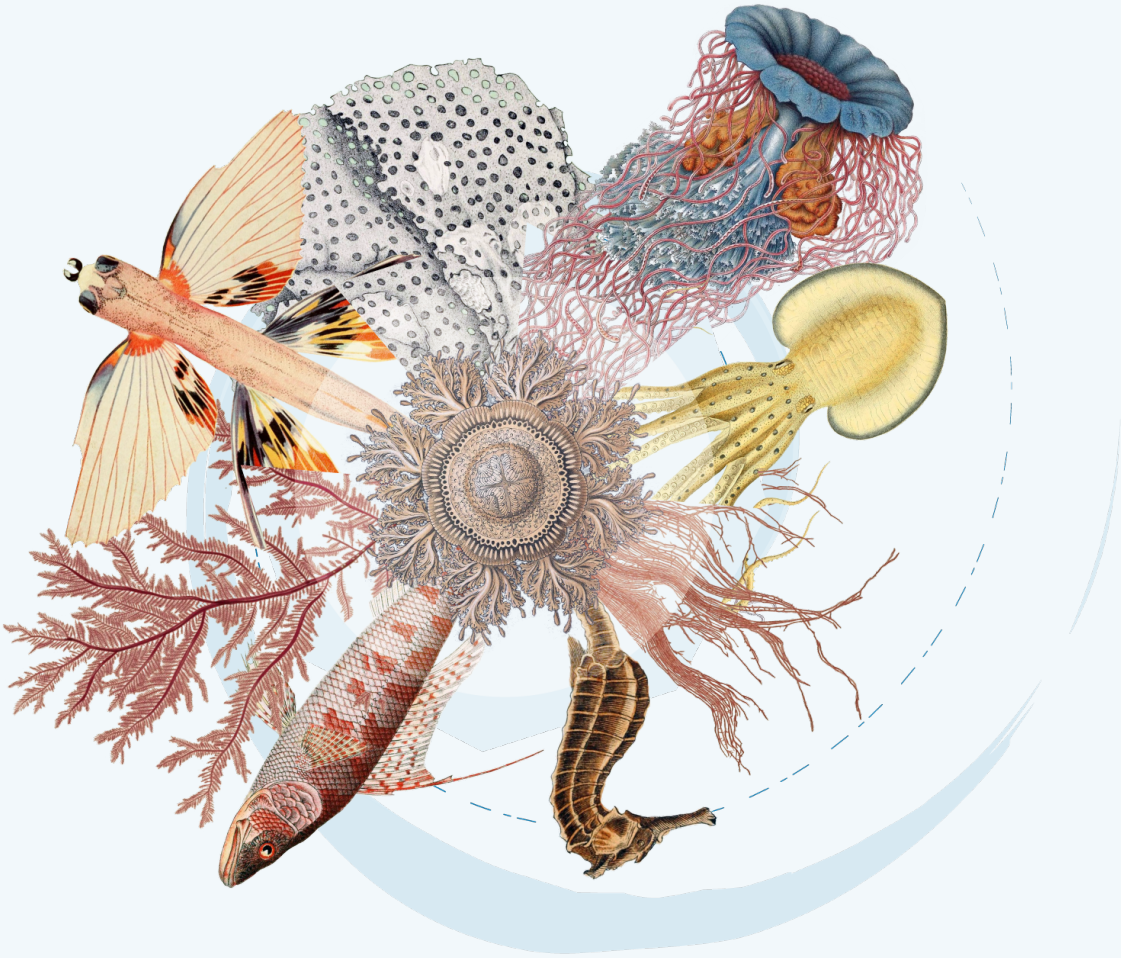

Publicación de datos derivados de ADN a través de plataformas de datos sobre biodiversidad

Kessy Abarenkov • Anders F. Andersson • Andrew Bissett • Anders G. Finstad • Frode Fossøy • Marie Grosjean • Michael Hope • Thomas S. Jeppesen • Urmas Kõljalg • Daniel Lundin • R. Henrik Nilsson • Maria Prager • Pieter Provoost • Dmitry Schigel • Saara Suominen • Cecilie Svenningsen • Tobias Guldberg Frøslev

Versión 1.3.3, 27 Feb 2025

GBIF Global Biodiversity Information Facility

OBIS OCEAN BIODIVERSITY INFORMATION SYSTEM

Tabla de Contenido

Colofón

Citación sugerida

Abarenkov K, Andersson AF, Bissett A, Finstad AG, Fossøy F, Grosjean M, Hope M, Jeppesen TS, Kõljalg U, Lundin D, Nilsson RN, Prager M, Provoost P, Schigel D, Suominen S, Svenningsen C & Frøslev TG (2023) Publicación de datos derivados de ADN a través de plataformas de datos sobre biodiversidad, v1.3. Copenhagen: GBIF Secretariat. https://doi.org/10.35035/doc-vf1a-nr22.

