
                          var = var.data.list[[k]][,t])
   ord.lim <- max(abs(ord.list[[k]][,2]), abs(ord.list[[k]][,3]))
   xy.lim <- xylim(-1.2*ord.lim, 1.2*ord.lim)

   if (any(is.na(ord.data[,'var']))){
    ord.data <- ord.data[-which(is.na(ord.data[,'var'])),]
   }

   # Variabilidad asociada a cada eje
   eig <- eig.list[[k]]
   var.eje <- as.character(round(eig[1:2]/sum(eig)*100, digits = 2))
   var.eje <- paste0('[',var.eje,' %',']')

   p <- ggplot(ord.data,aes( x = XX, y = YY,
                             fill = var)) +
    geom_hline(yintercept = 0, color = 'red', linetype = 2) +
    geom_vline(xintercept = 0, color = 'red', linetype = 2) +
    geom_point( color = 'black',shape = 21, na.rm = T, size = 4) +
    scale_fill_gradient(low = col.grad[1,k], high = col.grad[2,k],
                        na.value='black',
                        limits = vlim[,k],
                        breaks = vlim[,k],
                        labels = vlim[,k],
                        guide = guide_colorbar(title = auxivar[t,'var.uni'],
                                               title.position = 'top',
                                               title.hjust = 0.5,
                                               label = T,
                                               draw.ulim = T,
                                               draw.llim = T,
                                               frame.colour = 'black',
                                               ticks = T,
                                               ticks.colour = 'black',
                                               barwidth = 15,
                                               barheight = 0.5,
                                               label.position = 'bottom',
                                               direction = 'horizontal')) +
    guides(size = 'none') +
    geom_text_repel( aes(label = Muestra), size = 3 ,color = 'black') +
    scale_x_continuous(limits = xy.lim,
                       expand = expansion(mult = 0 , add = 0)) +
    scale_y_continuous(limits = xy.lim,
                       expand = expansion(mult = 0 , add = 0)) +
    xlab(paste(ttls[q][1,],var.eje[1])) +
    ylab(paste(ttls[q][2,],var.eje[2])) +
    ggtitle(paste(ttls[q][3,],k)) + theme_bw() + coord_fixed() +
    theme(plot.title = element_text(hjust = 0.5),
          legend.position = 'bottom')

   cor.tab <- cor.tab.list[[k]]

   if(any(auxivar[t,'Variable'] == cor.tab[,1])){
    nn <- which(cor.tab[,1] == auxivar[t,'Variable'])
    p <- p + geom_segment(data = cor.tab[nn,],
                          aes(x = 0, xend = rho.x, y = 0, yend = rho.y),
                          color = 'darkgreen', linejoin = 'round',
                          arrow = arrow(type = 'closed',
                                        length = unit(0.01,'npc')),
                          inherit.aes = F) +
     geom_text_repel(data = cor.tab[nn,],
                     aes(x = 1.2*rho.x, y = 1.2*rho.y,
                         label = comp, fontface = 'bold'),
                     color = 'darkgreen', min.segment.length = Inf,
                     force_pull = 1, force = 0.001, size = 2.5,
                     max.overlaps = 50, max.time = 2,
                     inherit.aes = F) +
    annotate('text', x = 0.95*xy.lim[2], 0.95*xy.lim[2],
             label = paste0('Ke = ', round(K.esc, digits = 2)),
             hjust = 1, vjust = 1, size = 3.5)
   }
```

```
      p.list[[k]] <- p

    }

    p.list$Lecho + p.list$Agua + plot_layout(width = c(1,1)) +
     plot_annotation(title = paste0('Variable ambiental: ',
                                    auxivar[t,'Variable']),
                     tag_levels = 'A', tag_suffix = '.') &
     theme(plot.title = element_text(hjust = 0.5),
           plot.tag.position = c(0,0.98))

    if (GUARDAR == T){
     ggsave(paste0('Gradiente_',gsub('/','div',
                                     auxivar[t,'Variable']),'.png'),
            plot = last_plot(), path = dir.aux2, scale = 1.4,
            width = 20, height = 10, units = 'cm', dpi = 500, limitsize =
T)

