3.1 Protocolos de genómica para monitoreo ambiental asociado a acciones de respuesta por impacto o contingencia ambiental formalizados y listos para ser transferidos a usuarios interesados

Paola Montoya Valencia
Nicolás D. Franco-Sierra
María Claudia González
Nathalie Baena-Bejarano
Paola Pulido-Santacruz
Alejandro Salazar Villegas
Eduardo Tovar Luque
Mailyn Gonzalez Herrera

Línea de Gestión de los Recursos Genéticos,
Programa de Ciencias de la Biodiversidad

Instituto de Investigación de Recursos Biológicos Alexander von Humboldt
Bogotá, D.C., 2020

Producto 3 del Plan Operativo Anual (POA) 2020
Resolución 0041 de 2020

Catalogación de la fuente

Montoya Valencia, Paola

3.1 Protocolos de genómica para monitoreo ambiental asociado a acciones de respuesta por impacto o contingencia ambiental formalizados y listos para ser transferidos a usuarios interesados / Paola Montoya Valencia – Bogotá: Instituto de Investigación de Recursos Biológicos Alexander von Humboldt, 2020.

69 p.

Incluye bibliografía, tablas, mapas, fotos a color

1. Metabarcoding. – 2. Diseño experimental. – 3. Protocolos para Monitoreo. – 4. Hidrocarburos. – 5. Insectos. – 6. Estudio de Caso. – 7. Diversidad. I. Instituto de Investigación de Recursos Biológicos Alexander von Humboldt II. Considerations for the development of an information policy in relation to the Final Technical Report.

Catalogación en la fuente – Biblioteca Francisco Matís – Diana Bejarano.

Como citar este documento:

Montoya Valencia, P., Franco-Sierra, N. D., González, M. C., Baena-Bejarano, N., Pulido-Santacruz, P., Salazar Villegas, A., Tovar Luque, E. & Gonzalez Herrera, M. (2021). 3.1 Protocolos de genómica para monitoreo ambiental asociado a acciones de respuesta por impacto o contingencia ambiental formalizados y listos para ser transferidos a usuarios interesados. Bogotá: Instituto de Investigación de Recursos Biológicos Alexander von Humboldt.

Contenido:

Introducción 4

3.1.1. Flujo de trabajo de metabarcoding y diseño experimental 6

i. Selección de primers 7

ii. Réplicas de PCR 7

iii. Controles negativos de extracción y PCR 8

iv. Controles positivos 8

v. Controles de secuenciación 8

iii. Umbrales mínimos de número de copias 8

iv. Umbrales de agrupamiento 9

3.1.2. Propuesta de protocolo para la implementación del metabarcoding en estudios de monitoreo 10

3.1.2.1 Toma de muestra 12

3.1.2.2 Extracción de ADN 18

3.1.2.3. Preparación de librerías 20

3.1.2.4 Procesamiento bioinformático para metabarcoding 22

Alineamiento de las secuencias 23

Remoción de secuencias no correctamente alineadas 24

Asignar secuencias a muestras (demultiplex) 24

Remoción de secuencias con bases ambiguas 25

Dereplicacion y eliminación de singletons 25

Filtrado por tamaño y remoción de secuencias poco soportadas 26

Asignación taxonómica de las secuencias 27

Clustering (generación de OTUs) 27

Generación de tabla de comunidades 27

Remoción de OTUs no identificados 28

Identificación de artefactos o contaminantes 29

Identificación de cambios de etiqueta o tag switching y remoción de su efecto 32

Eliminando potenciales errores y contaminantes 35

Eliminación de PCRs fallidas 36

Construyendo la matriz final de comunidades 39

Construcción de bases de datos 41

3.1.3 Estudio de caso: Diseño e implementación de un programa de monitoreo para evaluar el estado de la biodiversidad presente en las áreas afectadas por derrames de hidrocarburos generados por terceros y por eventos operacionales en el corredor vial Puerto Vega – Teteyé – Bloque suroriente, y el efecto del programa de limpieza y

recuperación realizado por la empresa Gran Tierra Energy Colombia o sus contratistas 45

3.1.4 Uso de metabarcoding en el monitoreo ambiental con insectos 58

Literatura citada 61

Introducción

La transformación del paisaje por actividades antrópicas ha aumentado en las últimas décadas, con consecuencias en la biodiversidad, en múltiples ocasiones, desconocidas (Steffen et al. 2015, Correa et al. 2020). Esto genera la necesidad, no sólo de conocer el estado de la biodiversidad actual, si no también de evaluar la influencia de la actividad humana en las comunidades bióticas.

El monitoreo de la diversidad, la evaluación periódica de la misma en un ecosistema, genera información que puede ser considerada en las decisiones políticas que van dirigidas a reducir la presión sobre la biodiversidad y a salvaguardar los ecosistemas, especies y diversidad genética (Metas B, C AICHI, Proença et al. 2017, Dallmeier 1996). El metabarcoding es una técnica que, usando ADN y tecnología de secuenciación masiva, permite caracterizar la biodiversidad en un amplio rango de grupos biológicos en diferentes taxones y ambientes, de una forma más costo eficiente en términos económicos y temporales por muestra que técnicas tradicionales de monitoreo (Valentini et al. 2016, Deiner et al. 2107, Bush et al. 2019, Zinger et al. 2019).

El uso del metabarcoding se ha extendido en la última década en múltiples campos, desde la reconstrucción de comunidades históricas (p.e. Willerslev et al. 2014, Parducci et al. 2017), el monitoreo de diferentes sistemas (p.e. Watts et al. 2019, Lobo et al. 2017, Oliverio et al. 2018, Pawlowski et al. 2016, Valentini et al. 2016), determinación de dietas (p.e. Kartzinel et al. 2015, Lopes et al. 2015, Berry et al. 2017, Kerley et al. 2018) o detección de especies invasoras (p.e. Brown et al. 2016, Harrrelms-Tuohy et al. 2016, Bol et al. 2017, Mychek-Londer et al. 2020). Así también, se ha usado para caracterizar las comunidades bióticas en regiones con influencia de múltiples actividades humanas como actividades agrícolas (p.e. Pawlowski et al. 2014, Chen et al. 2018, Treonis et al. 2018), actividades de la industria de hidrocarburos (p.e. Cluff et al. 2014, Lanzén et al. 2016, Daly et al. 2016, Laroche et al. 2018, Hull et al. 2018) o polución (p.e. Chariton et al. 2015, Li et al. 2018, Valentin et al. 2019, Sun et al. 2019). En Colombia, sin embargo, a nuestro conocimiento los estudios de biodiversidad usando metabarcoding son escasos, pese a la alta tasa de transformación de los ecosistemas que vivimos actualmente (Correa et al. 2020).

En este documento se describen los aspectos más relevantes para aplicar la técnica de metabarcoding en estudios de biodiversidad. En la primera sección se realiza un recuento de los aspectos relevantes en el diseño experimental propios a la técnica. En la segunda se plantea una propuesta de protocolo para la aplicación del metabarcoding, incluyendo la toma de muestra, el procesamiento molecular y el procesamiento o depuración bioinformática. El diseño de muestreo y los análisis ecológicos no son incluidos desde que estos son totalmente contexto-dependientes. En la tercera sección se describe un estudio de caso usando metabarcoding en la caracterización de microorganismos de una zona altamente contaminada por actividades relacionadas a la extracción convencional de hidrocarburos. En la cuarta parte se realiza una revisión del potencial uso del metabarcoding en el monitoreo ambiental a través insectos. Finalmente, en la quinta y última parte se exponen algunas de las perspectivas y limitaciones del uso de la técnica en Colombia.

El objetivo de este documento es fomentar el uso de esta herramienta en escenarios colombianos, generando datos cuantitativos de diversidad que puedan ser incluidos en la toma de decisiones.

3.1.1. Flujo de trabajo de metabarcoding y diseño experimental

El desarrollo y avance de las tecnologías de secuenciación masiva ha permitido la captura simultánea de múltiples secuencias genéticas a partir de muestras ambientales complejas (Taberlet et al. 2018). El metabarcoding, una herramienta que hace parte de esta innovación, se basa en la captura de una región específica del genoma de múltiples organismos a partir de una muestra compleja, como lo son fracciones de suelo, agua, heces, contenido estomacal, polen, etc. (Taberlet et al. 2012, 2018).

En términos generales, el metabarcoding es una herramienta que se compone de siete pasos principales: 1) recolección de una muestra compleja o ambiental (agua, sedimento, suelo, heces, polen, hojarasca, etc.); 2) extracción y captura del ADN (extra o intracelular) que se encuentra en la muestra recolectada; 3) preparación de librerías (amplificación de una región de interés en el ADN extraído); 4) secuenciación de la región amplificada; 5) depuración y procesamiento bioinformático de las secuencias obtenidas; 6) agrupamiento de las secuencias por similitud y asignación taxonómica - este último opcional de acuerdo con las bases de datos de referencia existentes; y 7) análisis ecológicos de interés (figura 1).

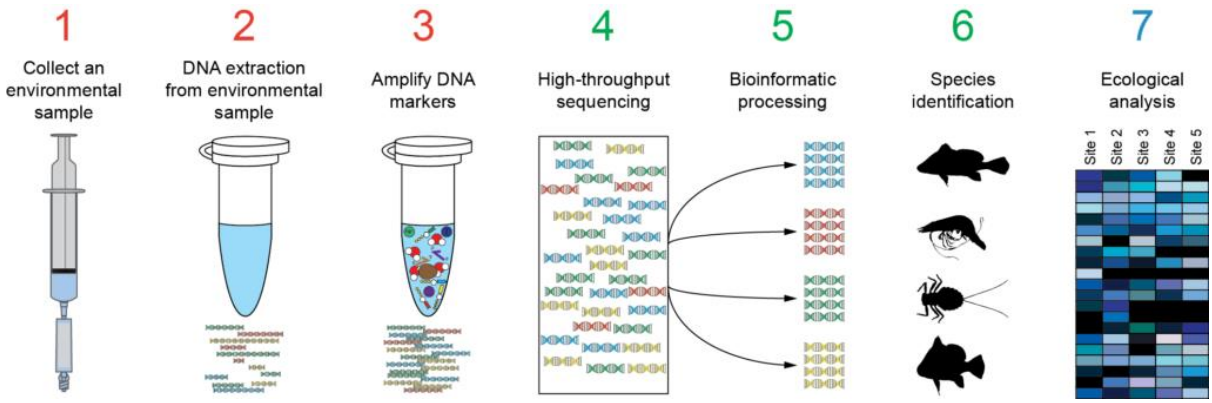

Figura 1. Flujo de trabajo de metabarcoding. Imagen tomada de Nature Metrics

El primer paso para desarrollar un estudio de metabarcoding, al igual que para casi cualquier investigación, es definir la pregunta e hipótesis de investigación. Esto permitirá proceder a plantear un diseño de muestreo que responda al objetivo de investigación. El diseño definirá el número de puntos de muestreo, número de muestras por punto de muestreo y la forma de muestreo. En este contexto, es importante resaltar que el diseño de muestreo es contexto-dependiente y que el metabarcoding constituye una herramienta, por lo que el proceso de diseño de muestreo debe seguir los mismos lineamientos que cualquier estudio de diversidad o ecológico (Magurra & McGill 2011). Una vez definido el diseño de muestreo (dónde y cuántas muestras tomar), es necesario analizar una serie de aspectos propios a la técnica que pueden llevar a generar resultados o conclusiones sesgadas (Zinger et al. 2019, Mendoza et al. 2015). En cada caso, se deben evaluar las posibles medidas que puedan ser incluidas en el diseño para controlar los desafíos propios del metabarcoding.

A través de diferentes estudios se han identificado diferentes aspectos claves en el diseño experimental que pueden repercutir en los datos generados por metabarcoding, y que deben ser

tomados en cuenta en el diseño experimental y en el procesamiento de datos(Alberdi et al. 2018, Beentjes et al. 2019, Calderón-Sanou et al. 2020):

i. Selección de marcadores y primers
ii. Réplicas de PCR
iii. Controles negativos de extracción y PCR
iv. Controles positivos
v. Controles de secuenciación
vi. Umbrales de agrupamiento
vii. Número mínimo de lecturas

i. Selección de primers

Un aspecto importante en metabarcoding es la selección de marcadores. Los marcadores deben ser cortos (menor a 300 pares de bases), con una buena resolución taxonómica y con *primers* disponibles de amplio espectro. Los *primers* son una cadena de bases nucleotídicas que sirve como ancla para iniciar el proceso de amplificación, siendo complementario a la cadena de un fragmento de ADN ubicado en los flancos de la región de interés (Hebert et al. 2003). Aunque diferentes marcadores y *primers* han sido sugeridos por diferentes autores, debe tomarse en cuenta la disponibilidad de bases de referencias, ya que de esto dependerá que se realice una buena asignación taxonómica.

En el caso particular de los *primers*, es importante su especificidad al momento de su implementación para metabarcoding. *Primers* muy específicos restringirán los grupos biológicos que puedan ser recuperados de la muestra; por el contrario, *primers* muy poco específicos, aunque permitirán un mayor número de grupos, podrían generar mayor número de errores y la amplificación de ADN contaminante (Op De Beeck et al. 2014).

Dado que esta selección debe ajustarse al objetivo del proyecto y al conocimiento que se tenga sobre el sistema que se trabajará, una opción recomendada es sintetizar primers propios. Para esto, diferentes programas permiten realizar amplificaciones o PCR (reacción en cadena de la polimerasa o polymerase chain reaction) digitales a través de bases de datos de referencia, evaluando la efectividad y especificidad de los potenciales *primers* previo a su implementación real.

ii. Réplicas de PCR

Uno de los aspectos más importantes al plantear un estudio de metabarcoding, es realizar múltiples réplicas de PCR (Ficetola et al. 2015). El número de PCRs recomendado es variable y depende enteramente de cada sistema y de los recursos con los que se cuenten. Se recomiendan al menos tres réplicas por extracción de ADN, aunque el escenario ideal podría estar en realizar pruebas pilotos para determinar el número apropiado de réplicas de PCR. Esto, sin embargo, es poco viable dado el costo. La sensibilidad de la tecnología de secuenciación, sumada a la aleatoriedad de la amplificación generan un alto número de errores. En diferentes estudios usando metabarcoding se reporta menos del 20% de unidades moleculares (OTU) compartidos a través de las réplicas de PCR (Wen et al. 2020, Alberdi et al. 2018). Contar con estas réplicas permite aplicar diferentes estrategias de filtro que aumentan la confiabilidad en las unidades moleculares obtenidas para cada muestra.

iii. Controles negativos de extracción y PCR

Por cada evento de extracción de ADN que se realice es necesario realizar un control negativo. Esto hace referencia a una muestra estéril, sin ADN de ningún tipo, que incluya todos los reactivos usados en el proceso de extracción. Dependiendo del tipo de muestra, puede consistir en agua destilada, suelo o sedimento esterilizado previamente o algún buffer usado durante el proceso. Así también, se deben incluir controles negativos de PCR, donde al igual que los controles de extracción, se adicionan todos los reactivos usados en la PCR y agua, u otro tipo de solución estéril, en lugar de ADN. Desde que no se adiciona ADN en los controles negativos, estos ayudarán a identificar potenciales contaminantes en las muestras o reactivos, y removerlos de manera eficiente de todas las muestras trabajadas.

iv. Controles positivos

Los controles positivos consisten en una comunidad de organismos del grupo de interés, con identidades y secuencias conocidas en concentraciones conocidas. Idealmente, el control positivo debe estar conformado por entidades que no se encuentren naturalmente en el área de estudio, ya que de esta forma se permite rastrear y detectar en otras muestras donde pueda actuar como contaminante. Estos controles, en concentraciones diferentes, permiten analizar el comportamiento de los diferentes procesos, como la tasa de error en la amplificación y en la secuenciación, detección de potenciales contaminantes, de afinidades del primer y profundidad de secuenciación de las muestras.

v. Controles de secuenciación

Finalmente, el último control recomendado es el de secuenciación. Estos consisten en pozos vacíos (sin ADN ni reactivos) en la placa de amplificación, que representarán combinaciones de etiquetas sin usar, en un sistema de doble etiqueta. Normalmente, se recomienda que los pozos se ubiquen en la diagonal de la placa. De esta forma, analizando las secuencias obtenidas en estos pozos, que idealmente deberían ser ninguna, se pueden detectar quimeras (secuencias de ADN de dos organismos diferentes que se recombinan), salto de secuencias entre los pozos (tag-jumping o tag switching), secuencias productos de errores generados en el proceso y potenciales contaminantes.

iii. Umbrales mínimos de número de copias

No existe un número mínimo de copias de lecturas por secuencia que deba esperarse en muestras procesadas por secuenciación masiva de alto rendimiento, por lo que es un tema de gran discusión e incertidumbre en metabarcoding. Dada la naturaleza y sensibilidad de la técnica (amplificación y secuenciación) se espera que cada secuencia tenga un alto volumen de lecturas. La complicación en metabarcoding, es que *a priori* no se conoce el número de organismos y secuencias presentes en la muestra y por tanto no es posible estimar el número de lecturas esperados por cada una. Diferentes protocolos incluyen la remoción de *singletones* (secuencias soportadas por una lectura) y algunos la eliminación de *doubletones* (secuencias soportadas por dos lecturas), pero esto implica reconocer como secuencias válidas a aquellas soportadas con al menos tres lecturas en adelante. Es por esto, que a través de diferentes aproximaciones se ha evaluado el efecto de remover secuencias soportadas por diferente número de lecturas. El principal efecto ha sido encontrado al remover los *singletones*, ya que un alto porcentaje de las lecturas obtenidas en metabarcoding corresponden a lecturas únicas. El efecto de remover secuencias con otro número de copias parece depender directamente del filtro que se haya usado para seleccionar los OTUs a partir de las réplicas de PCR. Cuando este

filtro ha sido estricto o incluso cuando solo se usan los OTUs compartidos en al menos dos réplicas de PCR, gran parte de las secuencias con bajo número de lecturas son eliminadas. Remover secuencias soportadas por pocas lecturas, y realizar filtros de OTUs a partir de las réplicas de PCR, influyen directamente en otros pasos bioinformáticos como la eliminación de quimeras y contaminantes.

iv. Umbrales de agrupamiento

En los protocolos de procesamiento de bioinformático se incluye agrupamiento de secuencias a OTUs o MOTUs (unidades taxonómicas operacionales moleculares). Éstos agrupan las secuencias de acuerdo con su similitud a través de umbrales que son definidos por el investigador, y es aquí donde los resultados pueden variar considerablemente. Tradicionalmente se ha usado un umbral de agrupamiento de 97%, sin embargo, a través de diferentes estudios se ha cuestionado su uso de forma generalizada entre grupos. El umbral de 97% de similitud parece funcionar muy bien para bacterias, donde la similitud entre especies se ha estimado en el 97%, sin embargo, este no es el estimado para otros grupos biológicos donde puede ser muy variable dependiendo del grupo específico del que se esté hablando. El impacto en usar un umbral u otro es alto en la diversidad encontrada, y este también se ve influenciado por el filtro de OTUs con las réplicas de PCR. De forma general, entre más bajo sea el umbral de similitud para agrupamiento (por ejemplo, 99%) menor diversidad se reporta ya que se requiere menor similitud para generar un grupo. Las recomendaciones se enfocan en usar diferentes umbrales y evaluar el impacto en los resultados, una vez ya se han realizado otro tipo de procedimientos y filtros de depuración, ya que el mejor umbral es específico a cada grupo y sistema. Existen varias propuestas al respecto, como

También se han identificado otros pasos aspectos, principalmente en el procesamiento bioinformático, que requieren la toma de decisiones pero que, de acuerdo con las evaluaciones, no tienen un impacto alto en los resultados o son resueltos y/o minimizados a través de los puntos que se han mencionado anteriormente. Algunos de ellos son: la profundidad de secuenciación, eliminación de quimeras, bases de datos de referencia, y porcentajes de asignación taxonómica.

Con un diseño de muestreo dirigido a responder la pregunta de interés planteado, y tomando en cuenta los diferentes aspectos anteriormente mencionados, se puede iniciar un estudio usando el metabarcoding. La siguiente sección plantea una propuesta del flujo de trabajo de metabarcoding (figura 1), sin incluir diseño ni análisis de datos ya que estos son dependientes del tipo de investigación. Este planteamiento es sólo una de las múltiples formas que existen para llevar a cabo un estudio de metabarcoding, por lo que invitamos a todos aquellos interesados en el tema a revisar literatura científica con otros planteamientos.

3.1.2. Propuesta de protocolo para la implementación del metabarcoding en estudios de monitoreo

Como se mencionó en la sección anterior, al iniciar un proyecto con metabarcoding, es necesario definir el diseño de muestreo apropiado al objetivo o pregunta de estudio. El diseño de muestreo no responde al objetivo de este documento, sin embargo, la claridad de algunos términos usados en el planteamiento, facilitarán el seguimiento del documento y su aplicación a cada caso:

- *Unidad o punto de muestreo*: Unidad mínima de observación de la que se obtendrá información de las variables de interés. Cada unidad de muestreo representará un punto en los análisis ecológicos. Para el caso exclusivo de este documento, la unidad de muestreo estará conformada por tres muestras de sustrato (por sustrato se entiende agua, sedimento o suelo) (figura 2).

- *Muestra*: Corresponde a cada evento de toma de material, sea esta agua, sedimento o suelo. Las muestras pueden ser simples o compuestas. Estas últimas corresponden a muestras compuestas de la mezcla de múltiples sub-muestras tomadas en el área de la muestra. Esto último con el objetivo de capturar la variabilidad natural del micrositio donde se toma la muestra.

Antes de tomar una muestra ambiental es necesario tener dos aspectos en cuenta: 1) la esterilización de los implementos que serán usados y 2) la preservación de la muestra.