Autores

- Kessy Abarenkov, kessy.abarenkov@ut.ee, Natural History Museum and Botanical Garden, University of Tartu, 46 Vanemuise Street, 51003 Tartu, Estonia
- Anders F. Andersson, anders.andersson@scilifelab.se, Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, 17121 Stockholm, Sweden
- Andrew Bissett, Andrew.Bissett@csiro.au, CSIRO O&A, GPO box 1533, Hobart, Tasmania, 7000, Australia
- Anders G. Finstad, anders.finstad@ntnu.no, Departamento de Historia Natural, Centro de Dinámica de Biodiversidad, Universidad Noruega de Ciencia y Tecnología, Trondheim, Noruega
- Frode Fossøy, Frode.Fossoy@nina.no, Centre for Biodiversity Genetics (NINAGEN), Norwegian institute for nature research (NINA), P.O. Box 5685 Torgarden, NO-7485 Trondheim, Noruega
- Marie Grosjean, mgrosjean@gbif.org, Global Biodiversity Information Facility, Universitetsparken 15, 2100 København Ø, Dinamarca
- Michael Hope, Michael.Hope@ga.gov.au, Atlas of Living Australia, CSIRO National Collections & Marine Infrastructure, GPO Box 1700, Canberra ACT 2601, ealia.
- Thomas S. Jeppesen, tsjeppesen@gbif.org, Global Biodiversity Information Facility, Universitetsparken 15, 2100 København Ø, Denmark
- Urmas Kõljalg, urmas.koljalg@ut.ee, Museo de Historia Natural y Jardín Botánico, Universidad de Tartu, calle 46 Vanemuise, 51003 Tartu, Estonia.
- Daniel Lundin, daniel.lundin@lnu.se, Centro de Ecología y Evolución en Sistemas de Modelo Microbiano - EEMiS, Universidad de Linnaeus, SE-39182 Kalmar, Suecia
- R. Henrik Nilsson, henrik.nilsson@bioenv.gu.se, Universidad de Gothenburg, Departamento de Ciencias Biológicas y Ambientales, Box 461, 405 30 Goëteborg, Suecia
- Maria Prager, maria.pragmer@scilifelab.se, Science for Life Laboratory, Department of Ecology, Environment and Plant Sciences, Stockholm University; Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet
- Pieter Provoost, p.provoost@unesco.org, Ocean Biodiversity Information Systemem, Jacobsenstraat 1, 8400 Oostende, Bélgica
- Dmitry Schigel, dschigel@gbif.org, Global Biodiversity Information Facility, Universitetsparken 15, 2100 København Ø, Dinamarca
- Saara Suominen, s.suominen@unesco.org, Ocean Biodiversity Information System, Jacobsenstraat 1, 8400 Oostende, Bélgica
- Cecilie Svenningsen, csvenningsen@gbif.org, Global Biodiversity Information Facility, Universitetsparken 15, 2100 København Ø, Dinamarca
- Tobias Guldberg Frøslev, tfroeslev@gbif.org, Global Biodiversity Information Facility,

Universitetsparken 15, 2100 København Ø, Denmark

Colaboradores

Las valiosas discusiones con miembros de las redes ELIXIR, iBOL, GGBN, GLOMICON, y la red de OBIS contribuyeron a la compilación de este borrador. Estamos especialmente agradecidos por los aportes y la iniciativa de Andrew Bentley, Matt Blissett, Pier Luigi Buttigieg, Kyle Copas, Camila A. Plata Corredor, Gabriele Dröge, Torbjørn Ekrem, Birgit Gemeinholzer, Quentin Groom, Tim Hirsch, Donald Hobern, Hamish Holewa, Corinne Martin, Raissa Meyer, Chris Mungall, Daniel Noesgaard, Corinna Paeper, Tim Robertson, Maxime Sweetlove, Andrew Young, John Waller, Ramona Walls, John Wieczorek, Lucie Zinger quienes han contribuido en el proceso de revisión de la comunidad de GBIF.

Licencia

URI persistente

https://doi.org/10.35035/doc-vf1a-nr22

Control del documento

Versión 1.3.3 liberada en 27 Feb 2025.

Esta versión añade un párrafo sobre los conjuntos de datos marinos y el Sistema de Información sobre la Biodiversidad del Océano (OBIS) junto con algunas ediciones menores de texto.

Actualización menor (febrero de 2025), agrega información y enlaces a Kit de herramientas de datos de Metabarcoding

Resumen

Cuando se utiliza información genética para describir o clasificar un taxón, la mayoría de los usuarios preveerán su uso en el contexto de la ecología molecular o la investigación genética. Es importante darse cuenta de que una secuencia con coordenadas y una marca de tiempo, es una valiosa ocurrencia sobre biodiversidad que es útil en un contexto mucho más amplio que su propósito original. Para evidenciar este potencial, es necesario que los datos derivados de ADN se puedan rastrear a través de plataformas de datos sobre biodiversidad. Esta guía le enseñará los principios y enfoques para disponer “secuencias con fechas y coordenadas” en el contexto más amplio de los datos sobre biodiversidad. La guía cubre opciones de esquemas y términos particulares, fallas comunes y buenas prácticas, sin entrar en detalles específicos de las plataformas. Esto beneficiará a cualquiera que esté interesado en una mejor exposición de los datos derivados de ADN mediante plataformas de datos sobre biodiversidad, incluidos los portales nacionales.

Prefacio

El trabajo sobre esta guía comenzó a partir de discusiones en la biodiversity_next conference en 2019, consolidando insumos de varios recursos como:

- Informe final sobre el proyecto de la plataforma de ciencia futura ambiental

- Nota de Blog sobre registros de eDNA disponibles en ALA
- ADN ambiental (eDNA) en ALA
- ALA eDNA data template
- Norwegian Criteria for depositing eDNA samples and data, including vouchered specimens
- Molecular biodiversity data in SBDI, Sweden
- GBIF resources (How) can I publish molecular/sequence/DNA based data to GBIF?
- Molecular data in GBIF
- GBIF quick guide to publishing data and detailed guides to publishing data
- Cómo publicar datos a través de GBIF, así como una visión general del DwC/extensiones.
- Genomic Biodiversity Interest Group