    }

   }

   # Prueba Mantel
   Mantel.list <- list()
   for (k in c('Lecho','Agua')){
    # Matriz disimilitud
    if(q == 'vst'){
     DD.com <-  vegdist(otu.vst.list[[i]][[k]])
    } else {
     DD.com <-  dist(otu.com.list[[i]][[k]])
    }

    ite1 <- T
    for (t in sc.var) {
     man.var <- var.data.list[[k]][,t,drop = F]
     DD.var <- dist(man.var,'euclidian')
     DD.com2 <- DD.com
     if (is.na(sum(man.var))==T){
      eli.var <- which(is.na(man.var))
      DD.var <- as.dist(as(DD.var,'matrix')[-eli.var,-eli.var])
      DD.com2 <- as.dist(as(DD.com2,'matrix')[-eli.var,-eli.var])
     }
     set.seed(1996)
     Mt <- mantel(DD.com2, DD.var, permutations = 999, method =
'spearman')
     Mantel <- data.frame(Variable = auxivar[t,'Variable'],
                          Estadistico = Mt$statistic,
                          p.val = Mt$signif)

     if (ite1 == T) {
      Mantel.tab <- Mantel
      ite1 <- F
     } else {
      Mantel.tab <- rbind(Mantel.tab,Mantel)
     }
    }

    Mantel.tab$p.adj <- p.adjust(Mantel.tab$p.val,'BH')
    Mantel.tab$sig <- ifelse((Mantel.tab$p.adj < 0.05 |
                              Mantel.tab$p.val < 0.01),'*','')
    Mantel.list[[k]] <- Mantel.tab

    if (GUARDAR == T) {
     if(q == 'vst'){
      fn <- paste('/Mantel_VST_',k,'.csv')
     } else {
      fn <- paste('/Mantel_COMP_',k,'.csv')}
     write.csv(Mantel.tab, paste0(dir.aux1,fn))
    }

   }

  }

}
```

```
# Variables DGA y taxones ---------------------------------------------------

for (i in 1:2){
 # Directorios
 dir.auxi <- paste0(dir.taxv,'/0',i,' - ',TAXA[i])
 if (file.exists(dir.auxi)==F){dir.create(dir.auxi)}
 dir.aux1 <- paste(dir.auxi,'01 - TABLAS',sep = '/')
 if (file.exists(dir.aux1)==F){dir.create(dir.aux1)}
 dir.aux2 <- paste(dir.auxi,'02 - GRÁFICOS',sep = '/')
 if (file.exists(dir.aux2)==F){dir.create(dir.aux2)}

 # Regresiones
 n = 1
 for (o in c('Lecho','Agua')){
  dir.aux3 <- paste0(dir.aux2,'/0',n,' - ',o)
  if (file.exists(dir.aux3)==F){dir.create(dir.aux3)}

  ps <- subset_samples(Taxa.list[[i]],Origen == o) %>%
   filter_taxa(function(x) sum(x)>0,prune = T)

  # Dataframe Variables
  df.vari <- data.frame(sample_data(ps))
  sc.vari <- paste0('V',c(1,2,4,10,16,18,20,21,22,23,27))
  df.vari <- cbind(df.vari[1:3],df.vari[,sc.vari])[-2,] # P2 sin datos en
DGA.