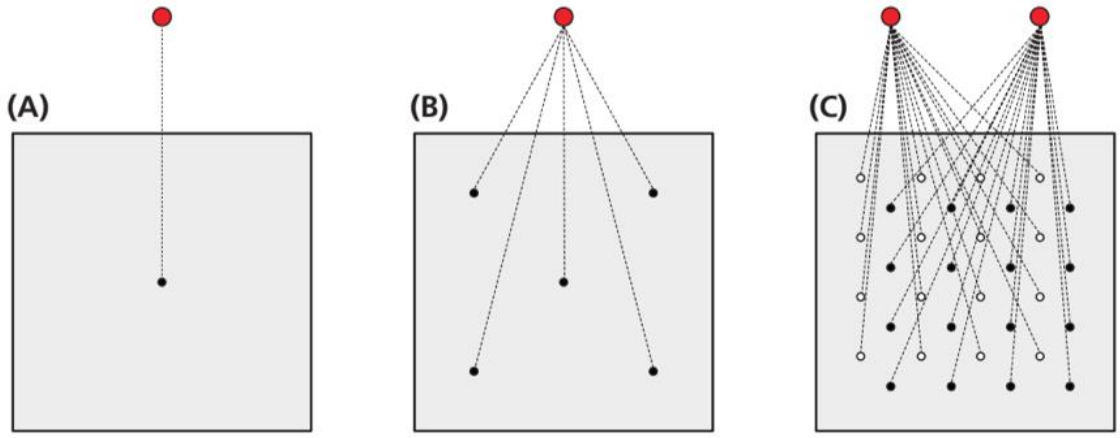

Figura 2. Esquema de unidad de muestreo (rectángulo), muestra (círculos rojos) y submuestra (círculos negros y blancos). A) Unidad de muestreo conformada por una única muestra simple; B) unidad de muestreo conformada por una muestra compuesta (con múltiples submuestras); C) unidad de muestreo conformada por múltiples muestras compuestas. Imagen tomada de Taberlet et al. 2018.

Esterilización – Riesgo de contaminación

Una muestra ambiental contiene gran cantidad de fragmentos de ADN de diferentes organismos. Este ADN puede estar libre (fuera de la célula) y, por lo general, fragmentado y

degradado. Es fundamental evitar la contaminación de las muestras ambientales con ADN de la persona que toma la muestra o de las superficies de trabajo. El riesgo de contaminación es alto debido a que el ADN externo a la muestra puede ser recuperado, amplificado y secuenciado en una alta proporción, dando resultados errados. Este riesgo es más alto aún dado el principio de la amplificación del ADN por PCR (Polymerase Chain Reaction o Reacción en cadena de la polimerasa) y de la sensibilidad de la secuenciación masiva de alto rendimiento. Es por esto que durante la manipulación de la muestra, desde su toma hasta su secuenciación, es necesario tomar medidas que minimicen al máximo el riesgo de contaminación y contar con múltiples controles negativos y positivos, que permitan detectar potenciales contaminantes. Durante la fase de toma de muestras, se recomienda esterilizar regularmente las herramientas utilizadas en su manipulación.

Tenga en cuenta las siguientes recomendaciones antes de procesar cada muestra:

i. Esterilice todos los utensilios o materiales que serán usados. Para esto, los utensilios o materiales pueden ser sumergidos en una mezcla de hipoclorito al 10%, autoclavados o puestos en una cámara UV, dependiendo del material y disponibilidad de equipos.
ii. En el caso de materiales desechables como guantes, tapabocas, filtros de agua, entre otros, deben ser cambiados entre cada muestra. Si durante la manipulación se considera que estos han tocado materiales, superficies o elementos que pueden contaminar la muestra, se deben descartar apropiadamente y reemplazar.
iii. En el caso de palas, pinzas, tijeras u otros materiales metálicos, una alternativa es usar flameo tras ser cubiertos con alcohol al 96%. Se debe tener particular cuidado usando la técnica de flameo para evitar accidentes.
iv. Los recipientes donde se depositan las muestras deben estar esterilizados y bien sellados antes de ser usados.

Preservación de la muestra

El metabarcoding es un método que permite analizar la composición de la comunidad que se encuentra en la muestra y su alrededor, proporcionando una fotografía del sistema o ambiente de interés. Es por esto que una vez se toma la muestra ambiental, se hace necesario conservar las condiciones que responden a ese instante deteniendo actividades biológicas como el crecimiento celular o la degradación de ADN (Delavaux et al. 2020). No hacerlo puede llevar a que se degrade el ADN y no sea posible recuperarlo o, a que la comunidad cambie en respuesta al proceso de recolección y que los resultados obtenidos representen los de una comunidad diferente a la muestreada. Es importante conservar la muestra a bajas temperaturas desde el momento de recolección hasta que es procesada. Lo más recomendado es realizar las extracciones de ADN las horas siguientes a la recolección, no obstante, si esto no es posible es recomendado usar un método de preservación que asegure la estabilidad del ADN desde el momento de toma hasta el proceso de extracción de ADN. Existen diferentes métodos para esto, algunos de ellos son:

i. Preservación de la muestra en un contenedor a bajas temperaturas: entre 4º (nevera con hielo) y -196ºC (nitrógeno líquido). Si la muestra es preservada en neveras con hielo (aproximadamente 4ºC) es necesario procesarla en las siguientes horas (Delavaux et al. 2020, Hinlo et al. 2017).
ii. Reactivos comerciales estabilizadores del ADN, que permiten conservar el ADN a temperatura ambiente (Ficetola et al. 2008, Longmire et al. 1997, Wegleitner et al. 2015).

iii. Preservación en seco. Este método es particularmente útil para muestras de suelo o filtros de nitrato de celulosa usados en la filtración de muestras de agua, embebiendo las muestras en silica gel, sin que la muestra entre directamente en contacto con la misma (Taberlet et al. 2018).

Antes de tomar la muestra ambiental defina el método de preservación más apropiado y viable para su caso, en dos momentos del proceso: 1) desde la toma de muestra hasta su procesamiento y 2) desde su procesamiento hasta la extracción de ADN.

3.1.2.1 Toma de muestra

Una vez definido un diseño apropiado para responder la pregunta asociada al estudio, se procede a realizar la toma de muestras ambientales. En este caso, se realizará la descripción de un método para tomar muestras en agua, suelo y sedimento, que ha sido puesta en práctica por el Instituto Humboldt; sin embargo, se resalta que esta es sólo una de las formas posibles, y que otras formas pueden ser encontradas en literatura científica.

Recuerde antes de tomar la muestra, rotular con marcador o bolígrafo resistente al agua cada uno de los recipientes donde la muestra será almacenada. Es recomendable usar además del código de la muestra, información adicional que permita corroborar la identidad de la muestra como coordenadas, localidad o fecha. Tome las coordenadas del punto de muestreo y realice una descripción general del sitio de forma escrita o con registro fotográfico. Tenga presente que en el punto de muestreo tomará la muestra y las observaciones de las variables de su interés que serán relacionadas a la diversidad o composición de la comunidad.

Agua

El método aquí descrito está enfocado en la toma de aguas superficiales. Los materiales necesarios son ilustrados en la figura 3. Se propone una unidad de muestreo compuesta de tres muestras simples, tomadas a lo largo de un transecto en el cuerpo de agua. La longitud del transecto dependerá enteramente de la escala y el objetivo del estudio. Si dado el área de su estudio selecciona un transecto de al menos 150 metros de longitud, cada muestra de agua debería ser tomada cada 50 metros (figura 4).

En el caso de sistemas lóticos, las muestras deben ser tomadas desde en dirección contraria al flujo del agua, para evitar perturbar el agua y modificar las condiciones de la muestra. Para la toma de cada muestra, tome al menos 1 litro de agua superficial usando recipientes previamente esterilizados y usando guantes y tapaboca limpios. Almacene inmediatamente a baja temperatura, al menos 4ºC, usando una nevera con hielo, un refrigerador o un tanque de nitrógeno líquido. Entre la toma de cada muestra, cambie los guantes y tapabocas.

Figura 3. Listado de materiales necesarios para recolectar muestras de agua.

En el caso de sistemas lénticos, tome las tres muestras tratando de cubrir el área del cuerpo de agua, siendo una de las muestras tomada hacia el centro del cuerpo de agua. Si esto no es posible, tome las tres muestras en tres puntos equidistantes en las orillas. Dependiendo de la extensión del cuerpo de agua, considere realizar más de una unidad de muestreo por cuerpo de agua. Recuerde que la localización de las muestras debe corresponder al interés de la investigación. Por ejemplo, en el caso de caracterizaciones biológicas, es recomendable seleccionar áreas que representen la heterogeneidad del cuerpo de agua.

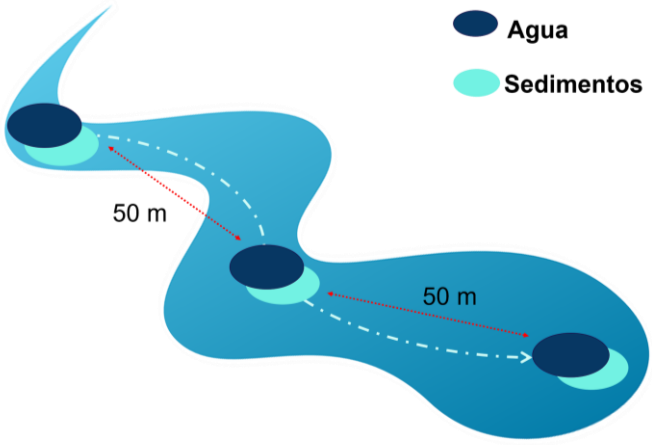

Figura 4. Esquema de toma de muestras de aguas superficiales y sedimentos en un cuerpo lótico siguiendo un método de transecto.

Una vez haya tomado las muestras de agua de la unidad de muestreo diríjase al lugar de procesamiento de las muestras o estación de trabajo. El procesamiento debe llevarse a cabo

dentro de las 12 horas siguientes a la toma de muestra, siempre y cuando se asegure la preservación de la muestra con alguno de los métodos mencionados anteriormente.

Para metabarcoding se recomienda filtrar las muestras de agua usando filtros de nitrato de celulosa con un tamaño de poro de al menos 0.45 µm, los cuales pueden ser ajustables a embudos recolectores o estar en unidades de filtración (por ejemplo, Sterivex^{TM}). Existen diferentes sistemas de filtración como son los sistemas de filtración por bomba peristáltica o por bomba de vacío, que pueden usarse directamente desde la fuente de agua o en la estación de trabajo. En este documento describiremos el procedimiento usando un sistema de filtración por bomba de vacío en la estación de trabajo (figura 5).

Figura 5. Ejemplo de sistema de filtración por bomba de vacío. Fotografía de Paola Pulido-Santacruz - IAVH

Este sistema de filtración cuenta con múltiples embudos recolectores, en los cuales se ubican los filtros. Los embudos deben estar previamente esterilizados con sumersión completa en una solución de hipoclorito al 10% por al menos 30 min. Es muy importante que, finalizados los 30 minutos, enjuague con abundante agua destilada. En los casos en los que sea posible, también puede usarse autoclave o exposición a UV.

Con el sistema de embudo recolectores cerrado, encienda la bomba de vacío, instale los filtros usando pinzas esterilizadas por flameo o hipoclorito, vierta las muestras de agua y abra

el sistema de paso de aire. Dependiendo de la cantidad de sólidos disueltos en el agua, puede ser necesario cambiar los filtros constantemente. Almacene los filtros que se saturen en un tubo de 50 ml esterilizado, y use un solo tubo por muestra. Use pinzas esterilizadas para la manipulación de los filtros y durante todo el proceso use guantes y tapabocas. Por cada evento de filtración genere un control negativo de filtrado, usando agua destilada en lugar del agua superficial tomada de las unidades de muestreo. En los posteriores procesos, este control deberá ser tratado como una muestra más.

Una vez terminado el proceso de filtración preserve el tubo de 50 ml con los filtros a -20ºC o en seco con sílica gel sin que esta entre en contacto directo con la muestra.

Sedimento

Figura 6. Listado de materiales para toma de muestras de sedimento superficial. La selección entre el uso de una draga (6) o una pala (9) dependerá del cuerpo de agua.

A continuación, se describe un método para tomar muestras de sedimento superficial (alrededor de 10 cm de profundidad), siendo los materiales ilustrados en la figura 6. Al igual que para aguas, se propone una unidad de muestreo compuesta de tres muestras simples, tomadas a lo largo de un transecto en el cuerpo de agua, cuya longitud dependerá enteramente de la escala y el objetivo del estudio. Se recomienda tomar las muestras de sedimento en el mismo lugar de toma de muestras de aguas superficiales. Esta recomendación responde a la diferente información que cada tipo de muestra puede proporcionar (Turner et al. 2015, Deiner et al. 2017). Dependiendo de la profundidad del cuerpo de agua, en algunas ocasiones no es posible tomar la muestra de sedimento y agua exactamente en el mismo lugar. En estos casos, se recomienda tomar la muestra de sedimento lo más cerca posible al de agua hacia una zona de menor profundidad. En caso de tomar las muestras de agua y sedimento de forma simultánea, tome primero la de agua, de forma que evite que la toma de sedimento perturbe las condiciones del sitio, modificando las condiciones de la muestra de agua.

Si dada el área de su estudio selecciona un transecto de al menos 150 metros de longitud, tome las muestra de sedimento cada 50 metros. Las muestras de sedimento pueden ser tomadas usando diferentes utensilios, como palas, dragas o cilindros. En el lugar seleccionado tome sedimento superficial de al menos 10 cm de profundidad con cualquiera de los equipos mencionados, estando estos limpios y esterilizados en el caso de las palas o los cilindros. En el caso de la draga, use aquellas con compuertas superiores, de forma que usando una espátula o cuchara esterilizada pueda tomar la muestra del centro de la porción recolectada por la draga, la cual no ha entrado en contacto con la superficie y paredes de la draga y cuenta con una menor probabilidad de contaminación. No incluya en su muestra material orgánico como plantas, algas o raíces; retírelos de ser necesario. Almacene la muestra recolectada en una bolsa hermética o frasco, previamente esterilizado, e inmediatamente presérvela a bajas temperaturas usando una nevera con hielo, refrigerador o nitrógeno líquido y diríjase al lugar de procesamiento de las muestras o estación de trabajo. El procesamiento debe llevarse dentro de las 12 horas siguientes a la toma de muestra, siempre y cuando se asegure la preservación del ADN de la muestra con alguno de los métodos mencionados anteriormente.

Una vez en la estación de trabajo, homogenice, mezclando bien cada una de las muestras. De cada muestra, tome 15 g de sedimento para el protocolo de extracción de ADN extracelular y deposítelos en un tubo de 50 ml usando una espátula o cuchara previamente esterilizada. Para la extracción de ADN intracelular tome 250-500 mg y deposítelos en un tubo esterilizado de 2 ml. Cada muestra tomada en el punto o unidad de muestreo tendrá su correspondiente tubo de 50 ml o de 2 ml, a cada uno de los cuales se les realizará la extracción de ADN. Conserve los tubos de 50 ml o 2 ml con los 15 g o 250-500 mg correspondientes de sedimento a -20ºC hasta el momento de la extracción de ADN.

Suelo

Para el suelo, se describe un método para la toma de suelos superficiales (alrededor de 10 cm de profundidad). Los materiales son ilustrados en la figura 7. Se propone una unidad de muestreo con tres muestras compuestas, tomadas cada una entre cuadrantes dentro del área estipulada como unidad muestreal. El área de cada cuadrante dependerá enteramente de la escala y el objetivo del estudio.

Si dado el área de su estudio selecciona una unidad de muestreo de 1 ha, realice cuadrantes de alrededor de 30 m x 30 m que no se sobrepongan entre ellos. En cada cuadrante, genere una cuadrícula de tres puntos equidistantes, a lo largo y ancho, iniciando y terminando en los vértices del cuadrante. En este caso, serían 9 puntos separados cada 15 metros (figura 8). Usando guantes, tapabocas y una pala o barreno previamente esterilizado, obtenga una sub-muestra de suelo de cada uno de los puntos y deposítela en una bolsa plástica hermética. Las submuestras de cada cuadrante serán depositadas en la misma bolsa, que será posteriormente almacenada en una nevera con hielo, refrigerador o nitrógeno líquido, para preservarla adecuadamente hasta su procesamiento. Una vez finalice de tomar las submuestras de cada cuadrante diríjase a la estación de trabajo.

Figura 7. Materiales necesarios para la toma de muestra de suelos superficiales.

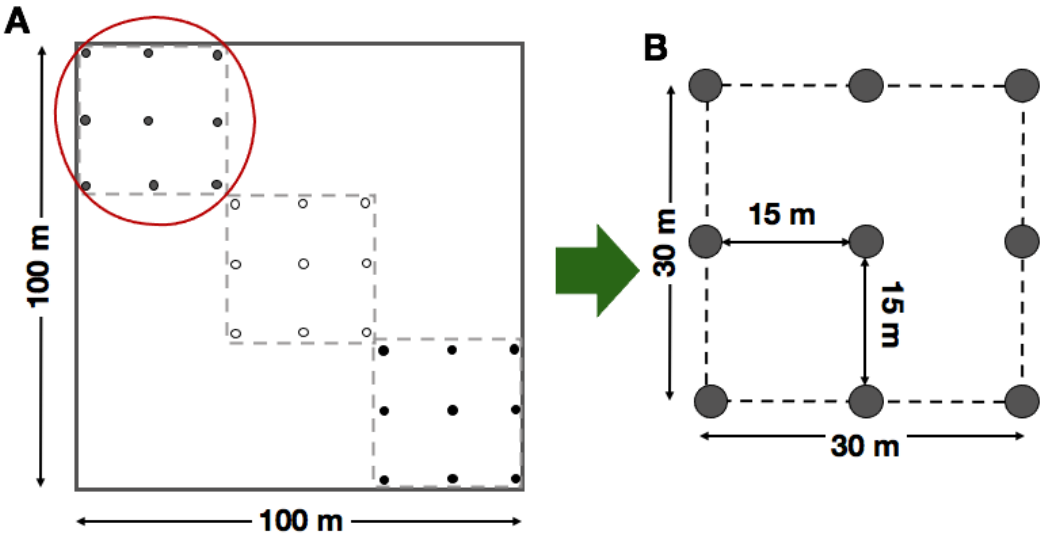

Figura 8. Esquema de lugares de toma de sub-muestras en parcelas de suelo. A) Unidad de muestreo (línea sólida) de 1 ha (100 m x 100 m) con un ejemplo de la ubicación de los tres cuadrantes (línea punteada) de 30 m x 30 m instalados en su interior sin sobreposición. La ubicación de los cuadrantes puede corresponder a la heterogeneidad del área. De cada cuadrante se obtendría una muestra compuesta. B) Distribución de la posición de las submuestras (círculos) dentro de un cuadrante. Se deben recolectar 9 submuestras de suelo distribuidas uniformemente. En este ejemplo, están separadas por 15 m.

En la estación de trabajo, homogenice mezclando bien cada una de las tres muestras. Para el protocolo de ADN extracelular de cada muestra, obtenga 15 g de suelo y deposítelo en un tubo de 50 ml usando una espátula o cuchara previamente esterilizada. Para el protocolo de ADN intracelular de cada muestra, obtenga 250-500 mg de suelo y deposítelo en un tubo de 2 ml usando una espátula o cuchara previamente esterilizada. Cada muestra tomada en el punto o unidad de muestreo tendrá su correspondiente tubo de 50 ml o 2 ml, dependiendo del protocolo, los cuales serán usados para la extracción de ADN. Conserve los tubos de 50 ml con los 15 g de suelo o los tubos de 2 ml con 250-500 mg de suelo a -20ºC o en seco con sílica gel, hasta el momento de la extracción de ADN. Recuerde que la sílica gel no debe entrar en contacto directo con la muestra.

Con las muestras tomadas y procesadas el paso siguiente es la extracción de ADN. Si este procedimiento no puede ser realizado en los días siguientes a la recolección de la muestra, es recomendable conservar la muestra al menos a -80ºC. Si la muestra está siendo conservada en seco, revise constantemente la sílica para evaluar si es necesario su cambio. En cualquier caso, evite congelar y descongelar el material más de una vez, y consérvelo en un lugar oscuro y sin exponerlo a constantes cambios de luz.

3.1.2.2 Extracción de ADN

Existen en el mercado diferentes kits comerciales especiales para extraer ADN de muestras ambientales, que han sido usados a través de diferentes estudios con metabarcoding. Generalmente, estos hacen distinción entre agua y suelos, este último también funcional en sedimento.

El Instituto Alexander von Humboldt ha realizado extracciones de ADN usando el kit comercial NucleoSpin™ Soil de Macherey-Nagel™ con modificaciones en el protocolo para extraer ADN extracelular, siguiendo a Taberlet et al. 2018. Este kit ha sido usado para los tres sustratos con resultados satisfactorios. Si el objetivo de la investigación es obtener ADN intracelular, se puede seguir el protocolo entregado por Macherey-Nagel™ en el kit. En este caso, es necesario tomar en cuenta el tipo de suelo para usar uno de los dos buffers incluidos. También, si se desea usar para extraer ADN de muestras de agua, es necesario pasar los filtros en una solución de buffer fosfato antes de usar el kit, y usar el sobrenadante como material de entrada para el proceso de extracción del ADN intracelular.

Sea cual sea el kit seleccionado, se recomienda que en un estudio de monitoreo o un estudio donde se realizarán comparaciones temporales, sea usado el mismo kit de extracción de ADN a través de todo el estudio.

A continuación se describe el protocolo usado por el Instituto.

Protocolo de extracción de ADN extracelular

El protocolo aquí descrito está dirigido a extraer ADN extracelular usando el kit NucleoSpin Soil de Macherey-Nagel (Taberlet et al. 2012), adaptado al trabajo en campo.

1. Preparación de la muestra.

1.1. Pesar 15 g de suelo o sedimento usando cucharas y recipientes descartables. En el caso de agua, tome todos los filtros usados durante el proceso de filtración.

1.2. Transfiera a un tubo de 50 ml.

2. Preparación del Buffer de Fosfato de Sodio (FS)

2.1. Pesar 1.97 g de fosfato de sodio monobásico (NaH2PO4) y 14.7 g de fosfato de sodio dibásico (Na2HPO4).

2.2. Lleve a volumen de 1 L con ddH2O (agua doblemente destilada)

2.3. Mida el pH, el cual debería estar cercano a 8, de lo contrario baje el pH usando NaOH o KOH. Para un pH cercano a 8 agregue 50 gotas de solución NaOH 5M.

2.4. Almacene a 4°C y use dentro de los 2-3 días siguientes.

3. Preparación muestra control

3.1. Agregue a un tubo de 1.5 ml 800 µl de buffer FS y 400 µl de buffer SB (incluido en el kit).

3.2. Mezcle por pipeteo o agitación por 5 s.

4. Liberación ADN

4.1. Agregue 15 ml de buffer FS a cada tubo de 50 ml.

3.1. Mezcle con agitación fuerte por 15 min.

3.3. Para muestras de suelos y sedimentos: tome 1.9 ml de sobrenadante (dos veces 950 µl) con puntas de 1000 previamente cortadas en la punta, y transferir a un tubo de 2ml. Centrifugue por 3 min a las máximas rpm. Luego transfiera 800 µl a un tubo de 2 ml.

Para muestras de agua: transfiera directamente 800 µl del sobrenadante a un tubo de 2 ml.