1. Introducción

1.1. Justificación

Los últimos 20 años han traído una mayor comprensión del inmenso poder de los métodos moleculares para documentar la diversidad de la vida en la Tierra. Los sustratos aparentemente sin vida y mundanos, como el suelo y el agua de mar, resultan estar llenos de vida, aunque tal vez no de una manera que el observador casual pueda apreciar de inmediato. Estudios basados en el ADN, han demostrado que grupos de organismos como hongos, insectos, oomycetes, bacterias y archaeas están en todas partes, aunque a menudo no podemos observarlos físicamente (Debroas et al. 2017). Los beneficios de los métodos moleculares no se limitan al mundo microscópico: hay muchos organismos, como algunas especies de peces, que al menos teóricamente se puede observar físicamente, pero muy costosos en términos de intensidad y esfuerzo, y tal vez, los métodos para hacerlo sean invasivos (Boussarie et al. 2018). En tales situaciones, los datos del ADN nos permiten registrar la presencia (y la presencia pasada) de estos órganismos de forma no invasiva y con un esfuerzo mínimo. Estos desarrollos significan que no siempre necesitamos manifestaciones físicas tangibles de todos los órganismos presentes en algún lugar para registrarlos. Todos los organismos, sean o no físicamente observables, pueden ser importantes a la hora de entender la biodiversidad, la ecología y la conservación biológica.

Los datos derivados de ADN nos permiten registrar taxones apenas visibles o inobservables que caen por fuera del radar de los protocolos para el trabajo de campo, listas de chequeo, depósitos en colecciones científicas, etc. La madurez actual de las metodologías de ADN nos permite registrar la presencia de estos organismos con un nivel de detalle que supera el de las observaciones macroscópicas de los organismos en general. Sin embargo, teniendo en cuenta que las metodologías del ADN vienen con sus propios problemas y sesgos, es importante aprovechar este momento para definir y acordar cómo debemos registrar e informar sobre un organismo presente en algún substrato o localidad a través de datos moleculares. Hacerlo ayudará a evitar ineficiencias significativas que han sido reportadas en otros dominios, en los cuales, la falta de estándares y directrices ha llevado a tener estructuras de datos muy heterogéneas e incomparables en gran medida (Berry et al. 2021; Leebens-Mack et al. 2006; Yilmaz et al. 2011; Nilsson et al. 2012; Shea et al. 2023). Además, una documentación clara del procesamiento computacional desde la lectura de la secuencia en bruto hasta la deducción de la observación de una especies, permitirá re-analizarlas cuando aparezcan métodos mejorados.

Los registros de especies derivadas de ADN deben ser tan estandarizados y reproducibles como sea posible, independientemente de que las especies detectadas tengan o no nombres científicos formales. En algunos casos, tales registros de ocurrencia apuntarán a propiedades geográficas y ecológicas de las especies que eran desconocidas, enrriqueciendo el conocimiento sobre estos

taxones. En otros casos, los datos pueden permitirnos amalgamar y visualizar información sobre especies no descritas actualmente, lo que potencialmente acelera su posible descripción formal. La capacidad de recopilar datos utilizables incluso para especies no nombradas contribuye significativamente a las muchas maneras en que GBIF y otras plataformas de datos sobre biodiversidad indexan el mundo vivo, y ponen este conocimiento a disposición de todos y para una variedad de propósitos, incluyendo la conservación de la biodiversidad. Las estimaciones recientes sugieren que al menos el 85 por ciento de todas las especies existentes no están descritas (Mora et al. 2011; Tedesco et al. 2014). Los estándares de datos existentes han sido diseñados para la minoría de taxones que han sido descritos. Las buenas prácticas para tratar con datos derivados de ADN ayudarán a caracterizar las ocurrencias de todos los organismos, ya sean descritos o no.

Esta guía describe las formas en las que se deben reportar los datos de ocurrencia derivados del ADN para su inclusión estandarizada en GBIF y otras plataformas de datos sobre biodiversidad. Esta guía no expresa ninguna opinión sobre el acceso y los beneficios de compartir información de secuencias digitales, un tema ampliamente discutido a través del Convenio sobre la Diversidad Biológica (CDB). Sin embargo, vale la pena señalar que los códigos de barras genéticos y el metabarcoding son típicamente genes o fragmentos de ADN no codificante, que no son adecuados para la explotación comercial. Tal como la documentación de secuencias a través de International Nucleotide Sequence Database Collaboration (INDSC) es una norma generalizada en la investigación basada en secuencias, la publicación de datos de ocurrencia originados a través de secuencias no implica la publicación de nuevas secuencias. En la mayoría de los casos ya se han colocado en un depósito genético público. Por lo tanto, esta guía trata el valor añadido de derivar datos de ocurrencia espacio-temporal y nombres basados en ADN de datos derivados de ADN, más que el valor de la propia información genética. Además de tratar los datos derivados de secuencias, esta guía también incluye sugerencias para publicar datos de ocurrencia de especies derivadas de los análisis qPCR o ddPCR.

Reportar las ocurrencias derivadas de DNA-de una manera abierta y reproducible trae muchos beneficios: en particular, aumenta la citabilidad, destaca los taxones afectados en el contexto de la conservación biológica y contribuye al conocimiento taxonómico y ecológico. Además, también proporciona un mecanismo para almacenar registros de ocurrencia de especies no descritas. Cuando estos taxones, aún por describir, finalmente se vinculen a un nuevo nombre linneano, todos estos registros de ocurrencia estarán disponibles inmediatamente. Cada uno de estos beneficios proporciona una fuerte justificación para que los profesionales adopten las prácticas esbozadas en esta guía, ayudándolos a destacar una proporción significativa de la biodiversidad existente, acelerando su descubrimiento e integrándolo en la conservación biológica y la elaboración de políticas.