  # Dataframe taxones (top5)
  df.taxa <- data.frame(tax_table(ps)[,TAXA[i]])
  df.taxa <- data.frame(tax = df.taxa[,,drop = T],
                        otu_table(ps) %>%
                         apply(.,2,function(x){x/sum(x)}))
  df.taxa <- df.taxa[!(df.taxa$tax=='N/A'),]
  df.taxa <- df.taxa[order(apply(df.taxa[-1],1,mean),
                           decreasing=T),][1:5,]
  rownames(df.taxa) <- df.taxa[,1]
  df.taxa <- t(df.taxa[-1])[-2,]
  colnames(df.taxa) <- gsub('-','_',colnames(df.taxa))
  colnames(df.taxa) <- gsub(' ','__',colnames(df.taxa))
  df.taxa <- data.frame(df.vari[1:3], df.taxa, check.names = F)

  # Dataframe Variables - taxones
  df.vtaxa <- left_join(df.vari, df.taxa, by =
c('Muestra','Sitio','Origen'))
  col.ori = data.frame(Lecho = 'firebrick4', Agua = 'dodgerblue4')

  ite1 <- T
  for (t in colnames(df.taxa)[-c(1,2,3)]){
   for (v in colnames(df.vari)[-c(1,2,3)]){
    ct.s <- cor.test(df.vtaxa[,t],df.vtaxa[,v], 'spearman',
                     alternative = 'two.sided', exact = F, na.rm = T)
    ct.p <- cor.test(df.vtaxa[,t], df.vtaxa[,v], 'pearson',
                     alternative ='two.sided',na.rm = T)
    # Fila variable-taxa
    fila <- data.frame(var = auxivar[v,'Variable'], tax = t, Origen = o,
                       sp.r = ct.s$estimate, sp.p = ct.s$p.value,
                       pe.r = ct.p$estimate, pe.p = ct.p$p.value)

    # regresión Lineal
    rl <- lm(as.formula(paste(t,v,sep='~')),df.vtaxa)
    s.rl <- summary(rl)
    a.rl <- anova(rl)

    fila <- data.frame(fila,
                       B0 = s.rl$coefficients[1,1],
                       B0.p = s.rl$coefficients[1,4],
                       B1 = s.rl$coefficients[2,1],
                       B1.p = s.rl$coefficients[2,4],
                       r2 = s.rl$r.squared,
                       r2.adj = max(s.rl$adj.r.squared,0),
                       Fval = a.rl$`F value`[1],
                       p = a.rl$`Pr(>F)`[1])

    tax.titu <- gsub('_','-',gsub('__',' ',t))
```

```
      # Gráfico
      ggplot(df.vtaxa, aes(x = .data[[v]], y = .data[[t]])) + theme_bw()
+
        geom_point(color = col.ori[,o], na.rm = T) +
        geom_text_repel(aes(label = Muestra), size = 4, na.rm = T) +
        geom_smooth(method='lm', formula = y~x, se = F,na.rm = T,
                    linetype='solid', color='black', linewidth = 0.5) +
        scale_y_continuous(labels = percent,
                           expand = expansion(mult = c(0.05,0.2))) +
        xlab(auxivar[v,'var.uni']) +
        ylab(paste0('Abundancia relativa<br>',
                    TAXA[i],' *',gsub('_','-',gsub('__',' ',t)),'*')) +
        ggtitle(paste0('*',gsub('_','-',gsub('__',' ',t)),'*',' v/s ',
                       auxivar[v,'Variable'])) +
        theme(plot.title = element_markdown(hjust = 0.5, size = 12),
              axis.title.x = element_markdown(size = 10),
              axis.title.y = element_markdown(size = 10)) +
        stat_poly_eq(formula =y~x, eq.with.lhs = "italic(hat(y))~`=`~",
                     aes(label = paste(..eq.label.., ..adj.rr.label..,
                                       sep = "'*plain(\",\")~")),
                     parse = TRUE, label.x=0.95,label.y = 0.98, na.rm = T)