5. Extracción con kit NucleoSpin™ Soil

5.1. Agregue 400 µl de buffer SB a la muestra, agite por 5 s, y centrifugue 90 s a máx. rpm.

5.2. Ponga una columna (anillo verde) sobre un tubo recolector y agregue 600 µl de la mezcla anterior.

Nota: reutilice el tubo recolector para todo el procedimiento.

5.3. Centrifugue a máx. rpm por 90 s y descartar el filtrado.

5.4. Repita desde el paso 5.2 usando la misma columna (anillo verde) del paso anterior.

5.5. 1er lavado: agregue 500 µl de buffer SB. Centrifugue a máx. rpm por 90 s y descartar el filtrado.

5.6. 2do lavado: agregue 550 µl de buffer SW1. Centrifugue a máx. rpm por 90 s y descarte el filtrado.

5.7. 3er lavado: agregue 680 µl de buffer SW2. Agite 2 s por inversión. Centrifugue a máx. rpm por 90 s y descartar el filtrado.

5.8. 4to lavado: repita el paso anterior.

5.9. Centrifugue por 5 min a máx. rpm para secar la membrana.

5.10. Pase la columna a un tubo nuevo de 1.5 ml, con la tapa previamente removida. Selle con papel parafilm y almacene en seco con silicagel.

Recuerde incluir controles negativos de extracción por cada evento de extracción de ADN. Para esto siga el proceso anteriormente descrito sin incluir muestra alguna.

Protocolo de extracción de ADN intracelular

En el caso del ADN intracelular, siga las recomendaciones del fabricante, en este caso Macherey-Nagel™. La principal diferencia de este al protocolo anterior se basa en la cantidad

del material base. Mientras que para el protocolo de ADN extracelular se usan 15 g de suelo o sedimento, para ADN intracelular el material de base es menor 500 mg. Adicionalmente, este protocolo exige un paso de lisis de las células que se logra cambiando el buffer fosfato por el buffer inicial de lisis incluido en el kit, usado en un tubo con perlas para fragmentación que, a través de la agitación por vortex, facilita el rompimiento de las células. Así también, incluye una etapa de remoción de inhibidores (columna roja incluida en el kit). El volumen inicial del buffer de lisis del kit es de 700 µl, por lo cual para la extracción en agua es importante realizar una resuspensión del filtrado de los filtros con 15 ml del buffer fosfato y agitación por 15 minutos, y posteriormente tomar los 800 µl para seguir el procedimiento del kit con el protocolo de ADN intracelular.

3.1.2.3. Preparación de librerías

Los pasos siguientes a la extracción de ADN, son la amplificación de marcadores y la secuenciación. Estas fases en conjunto son conocidas como preparación de librerías.

Amplificación de marcadores

Durante la extracción de ADN, sea este intra o extracelular, se capturan todos los fragmentos de ADN que se encuentran en la muestra ambiental, sin embargo, no todos son útiles para la identificación de los organismos o taxones presentes. Es por esto que, a través de la reacción en cadena de la polimerasa o PCR (polymerase chain reaction), la cual genera millones o miles de millones de copias de la región de interés en pocas horas (Mullis 1990, figura 9), se amplifican los fragmentos de interés.

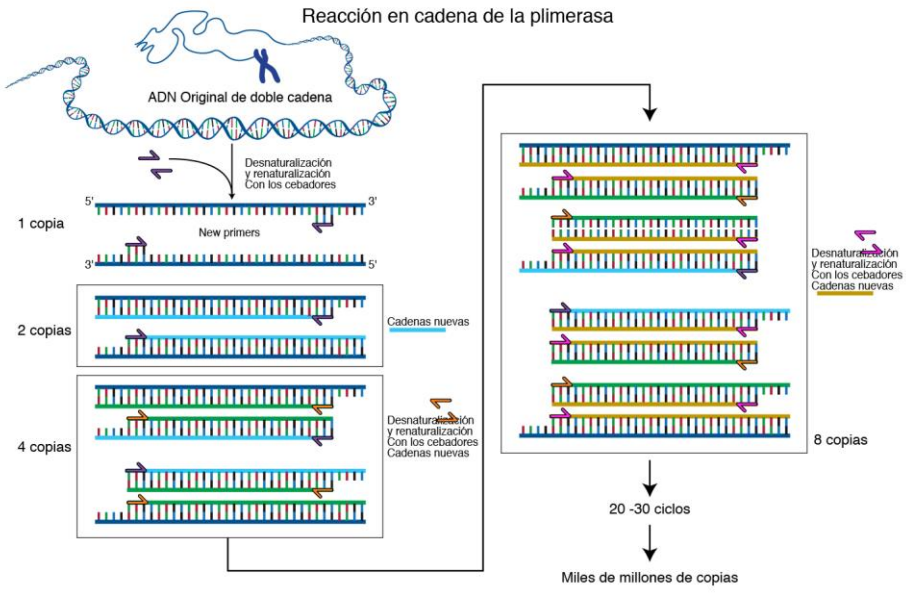

Figura 9. Esquema de la reacción en cadena de la polimerasa o PCR. Imagen tomada de National Human Genome Research Institute

En esta fase, uno de los puntos más importantes a considerar es la selección de los marcadores o regiones de interés para metabarcoding.

Actualmente existen diferentes marcadores moleculares usados para metabarcoding. La selección del marcador está directamente relacionada a la disponibilidad de bases de referencia que permiten realizar el enlace taxón-secuencia, y a la disponibilidad de *primers* que permitan amplificar la región o marcador a través del grupo o grupos de interés. Para algunos marcadores con buena referencia no se cuenta con *primers* generales que funcionen a través de diferentes grupos, lo que no los hace apropiados para metabarcoding, ya que la amplificación puede estar sesgada hacia unos grupos biológicos específicos. Un caso que ayuda a ejemplificar esto es el marcador COI. Esta región es usada cómo códigos de barras genéticos, por lo que cuenta con relativas buenas bases de datos para muchos grupos biológicos a través de diferentes regiones geográficas. La región usada como códigos de barras tiene una longitud aproximada de 650 pb (pares de bases), por lo que en metabarcoding se ha propuesto usar una región más pequeña que cubre alrededor de 150 pb. Sin embargo, hasta el momento no se ha logrado un diseño exitoso de *primers*, que permitan amplificar esta región a través de diferentes grupos biológicos, y los *primers* que existen actualmente tienden a funcionar en grupos específicos (como diferentes órdenes dentro de insectos). Es por esto por lo que la selección de marcadores es uno de los aspectos claves en estudios de metabarcoding, y la cual es resaltada consistentemente a través de diferentes estudios. Una recomendación que surge a través de los análisis revisados es la necesidad de buscar información que permita hacer un diagnóstico de cuál es un marcador ideal (longitud en pares de bases y bases de datos de referencia) y cómo funcionan los *primers* para el grupo de interés (que no existan sesgos hacia grupos particulares). Para evaluar el funcionamiento del primer, existen diferentes programas para hacer PCRs en silico que permiten identificar potenciales sesgos taxonómicos y evaluar el funcionamiento general del primer (tasa de amplificación, por ejemplo). Otra alternativa planteada es la inclusión de pruebas piloto en el proyecto, o el uso de diferentes *primers* en el estudio. Esto, sin embargo, repercute directamente en el presupuesto destinado al proyecto, por lo que en muchos casos no es una alternativa viable. En la tabla 1 se registran algunos de los *primers* comúnmente usados en estudios de metabarcoding.

Tabla 1. Ejemplos de secuencias de *primers* usados comúnmente en estudios de metabarcoding.

Grupo biológico	**Nombre**	**Secuencia *forward***	**Secuencia *reverse***	**Referencia**
Dominio Bacteria - Parcialmente Dominio Archaea	Bact01	GGATTAGA TACCCTGGTAG T	CACGACA CGAGCTGACG	Fliegerova et al. 2014
Reino Fungi	Fung02	GGAAGTAAAAG TCGTAACAAGG	CAAGAGATCCG TTGYTGAAAGT K	White et al. 1990
Dominio Eukarya	Euka02	TTTGTCTGSTTA ATTSCG	CACAGACCTGT TATTGC	Guardiola et al. 2015

Multiplex

Una vez seleccionado el marcador y los primers a ser empleados en el estudio, la estrategia típicamente empleada es la de hacer uso de etiquetas de oligonucleótidos acopladas a los primers del marcador seleccionado para la obtención de los *metabarcodes* (e.g. marcador 16S para bacterias, ITS para hongos, 18S para otros eucariotas). Estas etiquetas se adicionan en el paso de amplificación por PCR en una combinación única por muestra. Múltiples muestras son combinadas y procesadas por corrida de secuenciación de ADN, con el fin de optimizar el recurso de secuenciación. Al proceso de combinar múltiples muestras en en una corrida de secuenciación, mediante el uso de etiquetas, se le denomina 'multiplex' y permite discriminar nuevamente cada muestra en la fase analítica computacional (Taberlet et al. 2018).

Secuenciación de ADN

Los fragmentos obtenidos, amplificados con el marcador de interés e identificados por muestra con las etiquetas, deben ser secuenciados. Este proceso, realizado en equipos llamados secuenciadores, permite obtener la información del orden de las bases nucleotídicas de fragmentos de ADN de interés, en este caso, de los amplicones.

Actualmente, en el mercado se encuentran diferentes plataformas de secuenciación de alto rendimiento, cada una de las cuales difiere en cuanto al modo de generar y analizar los datos. Sin embargo, todas las plataformas tienen el objetivo fundamental de determinar en paralelo la secuencia (orden de las bases nucleotídicas) de múltiples fragmentos de ADN provenientes de una o más muestras. Para análisis de ADN ambiental se ha documentado el uso de diferentes tecnologías de secuenciación de alto rendimiento, siendo la plataforma de lecturas cortas Illumina, con su equipo MiSeq, una de las más usadas y estandarizadas para los análisis bioinformáticos subsecuentes.

3.1.2.4 Procesamiento bioinformático para metabarcoding

De igual manera, para el análisis computacional de los datos secuenciados también existe una amplia gama de métodos disponibles. Cada uno cuenta con sus especificidades y podrán ser algunos más apropiados que otros teniendo en cuenta los procedimientos realizados y las naturalezas de los datos. Para *metabarcoding* se destacan paquetes como: MOTHUR (Schloss et al. 2009), USEARCH (Edgar 2010), QIIME (Caporaso et al. 2010), DADA2 (Callahan et al. 2016), OBITools (Boyer et al. 2016), SLIME (Dufresne et al. 2019), VSEARCH (Rognes et al. 2016), entre otros. La gran mayoría de estos son rutinas computacionales que integran diferente software y etapas analíticas iniciando desde los archivos crudos obtenidos del secuenciador, típicamente en formato FASTQ, y normalmente finalizando en una tabla de comunidades (OTUs: Operational Taxonomic Units o ASVs: Amplicon Sequence Variants) con sus respectivas abundancias relativas para cada punto de muestreo. Estos paquetes en su mayoría se ejecutan a través de líneas de comando en sistemas operativos tipo UNIX (e.g. Linux, MacOSX) y dependiendo del volumen de datos a ser analizados es más conveniente el uso de clusters computacionales con prestaciones suficientes de espacio en disco, RAM y procesador para poder hacer un análisis eficiente.

El procesamiento bioinformático puede dividirse en cuatro etapas principales: A) manipulación básica de las lecturas del secuenciador, B) clasificación de las secuencias (clustering y/o asignación taxonómica), C) comparaciones experimentales (depuraciones

teniendo en cuenta réplicas y controles), D) análisis ecológicos (seleccionados según la pregunta de investigación).

En este documento se describe el flujo de trabajo bioinformático partiendo de muestras amplificadas y multiplexadas con etiquetas, secuenciadas mediante plataforma Illumina y analizadas computacionalmente con el *pipeline* OBITools v 1.2.13 (Boyer et al. 2016) y funciones del software estadístico R para las etapas de depuración en las que se examinan las muestras para la detección de contaminantes a partir de los controles usados.

Procesamiento bioinformático con ObiTools para metabarcoding.

1. Alineamiento de las secuencias

Cuando la tecnología de secuenciación empleada es Illumina, regularmente se generan datos pareados o *paired-end*. Básicamente esto consiste en que cada fragmento de ADN en el secuenciador se secuencia en dos rondas; una por el extremo *forward* (R1) y otro por el *reverse* (R2). Es decir, para cada fragmento de ADN habrá dos lecturas de secuenciación asociadas. Dependiendo del tamaño promedio de los fragmentos, dado por la longitud del amplicón correspondiente al marcador de interés y de las etiquetas más adaptadores de Illumina, y del número de ciclos de secuenciación (bases leídas por lectura) programadas va a haber una región sobrelapante que debe ser unida computacionalmente (figura 10). A este proceso se le llama alineamiento.

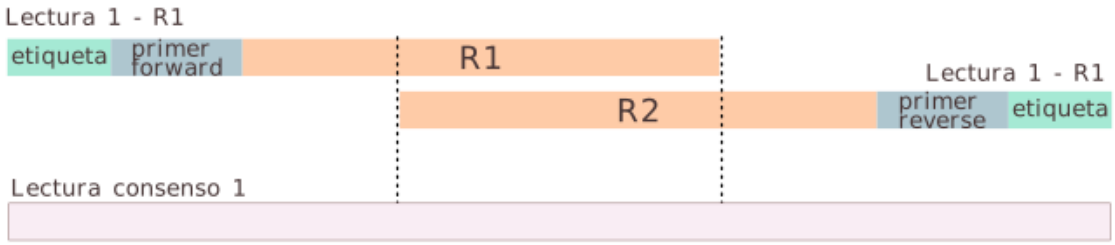

Figura 10. Esquema ilustrando el alineamiento de lecturas cortas pareadas.

El formato de archivo en el que vienen los datos desde el secuenciador es el FASTQ. Es un archivo de texto con múltiples líneas (4 por lectura) en la que están las secuencias de las lecturas generadas, los valores de calidad de cada base (escala Phred-score) y encabezados que contienen identificadores únicos para cada lectura.

De una secuenciación paired-end vendrán 2 archivos FASTQ con un número igual de lecturas entre ambos archivos, dado que están asociadas en pares. Entonces, el alineamiento tiene por objetivo unir la región sobrelapante de cada lectura entre los dos archivos de entrada.

El alineamiento se puede hacer especificando un puntaje. En caso de que no se especifique un valor, se recomienda hacer un histograma de puntajes (scores) de alineamiento y mantener sólo los que tengan puntajes altos.

Ejemplo:

```
illuminapairedend --score-min=40 -r data_R2.fastq data_R1.fastq > data.fastq
```

2. Remoción de secuencias no correctamente alineadas

Idealmente se espera que las lecturas R1 y R2 sobrelapen entre ellas para construir los consensos de cada una de las lecturas, sin embargo, en la práctica es posible que en algunos casos no se pueda realizar el sobrelapamiento de la forma apropiada; esto podría suceder por múltiples motivos, entre ellos errores de secuenciación o amplificación.

Con el fin de no introducir ruido al análisis, estos posibles errores deben ser removidos. Para esto se emplea la función *obigrep* con los parámetros de filtrado especificados.

La función *illuminapairedend*, del paso anterior, asigna una serie de etiquetas en el archivo a cada alineamiento o lectura, las etiquetas indican si el par de secuencias fue unido (‘joined’) y con qué nivel de calidad (score_norm) se realizó este alineamiento. Sólo se conservan las secuencias alineadas y con un score superior a ‘3.89’.

Ejemplo:

```
obigrep -p '(mode!="joined"") and (score_norm>3.89)' data.fastq > data.ali.fastq
```

3. Asignar secuencias a muestras (demultiplex)

Como se ha enunciado anteriormente, en una corrida de secuenciación para metabarcoding se suelen procesar juntas varias muestras a las que se le adicionan unas secuencias que funcionan como etiquetas moleculares (‘multiplex’). Para continuar el análisis es necesario discriminar las secuencias con la información del etiquetado molecular para determinar a qué muestras corresponden.

Para realizar este paso, se emplea la función *ngsfilter* y adicionalmente se debe tener un archivo de texto tipo tabla (especificado en la opción -t) con la relación de etiquetas y primers empleados durante la amplificación para cada muestra. Este archivo regularmente es enviado por la empresa que realiza la amplificación y secuenciación, o se puede construir con la información suministrada por la misma.

Nuevamente, por errores de secuenciación, entre otros, es posible que no todas las secuencias puedan ser asignadas a una muestra por problemas a la hora de reconocer las secuencias etiqueta, así que éstas sin asignar son descartadas en un archivo, con el nombre que se le indique en el parámetro -u.

Ejemplo:

```
ngsfilter -t data_ngsfilter.txt -u unidentified.fastq data.ali.fastq >   data.ali.assigned.fastq
```

4. Remoción de secuencias con bases ambiguas

Dentro de los pasos de depuración se considera la presencia de bases nucleotídicas ambiguas. Si bien, el ADN está conformado por cuatro nucleótidos (A: adenina, C: citosina, T: timina, G: guanina), a veces no hay certeza de la base exacta en una posición determinada, a estas se le denominan ambigüedades (i.e. R: guanina o adenina, Y: citosina o timina, N: nucleótido indeterminado, entre otros). Las secuencias que contengan al menos una base ambigua son removidas del análisis por considerarse no informativas. Para esto se emplea la función *obigrep*.

Ejemplo:

```
obigrep -s '^[acgt]+$' data.ali.assigned.fastq > data.na.fastq
```

5. Dereplicacion y eliminación de singletons

Un experimento de secuenciación puede generar cientos de miles, e incluso millones, de lecturas. El almacenamiento en disco es un recurso importante a tener en cuenta porque cada uno de los archivos generados tienen pesos de varios gigabytes. En este punto, en el archivo de trabajo tendremos muchas secuencias que son idénticas, así que con el fin de optimizar el recurso computacional se realiza un paso al que se le denomina ‘dereplicación’ empleando la función *obiuniq*. Consiste en dejar en el archivo una única secuencia de cada tipo y adicionar al archivo una etiqueta de conteo (‘count’) la cual tendrá el número de veces que esa misma secuencia se encontraba en el archivo original. Adicionalmente este atributo ‘count’ se emplea para los pasos subsecuentes de depuración.

Otro punto importante es la remoción de los singletons. Un singleton es una secuencia que está una única vez (count=1) en todo el experimento. Es muy poco probable que las secuencias con un soporte tan bajo correspondan a una realidad biológica así que en su mayoría su presencia se debe a errores introducidos durante la amplificación, secuenciación o formación del consenso. De modo que son removidas del análisis empleando la función *obigrep*.

Ejemplo:

```
obiuniq -m sample data.na.fastq > data.uniq.fastq
obigrep -p 'count>1' data.uniq.fastq > data.uniq.ns.fastq
```

6. Filtrado por tamaño y remoción de secuencias poco soportadas

El objetivo de las depuraciones es el de remover secuencias que no correspondan a la realidad biológica del sistema (e.g. errores de secuenciación, secuencias quiméricas, amplificaciones inespecíficas, entre otros). Un modo de controlar esto es verificando la distribución de tamaños de las secuencias generadas, así como su frecuencia y realizar la validación gráficamente a través de histogramas. ObiTools, a través de la función *obistats*, permite generar tablas con estadísticas que se pueden usar para construir gráficas (por ejemplo con R u otro software de graficación). Se realizaría de la siguiente manera:

```
obistat -c seq_length data.uniq.ns.fastq > stats_length.txt
obistat -c count data.uniq.ns.fastq > stats_count.txt
```

Con estas tablas se construyen gráficas como las siguientes:

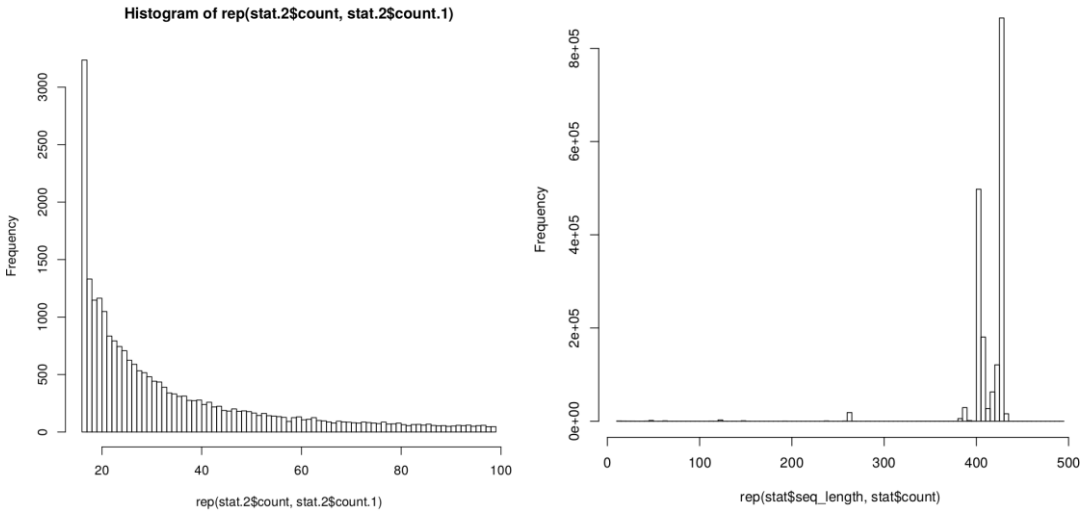

Figura 11. Ejemplo de histogramas creados en R a partir de los datos generados con la función *obistats* de ObiTools. Estas gráficas se usan para la visualización de estadísticos útiles en la depuración de datos.

Utilizando la función *obigrep*, basado en esta información y con el tamaño de amplicón esperado según los primers empleados, se puede establecer la longitud de tamaños de las secuencias que potencialmente no son errores y definir un umbral mínimo (parámetro -l) y máximo (parámetro -L) de secuencias a ser conservadas. Así como el valor mínimo del conteo en que se considerará como un 'soporte bajo' (especificado con 'count>X') para conservar las que tienen conteos mayores. Utilizando *obiclean* se realiza la identificación de posibles errores de PCR.

Ejemplo:

```
obigrep -l 380 -L 440 -p 'count>15' data.uniq.ns.fastq > data.filtlen.l380L440c15.fastq
obiclean -s merged_sample -r 0.05 -H data.filtlen.l380L440c15.fastq >
data.filtlen.l380L440c15.cleaned.fastq
```

7. Asignación taxonómica de las secuencias

Las secuencias depuradas en esta etapa pueden relacionarse con una base de datos previamente construida para hacer una asignación taxonómica (e.g. phylum, clase, orden, familia, género, especie). Para esto se emplea la función *ecotag* y un umbral de identidad especificado (similitud de la secuencia evaluada contra una de la base de datos con la que tenga correspondencia). La base de datos a emplear debe ser construida para el marcador de interés y los primers empleados en el experimento, para ello se emplea información pública que puede ser complementada con secuencias generadas por el investigador previamente. El proceso de construcción de las bases de datos para ObiTools se detalla en el numeral 16.