1.2. Público objetivo

Esta guía ha sido desarrollada para múltiples audiencias objetivo: estudiantes que planean un primer estudio basado en ADN, investigadores con secuencias antiguas y tablas de abundancia que quieren revivir o conservar, especialistas en datos sobre biodiversidad que son nuevos en los datos derivadas de ADN y bioinformáticos familiarizados con los datos de secuencia pero nuevos en las plataformas de datos sobre biodiversidad. La guía no se dirige directamente a los usuarios de datos moleculares en plataformas de datos sobre biodiversidad, pero estos usuarios pueden encontrar section 1.7 Resultados particularmente interesantes. La intención de los autores es proporcionar orientación sobre la publicación de datos y atributos asociados a las secuencias genéticas a través de plataformas de datos sobre biodiversidad en general.

El flowchart esboza los pasos de procesamiento involucrados en la publicación de datos de biodiversidad molecular derivados del amplicón, en repositorios como GBIF y plataformas nacionales de datos sobre biodiversidad, incluyendo aquellos construidos en la plataforma ALA. El enfoque de esta guía está principalmente en los pasos posteriores a la llegada de secuencias crudas FASTQ del paso de secuenciación. Al familiarizarse con el diagrama de flujo —y tener en cuenta cualquier paso

que parezca familiar o poco claro— los usuarios podrán navegar por el contenido incluido en la guía.

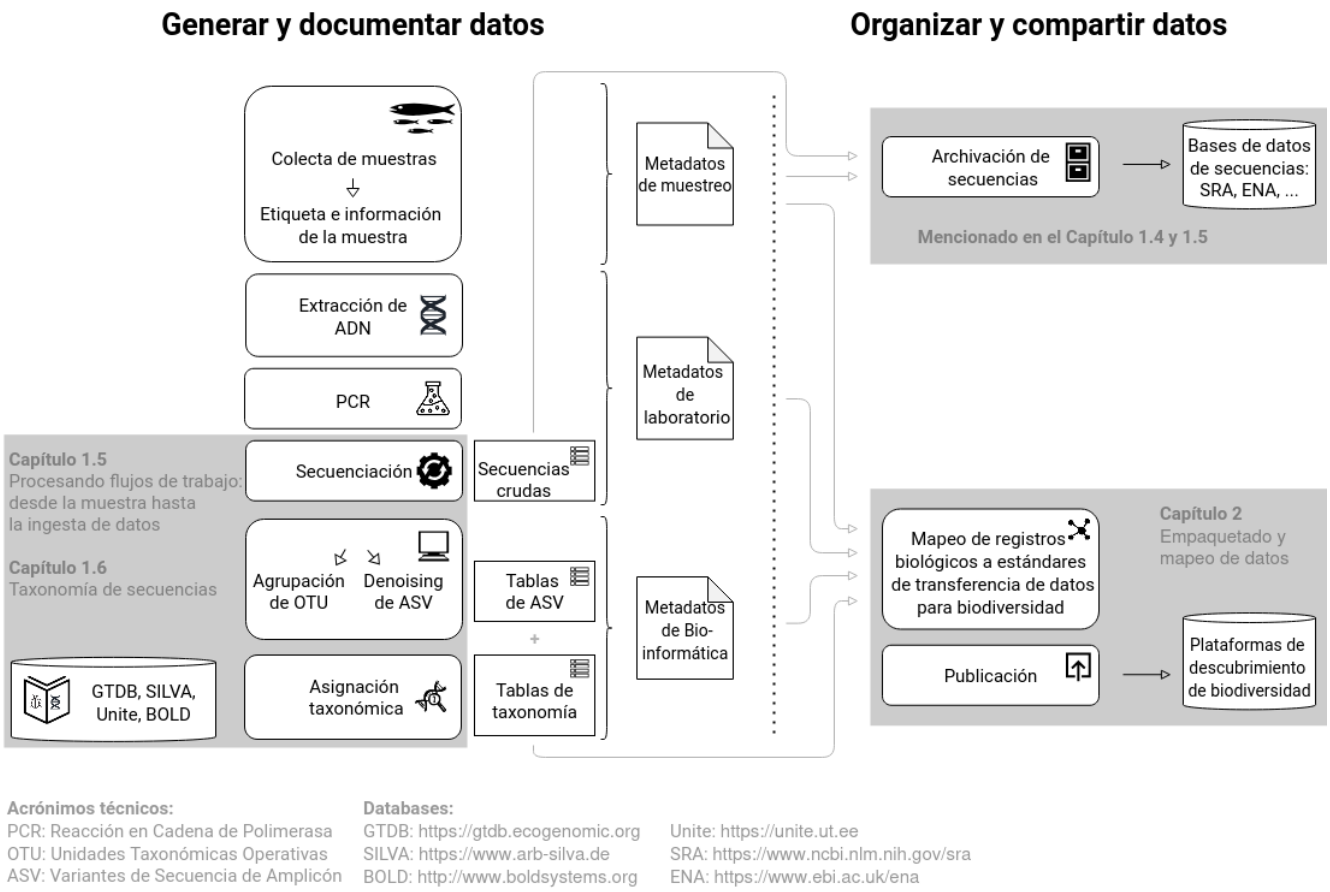

Figura 1. Flujo de trabajo general para los datos sobre biodiversidad derivados de secuencias de ADN, tal y como se describe en esta guía.

Hemos hecho todo lo posible por presentar la información en esta guía de forma que sea útil para cada una de las audiencias descritas anteriormente, pero una lectura a fondo (e.g. GBIF quick guide to data publishing) puede ser requerida en ciertos casos.