      if (GUARDAR==T){
        taxnom <- gsub('_','-',gsub('__',' ', abbreviate(t, minlength =
15)))
        if(abs(fila$pe.r) > 0.7 & fila$pe.p < 0.01 |
           abs(fila$sp.r) > 0.7 & fila$sp.p < 0.01 ) {
          ggsave(paste0('ÚTIL_',taxnom,'~',
                        gsub('/','_',auxivar[v,'Variable']),'.png'),
                 plot = last_plot(), path = dir.aux3, scale = 1.8,
                 width = 7.5, height = 6, units = 'cm', dpi = 500,
                 limitsize = TRUE)
        } else {
          ggsave(paste0('ÚTIL_',taxnom,'~',
                        gsub('/','_',auxivar[v,'Variable']),'.png'),
                 plot = last_plot(),path = dir.aux3, scale = 1.8,
                 width = 7.5, height = 6, units = 'cm', dpi = 500,
                 limitsize = TRUE)
        }
      }

      if (ite1 == T){
        vartax <- fila
        ite1 <- F
      } else {
        vartax <- rbind(vartax,fila)
      }

    }

  }

  vartax <- data.frame(vartax[1:5],
                       sp.padj = p.adjust(vartax$sp.p,'BH'), vartax[6:7],
                       pe.padj = p.adjust(vartax$pe.p,'BH'), vartax[8:15])

  if (o == 'Lecho'){
    vartax.tab <- vartax
  } else {
    vartax.tab <-rbind(vartax.tab,vartax)
  }

  n = n+1
  }


  if (GUARDAR == T){
    write.csv(vartax.tab,
              paste0(dir.aux1,'/vartax_',TAXA[i],'.csv'))
  }

  vartax.tab$var <- gsub('DGA','',vartax.tab$var)

  vartax.tab$tax <- paste0(vartax.tab$tax,' (',
                           substr(vartax.tab$Origen,1,1),')')

  sp.rho <- vartax.tab %>%
    dplyr::select(var,tax,sp.r) %>%
    pivot_wider(names_from = var, values_from = sp.r) %>%
    column_to_rownames(var = 'tax') %>%
    as.matrix()

  sp.pval <- vartax.tab %>%
    dplyr::select(var,tax,sp.p) %>%
    pivot_wider(names_from = var, values_from = sp.p) %>%
    column_to_rownames(var = 'tax') %>%
    as.matrix()

  sp.col <- ifelse(sp.pval < 0.01, 'black', '#1C00ff00')

  pe.rho <- vartax.tab %>%
    dplyr::select(var,tax,pe.r) %>%
    pivot_wider(names_from = var, values_from = pe.r) %>%
    column_to_rownames(var = 'tax') %>%
    as.matrix()

  pe.pval <- vartax.tab %>%
    dplyr::select(var,tax,pe.p) %>%
    pivot_wider(names_from = var, values_from = pe.p) %>%
    column_to_rownames(var = 'tax') %>%
    as.matrix()

  pe.col <- ifelse(pe.pval < 0.01, 'black', '#1C00ff00')


  if (GUARDAR == T) {
    # Gráfica Spearman
    png(paste(dir.auxi,'/spearman_',TAXA[i],'.png',sep=''),
        width = 15, height = 15, units = 'cm', res = 800)
    corrplot(sp.rho, p.mat = sp.pval, sig.level = 0.01, insig = 'pch',
             pch.cex = 4, pch.col = 'white', addCoef.col = sp.col,
             number.cex = 0.6, tl.col='black', tl.cex=0.8, tl.srt = 45)
    dev.off()

    # Gráfica Pearson
    png(paste(dir.auxi,'/pearson_',TAXA[i],'.png',sep=''),
        width = 15, height = 15, units = 'cm', res = 800)
    corrplot(pe.rho, p.mat = pe.pval, sig.level = 0.01, insig = 'pch',
             pch.cex = 4, pch.col = 'white', addCoef.col = pe.col,
             number.cex = 0.6, tl.col='black', tl.cex=0.8, tl.srt = 45)
    dev.off()
  }
}

# FIN ----------------------------------------------------------------------
```