La asignación taxonómica de las secuencias, empleando un umbral bajo (60%) sería el siguiente:

```
ecotag -m 0.6 -d DB_emblPro_Ago2020 -R proka16S_refuniqueAgo2020.fasta -r
data.filtlen.l380L440c15.cleaned.fastq > data.taxoassigned.fasta
```

8. Clustering (generación de OTUs)

Para los análisis de metabarcoding es necesario definir las entidades genéticas representativas con las que se trabajará. En este documento se emplean las Unidades Taxonómicas Operativas (OTUs, en inglés) las cuales se obtienen por un proceso de agrupamiento o clustering basadas en similitud de secuencia. Las secuencias genéticas se agrupan entre sí basándose en su similitud de acuerdo a un umbral definido cuya idoneidad dependerá del marcador empleado y el grupo biológico de estudio. El umbral más usado para realizar el agrupamiento en bacterias empleando 16S es de 97%, esto quiere decir que todas las secuencias dentro de un OTU (grupo) tienen un 97% de similitud, y sólo tienen un 3% de diferencias. Luego, se retiene la secuencia correspondiente al centro del cluster para definir cada OTU, usando el comando *obigrep*.

Ejemplo:

```
sumaclust -t 0.97 -p 4 -e -o data.taxoassigned.fasta > data.clustered97fastq
obigrep -p 'cluster_center==True' data.clustered97fastq > data.clustered97center.fasta
```

9. Generación de tabla de comunidades

Una vez realizadas todas las depuraciones de secuencias, la asignación taxonómica y el clustering se procede a formar la tabla de comunidades para seguir siendo filtrada en los análisis siguientes. La tabla se genera empleando la función *obitab*, y tiene un formato delimitado por tabulaciones, archivo que puede ser cargado y leído en programas como Excel o R para los subsecuentes análisis. En resumen, esta tabla contiene los OTUs en las filas y en las columnas los atributos (como los respectivos conteos por muestra, asignación taxonómica, entre otros).

Ejemplo:

```
obitab -d -o data.clustered97center.fasta > data.obitab.tab
```

Con la generación de la tabla de comunidades finaliza el procesamiento en el programa ObiTools.

Procesamiento bioinformático post-ObiTools para metabarcoding

Con la finalización del procesamiento en ObiTools, se continúa con la depuración de secuencias a partir de las comparaciones experimentales que tienen en cuenta los controles y las réplicas de PCR. Esta depuración está dirigida a detectar y descartar potenciales contaminantes, errores o artefactos. El procedimiento descrito a continuación se ejecutará usando diferentes librerías del software R v. 4.0.2. (2020), basado en el guión construido por el Dr. Frédéric Boyer (frederic.boyer@univ-grenoble-alpes.fr) y la Dra. Lucie Zinger (zinger@biologie.ens.fr).

Para realizar la depuración usando este guión en R, es necesario contar con una tabla que describa las características de las muestras. Es decir, que para cada muestra - incluyendo controles - describa los códigos nomenclaturales de las muestras, las réplicas biológicas, técnicas o de PCR, el tipo de muestra, y el tipo de control.

```
# 'samples' es el archivo que describe las características de todas las muestras, incluyendo el tipo de muestra (control de extracción, blancos, muestras, etc)
# Cargar la tabla en R
setdw('/Users/user/Google_Drive/postObitools')
samples <- read.delim('Samples.csv', header=T)
head(samples)

# La columna 'Codigo' es el código nomenclatural de las muestras
rownames(samples) <- samples$Codigo

# Cargar la tabla resultado de ObiTools en R
tmp <- read.csv('/MatricesObitools/Eukarya.txt', h=T, sep="\t", stringsAsFactors = F)
```

10. Remoción de OTUs no identificados

El primer paso comprende una manipulación básica de limpieza de la tabla, eliminando columnas y filas que se consideren innecesarias, facilitando su posterior procesamiento. Cada fila corresponde a un OTUs que cuenta con una serie de información, incluyendo su asignación taxonómica, el grado de similitud con el que se realiza la asignación, y su jerarquía taxonómica. Aunque la asignación se ejecute bajo un umbral mínimo, por ejemplo en el caso de este documento bajo un umbral del 60%, en todos se registra el porcentaje de similitud de la

asignación, siendo todos aquellos asignados con una similitud menor al umbral fijado o no asignados, registrados como sin rango (“no rank”).

```
# Remover las columnas innecesarias para facilitar la manipulación. En este caso todas las columnas resultantes del comando obiclean en ObiTools
tmp <- tmp[,grep("obiclean_",colnames(tmp),inv=T)]

# Remover OTUs no identificados
tmp <- tmp[which(tmp$rank!="no rank"),]

# Remover OTUs asignados con una similitud menor al 90%. La columna de porcentaje de asignación es nombrada de acuerdo con el nombre de la base de datos usada, sin embargo, normalmente inicia con el término "best_identity". Antes de aplicar esta línea es necesario verificar el nombre de la columna.
if(length(grep("best_identity",colnames(tmp)))>1) print("Stop: You need to identify the column with the identity percentage")

colnames(tmp)[grep("best_identity",colnames(tmp))] <- "best_identity"
hist(tmp$best_identity)

# El umbral seleccionado es de 90%
tmp <- tmp[which(tmp$best_identity>=0.9),]
```

En este paso, también se puede filtrar por un porcentaje o grado de asignación taxonómica mínimo. Este umbral dependerá enteramente del grupo biológico, el sistema en el que se desarrolla el estudio y la base de datos de referencia usada. Es muy recomendado explorar los datos antes de tomar una decisión.

11. Identificación de artefactos o contaminantes

Para la identificación de potenciales artefactos, errores o contaminantes se usan tres objetos: i) la tabla que describe las características de las muestras; ii) la matriz de comunidades de OTUs; y iii) una tabla con la información de cada uno de los OTUs.

i) Tabla que describe las características de las muestras: La tabla que describe las características de las muestras se usará como plantilla para crear una tabla que almacene la información del tipo de la muestra, es decir si es una muestra o control y el tipo de control.

```
# Crear una tabla donde se especifica el tipo de muestra. En este caso se denominará 'samp'
samp <- samples[match(rownames(reads), rownames(samples)),]
samp$Control <- "Sample"
samp[which(samp$Tipo.de.muestra=="Positivo"),"Control"] <- "Positivo"
samp[which(samp$Tipo.de.muestra=="Extraccion"),"Control"] <- "Negativo"
samp[which(samp$Tipo.de.muestra=="PCR"),"Control"] <- "Negativo"
samp[which(samp$Tipo.de.muestra=="Blanco"),"Control"] <- "Blanco"
```

ii) Matriz de comunidades de los OTUs: Como se mencionó anteriormente, la tabla generada por ObiTools es un archivo con los OTUs ubicados en las filas y los atributos (conteos o lecturas por muestra, asignación taxonómica, entre otros) en las columnas. Para que esta sea usada como una matriz de comunidades se deben entonces, remover toda la información que no esté asociada a las lecturas y transponer la tabla, es decir pasar los OTUs a las columnas y dejar en las filas a las muestras.

```
# Crear la matriz de comunidades. En este caso denominaremos la matrix 'reads'
reads <- t(tmp[,grep("sample\\.", colnames(tmp))])
rownames(reads) = gsub("sample\\.", "", rownames(reads))

# Los OTUs (columnas) son nombrados de acuerdo con el nombre del agrupamiento al cual pertenecen
colnames(reads) = tmp$cluster
head(reads)
reads <- reads[which(!is.na(match(rownames(reads),rownames(samples)))),]
```

iii) Tabla con la información de cada uno de los OTUs: La información asociada a cada OTUs se guardará en un objeto para llevar un registro sobre el potencial como contaminante, artefacto o error de cada uno, además de la información más relevante en cuanto a su uso posterior. Esta tabla se crea con el objetivo de contener la información asociada a cada OTU en un objeto de menor tamaño y filtrado, de forma que sea más fácil su manipulación.

```
# Crear la tabla con la información de los OTUs. En este caso se denominará como 'motus'
motus <- tmp[,grep("sample\\.", colnames(tmp), invert = T)]

# Filtrar la información que se desea conservar en la tabla
idx <- c("cluster","count","best_identity", "order_name","family_name",
"genus_name","scientific_name","seq_length","sequence","taxid")
motus <- motus[, idx]

# Se ordena la información (las columnas en el caso de la matriz de comunidades y las filas en el caso de la tabla de OTUs) por abundancia. Esto es opcional
idx <- order(colSums(reads), decreasing=T)
motus <- motus[idx,]
reads <- reads[,idx]
```

Una vez se cuente con estos objetos, se puede proceder a la identificación de los contaminantes o artefactos.

Una buena práctica al explorar y manipular datos de *metabarcoding* es visualizar los datos. Eso permite entender mejor los datos y el efecto de la remoción de las lecturas u OTUs.

```
# Visualización número de lecturas versus el número de OTUs por muestra
```

```
plot(rowSums(reads), rowSums(reads>0),
     col=as.factor(samp[rownames(reads), 'Tipo.de.muestra']),
     log='xy', pch=16, xlab='#Reads', ylab='#OTUs')

legend('topleft',
       legend = levels(as.factor(samp[rownames(reads), 'Tipo.de.muestra'])),
       col=1:nlevels(as.factor(samp[rownames(reads), 'Tipo.de.muestra'])),
       pch=16, cex=0.5)
```

Los contaminantes son identificados a partir de los controles usados durante el procesamiento molecular, es decir, los controles negativos de extracción, de PCR, los controles positivos, los blancos, y si aplica, los controles usados desde el procesamiento de las muestras como durante la filtración de aguas. Existen diferentes estrategias para decidir cuando un OTUs es probablemente un contaminante o un artefacto; en este caso, consideraremos que un OTU tiene una alta probabilidad de no ser biológicamente real cuando su abundancia máxima (número de lecturas) se encuentre en alguno de los controles.

```
# Identificar los OTU "problemáticos", es decir, con máxima abundancia en controles

stopifnot(all(rownames(samp)==rownames(reads)))

maxInExtractionCtrl <- apply(reads[samp$Tipo.de.muestra=='Extraccion',], MARGIN=2,
function(x) max(x,na.rm = T))
maxInPCRCtrl        <- apply(reads[samp$Tipo.de.muestra=='PCR',], MARGIN=2,
function(x) max(x,na.rm = T))
maxInSamples        <- apply(reads[samp$Control=="Sample",], MARGIN=2, function(x)
max(x,na.rm = T))
maxInBlancos        <- apply(reads[samp$Control=="Blanco",], MARGIN=2, function(x)
max(x,na.rm = T))
maxInPos            <- apply(reads[samp$Control=="Positivo",], MARGIN=2, function(x)
max(x,na.rm = T))

df <- data.frame(maxInExtractionCtrl, maxInPCRCtrl,  maxInBlancos, maxInPos,
maxInSamples)

# Determinar el tipo de control asociado a los OTUs "problemáticos"
motus$bias <- c('Extraction','PCR','Blanco','Positivo',NA)[apply(df, MARGIN=1,
FUN=which.max)]

# Adicionar información sobre el OTU
infosCols <- c("count", "best_identity", "family_name", "genus_name","scientific_name",
"bias")

df <- cbind(df, motus[,infosCols])
```

```
# Visualizar el número de lecturas versus el número de OTUs por muestra de OTUs problemáticos. 'df' es un objeto que solo contiene OTUs potencialmente contaminantes o resultados de artefactos o errores
df <- df[!is.na(df$bias),]
df <- df[order(df$maxInExtractionCtrl+df$maxInPCRCtrl, decreasing=T),]

stopifnot(all(colnames(reads)==motus$cluster))

plot(rowSums(reads[, !is.na(motus$bias)]), rowSums(reads[,!is.na(motus$bias)]>0),
    col=as.factor(samp[rownames(reads), 'Tipo.de.muestra']),
    log='xy', pch=16, xlab='#Reads', ylab='#OTUs')

legend('topleft',
      legend = levels(as.factor(samp[rownames(reads), 'Tipo.de.muestra'])),
      col=1:nlevels(as.factor(samp[rownames(reads), 'Tipo.de.muestra'])),
      pch=16, cex=0.5)
```

Los OTUs identificados como potenciales contaminantes son marcados en la tabla que contiene la información de OTUs (‘motus’ en el guión) en la columna denominada *bias*.

12. Identificación de cambios de etiqueta o *tag switching* y remoción de su efecto

La preparación de librerías es el paso del procesamiento molecular donde se amplifica la región de interés y se adicionan los adaptadores para la secuenciación y las etiquetas. Las etiquetas son cadenas cortas de bases nucleotídicas que serán adicionadas a las secuencias de cada muestra, y que marcarán a que muestra pertenece cada secuencia (Zinger et al. 2019, Schnell et al. 2015). Sin embargo, durante la preparación de las librerías o durante la secuenciación se pueden presentar “saltos de etiquetas” o “tag jumps” o “tag switching”, en la cual, la etiqueta asignada a una muestra en particular se recombina con secuencias perteneciente a otras muestras (Zinger et al. 2019, Taberlet et al. 2018). Así también, los índices asociados a los adaptadores de secuenciación pueden sufrir “saltos de índices” o “index jumps” durante la secuenciación, generando el mismo efecto que los tag jumps (Kircher Sawyer & Meyer 2012). Aunque la tasa de saltos puede disminuirse a partir de estrategias durante la preparación de librerías y la secuenciación, son propios al uso de la tecnología y por tanto se hace necesaria a nivel bioinformático aplicar procedimientos para detectarlos e intentar remover su efecto en los datos.

La estrategia bioinformática aquí planteada sigue un sistema de doble etiqueta en las secuencias y pozos actuando como blanco de secuenciación, es decir, pozos de las placas de PCR dejadas vacías (sin reactivos ni muestras) que representan combinaciones de etiquetas sin usar. Haciendo uso de las secuencias reportadas en los blancos de secuenciación, las cuales serán consideradas como fugas, se realiza una estimación de la tasa de fugas por *tag switching*. Esto parte de la hipótesis de que la tasa de error estará directamente relacionada a la abundancia, es decir, hay más oportunidad de que una secuencia o OTU esté involucrado con un evento de *tag switching* entre este sea más abundante.

```
blks <- samp$Tipo.de.muestra=='Blanco'
blks[is.na(blks)] <- FALSE

OTUsInBlks <- colSums(reads[blks,])>0

plot(colSums(reads[blks,OTUsInBlks]), colSums(reads[!blks,OTUsInBlks]),
    log='xy', pch=16, col='#00000080',
    xlab='Sum of abundance in blanks', ylab='Abundance in total')

plot(apply(reads[blks,OTUsInBlks], MARGIN=2, FUN=max),
colSums(reads[!blks,OTUsInBlks]), log='xy', pch=16, col='#00000080',
    xlab='Max abundance in blanks', ylab='Abundance in total')

# OTUs en los blancos deben contener altos conteos (o número de lecturas)
boxplot(list('OTUs in blks'=colSums(reads[,OTUsInBlks]),
          'OTUs not in blks'=colSums(reads[,-OTUsInBlks])), outpch = NA,
main='Abundance of OTUs found and not found\nin blanks', ylab='Abundance', log='y')

stripchart(list(OTUsInBlk=colSums(reads[,OTUsInBlks]),
             OTUsNotInBlks=colSums(reads[,-OTUsInBlks])), vertical = T, method="jitter",
pch=16, cex=0.4, add=T)

wilcox.test(x=colSums(reads[,OTUsInBlks]),y=colSums(reads[,-OTUsInBlks]), alternative
= "greater")

plot(colSums(reads>0), colSums(reads),
    log='xy', pch=16, col='#00000080',
    xlab='#samples the OTU has count > 0', ylab='Abundance in total')
```

Una vez se estima la tasa de “fuga”, se puede proceder a la remoción de su potencial efecto. La estrategia aquí aplicada, considera la existencia de una relación lineal entre la abundancia de OTUs y la probabilidad de ser parte de un evento de *tag switching*, y se basa en disminuir el número de lecturas de cada OTU de acuerdo con su abundancia y la tasa de fuga estimada.

```
# Determinación del porcentaje de contribución media de "fuga" para todos los OTUs de
acuerdo con la información encontrada en blancos

totalCountsDueToLeaking <- colSums(reads[blks,OTUsInBlks])/sum(blks)*nrow(reads)

ratios <- totalCountsDueToLeaking / colSums(reads[,OTUsInBlks])
# El histograma permite la visualización de los datos y el potencial efecto de remover un
número determinado de lecturas
hist(ratios[ratios>0], breaks=1000)
hist(ratios[ratios>0], breaks=seq(0,max(ratios)+0.0001,by=0.0001), xlim=c(0,1.5))
thrLeak <- c(0, 1/10000, 5/10000, 1/1000, 1/100, 2/100, 3/100, 4/100, 5/100)
```

```
abline(v=thrLeak, col='red')

correctedCountsForBlks <- lapply(thrLeak, function(thr) {
  r <- sweep(reads[blks,], MARGIN=2, STATS=ceiling(colSums(reads)*thr/nrow(reads)),
FUN='-')
  r[r<0] <- 0
  r
})

names(correctedCountsForBlks) <- thrLeak
boxplot(lapply(correctedCountsForBlks, function(x) rowSums(x>0)), ylab='Remaining
#OTUs', las=2, cex=0.5, outpch = NA, main='Effect of leaking removal\non #OTUs for
blanks')
stripchart(lapply(correctedCountsForBlks, function(x) rowSums(x>0)), vertical = T,
method="jitter", pch=16, cex=0.4, add=T)

boxplot(lapply(correctedCountsForBlks, function(x) rowSums(x)), ylab='Remaining
#Reads', las=2, cex=0.5, outpch = NA, main='Effect of leaking removal\non #Reads for
blanks')


stripchart(lapply(correctedCountsForBlks, function(x) rowSums(x)), vertical = T,
method="jitter", pch=16, cex=0.4, add=T)

# 'thr' será el objeto que albergará el valor de "fuga" estimado a partir de la visualización de
los datos
thr <- 1/100

# Conteo de lecturas para remover en cada una de las muestras
toRemove <- ceiling(colSums(reads)*thr/nrow(reads))

correctedCounts1 <- sweep(reads, MARGIN=2, STATS=toRemove, FUN='-')
correctedCounts1[correctedCounts1<0] <- 0

# Visualización del impacto de la remoción de lecturas
plot(rowSums(reads), rowSums(correctedCounts1), xlab='#Reads before cleaning for
leaking',
    ylab='#Reads after cleaning for leaking', main='Effect of leaking removal on #Reads')
abline(a=0,b=1)

plot(rowSums(reads>0), rowSums(correctedCounts1>0), xlab='#OTUs before cleaning for
leaking',
    ylab='#OTUs after cleaning for leaking', main='Effect of leaking removal on #OTUs')
abline(a=0,b=1)
```

13. Eliminando potenciales errores y contaminantes

En este paso se procede a eliminar todos aquellos OTUs marcados como potencialmente contaminantes y errores, a través de los procedimientos anteriores. Además, se eliminan los OTUs con muy baja abundancia ya que existe una alta probabilidad de que estos sean producto de errores de PCR o de secuenciación. Sin embargo, también es posible que estas secuencias de baja abundancia representen especies raras naturalmente de baja abundancia en el sistema estudiado. La decisión de eliminar estas secuencias debe ser tomada para cada juego de datos y debería depender del comportamiento de los datos y del objetivo de la investigación. En este guión se procede a eliminar los OTUs con una abundancia menor a un umbral determinado de acuerdo con el total de los datos recuperados. El umbral debe ser evaluado para cada caso.

```
# Removiendo OTUs con una abundancia menor al 5% (cuantil 0.05) del total de los datos. Estos, son los OTUs cuya abundancia esté por debajo del cuantil 0.05 del total de las lecturas por OTU.
motus[which(colSums(reads)<=quantile(colSums(reads),0.05)),"bias"] <- "LowAbund"
```

En el paso 11, en la columna *bias* de la tabla 'motus' se marcaron los OTUs problemáticos, es decir que podrían ser producto de contaminación. Haciendo uso de esta información se eliminarán estos OTUs de la matriz de comunidades.

```
# Conteo de OTUs problemáticos teniendo en cuenta su potencial origen
table(motus$bias)

# Ajustando el número de lecturas de los OTUs problemáticos a cero
correctedCounts2 <- correctedCounts
correctedCounts2[,!is.na(motus$bias)] <- 0

plot(rowSums(reads), rowSums(correctedCounts2), xlab='#Reads before cleaning for contaminants', ylab='#Reads after cleaning for contaminants', main='Effect of contaminants removal\non #Reads', col=as.factor(samp[rownames(reads), 'Tipo.de.muestra']), pch=16)
abline(a=0,b=1)

legend('topleft', legend = levels(as.factor(samp[rownames(reads), 'Tipo.de.muestra'])), col=1:nlevels(as.factor(samp[rownames(reads), 'Tipo.de.muestra'])), pch=16, cex=0.5)

plot(rowSums(reads>0), rowSums(correctedCounts2>0), xlab='#OTUs before cleaning for contaminants', ylab='#OTUs after cleaning for contaminants', main='Effect of contaminants removal\non #OTUs', col=as.factor(samp[rownames(reads), 'Tipo.de.muestra']), pch=16)
abline(a=0,b=1)

legend('topleft', legend = levels(as.factor(samp[rownames(reads), 'Tipo.de.muestra'])), col=1:nlevels(as.factor(samp[rownames(reads), 'Tipo.de.muestra'])), pch=16, cex=0.5)
```

14. Eliminación de PCRs fallidas

Para estudios de *metabarcoding* se recomienda tener múltiples réplicas de PCR para cada muestra (Taberlet et al. 2018). En este, el último paso de la depuración, se detectarán PCRs fallidas haciendo uso del sistema de múltiples réplicas de amplificación. Una PCR puede considerarse fallida de acuerdo con una muy baja abundancia de lecturas o una composición muy diferente a las demás réplicas de amplificación de su misma muestra. En este caso, se parte de un sistema de tres réplicas de PCR.