1.3. Introducción a datos de ocurrencias derivados de DNA

Los datos biológicos derivados de ADN incluyen información derivada de ADN de organismos individuales, pero también de ADN ambiental (eDNA, es decir, ADN extraído de muestras ambientales, Thomsen & Willerslev 2015) y de muestras masivas que componen a muchos individuos (e.g. Las muestras de plankton o muestras de trampas de Malaise que consisten en múltiples individuos de múltiples especies). En la actualidad, el mayor volumen de datos de ocurrencia derivados de la ADN privienen del eDNA. Desde que los métodos analíticos y los productos finales son en gran medida similares para todas las fuentes de muestras, la siguiente discusión se centrará en el eDNA (§ 2.1.1 y § 2.1.2), teniendo en cuenta que el contorno es aplicable a otras fuentes. Las investigaciones a menudo utilizan secuencias selectivas de marcadores genéticos taxonómicos e informativos, pero también pueden usarlos por ejemplo, enfoques basados en qPCR, que no resultan directamente en datos de secuencia de ADN (§ 2.1.3 y [mapping-ddpcr-qpcr-data]). Esta guía puede parecer pesada en términos relacionados con el ADN; si este es el caso, consulte el [glossary].

1.3.1. ADN ambiental como fuente de datos de ocurrencia derivados de DNA

El término ADN ambiental se ha utilizado desde 1987, cuando se uso por primera vez para describir el ADN de microbios encontrado en muestras de sedimento (Ogram et al. 1987). Actualmente, el eDNA se

utiliza más ampliamente para describir una compleja mezcla de ADN proveniente de diferentes organismos (Taberlet et al. 2018 and 2012). Este eDNA incluye todo el ADN obtenido de una muestra ambiental especifica, independientemente del tipo de sustrato o las especies que contenga. Se puede extraer de una amplia gama de fuentes, incluyendo las células de la piel y el cabello, saliva, suelo, heces, organismos vivos o muertos recientemente (Pietramellara et al. 2009). El ADN ambiental suele ser suficiente para representar todos los organismos encontrados dentro de una muestra tomada, sin embargo, en la práctica, la presencia del ADN en la muestra ambiental depende de la selección del hábitat de los organismos, tamaño del cuerpo, morfología y nivel de actividad. Además, de los métodos de muestreo utilizados para la capturar del ADN (Taberlet et al. 2018) y el estado de degradación del mismo.

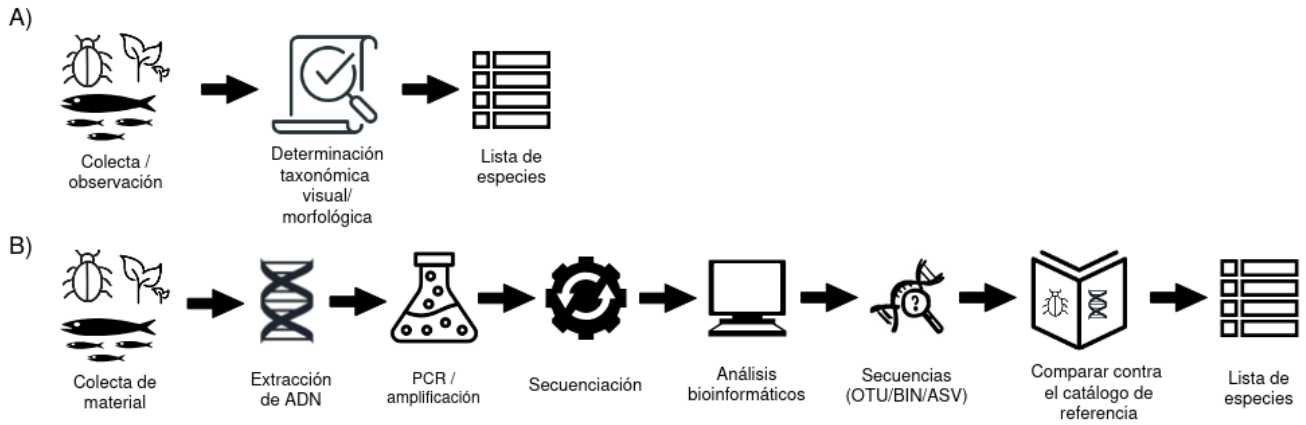

Figura 2. Representación de los procesos de muestreo que comparan la recopilación de datos por A) métodos tradicionales de muestreo ecológicos y de biodiversidad, y B) una representación simplificada de estudios basados en eDNA, ejemplificado por metabarcoding. Para eDNA, la mayoría de los pasos hasta la secuenciación incluyen réplicas técnicas o biológicas para identificar la contaminación de las muestras, así como falsos positivos y falsos negativos en los resultados, procesos que generan una estructura jerárquica en los datos y metadatos. Sin embargo, los estudios a menudo incluirán ambos tipos de muestreo. Por ejemplo, si el 'Catálogo de referencia' utilizado en B) no contiene todas las especies relevantes de un determinado grupo de organismos, será necesario volver a A). También puede ser que al "comparar contra el catálogo de referencia" se produjeran resultados inesperados o improbables, en cuyo caso se requerirán más estudios que utilicen la metodología tradicional para determinar si las especies identificadas por el análisis bioinformático pueden ser verificadas.

Por lo tanto, el eDNA es un tipo de muestra, no un método, que incluye el ADN derivado de cualquier muestra ambiental en lugar de la captura y secuenciación de un individuo específico. Tales tipos de muestras incluyen agua, suelo, sedimento y aire; pero también muestras de contenido intestinal y tejidos (planta/animal) donde el ADN del hospedero no es el objetivo (Taberlet et al. 2018). Existen una serie de métodos analíticos para el estudio del ADN ambiental, estos pueden dividirse en dos clases principales: 1) aquellos que tienen como objetivo detectar un organismo específico y 2) aquellos que describen un ensamblaje o una comunidad de una variedad de organismos. Diferentes métodos de análisis generarán diversos tipos y volúmenes de datos. La mayoría de las veces las concentraciones de ADN son bajas y las réplicas técnicas y biológicas deben utilizarse para validar la detección de especies.

Numerosos estudios señalan que para muestras de agua, los análisis basados en el eDNA pueden tener una mayor probabilidad de encontrar especies raras y difíciles de muestrear que los métodos convencionales (Thomsen et al. 2012; Biggs et al. 2015; Valentini et al. 2016; Bessey et al. 2020). Lo mismo puede aplicar en otros entornos. Por lo tanto, el eDNA puede ser adecuado para monitorear especies raras documentadas en listas rojas y especies invasoras indeseables que a menudo tienen bajas densidades y que son difíciles de detectar con métodos convencionales. Ya que se pueden detectar trazas de ADN, aunque el organismo ya no está presente en la zona. Los métodos de ADN ambiental son capaces de detectar organismos crípticos, especialmente aquellos que son pequeños y

no pueden ser detectados a simple vista (ej. bacterias y hongos). Además, el eDNA también se puede utilizar para la observación de muchas especies simultáneamente, y puede describir comunidades biológicas enteras o los componentes principales de estas (Ekrem & Majaneva 2019).

Algunos estudios muestran una relación entre la cantidad de ADN de una determinada especie en una muestra ambiental y la biomasa de la especie en el medio ambiente. Por lo tanto, se puede pensar también que el ADN ambiental, permite una estimación semi cuantitativa (objetivo indirecto) de la biomasa del organismo, tanto de muestras ambientales como de muestras en masa (Takahara et al. 2012; Thomsen et al. 2012; Andersen et al. 2012; Ovaskainen et al. 2013; Lacoursière-Roussel et al. 2016; Thomsen et al. 2016; Valentini et al. 2016; Fossøy et al. 2019; Yates et al. 2019; Doi et al. 2017). Sin embargo, otros estudios muestran poca correlación entre la cantidad de ADN ambiental y la densidad estimada de una población (Knudsen et al. 2019). Con frecuencia se debaten diferentes sesgos como la PCR, la cuantificación, la mezcla entre otros. Por ejemplo, las mudas, la reproducción y la muerte masiva de crustáceo pueden contribuir a aumentar su representatividad en el ADN ambiental en el agua, mientras que la turbidez y la mala calidad del agua reducen la cantidad de ADN ambiental detectable (Strand et al. 2019). Por lo tanto, animamos a los publicadores de datos a proporcionar tanto el recuento de lectura para cada OTU o ASV por muestra, como el recuento de lectura total por muestra, ya que esta información es necesaria para que los usuarios hagan sus propias conclusiones sobre la presencia/ausencia y abundancia (relativa).

1.3.2. ADN-metabarcoding: datos derivados de secuencias

La generación de datos derivados de secuencias actualmente está aumentando rápidamente debido al desarrollo de DNA-metabarcoding. Este método utiliza primers generales para generar miles a millones de cortas secuencias de ADN para un determinado grupo de organismos, con la ayuda de secuencias de alto rendimiento (HTS, alt. secuenciación de nueva generación (NGS)). Al comparar cada secuencia de ADN con una base de datos de referencia como GenBank (Benson et al. 2006), BOLD (Ratnasingham et al. 2007), or UNITE (Nilsson et al. 2019), cada secuencia se puede asignar a una especie o a un taxón de rango superior. DNA-metabarcoding se utiliza para muestras procedentes tanto de entornos terrestres como acuáticos, incluyendo agua, suelo, aire, sedimentos, biopelículas, plancton, muestras en masa y caras, identificando simultáneamente cientos de especies (https://doi. rg/10.1016/j.gecco.2019.e00547[Ruppert et al. 2019^]).

The identification and classification of organisms from sequence data and marker-based surveys depends on access to a reference library of sequences taken from morphologically identified specimens that are matched against the newly generated sequences. The efficacy of classification depends on the completeness (coverage) and the reliability of reference libraries, as well as the tools used to carry out the classification. These are all moving targets, making it essential to apply taxonomic expertise and caution in the interpreting results ([taxonomy-of-sequences]). Availability of of all verified amplicon sequence variants (Callahan et al. 2017) allow for precise reinterpretation of data, intra-specific population genetic analyses (Sigsgaard et al. 2019) and is likely to increase identification accuracy, and for this reason we recommend to share (unclustered) ASV data. In 2024, GBIF started the Metabarcoding Data Programme to facilitate the publication of eDNA metabarcoding data through GBIF.

1.3.3. Metagenómica: datos derivados de secuencias

Los datos de diversidad derivados de secuencias también pueden ser generados utilizando métodos metagenómicos libres de amplificación, mediante los cuales todo el ADN de una muestra está destinado a la secuenciación (Tyson & Hugenholtz 2005), en lugar de amplicones o códigos de barras específicos, como se describio anteriormente. Los datos de diversidad derivados de secuencias, obtenidos de secuenciación metagenómica pueden presentarse en forma de coincidencias de secuencias con bases de datos de genes anotadas (como se indicó anteriormente) o (casi) como genomas ensamblados con metagenomas completos (MAG). Mientras que los métodos de

metabarcoding siguen dominando en términos de información de diversidad derivados de secuencias, los datos de metagenómica se están volviendo más importantes como lo demuestra el creciente número de MAGS y su utilidad para informar la filogenia y la taxonomía (Parks et al. 2020); la discusión de la rápida evolución de los métodos asociados con el análisis del metagenoma está más allá del alcance de este documento. Este documento utiliza el metabarcoding como modelo de discusión sobre conceptos y métodos para publicar datos de diversidad derivados de secuencias y aunque las rutas bioinformáticas difieren de los datos metagenómicos, el resultado final (una secuencia, a menudo en forma de contig/ensamblaje) es congruente con los conceptos sugeridos para los datos de metabarcoding (es decir, muestras específicas, colecta de muestras, generación de datos y el procesamiento de los metadatos del flujo de trabajo deben capturarse).