El primer paso será detectar PCR fallidas de acuerdo a su abundancia. Para esto, es necesario seleccionar un umbral de número de lecturas mínimo bajo el cual se considerará que la PCR no funcionó de forma adecuada. Este umbral puede seleccionarse a partir de las recomendaciones realizadas en la literatura o del comportamiento de los datos. En este caso, se realizará una exploración de los datos para seleccionar el umbral.

```
l <- lapply(levels(as.factor(samp[rownames(correctedCounts2), 'Control'])), function(st) {
rowSums(correctedCounts2[samp[rownames(correctedCounts2), 'Control']==st,])})
names(l) <- levels(as.factor(samp[rownames(correctedCounts2), 'Control']))

boxplot(l, las=2, cex=0.25, outpch=NA, main='#reads before filtering')
stripchart(l, vertical = T,method="jitter", pch=16, cex=0.4, add=T)

# A partir de la exploración de los datos se seleccionó un umbral de 5%. El valor del cuantil
de 0.05 será almacenado en el objeto 'thrCount'
thrCount <- quantile(rowSums(correctedCounts2[samp[rownames(correctedCounts2),
'Control']=="Sample",]),0.05)
abline(h=thrCount, col='red')

# Marcando las PCR with muy pocas lecturas
samp$empty_PCR <- FALSE

stopifnot(all(rownames(samp)==rownames(correctedCounts2)))
samp$empty_PCR[rowSums(correctedCounts2)<thrCount] <- TRUE
```

El segundo paso será identificar y remover aquellas réplicas de PCR que sean muy diferentes a las demás réplicas de la misma muestra. Se pueden usar diferentes métodos para realizar este procedimiento. Aquí, nos basaremos en un análisis de correspondencia para calcular la distancia, de acuerdo con la composición de OTUs, entre réplicas de la misma muestra (intra-muestra) y respecto a réplicas de otras muestras (inter-muestra). Se descartarán aquellas réplicas de PCR que presenten una distancia inter-muestras menor a la intra-muestras, usando comparaciones con un modelo nulo. En este guión se usarán las distancias euclidianas de los dos primeros ejes del análisis de correspondencia realizado sobre datos transformados con raíz cuadrada. Esto, sin embargo, dependerá de cada investigador.

```
require(vegan)
```

```
require(ade4)

# Función para remover PCRs fallidas en una sub-matriz de distancias de las réplicas de una misma muestra.

## ----- Función construida y escrita por Frédéric Boyer (frederic.boyer@univ-grenoble-alpes.fr)

identifyBad <- function(subMat, thr) {
  toSuppress <- c()
  if (any(subMat>thr)) {
    if (nrow(subMat)==2) {
      toSuppress <- c(toSuppress, rownames(subMat))
    } else {
      toSuppress <- c(colnames(subMat)[which.max(colSums(subMat))],
                      identifyBad(subMat[-which.max(colSums(subMat)),
                                         -which.max(colSums(subMat))], thr))
    }
  }
  return(toSuppress)
}

# Las réplicas disímiles serán marcadas en la tabla 'samp' en la columna nonReplicating
samp$nonReplicating <- FALSE
# Se ejecuta la función y el marcaje para cada muestra
i <- 0
repeat {
  i <- i+1
  print(paste('Iteration',i))

  dataM <- correctedCounts3[samp$Control=='Sample' & ! samp$nonReplicating,]

  h <- ade4::dudi.coa(sqrt(dataM), scannf=F, nf=2)

  #---
  cols = 1:nlevels(as.factor(samp[rownames(h$li),'Tipo.de.muestra']))
  nbReads <- rowSums(correctedCounts3[rownames(h$li),])

  plot(h$li, col=cols[as.factor(samp[rownames(h$li),'Tipo.de.muestra'])], pch=16,
       main=paste('Correspondance analysis\non sqrt transformed data\nIteration',i),
       cex=nbReads/max(nbReads)*1)

  #---

  distM <- as.matrix(dist(h$li))
```

```
  replicates <- gsub(pattern='[a-c]$',",rownames(distM))

  withinReplicates <- outer(replicates, replicates, FUN="==") & upper.tri(distM)
  notWithinReplicates <- outer(replicates, replicates, FUN="!=") & upper.tri(distM)

  d1 <- density(distM[withinReplicates], from=0, to=max(distM), n=1000)
  d2 <- density(distM[notWithinReplicates], from=0, to=max(distM), n=1000)

  plot(d1$x, d1$y, type='l', xlab='Distances', ylab='Density',
       main=paste('Distances densities\nIteration',i))
  lines(d2, col='red')
  thrDist <- d2$x[min(which(d1$y<d2$y))]
  abline(v=thrDist, col='red')

 # Es necesario revisar la nomenclatura de las réplicas de PCR
  needToBeChecked <- unique(gsub('[a-c]$',",rownames(which(distM>thrDist &
withinReplicates,
                                                arr.ind=T))))
  if (length(needToBeChecked)>0) {
    for (s in needToBeChecked) {
      pattern <- paste0('^',s)
      subMat <- distM[grep(rownames(distM), pattern = pattern),
                      grep(colnames(distM), pattern = pattern)]
      samp$nonReplicating[rownames(samp) %in% identifyBad(subMat, thrDist)] <- TRUE
    }
  }
  else {
    break;
  }
}
```

Finalmente, una vez se han identificado estas réplicas de PCR, se visualiza su impacto en los datos, y se eliminan de la matriz de comunidades. Este paso finaliza el proceso de depuración.

```
# Eliminar aquellas réplicas marcadas con FALSE en la columna nonReplicating de 'samp':
¿Son ellas réplicas de la muestra? Falso
# Revisar la nomenclatura de las réplicas de PCR
tt <- table(table(gsub('[a-c]$',",rownames(samp)[!samp$nonReplicating &
samp$Control=='Sample'])))
barplot(tt, main = '#kept replicates')

samp[which(samp$nonReplicating==FALSE),]

cols = 1:nlevels(as.factor(samp[rownames(h$li),'Tipo.de.muestra']))
```

```
nbReads <- rowSums(correctedCounts3[rownames(h$li),])
plot(h$li, col=cols[as.factor(samp[rownames(h$li),'Tipo.de.muestra'])], pch=16,
main='Correspondance analysis\non sqrt transformed data\nfor the kept
replicates',cex=nbReads/max(nbReads)*1)
legend('topleft', legend = levels(as.factor(samp[rownames(h$li),'Tipo.de.muestra'])),
pch=16,
       col=cols, cex=0.75)

samp$RemovePCR <- FALSE
samp[which(samp$empty_PCR==TRUE | samp$nonReplicating==TRUE),"RemovePCR"]
<- TRUE

matrix.final <- correctedCounts2[!samp$RemovePCR & samp$Control=='Sample',]
matrix.final <- matrix.final[,which(colSums(matrix.final)>0)]
```

15. Construyendo la matriz final de comunidades

Tras realizar la depuración, queda finalizar la matriz de comunidades. Para esto es necesario ensamblar las muestras a partir de los OTUs presentes en las diferentes réplicas de PCR que quedaron a partir del filtro.

Una alta proporción de OTUs son únicos en cada réplica de PCR (Alberdi et al. 2018). Es por esto, que es necesario decidir si todos los OTUs son tomados en cuenta o si por el contrario se toman en cuenta OTUs que se encuentren en más de una de las réplicas por muestras o incluso en todas las réplicas. De manera similar, sucede con el número de lecturas, ya que estas pueden sumarse, promediarse, etc. Cada elección tiene diferentes implicaciones (Alberdi et al. 2018, Taberlet et al. 2018) y estas deben ser evaluadas por los investigadores.

En este guión se ensamblarán las muestras a partir de las réplicas de PCR siguiendo una estrategia semi-estricta, tomando en cuenta sólo aquellos OTUs que se encuentren en al menos dos de las tres réplicas de PCR, sumando el número de lecturas. Aquellas muestras que quedaron solo con una réplica de PCR después de la remoción de PCRs fallidas no serán tomadas en cuenta.

```
names.samples <- gsub('[a-c]$','',rownames(matrix.final))
unique.names.samples <- unique(names.samples)
new.matrix <- matrix(ncol=ncol(matrix.final),nrow=length(unique.names.samples))
for(i in 1:length(unique.names.samples)){
  matrix.i <- matrix.final[which(names.samples==unique.names.samples[i]),]
  if(any(class(matrix.i)=="matrix")){
    pres.new.matrix.i <- matrix.i
    pres.new.matrix.i[pres.new.matrix.i>0] <- 1
    sum.i <- colSums(matrix.i)
    sum.i[which(colSums(pres.new.matrix.i)<=1)] <- 0
    new.matrix[i,] <- sum.i
```

```
  }
  else new.matrix[i,] <- rep(0,ncol(new.matrix))
  print(i)
}
rownames(new.matrix) <- unique.names.samples
colnames(new.matrix) <- colnames(matrix.final)
new.matrix <- new.matrix[,which(colSums(new.matrix)>0)]
new.matrix <- new.matrix[which(rowSums(new.matrix)>0),]
hist(rowSums(new.matrix), breaks=100)
```

Una vez que la matriz de comunidades por muestra es generada, es importante explorar la distribución de lecturas por muestra. Si las lecturas están desbalanceadas, es decir existen grandes diferencias en el número de lecturas entre muestras, algunos autores recomiendan hacer una rarefacción, ya que el desbalance puede afectar algunos de los análisis e índices ecológicos normalmente usados. Sin embargo, se debe tener presente que al realizar una rarefacción se pierde una gran cantidad de información lo que podría llegar a afectar las inferencias y conclusiones obtenidas.

```
library(vegan)
hist(rowSums(new.matrix.hongos), breaks=100)
# Se realiza una rarefacción con base en el número de lecturas mínimo reportado entre
todas las muestras
std.matrix <-
t(apply(new.matrix.hongos,1,function(x){rmultinom(1,min(rowSums(new.matrix.hongos)),
as.matrix(x))}))
rownames(std.matrix) <- rownames(new.matrix.hongos)
colnames(std.matrix) <- colnames(new.matrix.hongos)
rowSums(std.matrix)

# El objeto std.matrix constituye la matriz final de comunidades

# Finalmente, la tabla 'motus' -en la que se guardó el registro de la información asociada a
cada OTU- puede ser filtrada dejando los sólo OTUs incluidos en la matriz final de
comunidades. Esta tabla puede ser usada para completar la información taxonómica o hacer
análisis por jerarquías taxonómicas particulares
rownames(motus) <- motus$cluster
otus <- colnames(std.matrix)
motus.final = motus[match(otus, rownames(motus)),]

############################### Fin ################################
```

Tras este procedimiento se obtiene la matriz final de comunidades, que puede ser usada para realizar diferentes análisis estadísticos o descriptivos.

16. Construcción de bases de datos

Hay varias estrategias para la construcción de las bases de datos. Específicamente, una base de datos de referencia es un conjunto de secuencias genéticas, correspondientes a la región de interés, de alta calidad y con una identificación plena del organismo al cual pertenece. Aunque idealmente la información no debe ser redundante, una mayor representación de diferentes variantes de la misma región para un mismo grupo biológico contribuirá a realizar una mejor aproximación sobre el organismo encontrado en la muestra ambiental.

Actualmente, existen bases de referencia construidas disponibles para su uso, las cuales son constantemente actualizadas y depuradas para diferentes grupos biológicos. Este es el caso de SILVA (Quast et al. 2013; https://www.arb-silva.de/), la cual comprende secuencias de alta calidad de las regiones 16S/18S y 23S/28S del ARN ribosomal de representantes de los dominios de Bacteria, Archaea y Eukarya. También es el caso de UNITE (Nilsson et al. 2018; https://unite.ut.ee/), una base de datos que alberga la región ribosomal nuclear ITS para el reino Fungi (hongos), o BOLD Systems (https://www.boldsystems.org/; Ratnasingham & Hebert 2007), repositorio de la región mitocondrial COI para un alto número de taxones. Estas bases de datos mencionadas son sólo algunos ejemplos de iniciativas interdisciplinarias globales, que usan información genética pública, depurada de acuerdo con diferentes estándares de calidad que incluyen la calidad de la secuencia, apropiada y completa anotación, asignación por expertos, y trazabilidad desde el espécimen del cual se obtiene el ADN (Nilsson et al. 2018, Taberlet et al. 2018, Quast et al. 2013, Ratnasingham & Hebert 2007).

Una alternativa al uso de bases de datos globales consolidadas es la construcción de bases de datos personalizadas y propias que estén dirigidas a cubrir los organismos de interés (Taberlet et al. 2018). Esto puede ser particularmente útil en casos donde las secuencias genéticas de los organismos presentes en el área de estudio son escasas. Por ejemplo, en Colombia donde la información genética para muchas especies no es conocida incluso en grupos altamente estudiados como vertebrados, pero aún más en especies de microorganismos, los cuales presentan una alta diversidad y pueden presentar una alta variación genética en distancias geográficas relativamente cortas.

La construcción de bases de datos puede realizarse de diferentes formas, como por ejemplo a través del uso de secuencias publicadas en diferentes repositorios como los pertenecientes a la Colaboración Internacional de Bases de datos de Secuencias de Nucleótidos (International Nucleotide Sequence Database Collaboration, INSDC). A esta colaboración pertenecen ENA (European Nucleotide Archive en EMBL, https://www.embl.de/), GenBank (en NCBI, https://www.ncbi.nlm.nih.gov/genbank/) y DNA DataBank of Japan (DDBJ, https://www.ddbj.nig.ac.jp/index-e.html), tres de los repositorios más grandes de secuencias genéticas. Dado que los repositorios son diariamente sincronizados, al acceder a uno de estos, se asegura tener acceso a una gran cantidad de información. La diferencia entre descargar la información de uno u otro puede radicar principalmente en los formatos de los archivos.

Por compatibilidad de formato, con OBITools se sugiere el uso de las bases de datos nucleotídicas a partir de la información depositada en el repositorio de EMBL. En el siguiente ejemplo se muestra cómo se construye la base de datos de bacterias con el marcador 16S, usando secuencias estándar (STD) de procariotas (PRO) de EMBL.

```
################################## Inicio ###################################
### Construcción de bases de datos para Bacterias
## Descarga de secuencias para procariotas (PRO) desde EMBL

wget "ftp://ftp.ebi.ac.uk/pub/databases/embl/release/std/rel_std_pro_*.dat.gZ"
gunzip rel_std_pro_*
# 699964 sequences
```

```
## Descarga de la información de taxonomía asociada a la base de datos. En el caso de los repositorios de EMBL es taxonomía asociada a NCBI (https://www.ncbi.nlm.nih.gov/taxonomy).
mkdir TAXO
cd TAXO
wget ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz
tar -zxvf taxdump.tar.gz
```

obiconvert es una función de OBITools que permite la conversión de formatos para filtrar las secuencias pertenecientes a la región de interés.

```
obiconvert --embl --skip-on-error ./rel_std_pro_* -t ~/databases/EMBL/TAXO/ --ecopcrdb-output=bactDB_2020
```

Una vez la base de datos se encuentra en un formato legible para ObiTools se usa la función *ecoPCR*, la cual realiza una PCR (reacción en cadena de la polimerasa) en sílico, donde se amplificará la región de interés en las secuencias disponibles usando la secuencia de los cebadores usando durante el trabajo de laboratorio. Esta operación permite filtrar la base de datos, dejando solo la región de interés y por tanto aumentando la precisión de asignación, y disminuye el tamaño de la base de datos, haciéndola de más fácil manipulación. Este paso, sin embargo, es suprimible. Si se filtra la base de datos, por cada juego de primers usados debe realizarse esta operación para restringir la base de datos, incluso si amplifican la misma región. Por ejemplo, para dos juegos de datos en donde se obtendrá la secuencia de un mismo gen por metabarcoding usando primers diferentes durante la preparación de librerías, se debe construir una base de datos para cada juego, realizando una PCR en sílico con cada par de primers. Esto aumentará la probabilidad de asignación de cada caso.

```
conda install ecopcr
conda update ecopcr
ecoPCR -d bactDB_2020 -e 3 -l 100 -L 1000 GGATTAGATACCCTGGTAGT CACGACACGAGCTGACG > bactDB_2020.ecopcr
```

Una vez los resultados de la PCR están listos, estos se convierten a formato FASTA que permite la edición y depuración de secuencias.

```
obiconvert --ecopcr bactDB_2020.ecopcr --fasta-output > bactDB_2020.fasta
```

Además de los archivos que conformarán la base de datos, se construirá un archivo de referencia donde se agrupará toda la variación disponible para cada taxón identificado. Se agrupan las secuencias por su identificador taxonómico. Este procedimiento implica la agrupación de secuencias por su identificador taxonómico, para posteriormente dejar sólo un representante por taxón, y eliminar las secuencias con identificadores inválidos o duplicados.

```
obiannotate -S idtax:'taxid' bactDB_2020.fasta > bactDB_2020_id.fasta
obiuniq -m taxid bactDB_2020_id.fasta > bactDB_2020_unique.fasta
obiannotate -R idtax:taxid bactDB_2020_unique.fasta > bactITS_2020_unique.fasta
obigrep -A taxid bactITS_2020_unique.fasta > bactITS_2020_uniqueID.fasta
obiannotate --uniq-id bactITS_2020_uniqueID.fasta > bactITS_2020__Ref.fasta

# bactITS_2020__Ref.fasta es el archivo final de referencia que será usado en la asignación taxonómica.
############################### Fin ###############################
```

Este procedimiento contribuye a generar bases de datos de referencia basados en información depositada en repositorios públicos. Sin embargo, de acuerdo con el sistema biológico de interés y de estar disponibles, esta base de datos debe alimentarse con información de secuencias generadas de organismos locales o cercanos.

Una base de datos con secuencias locales o propias puede ser construida fácilmente siempre y cuando la secuencia sea de buena calidad, tenga alta confiabilidad en la identificación, se encuentre soportada por un voucher -idealmente- y finalmente, que tenga un identificador único taxonómico o *taxid*.

El *taxid* es un número único que el consorcio NCBI asigna a cada uno de los taxones que ingresan a su base de datos. Este aplica para todas las jerarquías taxonómicas (dominio, filo, reino, familia, género, especie, entre otros) y es único e irrepetible, lo que permite evitar confusiones dadas por sinonimias, nombres mal escritos, incompletos, u homónimos entre diferentes grupos biológicos. La información relacionada a los *taxids* se encuentra en la página oficial de taxonomía de NCBI (https://www.ncbi.nlm.nih.gov/taxonomy). Esta información se encuentra relacionada en los archivos *taxdump*, que hacen parte de las bases de datos ya construidas por NCBI y sus aliados (EMBL y DDBJ), y son usadas para la construcción de bases de datos de acuerdo con el procedimiento descrito anteriormente.

Para la construcción de base de datos con secuencias locales, cada secuencia debe estar en formato fasta y debe tener asignado el número correspondiente a su taxón o *taxid* en el encabezado. Algunas herramientas como la función *obiaddtaxids* del programa OBITools (https://pythonhosted.org/OBITools/scripts/obiaddtaxids.html) pueden realizar la asignación y anotación de forma automática basado en el nombre científico. En este caso, es preciso verificar que los nombres científicos sigan la taxonomía de NCBI para evitar que las secuencias no sean asignadas o se presenten asignaciones erróneas. También puede realizarse de forma manual.

En ambos casos, es importante tener presente que el encabezado de las secuencias debe contener en el primer argumento y diferenciado por el nombre 'taxid' el número único de NCBI, de lo contrario el programa no realizará la relación. El resto de los argumentos del encabezado deben ser separados por punto y coma (;).

A continuación se ilustra un ejemplo de secuencia en el formato requerido para la construcción de la base de referencia con secuencias locales:

```
>BBIHF001-19 taxid=1715255; Xylaria
GTCTCCGTTGGTGAACCAGCGGAGGGATCATTAAAGAGTTTTAAAACTCCCATA
CCCTGTGAATATACCTATACGTTGCCTCGGCAGGCTGTATCTCCCTTGCTATAA
ACTCGGTAAAAGGTGACTTTTACTGAGCTAGATGCCTGGGAGGTGCCCTGATG
GTCTGCCGGCGGCCCTTTAACTCTGTTTTATATTTAAATTCTGAGGCCCTAGTAA
ATATATTAAAACTTTCAACAACGGATCTCTTGGCTCTGGCATCGATGAAGAACG
CAGCGAAATGCGATACGTAATGTGAATTGCAGAAGTTAGTGAATCATCGAATC
TTTGAACGCACATTGCGCCCGCTAGTATTCTAGCAGGCATGCCTGTTCGAGCGT
CATTTCAACCCTTCAGCCTTTGTTGCTGAGTGTTGGGAGGCTATGCGTCAGCGT
ATCTCCTGAAATGCAGTGGCGGAGTTCGGTCAGCTCTAGACGTAGTAAAATCTT
TTATATCGCCTGTGAGTTAGACGGTCTACGGCCATAAAATTCCCATATTTTTAA
AGGTTGACCTCGAATCAGGTAGGGTTACCCGCTGAACTTAAGCATATCAATAA
```

Para unir las secuencias al archivo generado a partir de la información encontrada en EMBL, las secuencias en formato fasta deben convertirse al mismo formato. Para esto, sobre el archivo fasta con las secuencias locales se pueden correr las funciones *obiconvert* y *ecoPCR* del programa OBITools, tal cual se describió en el procedimiento anterior. El archivo fasta resultante se concatena al archivo que contiene las secuencias obtenidas de EMBL en formato OBITools, y se continúa con el procedimiento de limpieza y la obtención del archivo de referencia.

3.1.3 Estudio de caso: Diseño e implementación de un programa de monitoreo para evaluar el estado de la biodiversidad presente en las áreas afectadas por derrames de hidrocarburos generados por terceros y por eventos operacionales en el corredor vial Puerto Vega – Teteyé – Bloque suroriente, y el efecto del programa de limpieza y recuperación realizado por la empresa Gran Tierra Energy Colombia o sus contratistas

Sobre el corredor vial Puerto Vega – Teteyé (Puerto Asís. Putumayo), desde 1987 se han venido desarrollando procesos de exploración, perforación y establecimientos de campos petroleros. Actualmente, en el corredor existen tres campos de explotación, que están siendo operados por la petrolera Gran Tierra Energy (GTE). El crudo que se produce en estos pozos es transportado por carrotanques desde los puntos de carga de los campos del corredor, hasta la estación 1 de Ecopetrol localizada en Orito – Putumayo desde donde se bombea por el oleoducto trasandino Orito-Tumaco. Entre los años 2005 y 2015 se presentaron diferentes eventos de derrames de crudo en el corredor vial, afectando de manera directa el suelo y los cuerpos de agua cercanos a los derrames.

En un primer diagnóstico realizado por GTE, se identificaron y delimitaron 27 áreas de suelo contaminado que incluyen 12 humedales, 4 pocetas, 4 bajos inundables, 3 lagos, 6 caños y 5 quebradas. Gran Tierra Energy realizó la implementación de un proceso de limpieza y recuperación de las zonas afectadas, con el que espera ejecutar un programa de monitoreo de la biodiversidad que pueda dar luces sobre la efectividad de las medidas implementadas en la zona.