1.3.4. qPCR/ddPCR: datos de los registros de especies

Para la detección de especies específicas en muestras de eDNA, la mayoría de los análisis incluyen primers, qPCR (Reacción en cadena de polimerasa cuantitativa) o ddPCR (Reacción en cadena de polimerasa digital en gota) específicos de cada especie. Estos métodos no generan secuencias de ADN, y los datos de registros de especies dependen completamente de la especificidad de los primers/ensayos. Por lo tanto, hay recomendaciones estrictas para validar dichos ensayos y los requisitos para publicar los datos (Bustin et al. 2009, Huggett et al. 2013), así como la preparación para las pruebas en el seguimiento de rutinas (Thalinger et al. 2020). Los análisis de muestras eDNA usando qPCR requieren pocos recursos y pueden hacerse en la mayoría de los laboratorios de ADN. El primer ejemplo de uso de muestras de agua de eDNA utilizó qPCR para detectar la especie invasora de rana toro americana (*Rana catesbeiana*) (Ficetola et al. 2008). Los análisis de qPCR de muestras de agua eDNA se utilizan regularmente para detectar especies específicas de peces, anfibios, moluscos, crustáceos, entre otros, así como sus parásitos (Hernandez et al. 2020, Wacker et al. 2019, Fossøy et al. 2019, Wittwer et al. 2019). La detección de eDNA mediante qPCR genera datos importantes de registros de especies individuales.

1.4. Introducción a la publicación sobre biodiversidad

Publicar datos sobre biodiversidad es en gran medida un proceso de hacer que los datos de registros biológicos de las especies sean encontrables, accesibles, interoperables y reutilizables, de acuerdo con los principios FAIR (Wilkinson et al. 2016). Las plataformas de datos sobre biodiversidad ayudan a exponer y descubrir datos de secuencias genéticas como registros de ocurrencias de biodiversidad junto con otros tipos de datos, tales como muestras de colecciones de museos, observaciones científicas ciudadanas y las clásicas encuestas de campo. La estructura, gestión y almacenamiento de cada fuente de original de datos variará de acuerdo con las necesidades de cada comunidad. Las plataformas de datos sobre biodiversidad soportan el descubrimiento, acceso y reutilización de datos, al hacer que estos conjuntos de datos individuales sean compatibles entre sí, abordando inconsistencias taxonómicas, espaciales y de otro tipo en los datos de biodiversidad disponibles. Hacer que los datos estén disponibles a través de puntos únicos de acceso respalda la investigación, gestión y política en datos a gran escala.

Se utilizan una serie de estándares para datos generales sobre biodiversidad (https://www.gbif.org/standards), y un conjunto separado de estándares para los datos de secuencias genéticas (see MIxS and GGBN). Esta guía refleja algunos de los esfuerzos en curso por incrementar la compatibilidad entre los estándares para la biodiversidad general y los datos genéticos. A menudo, los estándares resaltan los subconjuntos de elementos más importantes o los que son más frecuentemente aplicables, estos pueden ser referenciados como "cores". El formato preferido para publicar datos en las redes de GBIF y ALA es actualmente el Darwin Core Archive (DwC-A) usando el estándar de datos Darwin Core (DwC). En la práctica, esta es una carpeta comprimida (archivo zip) que contiene archivos de datos en formato de texto estándar delimitado por comas o tabulaciones, un archivo de metadatos (eml.xml) que describe el recurso de datos y un archivo meta (meta.xml) que especifica la

estructura de los archivos y los elementos que lo incluyen. El empaquetamiento estandarizado garantiza que los datos puedan viajar entre sistemas utilizando protocolos de intercambio de datos específicos. La Section 2 de esta guía proporciona recomendaciones para el mapeo de los archivos de datos, mientras que las pautas y herramientas para construir los archivos xml pueden ser encontradas aquí: TDWG, GBIF, y ALA.

Una parte central del proceso de estandarización es el mapeo de los elementos, que es requerido para transformar la estructura original del elemento (columna) en una exportación de datos fuente a una estructura de campo estándar. La estandarización también puede afectar individualmente el contenido de los elementos dentro de cada registro, por ejemplo, al recalcular coordenadas a un sistema común, reorganizando los elementos de las fechas o mapeando el contenido de los elementos a un conjunto de valores estándar, a menudo llamado vocabulario. El proceso de estandarización también proporciona una oportunidad de mejorar la calidad de los datos, por ejemplo, rellenando omisiones, corrigiendo errores tipográficos y espacios extra, y manejando el uso inconsistente de los elementos. Tales mejoras contribuyen al aumento en la calidad de los datos e incrementan su idoneidad para la reutilización, pero al mismo tiempo, los datos publicados en cualquier estado son mejores que los datos que permanecen sin publicar o son inaccesibles. La estandarización es típicamente aplicada a una copia o a una exportación desde la fuente de datos, dejando el original intacto.