Como estrategia para el monitoreo de la efectividad de las actividades de limpieza, se realizó la caracterización de hongos y bacterias, debido a que tienen una alta capacidad de respuesta a los cambios ambientales, son altamente idóneos para evaluar y monitorear el estado de los ecosistemas (van Bruggen & Semenov 2000). La caracterización de las comunidades microbianas y de hongos se realizó a través de la técnica de metabarcoding, con la que se realizó la caracterización de la biodiversidad en muestras de agua, sedimentos y suelo mediante la secuenciación masiva de marcadores genéticos (Taberlet et al. 2012; Ji et al. 2013).

El proyecto consta de dos fases, la fase I se llevó a cabo en noviembre 2019, previa a las actividades de limpiezas por parte de GTE, y la fase II se llevará a cabo en el 2021, posterior a la limpieza (figura 12).

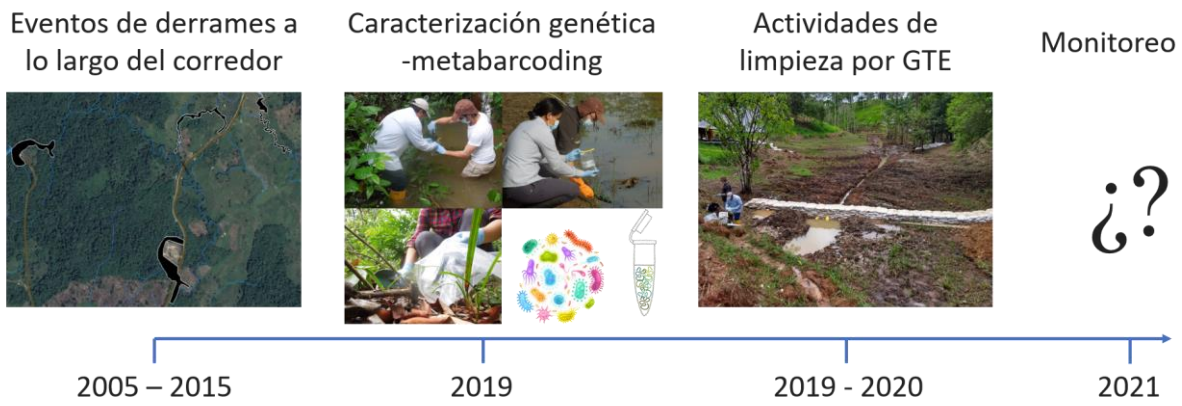

Figura 12. Línea del tiempo, 2005-105 se presentaron eventos de derrame; 2019 se estableció la línea base para microorganismos (hongos y bacterias), 2019-2020 GTE realizó las actividades de limpieza; 2021 primer monitoreo con microorganismos.

En cada una de las fases se va a realizar la caracterización genética de hongos y bacterias en muestras de agua, suelo y sedimento, en 39 puntos previamente seleccionados por GTE (figura 13, tabla 2).

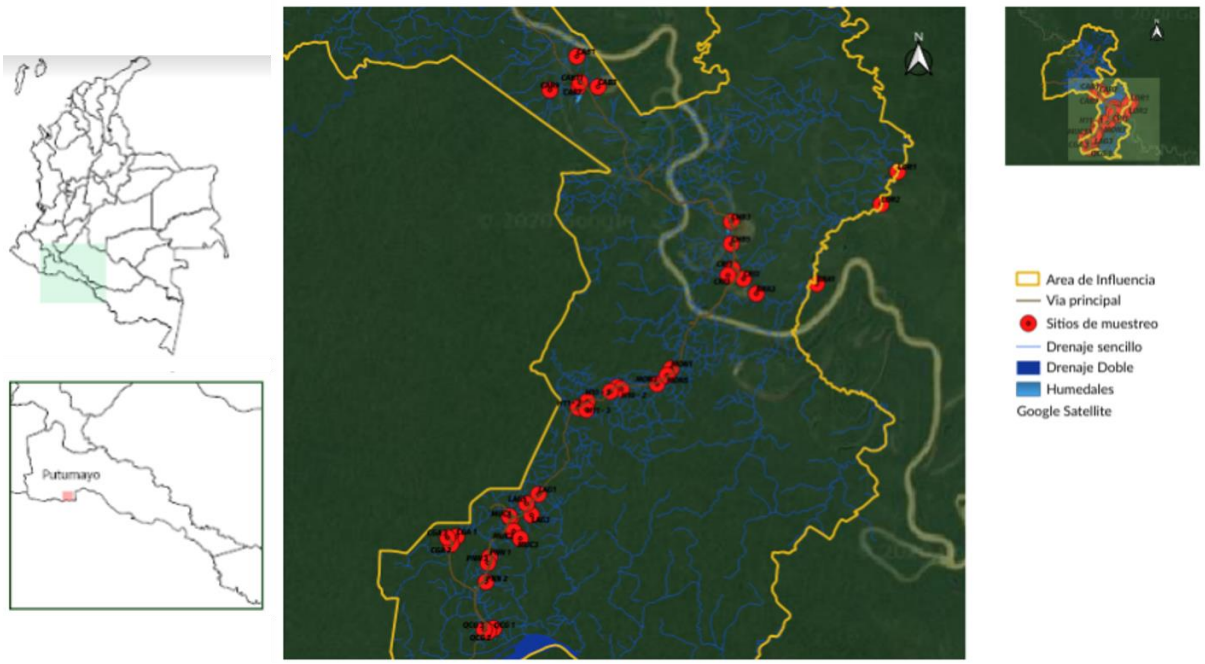

Figura 13. Mapa general de la zona de estudio (corredor vial). Los círculos rojos corresponden a los 39 puntos de muestreo.

Tabla 2. Puntos de muestreo, definidos por GTE

Puntos de muestreo			
QCG 1	H11 - 2	CAB1	BRA1
QCG 2	H11 - 3	CAB3	BRA3
QCG 3	H10 - 1	CAB7	MUC1
CGA 1	H10 - 2	CAB9	MUC1A
CGA 2	H10 - 3	CAB11	MUC2
CGA 3	MON1	CRI1	MUC3
PNN 1	MON3	CRI2	LAG1
PNN 2	MON5	CRI3	LAG3
PNN 3	CHB3	LOR1	LAG5
H11 - 1	CHB5	LOR2	

El muestreo consistió en tomar 3 muestras de agua, 3 de sedimentos y una muestra compuesta de suelos (Figura 14) en cada uno de los 39 puntos, para un total de 117 muestras por sustrato.

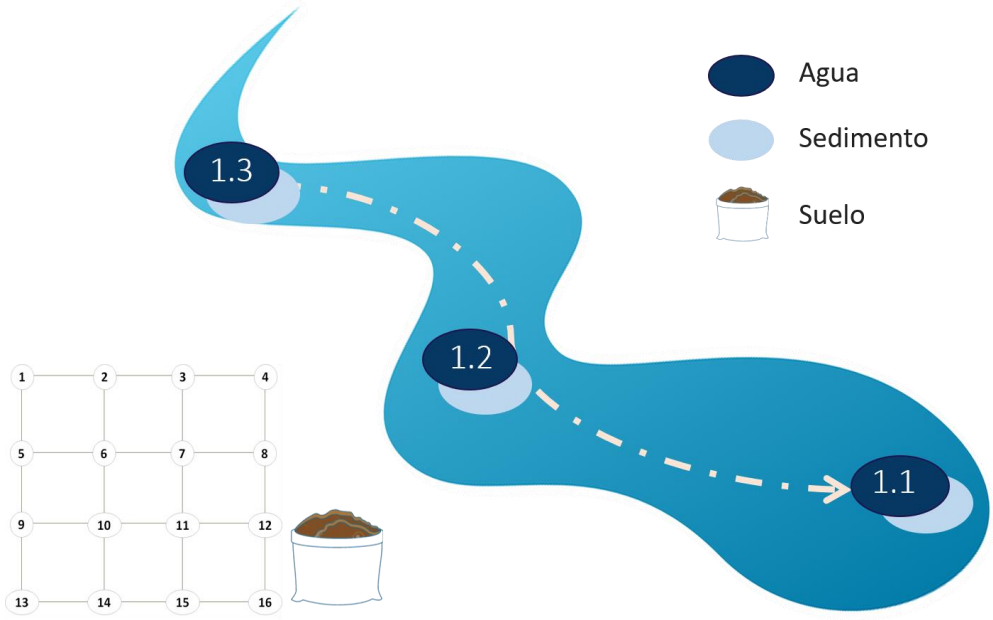

Figura 14. Muestreo realizado en cada uno de los 39 puntos.

El procesamiento de las muestras como el filtrado del agua y la extracción de ADN de las mismas se realizó en un laboratorio instalado en campo. La extracción del ADN extracelular para los tres tipos de muestras se llevó a cabo utilizando el kit NucleoSpinSoil™ de la marca Macherey-Nagel, y siguiendo un protocolo modificado e híbrido entre el del fabricante y el propuesto por Taberlet et al. 2012. Por cada ronda de extracción se incluyó una muestra control negativo, la cual consistió únicamente de la solución de fosfato de sodio, es decir sin ADN. De esta forma, durante los 12 días que tomó procesar todas las muestras, se obtuvieron 12 muestras control. El ADN extraído se guardó en criocajas y estas se guardaron en bolsas plásticas con silicagel, para evitar concentraciones de humedad.

Posteriormente para el proceso de amplificación y secuenciación, las muestras de ADN se enviaron a un laboratorio externo al Instituto. El ADN se amplificó por el método de Reacción en Cadena de la Polimerasa (PCR) a partir de los extractos obtenidos de la extracción, se emplearon cuatro controles negativos de PCR y tres réplicas de PCR por muestra. Con las réplicas de PCR de cada muestra se evaluó la confiabilidad de las secuencias que se mantienen posterior a aplicar los análisis bioinformáticos, para lo cual nos basamos en el supuesto de que las secuencias que aparecen en múltiples réplicas de PCR tienen más probabilidades de ser reales en lugar de secuencias generadas al azar por error de secuenciación. Para las PCR se usaron cebadores dirigidos a sitios de cebado comunes al clado de interés, en este caso cebadores universales para bacterias y cebadores específicos para hongos (Tabla 3) pero que flanquean regiones suficientemente variables para discriminar taxones. En este estudio, estos pares de cebadores fueron marcados con etiquetas cortas de combinación única para cada muestra. Estas etiquetas permiten recuperar, en el análisis bioinformático posterior, la muestra de origen.

Tabla 3. Cebadores utilizados para el procedimiento de metabarcoding.

Marcador genético	**Cebador**	**Secuencia**	**Referencia**
ITS (hongos)	5.8SF	5'-CAAGAGATCCGTTGTTGAAAGTK-3'	White et al. 1990
	ITS5	5'-GGAAGTAAAAGTCGTAACAAGG-3'	Epp et al. 2012
16S rRNA v3-v4 (bacterias)	341F	5'-CCTACGGGAGGCAGCAG-3'	Muyzer et al. 1993
	806R	5'-GGACTACHVGGGTWTCTAAT-3'	Caporaso et al. 2011

Además de los controles negativos de extracción y PCR, se tuvo en cuenta controles positivos que consistieron en una muestra de ADN de una comunidad con identidades conocidas, para este proyecto, se utilizaron comunidades de hongos de bosque alto-andino del norte de la cordillera oriental, y comunidades de bacterias seleccionadas del cepario de la Corporación CorpoGen. Los controles son tratados como muestras adicionales, y permiten analizar el comportamiento de los procesos de PCR y secuenciación, siendo usados en el proceso de limpieza bioinformática.

Después de las amplificaciones por PCR, se preparó una biblioteca de secuenciación donde todos los productos o amplicones de las diferentes muestras se combinaron y ligaron a cebadores de secuenciación. Las bibliotecas de secuenciación se secuenciaron utilizando una química de extremos emparejados de 250 pb en cartuchos V2 para Illumina MiSeq, realizando 3 corridas para bacterias y 3 para hongos con el objetivo de obtener millones de lecturas por biblioteca (25.000 lecturas / réplica de PCR / muestra).

Una vez obtenidas las secuencias se realizó el análisis bioinformático, iniciando con el alineamiento de hebras pareadas de ADN, demultiplexado y dereplicación de secuencias únicas, utilizando el programa OBITools 1.2.13 (Boyer et al. 2014). Las secuencias que presentaban nucleótidos ambiguos (N), longitudes por fuera de los límites esperados y que estuvieran soportadas por muy pocas secuencias, fueron descartadas. Las secuencias se agruparon en Unidades Taxonómicas Operativas Moleculares (MOTUs) utilizando Sumaclust (Mercier et al. 2013), y se generaron tablas de OTU que se analizaron contra las bases de datos genéticos estándar (STD) de procariotas (PRO) y hongos (FUN) de EMBL-EBI (ftp://ftp.ebi.ac.uk/pub/databases/embl/release/std) utilizando el algoritmo ecotag del paquete OBIToolsblastn. La taxonomía para cada OTU se asignó en función de la base de datos taxonomy de NCBI (https://www.ncbi.nlm.nih.gov/taxonomy).

Posteriormente, usando diferentes funciones base en R (2020), se eliminaron los OTUs que presentaron mayores abundancias en controles negativos de extracción o PCR, controles positivos o blancos de secuenciación. Para cada muestra, se descartaron aquellas réplicas de PCR que eran diferentes a las demás réplicas de su muestra, de acuerdo con análisis de correspondencia, y se conservaron sólo los OTUs que se registraron en al menos dos de las réplicas de PCR. Debido a que el número de lecturas en cada muestra era altamente variable, se realizó una rarefacción al cuantil 25 (25% de la distribución del número de lecturas) para todas las muestras depuradas. Finalmente, para evitar sesgos generados por las amplias diferencias en abundancias se estimaron distancias de Hellinger (Legendre & Gallagher 2001) por OTU y entre los puntos de muestreo, usando funciones del paquete 'vegan' (Oksanen et al. 2019).

Tras el proceso de depuración y limpieza de las secuencias, varias de las réplicas biológicas fueron excluidas. Esta exclusión permite obtener estimados confiables de diversidad y de los diferentes análisis ecológicos, y puede generarse por diferentes factores, desde presencia de inhibidores de amplificación en la muestra tomada, afinidad de los cebadores, contaminación, entre otros (Taberlet et al 2018). Las matrices finales con distancias de Hellinger usadas para los análisis ecológicos incluyeron para hongos 27 estaciones de muestreo en agua, 35 estaciones de muestreo para sedimentos y las 39 estaciones para suelo, para bacterias todas las estaciones quedaron incluidas.

Para cada punto de muestreo se estimó la diversidad alfa, beta y se hizo una descripción de la composición taxonómica a través de diferentes jerarquías. Los análisis de diversidad alfa se realizaron a partir del índice de Shannon (H'). Este índice se basa en la incertidumbre al predecir la identidad de un individuo desconocido, tomando en cuenta tanto el número de OTUs como sus abundancias relativas. Por ejemplo, en una comunidad altamente diversa con OTUs uniformemente distribuidos, la incertidumbre sobre la identidad de un individuo desconocido es alta ya que el individuo desconocido podría pertenecer a cualquier OTU. Por el contrario, en una comunidad con igual número de OTUs pero donde algunos de ellos son dominantes, podría ser más fácil predecir a qué especie pertenece el individuo, llevando a una menor incertidumbre y por tanto a una menor diversidad (Shannon 1948). El índice de Shannon se calculó para cada réplica biológica usando la función diversity del paquete 'vegan' en R. La diversidad alfa de cada punto de muestreo fue estimada como el valor promedio de cada una de las réplicas biológicas tomadas.

En este mismo proyecto y como parte de la estrategia de monitoreo por parte de GTE se realizó por parte de un grupo de especialistas la evaluación de la presencia de macroinvertebrados acuáticos (muestreados en los mismos puntos donde se tomaron las muestras para análisis genéticos). Parte de los resultados obtenidos por este grupo fue el cálculo del índice BMWP/Col, que evalúa la calidad del ambiente según la presencia de ciertas familias donde a cada una se le da una puntuación según su nivel de tolerancia a los ecosistemas eutrofizados, dando una categoría de calidad a cada uno de los sitios evaluados (Tabla 4) (Carrera & Fierro 2001).

Tabla 4. Puntajes BMWP/Col para los 39 puntos de muestreo

Puntos de muestreo			
QCG 1	H11 - 2	CAB1	BRA1
QCG 2	H11 - 3	CAB3	BRA3
QCG 3	H10 - 1	CAB7	MUC1
CGA 1	H10 - 2	CAB9	MUC1A
CGA 2	H10 - 3	CAB11	MUC2
CGA 3	MON1	CRI1	MUC3
PNN 1	MON3	CRI2	LAG1
PNN 2	MON5	CRI3	LAG3
PNN 3	CHB3	LOR1	LAG5
H11 - 1	CHB5	LOR2	

Categoría según calidad
Buena
Aceptable
Dudosa
Crítica
Muy crítica

Los resultados de diversidad alfa fueron corelacionados con la clasificación de calidad del punto, de acuerdo con el índice de calidad que se obtuvo de los puntajes obtenidos para el índice BMWP/Col para macroinvertebrados (Tabla 4). Para esto se realizaron pruebas de Kruskall-Wallis (Kruskall & Wallis 1952) y Pruebas de Dunn (Dunn 1964) usando la función base de R kruskal.test y dunnTest del paquete 'FSA' (Ogle et al. 2020).

La diversidad beta calcula la similitud o disimilitud de OTUs entre estaciones de muestreo, es decir, el número OTUs que son compartidos entre estas. Para estimar el grado de similitud entre puntos de muestreo se usó el índice de disimilitud de Bray-Curtis (Bray & Curtis 1957), el cual toma en cuenta tanto la presencia de OTUs como su abundancia. Esta estimación se realizó usando la función *vegdist* del paquete 'vegan' en R. El índice de Bray-Curtis se obtiene en comparaciones pareadas, por lo que para obtener un estimado de cada sustrato, se promedió el valor del índice a través de todas las comparaciones pareadas por grupo biológico. Como segunda estimación de diversidad beta se estimó la similitud entre las comunidades de los diferentes puntos de muestreo a través de un análisis de escalamiento multidimensional no métrico o NMDS (Nonmetric multidimensional scaling). Este es un método de ordenación basado en rangos que intenta ordenar los objetos o muestras de acuerdo con su distancia, en

este caso, estimada de acuerdo con su similitud en composición de OTUs. El análisis NMDS se ejecutó con la función *metaMDS* del paquete 'vegan' en R, estimando la distancia entre puntos de muestreo con el índice de disimilitud de Bray-Curtis. Al igual que con la diversidad alfa, se evaluó la correspondencia cualitativa entre la diversidad beta y las categorías de calidad de acuerdo con los índices obtenidos de los Puntajes BMWP/Col de macroinvertebrados (Tabla 4). Adicionalmente, se evaluó la correlación entre la beta diversidad y la distancia geográfica entre estaciones de muestreo, a través una prueba de Mantel (Mantel 1967) usando la función *mantel.rtest* del paquete 'ade4' (Bougeard & Dray 2018) en R.

En el Análisis de composición taxonómica Se examinó la estructura de las comunidades de hongos y bacterias de acuerdo a la asignación taxonómica de los OTUs obtenidos y las matrices con distancias de Hellinger de la siguiente manera: 1) composición taxonómica a nivel de clase, para hongos y bacterias, agrupando por sustrato evaluado (agua, sedimento, suelo); 2) composición taxonómica a nivel de clase, para hongos y bacterias, agrupando por cada estación de muestreo analizando cada sustrato de forma independiente; 3) composición taxonómica a nivel de clase, para hongos y bacterias, agrupando los puntos de muestreo de acuerdo a la categoría de calidad de agua obtenida mediante la evaluación de macroinvertebrados acuáticos, analizando cada sustrato también de forma independiente. Para realizar los cálculos y agrupaciones por categoría en estos análisis se utilizó el paquete 'phyloseq' de Bioconductor en R (McMurdie & Holmes, 2013), y para la elaboración de los gráficos se utilizó 'ggplot2' (Wickham, 2016), con los esquemas de colores de 'randomcoloR' (Ammar, 2019).

En cuanto a los resultados, en general se observó que tanto para hongo como para bacterias la diversidad fue altamente variable a través de las estaciones de muestreo, al hacer la relación de la diversidad alfa con el índice de calidad según BMWP/Col, en hongos si observamos una diferencia significativa en aquellas zonas más afectadas (categoría muy crítica según BMWP/Col) comparado al resto de las categorías. En cuanto a la composición se observó dominancia de una clase. Particularmente en suelos no se observó una señal respecto al índice de calidad (figura 15).

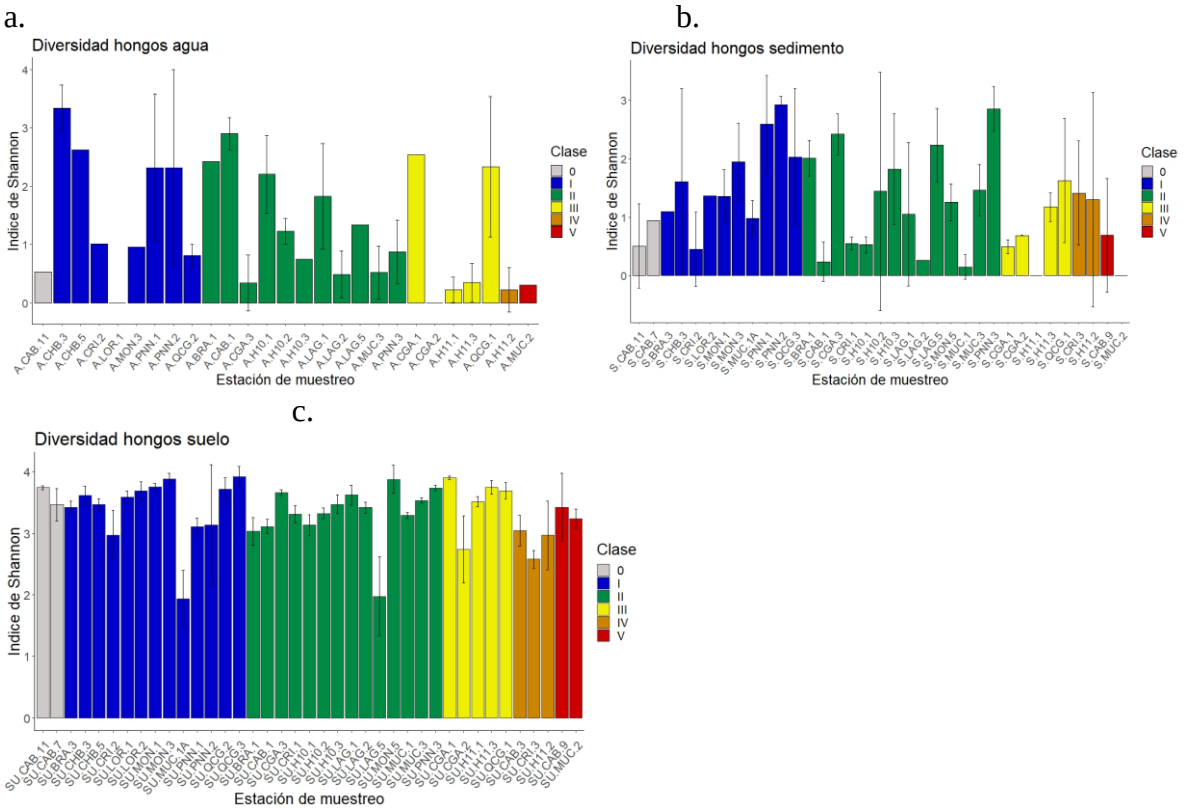

Figura 15. Diversidad alfa de hongos en a. agua, b. sedimento y c. suelo. El tamaño de las barras representa la diversidad alfa para cada uno de los sitios con su desviación estándar y los colores corresponden a la categoría de calidad del ambiente según el índice BMWP/Co de macroinvertebrados, donde 0 corresponde a estaciones que no les fueron asignadas una categoría, I es buena, II aceptable, III dudosa, IV crítica y V muy crítica

En cuanto a bacterias no se encontraron diferencias significativas al comparar con los el índice de calidad de agua, por el contrario, las señales parecen ser independientes, respecto al índice de calidad (figura 16).

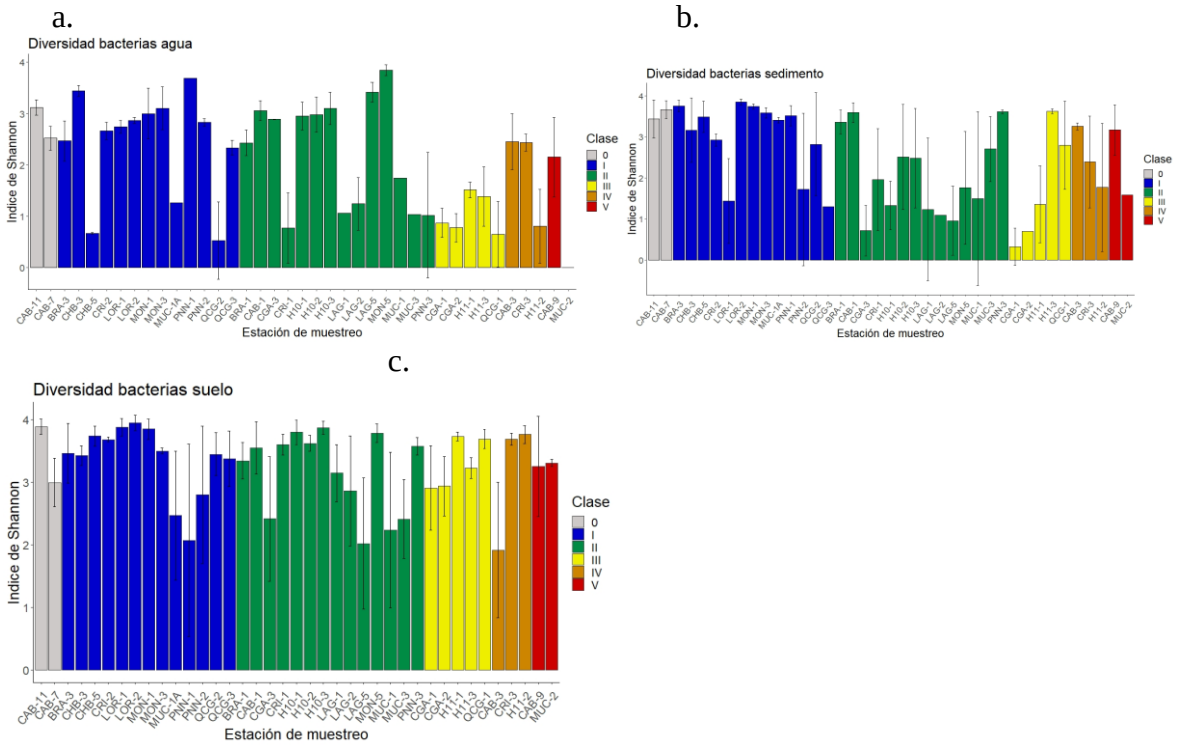

Figura 16. Diversidad alfa de bacterias en las muestras de a. agua, b. sedimento y c. suelo. El tamaño de las barras representa la diversidad alfa para cada uno de los sitios con su desviación estándar y los colores corresponden a la categoría de calidad del ambiente según el índice BMWP/Co de macroinvertebrados, donde 0 corresponde a estaciones que no les fueron asignadas una categoría, I es buena, II aceptable, III dudosa, IV crítica y V muy crítica.

En cuanto a la composición taxonómica, en total para la zona de estudio se encontraron 3737 OTUs de Bacterias, en muestras de agua 568, en sedimento 983 y en suelo 1026 y para hongos se encontraron 1723 OTUs para las muestras de agua 286, en sedimentos 306 y en suelos 1408, con diferente porcentaje de asignación taxonómica a nivel de filo y de clase (Tabla 5).

Tabla 5. Porcentajes de asignación taxonómica a nivel de filo y clase para los tres sustratos en hongos y bacterias.

	Sustrato	**filo**	**clase**
Hongos	Agua	69.6 %	58.1 %
	Sedimentos	72.5 %	41.8 %
	suelo	65.4 %	46.6 %
Bacterias	Agua	54.4 %	40.3 %
	Sedimentos	60.9 %	52.5 %
	suelo	63.3 %	52.9 %

Tanto de hongos (Figura 17) como de bacterias (Figura 18), vemos que en agua y sedimento es similar, en suelos cambia un poco.

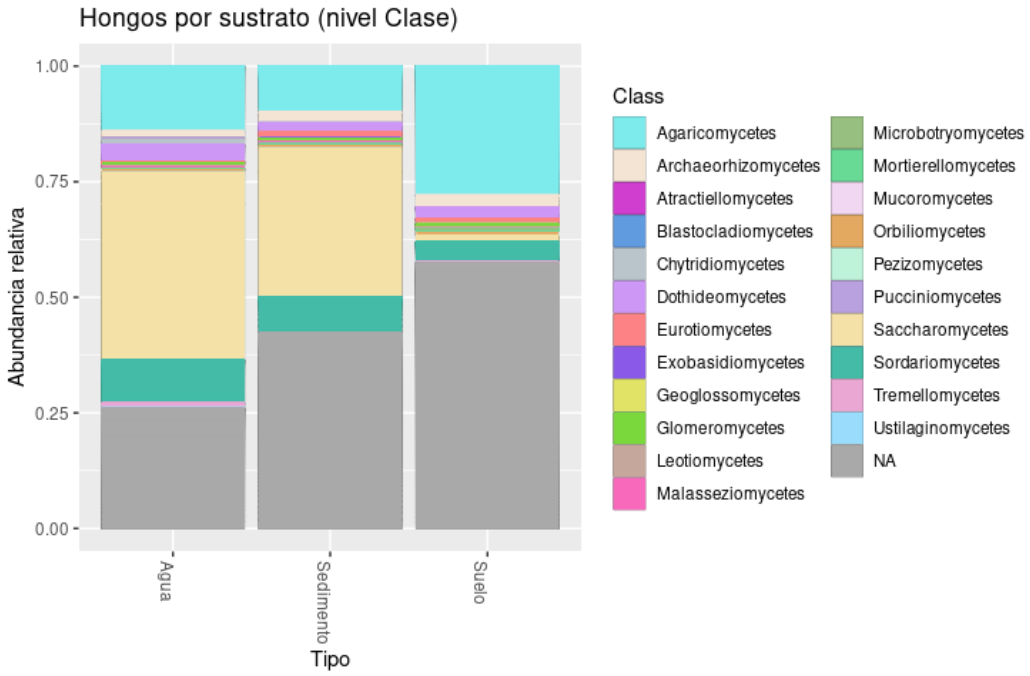

Figura 17. Composición taxonómica a nivel de clase de las comunidades de hongos por sustrato evaluado (agua, sedimento, suelo). Las barras representan las abundancias relativas por sustrato y cada color representa una clase (categoría taxonómica).

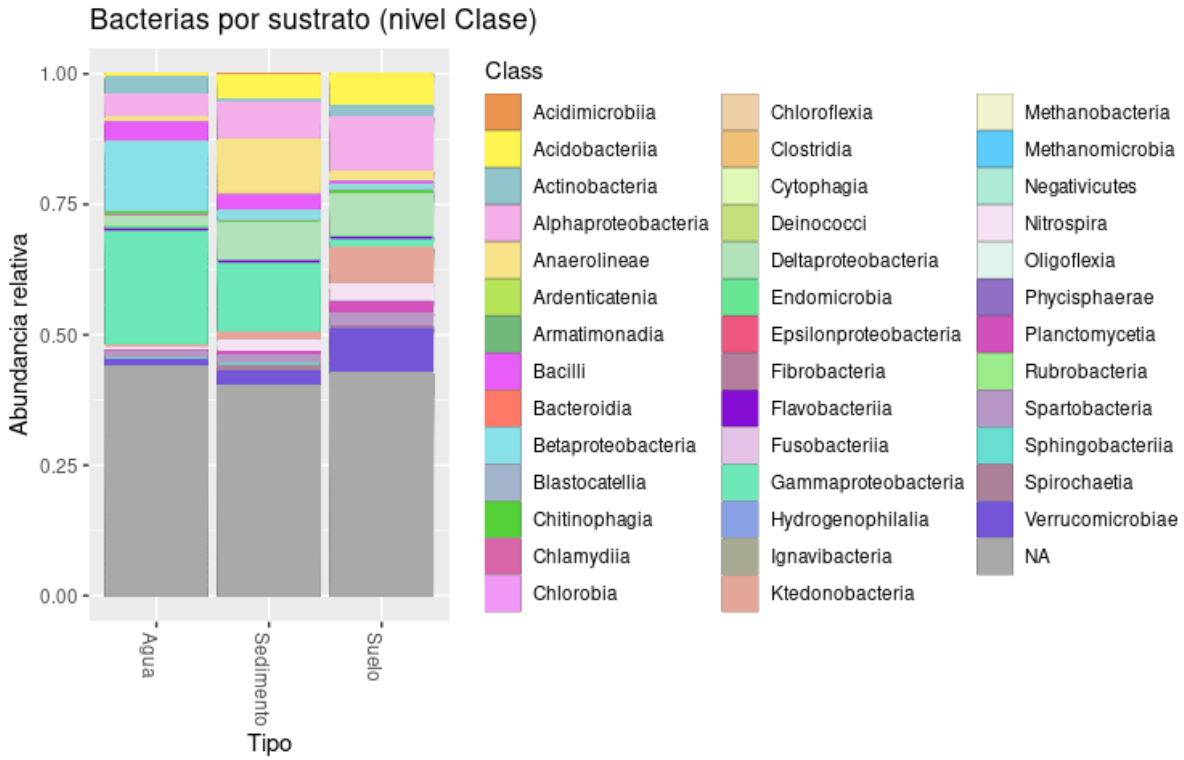

Figura 18. Composición taxonómica a nivel de clase de las comunidades de bacterias por sustrato evaluado (agua, sedimento, suelo). Las barras representan las abundancias relativas por sustrato y cada color representa una clase (categoría taxonómica).

Adicionalmente para los tres sustratos observamos organismos que en la literatura se han reportado como degradadores de crudo. Por ejemplo en bacterias , se evidenció presencia de las clases Betaproteobacteria, las Actinobacteria, Flavobacteriia, este último principalmente en las muestras de suelo, Deltaproteobacteria en los sustratos de sedimentos y agua, la clase Sphingobacteriia que predomina en las muestras de agua y sedimentos, Alphaproteobacteria presente en mayor proporción en las muestras de suelo y sedimento y bacterias de la clase Bacilli que aunque se encuentra presente en los tres sustratos, en suelos se encuentra en menor proporción (Ortiz & Chiappa 2013).

En cuanto a hongos se identificó en los tres sustratos la presencia del orden mortierella de la clase Mortierellomycetes, especies de este género tienen la capacidad acumular grasos insaturados y son asociados a suelos contaminados con hidrocarburos (Wagner et al, 2013). Reyes y Ortíz (2017) realizaron una recopilación de estudios que identifican algunas clases de hongos con capacidad para degradar el petróleo, algunos identificados en el presente estudio, presentes en bajas proporciones en los tres sustratos, entre los cuales se encuentran las clases, Sordariomycetes, Dothideomycetes, Glomeromycetes, Eurotiomycetes y Mucoromycota.

En esta primer fase del proyecto previa a las labores de limpieza, para la caracterización de microorganismos a partir de metabarcoding se pudo apreciar que no hay una diferencia entre las estaciones de muestreo en cuanto a su diversidad, podemos inferir que en parte puede deberse a que el sistema es muy homogéneo, se encuentra caracterizado por un paisaje

altamente intervenido, donde predominan los potreros y cultivos, además de ser una zona que históricamente ha estado expuesto a derrames de petróleo.

Sin embargo se ha demostrado que a largo plazo la contaminación por hidrocarburos genera cambios en la estructura y sucesión de las comunidades microbianas, debido a que la composición de los compuestos del petróleo crudo y su toxicidad cambian con el tiempo (Jeong et al., 2015), evaluar la asociación de la contaminación de las comunidades microbianas por hidrocarburos y su efecto en el tiempo y posterior a las actividades de limpieza, se espera poder tener una caracterización más completa del estado y las tendencias de las comunidades con el fin de evaluar la capacidad de respuesta posterior a la limpieza de las zonas que estuvieron expuestas a derrame de crudo.

3.1.4 Uso de metabarcoding en el monitoreo ambiental con insectos

Los insectos son organismos diversos y cosmopolita. Corresponden a algunos de los primeros animales en colonizar ecosistemas terrestres y de agua dulce (Misof et al. 2014). La abundancia y dominancia de estos organismos representa más del 50% de los seres vivos del planeta, donde la biomasa sólo de las hormigas constituye cerca de un cuarto de la biomasa animal total (Wilson 1992, Robinson et al. 2011). La importancia de los insectos es incalculable, así como aportan a los humanos y al ecosistema con sus numerosas funciones (polinizadores, ciclado de nutrientes, descomposición de materia orgánica, aireación del suelo, control biológico, fuente de alimento, entre otros), también hay porcentajes de especies que afectan negativamente la economía, en particular en áreas agropecuarias y en salud (Hill 1997, McIntyre et al. 2001, Cannings 2007, Song et al. 2015, Steward et al. 2017). Durante décadas varios grupos de insectos se han utilizado en estudios ambientales, pues se han encontrado en ellos agrupaciones idóneas para evaluar y monitorear cambios en los ecosistemas, debido a su diversidad y la variación en sensibilidad y respuestas a factores ambientales. Es así que, algunos insectos se utilizan como indicadores en la evaluación y monitoreo de la biodiversidad, calidad del agua y de suelos, degradación y recuperación de la estructura vegetal, estados sucesionales y fragmentación (Hill 1997, Roldán-Pérez 2016, Fernández et al. 2019). Se resalta en particular en estudios ambientales el uso de las mariposas, grillos y saltamontes, cucarrones y escarabajos, las hormigas y avispas, chinches acuáticos o en ocasiones agrupaciones amplias de varios órdenes y grupos de animales, como los macroinvertebrados acuáticos, entre los que se encuentran los efemerópteros, plecópteros y tricópteros que son insectos con estados de vida acuática (van Swaay 2008, Bicknell et al. 2014, Hochkirch et al. 2016, Roldán-Pérez 2016, Fernández et al. 2019).

Históricamente, el uso de los insectos para la caracterización de los ecosistemas consiste de manera general en la visita al sitio, observación y recolección de organismos que luego son identificados a partir de su morfología en un laboratorio, luego se analizan los datos y se proveen interpretaciones y conclusiones (Leese et al. 2018). El uso de morfología continua vigente, sin embargo, nuevas herramientas moleculares se han puesto a prueba y han demostrado su potencial para la caracterización y monitoreos con insectos a corto y largo plazo (Hering et al. 2018, Piper et al. 2019). Entre las aproximaciones más recientes está el uso de metabarcoding, que ha ido tomando fuerza a nivel mundial, en ecosistemas terrestres se realiza el monitoreo masivo de insectos con trampas Malaise, la cual corresponde a una trampa de interceptación de vuelo donde principalmente se capturan insectos de los órdenes Diptera (moscas), e Hymenoptera (abejas y avispas), pero frecuentemente se capturan otros órdenes Coleoptera (cucarrones y escarabajos), Hemiptera (chiches y saltahojas), Lepidoptera (mariposas y polillas) (Sheikh et al. 2016). Este monitoreo es de uso cotidiano a nivel mundial, tal vez por la facilidad de manipulación de la trampa que puede ser dejada por largos periodos de tiempo y visitada según la frecuencia de interés (días, semanas o algunos meses).

En diferentes países se han desarrollado proyectos para monitoreo de insectos con este tipo de trampas con la participación de voluntarios en comunidades remotas y en colegios quienes se encargan de hacer el montaje de trampas y de recolectar y enviar las muestras para análisis (Geiger et al. 2016, Karlsson et al. 2020). Las muestras en la trampa están preservadas en etanol para así garantizar la recuperación de ADN proveniente de los organismos. Al analizar las

muestras con técnicas de nueva generación se logra la secuenciación de ADN de organismos de forma masiva y en paralelo generando inventarios y logrando la identificación de los organismos a partir de códigos de barras (Piper et al. 2019). Esto permite disminuir costos y tiempos de procesamiento de datos; sin embargo, no permite asociar los especímenes recolectados con los códigos, para asociarlos tocaría procesar los especímenes de forma individual (Karlsson et al. 2020). El procesamiento individual (barcoding) puede traer consigo retos como altos costos y necesidad de especialistas en la identificación taxonómica, aunque es un proceso que ayuda a crear bases de datos de referencia. Al combinar un método de captura masivo con métodos de análisis masivos como el metabarcoding se logra incrementar la velocidad de conocimiento de la biodiversidad de las regiones, con esto se pueden crear nuevos estándares que faciliten los monitoreos y promuevan un entendimiento comprehensivo de la biodiversidad y tendencias en la pérdida o alteraciones de la misma.

Por otra parte, en ecosistemas acuáticos se ha encontrado que el ADN de insectos puede ser recuperado en muestras de agua, sin necesidad de hacer captura de los organismos (Uchida et al. 2020). El uso de metabarcoding en la detección de ADN ambiental (eDNA) es una técnica esencial y en creciente auge para el uso de ecosistemas acuáticos (Leese et al. 2018). El ADN ambiental corresponde a fragmentos de ADN o rastros de ADN que dejan los organismos en el ambiente, la detección de este rastro representa una oportunidad para determinar la presencia de un organismo en la fuente hídrica e incluso la implementación de monitoreo en un tiempo determinado. Este método parte de la premisa que para detectar el ADN de un organismo este se encuentra presente, reconociendo algunas falsas detecciones debidas a contaminación por depósito de excrementos de animales, depósito de otros organismos por el arrastre del viento, corrientes y afluentes (Taberlet et al.2018). Entre los mayores potenciales del uso del eDNA se encuentran la eliminación de la perturbación que traen los muestreos a los ecosistemas (Taberlet et al.2018, Uchida et al. 2020), en particular al tomar muestras abióticas en lugar de especímenes, y reducción del sesgo que puede generar el muestreo de los investigadores y el acceso a los sitios (Uchida et al.2020).

Se ha descrito variación en la detección de muestras de ADN con referencia al tiempo y a la distancia entre los organismos con el sitio de toma de muestra en el agua (Taberlet et al. 2018). El potencial de esta herramienta en monitoreos está en la persistencia del ADN en el ecosistema que puede variar entre días, semanas hasta menos de un mes y también puede detectar especies a cierta distancia, esta varía según el organismo en estudio; no se conoce en insectos, pero ejemplo en vertebrados como salamandras la distancia medida fue de 5 metros, mientras en peces se lograron detecciones de 2-3 km (Taberlet et al. 2018). La detección se hace con el análisis de las muestras de agua en las cuales se obtiene secuencias correspondientes a la región de códigos de barras con ayuda de *primers* específicos para estas regiones (procedimiento similar al que se realiza en trampas Malaise) (figura 1). En el caso de grupos macroinvertebrados dentro de los cuales se encuentran los insectos, se utiliza en gran medida la región de ADN mitocondrial del gen citocromo oxidasa I (COI), aunque otros genes también se han utilizado en la identificación de insectos (COI, 16S) (Sonet et al. 2013, Marquina et al. 2019). La región código de barras en varios casos se considera especie-especifica de forma que facilita la identificación de especies de los organismos de interés. Además, los códigos de barras permiten asignar números, conocidos como BIN (Barcode Index Number), a potenciales especies nuevas y crípticas o el tratamiento de estos números de forma similar a morfoespecies que puede aplicarse en análisis ecológicos (Ratnasingham & Hebert 2007, Paz et al. 2011).

Estudios recientes se han enfocado en la exploración del uso de índices de calidad del agua con eDNA de macroinvertebrados (Fernández et al. 2019, Mächler et al. 2019, Uchida et al. 2020). El uso de bioindicadores de calidad del agua a través de índices es una práctica que viene ocurriendo desde los años 50's y con más fuerza desde los 70's (Hilsenhoff 1987, Roldán-Pérez 2016). Los macroinvertebados acuáticos se establecieron para la evaluación del estado ambiental y ecológico de fuentes hídricas, pues estos grupos son abundantes, diversos, fáciles de recolectar y sensibles a los cambios ambientales. Para la medición de estos índices se requiere de la recolección de individuos; sin embargo, los estudios apuntan a evaluar si el uso de índices tradicionalmente estudiados con morfología pueden ser analizados a partir de eDNA.

Puesto que aún no se tiene un entendimiento de cómo se comparan las aproximaciones de morfología clásica y los métodos de eDNA. Fernández et al. (2019) compararon datos de macroinvertebrados a partir de morfología y eDNA, encontrando que ambos tipos de datos se correlacionan de forma positiva y dan respuestas similares a los índices de calidad de hábitat evaluados, por lo tanto sugieren que el eDNA puede ser utilizado para el cálculo de índices bióticos. Por su parte, Mächler et al. (2019) compararon los registros de insectos de Ephemeroptera, Plecoptera y Trichoptera (EPT) a partir de tres tipos de datos: registros históricos, ADN ambiental y morfológicos de especímenes capturados con red acuática (*Kick net*). En general los registros morfológicos y de eDNA fueron proporcionalmente similares y compartieron un 62% de los géneros identificados. En contraste, Uchida et al. (2020) al comparar registros morfológicos con red acuática (Surber) y de eDNA recuperaron casi el doble de datos para familias y géneros de EPT y otros órdenes de insectos como Diptera y Hemiptera con eDNA. En estos grupos acuáticos se conocen diferencias estacionales marcadas que fueron detectadas con ambos métodos, además el índice biótico evaluado a nivel de género mostró una mayor sensibilidad a contaminación orgánica en los datos de eADN. La aplicación del ADN ambiental para la detección de macroinvertebrados acuáticos también ha mostrado potencial registrando especies exóticas o no-nativas (Marchall y Stepien 2020). Marchall y Stepien (2020) detectaron con eDNA y morfología igual número de especies, aunque el eDNA obtuvo un mayor número de ocurrencias en comparación con morfología. Es entonces que aproximaciones con morfología clásica y eDNA apuntan a responder las preguntas de forma similar, aunque no de forma idéntica, esto aún permite considerar su implementación a futuro y lograr un uso más amplio de estas técnicas novedosas de generación de datos masivos.

Literatura citada

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods in Ecology and Evolution, 9(1), 134–147.

Ammar R (2019). randomcoloR: generate attractive random colors. https://github.com/ronammar/randomcoloR.

Beentjes, K. K., Speksnijder, A. G., Schilthuizen, M., Hoogeveen, M., & van der Hoorn, B. B. (2019). The effects of spatial and temporal replicate sampling on eDNA metabarcoding. PeerJ, 7, e7335.

Berry, T. E., Osterrieder, S. K., Murray, D. C., Coghlan, M. L., Richardson, A. J., Grealy, A. K., ... & Bunce, M. (2017). DNA metabarcoding for diet analysis and biodiversity: A case study using the endangered Australian sea lion (Neophoca cinerea). Ecology and Evolution, 7(14), 5435-5453.

Bougeard S. & S. Dray (2018). Supervised Multiblock Analysis in R with the ade4 Package. Journal of Statistical Software, 86(1), 1–17. doi: 10.18637/jss.v086.i01 (URL: https://doi.org/10.18637/jss.v086.i01).

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. Ecological monographs, 27(4), 325–349.

Borrell, Y. J., Miralles, L., Do Huu, H., Mohammed-Geba, K., & Garcia-Vazquez, E. (2017). DNA in a bottle—Rapid metabarcoding survey for early alerts of invasive species in ports. PloS one, 12(9), e0183347.

Boyer F, Mercier C, Bonin A, Le Bras Y, Taberlet P, Coissac E. (2016) Obitools: a unix-inspired software package for DNA metabarcoding. Mol Ecol Resour.; 16:176–182.

Brown, E. A., Chain, F. J., Zhan, A., MacIsaac, H. J., & Cristescu, M. E. (2016). Early detection of aquatic invaders using metabarcoding reveals a high number of non-indigenous species in C anadian ports. Diversity and Distributions, 22(10), 1045-1059.

Bush, A., Compson, Z. G., Monk, W., Porter, T. M., Steeves, R., Emilson, E. J., ... & Baird, D. (2019). Studying ecosystems with DNA metabarcoding: Lessons from biomonitoring of aquatic macroinvertebrates. Frontiers in Ecology and Evolution, 7, 434.

Calderón-Sanou, I., Münkemüller, T., Boyer, F., Zinger, L., & Thuiller, W. (2020). From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices?. Journal of Biogeography, 47(1), 193–206.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016). "DADA2: High-resolution sample inference from Illumina amplicon data." Nature Methods, 13, 581-583. doi: 10.1038/nmeth.3869.

Cannings, R. A. (2007). Recent range expansion of the Praying Mantis, Mantis religiosa Linnaeus (Mantodeaz Mantidae), in British Columbia. Journal of the Entomological Society of British Columbia, 104, 73–80.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat Methods, 7:335–336.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A 108(Suppl):4516–4522. doi:10.1073/pnas.1000080107.

Carrera, C., & Fierro, K. (2001). Manual de monitoreo: los macroinvertebrados acuáticos como indicadores de la calidad del agua. Quito: EcoCiencia.

Chariton, A. A., Stephenson, S., Morgan, M. J., Steven, A. D., Colloff, M. J., Court, L. N., & Hardy, C. M. (2015). Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. Environmental pollution, 203, 165-174.

Chen, W., Wilkes, G., Khan, I. U., Pintar, K. D., Thomas, J. L., Lévesque, C. A., ... & Lapen, D. R. (2018). Aquatic bacterial communities associated with land use and environmental factors in agricultural landscapes using a metabarcoding approach. Frontiers in microbiology, 9, 2301.

Cluff, M. A., Hartsock, A., MacRae, J. D., Carter, K., & Mouser, P. J. (2014). Temporal changes in microbial ecology and geochemistry in produced water from hydraulically fractured Marcellus shale gas wells. Environmental science & technology, 48(11), 6508-6517.

Correa, C. A., Etter, A., Díaz-Timote, J. J., Rodríguez, S., Ramirez, W., & Corzo, G. (2020). Spatiotemporal Evaluation of The Human Footprint in Colombia: Four Decades of Anthropic Impact in Highly Biodiverse Ecosystems. Ecological Indicators, 117, 106630.

Dallmeier, F. (1996). Biodiversity inventories and monitoring: essential elements for integrating conservation principles with resource development projects. — En : Szaro, R. C. & Johnston, D. W. (eds), Biodiversity in managed landscapes: theory and practice: 221–236. Oxford University Press, New York.

Daly, R. A., Borton, M. A., Wilkins, M. J., Hoyt, D. W., Kountz, D. J., Wolfe, R. A., ... & Wrighton, K. C. (2016). Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. Nature Microbiology, 1(10), 1-9.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Molecular ecology, 26(21), 5872–5895.

Delavaux, C. S., Bever, J. D., Karppinen, E. M., & Bainard, L. D. (2020). Keeping it cool: Soil sample cold pack storage and DNA shipment up to 1 month does not impact metabarcoding results. Ecology and Evolution.

Dufresne, Y., Lejzerowicz, F., Perret-Gentil, L. A., Pawlowski, J., & Cordier, T. (2019). SLIM: a flexible web application for the reproducible processing of environmental DNA metabarcoding data. BMC bioinformatics, 20(1), 1-6. Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26(19), pp.2460–2461.

Dunn, O.J. (1964). Multiple comparisons using rank sums. Technometrics 6:241–252.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26(19), pp.2460–2461.

Epp, L. S., Boessenkool, S., Bellemain, E. P., Haile, J., Esposito, A., Riaz, T., Brochmann, C. (2012). New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. Molecular Ecology, 21(8), 1821–1833.

Fernández. F., Flórez, Guerrero, R. J., & Delsinne, T. (2019). Hormigas de Colombia. Primera edición. Universidad Nacional de Colombia. Facultad de Ciencias. Instituto de Ciencias Naturales. Bogotá, 1200p.

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., … Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. Molecular Ecology Resources, 15(3), 543–556. https://doi.org/10.1111/1755-0998.12338

Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. Biology letters, 4(4), 423–425.

Geiger, M. F., Moriniere, J., Hausmann, A., Haszprunar, G., Wägele, W., Hebert, P. D., & Rulik, B. (2016). Testing the Global Malaise Trap Program–How well does the current barcode reference library identify flying insects in Germany?. Biodiversity data journal, 4: e10671.

Harms-Tuohy, C. A., Schizas, N. V., & Appeldoorn, R. S. (2016). Use of DNA metabarcoding for stomach content analysis in the invasive lionfish Pterois volitans in Puerto Rico. Marine Ecology Progress Series, 558, 181-191.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proc R Soc B - Biol Sci 270: 313–321

Hering, D., Borja, A., Jones, J. I., Pont, D., Boets, P., Bouchez, A., ... & Kelly, M. (2018). Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. Water Research, 138, 192–205.

Hill, D. S. (1997). The economic importance of insects. Springer Netherlands. 395p.

Hinlo, R., Gleeson, D., Lintermans, M., & Furlan, E. (2017). Methods to maximise recovery of environmental DNA from water samples. PloS one, 12(6), e0179251.

Hull, N. M., Rosenblum, J. S., Robertson, C. E., Harris, J. K., & Linden, K. G. (2018). Succession of toxicity and microbiota in hydraulic fracturing flowback and produced water in the Denver–Julesburg Basin. Science of the total environment, 644, 183-192.

Kruskal, K.W and Wallis, W.A. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47(260): 583–621.

Jeong, H.J., Lee, H.J., Hong, S., Khim, J.S., Shim, W.J., Kim, G.B., (2015). DNA damage caused by organic extracts of contaminated sediment, crude, and weathered oil and their fractions recovered up to 5 years after the 2007 Hebei Spirit oil spill off Korea. Mar. Pollut. Bull. 95 (1) 452e–457.

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., ... & Larsen, T. H. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecology letters, 16(10), 1245–1257.

Kartzinel, T. R., Chen, P. A., Coverdale, T. C., Erickson, D. L., Kress, W. J., Kuzmina, M. L., ... & Pringle, R. M. (2015). DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. Proceedings of the National Academy of Sciences, 112(26), 8019-8024.

Karlsson, D., Hartop, E., Forshage, M., Jaschhof, M., & Ronquist, F. (2020). The Swedish Malaise trap project: a 15 year retrospective on a countrywide insect inventory. Biodiversity Data Journal, 8. e47255.

Kerley, G. I., Landman, M., Ficetola, G. F., Boyer, F., Bonin, A., Rioux, D., ... & Coissac, E. (2018). Diet shifts by adult flightless dung beetles Circellium bacchus, revealed using DNA metabarcoding, reflect complex life histories. Oecologia, 188(1), 107-115.

Lanzén, A., Lekang, K., Jonassen, I., Thompson, E. M., & Troedsson, C. (2016). High-throughput metabarcoding of eukaryotic diversity for environmental monitoring of offshore oil-drilling activities. Molecular Ecology, 25(17), 4392-4406.

Laroche, O., Wood, S. A., Tremblay, L. A., Ellis, J. I., Lear, G., & Pochon, X. (2018). A cross-taxa study using environmental DNA/RNA metabarcoding to measure biological impacts of offshore oil and gas drilling and production operations. Marine Pollution Bulletin, 127, 97-107.

Leese, F., Bouchez, A., Abarenkov, K., Altermatt, F., Borja, A., Bruce, K., ... & Weigand, A. M. (2018). Why we need sustainable networks bridging countries, disciplines, cultures and generations for aquatic biomonitoring 2.0: a perspective derived from the DNAqua-Net COST action. In Advances in ecological research (Vol. 58, pp. 63–99). Academic Press.

Legendre, P. & E. D. Gallagher. (2001) Ecologically meaningful transformations for ordination of species data. Oecologia 129, 271–280.

Li, F., Peng, Y., Fang, W., Altermatt, F., Xie, Y., Yang, J., & Zhang, X. (2018). Application of environmental DNA metabarcoding for predicting anthropogenic pollution in rivers. Environmental science & technology, 52(20), 11708-11719.

Lobo, J., Shokralla, S., Costa, M. H., Hajibabaei, M., & Costa, F. O. (2017). DNA metabarcoding for high-throughput monitoring of estuarine macrobenthic communities. Scientific reports, 7(1), 1-13.

Longmire, J. L., Maltbie, M., & Baker, R. J. (1997). Use of" lysis buffer" in DNA isolation and its implication for museum collections. Museum of Texas Tech University.

Lopes, C. M., De Barba, M., Boyer, F., Mercier, C., da Silva Filho, P. J. S., Heidtmann, L. M., ... & Taberlet, P. (2015). DNA metabarcoding diet analysis for species with parapatric vs sympatric distribution: a case study on subterranean rodents. Heredity, 114(5), 525-536.

Mächler, E., Little, C. J., Wüthrich, R., Alther, R., Fronhofer, E. A., Gounand, I., ... & Altermatt, F. (2019). Assessing different components of diversity across a river network using eDNA. Environmental DNA, 1(3), 290–301.

Magurran, A. E., & McGill, B. J. (Eds.). (2011). Biological diversity: frontiers in measurement and assessment. Oxford University Press.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. Cancer Research 27 (2): 209–220.

Marquina, D., Esparza-Salas, R., Roslin, T., & Ronquist, F. (2019). Establishing arthropod community composition using metabarcoding: Surprising inconsistencies between soil samples and preservative ethanol and homogenate from Malaise trap catches. Molecular ecology resources, 19(6), 1516–1530.

Marshall, N. T., & Stepien, C. A (2020). Macroinvertebrate community diversity and habitat quality relationships along a large river from targeted eDNA metabarcode assays. Environmental DNA. 00:1–15.

McIntyre, N. E., Rango, J., Fagan, W. F., & Faeth, S. H. (2001). Ground arthropod community structure in a heterogeneous urban environment. Landscape and urban planning, 52(4), 257–274.

McMurdie PJ, Holmes S (2013). “phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data.” PLoS ONE, 8(4), e61217. http://dx.plos.org/10.1371/journal.pone.0061217.

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., ... & Zhou, X. (2014). Phylogenomics resolves the timing and pattern of insect evolution. Science, 346(6210), 763–767.

Mullis KB. (1990). The unusual origin of the polymerase chain reaction. Scientific American, 262(4):56–61. 64–5

Muyzer G, de Waal EC, UitterlindenAG. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Appl. Environ. Microbiol. 59:695–700.

Mychek-Londer, J. G., Chaganti, S. R., & Heath, D. D. (2020). Metabarcoding of native and invasive species in stomach contents of Great Lakes fishes. Plos one, 15(8), e0236077.

Nilsson RH, Larsson K-H, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L, Saar I, Kõljalg U, Abarenkov K. (2018). The

UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Research, DOI: 10.1093/nar/gky1022

Ogle, D.H., P. Wheeler, and A. Dinno. (2020). FSA: Fisheries Stock Analysis. R package version 0.8.30, https://github.com/droglenc/FSA.

Oksanen, J. Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E. & H. Wagner (2019). vegan: Community Ecology Package. R package version 2.5-6. https://CRAN.R-project.org/package=vegan

Oliverio, A. M., Gan, H., Wickings, K., & Fierer, N. (2018). A DNA metabarcoding approach to characterize soil arthropod communities. Soil Biology and Biochemistry, 125, 37-43.

Op De Beeck M, Lievens B, Busschaert P, Declerck S, Vangronsveld J, Colpaert JV (2014) Comparison and Validation of Some ITS Primer Pairs Useful for Fungal Metabarcoding Studies. PLoS ONE 9(6): e97629. https://doi.org/10.1371/journal.pone.0097629

Ortiz, L & Chiappa, X. (2017). Microbiología ambiental en México Diagnóstico, tendencias en investigación y áreas de oportunidad. Consejo Nacional de Ciencia y Tecnología (Conacyt).

Parducci, L., Bennett, K. D., Ficetola, G. F., Alsos, I. G., Suyama, Y., Wood, J. R., & Pedersen, M. W. (2017). Ancient plant DNA in lake sediments. New Phytologist, 214(3), 924-942.

Pawlowski, J., Esling, P., Lejzerowicz, F., Cedhagen, T., & Wilding, T. A. (2014). Environmental monitoring through protist next-generation sequencing metabarcoding: Assessing the impact of fish farming on benthic foraminifera communities. Molecular Ecology Resources, 14(6), 1129-1140.

Pawlowski, J., Esling, P., Lejzerowicz, F., Cordier, T., Visco, J. A., Martins, C. I., ... & Cedhagen, T. (2016). Benthic monitoring of salmon farms in Norway using foraminiferal metabarcoding. Aquaculture Environment Interactions, 8, 371-386.

Paz, A., Gonzalez, M., & Crawford, A. J. (2011). Códigos de barras de la vida: introducción y perspectiva. Acta Biológica Colombiana, 16(3), 161–175.

Piper, A. M., Batovska, J., Cogan, N. O., Weiss, J., Cunningham, J. P., Rodoni, B. C., & Blacket, M. J. (2019). Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. GigaScience, 8, 1–22.

Proença, V., Martin, L. J., Pereira, H. M., Fernandez, M., McRae, L., Belnap, J., ... & van Swaay, C. A. (2017). Global biodiversity monitoring: from data sources to essential biodiversity variables. Biological Conservation, 213, 256-263.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Opens external link in new windowNucl. Acids Res. 41 (D1): D590–D596

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.Rproject.org/

Ratnasingham, S, Hebert, PDN. (2007). BOLD: The Barcode of Life Data System (www.barcodinglife.org). Molecular Ecology Notes 7, 355–364. DOI: 10.1111/j.1471-8286.2006.01678.x.

Reyes, M & Arena, L. 2017. Bacterias y hongos con potencial biodegradador de hidrocarburos en diversos ambientes en Microbiología ambiental en México Diagnóstico, tendencias en investigación y áreas de oportunidad. Ed Salazar M.A. Consejo Nacional de Ciencia y Tecnología (Conacyt).

Robinson, G. E., Hackett, K. J., Purcell-Miramontes, M., Brown, S. J., Evans, J. D., Goldsmith, M. R., Lawson, D., Okamuro, J, Robertson, HM & Schneider, D. J. (2011). Creating a buzz about insect genomes. Science, 331(6023), 1386.

Rognes, T. et al., (2016). VSEARCH. Available at: https://github.com/torognes/vsearch .

Roldán-Pérez, G. (2016). Los macroinvertebrados como bioindicadores de la calidad del agua: cuatro décadas de desarrollo en Colombia y Latinoamerica. Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales, 40(155), 254–274.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol., 75:7537–7541.

Shannon, C. E. (1948). A mathematical theory of communication. Bell system technical journal, 27(3), 379–423.

Sheikh, A. H., Thomas, M., Bhandari, R., & Meshram, H. (2016). Malaise trap and insect sampling: Mini Review. Bio Bulletin, 2(2), 35–40.

Sonet, G., Jordaens, K., Braet, Y., Bourguignon, L., Dupont, E., Backeljau, T., ... & Desmyter, S. (2013). Utility of GenBank and the Barcode of Life Data Systems (BOLD) for the identification of forensically important Diptera from Belgium and France. Zookeys, (365), 307.

Song, H., Amédégnato, C., Cigliano, M. M., Desutter-Grandcolas, L., Heads, S. W., Huang, Y., ... & Whiting, M. F. (2015). 300 million years of diversification: elucidating the patterns of orthopteran evolution based on comprehensive taxon and gene sampling. Cladistics, 31(6), 621–651.

Steffen, W., Broadgate, W., Deutsch, L., Gaffney, O., & Ludwig, C. (2015). The trajectory of the Anthropocene: the great acceleration. The Anthropocene Review, 2(1), 81-98.

Steward, A.L., Langhans, S.D., Corti, R. & Datry, T. (2017) The biota of intermittent rivers and ephemeral streams: terrestrial and semiaquatic invertebrates. In: Datry, T., Bonada, N. & Boulton, A. (Eds.), Intermittent Rivers and Ephemeral Streams. Academic Press, London, pp. 245–271.

Sun, Z., Majaneva, M., Sokolova, E., Rauch, S., Meland, S., & Ekrem, T. (2019). DNA metabarcoding adds valuable information for management of biodiversity in roadside stormwater ponds. Ecology and evolution, 9(17), 9712-9722.

Taberlet, P., Bonin, A., Coissac, E., & Zinger, L. (2018). Environmental DNA: For biodiversity research and monitoring. Oxford University Press.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular ecology, 21(8), 2045–2050.

Treonis, A. M., Unangst, S. K., Kepler, R. M., Buyer, J. S., Cavigelli, M. A., Mirsky, S. B., & Maul, J. E. (2018). Characterization of soil nematode communities in three cropping systems through morphological and DNA metabarcoding approaches. Scientific reports, 8(1), 1-12.

Turner, C. R., Uy, K. L., & Everhart, R. C. (2015). Fish environmental DNA is more concentrated in aquatic sediments than surface water. Biological Conservation, 183, 93–102.

Uchida, N., Kubota, K., Aita, S., & Kazama, S. (2020). Aquatic insect community structure revealed by eDNA metabarcoding derives indices for environmental assessment. PeerJ, 8, e9176.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., ... & Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. Molecular ecology, 25(4), 929-942.

Valentin, V., Frédéric, R., Isabelle, D., Olivier, M., Yorick, R., & Agnès, B. (2019). Assessing pollution of aquatic environments with diatoms' DNA metabarcoding: experience and developments from France Water Framework Directive networks. Metabarcoding and Metagenomics, 3, e39646.

Van Bruggen, A. H., & Semenov, A. M. (2000). In search of biological indicators for soil health and disease suppression. Applied Soil Ecology, 15(1), 13–24.

Wagner, L., B. Stielow, K. Hoffmann, T. Petkovits, T. Papp, C. Vágvölgyi, G.S. de Hoog, G. Verkley & K. Voigt. (2013). A comprehensive molecular phylogeny of the Mortierellales (Mortierellomycotina) based on nuclear ribosomal DNA. Persoonia (30), 77–93.

Watts, C., Dopheide, A., Holdaway, R., Davis, C., Wood, J., Thornburrow, D., & Dickie, I. A. (2019). DNA metabarcoding as a tool for invertebrate community monitoring: a case study comparison with conventional techniques. Austral Entomology, 58(3), 675-686.

Wegleitner, B. J., Jerde, C. L., Tucker, A., Chadderton, W. L., & Mahon, A. R. (2015). Long duration, room temperature preservation of filtered eDNA samples. Conservation Genetics Resources, 7(4), 789–791.

Wen C, Wu L, Qin Y, Van Nostrand JD, Ning D, Sun B, et al. (2017) Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform. PLoS ONE 12(4): e0176716. https://doi.org/10.1371/journal.pone.0176716

White, T., Bruns, T., Lee, S., Taylor, J., Innis, M., Gelfand, D., et al. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. in: PCR Protocols: a Guide to Methods and Applications ed. MA. Innis, (Cambridge: Academic Press), 315–322 doi: 10.1016/b978-0-12-372180- 8.50042-1.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

Willerslev, E., Davison, J., Moora, M., Zobel, M., Coissac, E., Edwards, M. E., ... & Taberlet, P. (2014). Fifty thousand years of Arctic vegetation and megafaunal diet. Nature, 506(7486), 47-51.

Wilson, E. O. (1992).The Diversity of Life. W.W. Norton, New York.

Zepeda Mendoza, M. L., Sicheritz-Ponten, T., & Gilbert, M. T. P. (2015). Environmental genes and genomes: understanding the differences and challenges in the approaches and software for their analyses. Briefings in Bioinformatics, 16(5), 745–758.

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., ... & Taberlet, P. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. Molecular ecology, 28(8), 1857–1862.