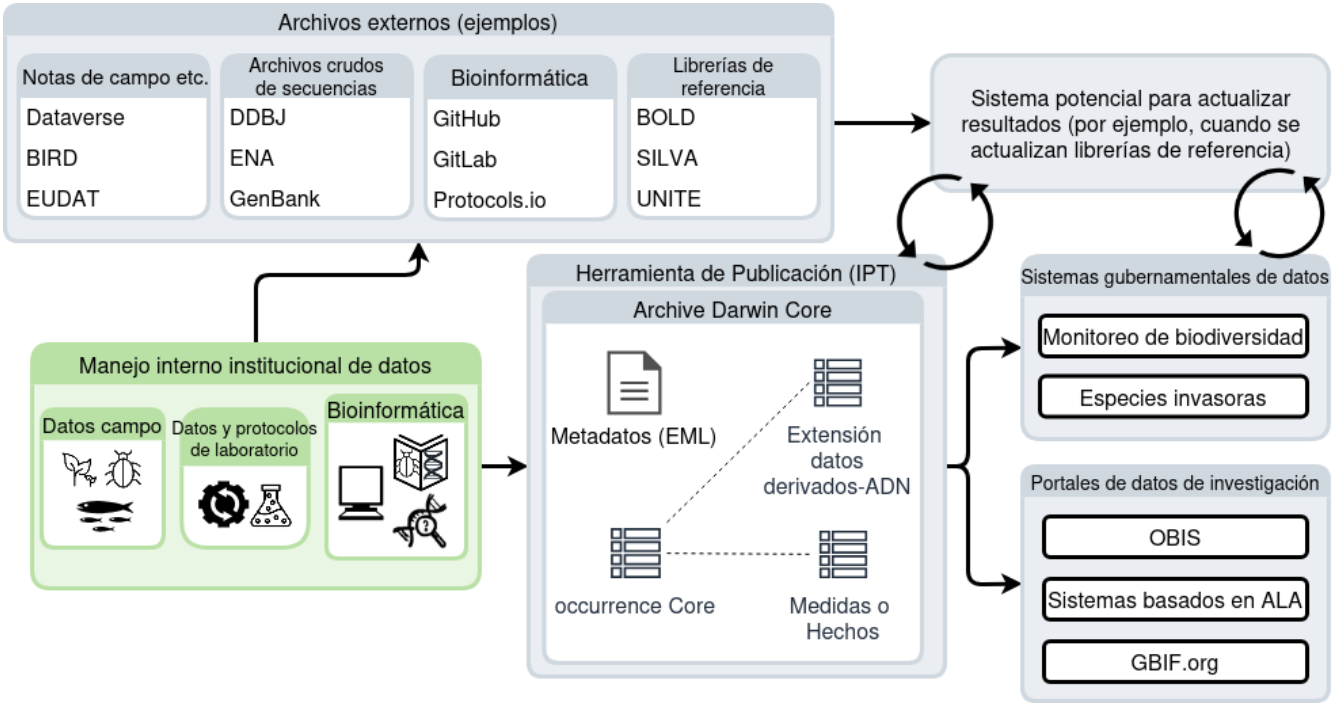

Figura 3. Esquema de una plataforma para informar y publicar secuencias de ADN y metadatos asociados (cuadro verde) basado en sistemas y estándares de datos existentes (cuadros grises). Un sistema previsto para la actualización regular de los resultados (basado en la lectura de datos de máquina a máquina, cuadro blanco) puede leer y actualizar el archivo Darwin Core o varios sistemas de administración. La transferencia de datos entre los distintos elementos (flechas negras) requerirá diversos grados de transformación y armonización de los datos y puede incluir una evaluación de la calidad mecánica o humana.

Una vez que un conjunto de datos ha pasado por estos procesos de estandarización y calidad de datos, debería ser puesto en una ubicación en línea accesible y ser asociado con metadatos relevantes. Los metadatos o información sobre el conjunto de datos incluyen parámetros clave que lo describen y mejoran aún más su capacidad de descubrimiento y reutilización. Los metadatos deberían incluir otros elementos importantes, tales como la autoría, los Identificadores de Objetos Digitales (DOI, por sus siglas en inglés) afiliaciones organizacionales y otra información de procedencia, así como información procedimental y metodológica de cómo se recopiló y curó el conjunto de datos. Recomendamos proporcionar una descripción de los detalles del flujo de trabajo y las versiones,

incluyendo el control de calidad en la methods section en el archivo EML.

Los conjuntos de datos y sus metadatos asociados son indexados por cada portal de datos: este proceso permite a los usuarios consultar, filtrar y procesar datos a través de los API's y portales web. A diferencia de las publicaciones en revistas, los conjuntos de datos pueden ser productos dinámicos que pasan por múltiples versiones con un número de registros que evoluciona y campos de metadatos mutables bajo el mismo título y DOI.

Tenga en cuenta que se espera que los poseedores de las secuencias genéticas carguen y archiven los datos sin procesar en repositorios tales como el de NCBI's SRA, EMBL's ENA o DDBJ. El tema de archivar la secuencia no se aborda aquí, pero a manera de ejemplo, Penev et al. (2017) proporciona una visión general sobre la importancia de presentación de los datos y directrices en relación con la publicación científica. Plataformas de datos sobre biodiversidad tales como ALA, GBIF y la mayoría de portales nacionales sobre biodiversidad no son archivos ni repositorios de lectura de secuencias sin procesar y sus archivos asociados. Sin embargo, destacamos la importancia de mantener vínculos entre esos datos primarios y registros biológicos derivados en la Section 2.

1.5. Procesando flujos de trabajo: desde la muestra hasta la ingesta de datos

Los datos de Metabarcoding se pueden producir a partir de diferentes plataformas de secuenciación (Illumina, PacBio, Oxford Nanopore, Ion Torrent, etc.) que se basan en diferentes principios para la lectura y la generación de datos que difieren con respecto a la longitud de lectura, si las secuencias son simples o emparejadas, etc. Actualmente la plataforma de lectura corta Illumina es la más ampliamente adoptada, y como tal es la base de las descripciones aquí. Sin embargo, el procesamiento bioinformático de los datos sigue los mismos principios generales (QC, denoising, clasificación) independientemente de la tecnología de secuenciación utilizada (Hugerth et al. 2017, Figura 2).

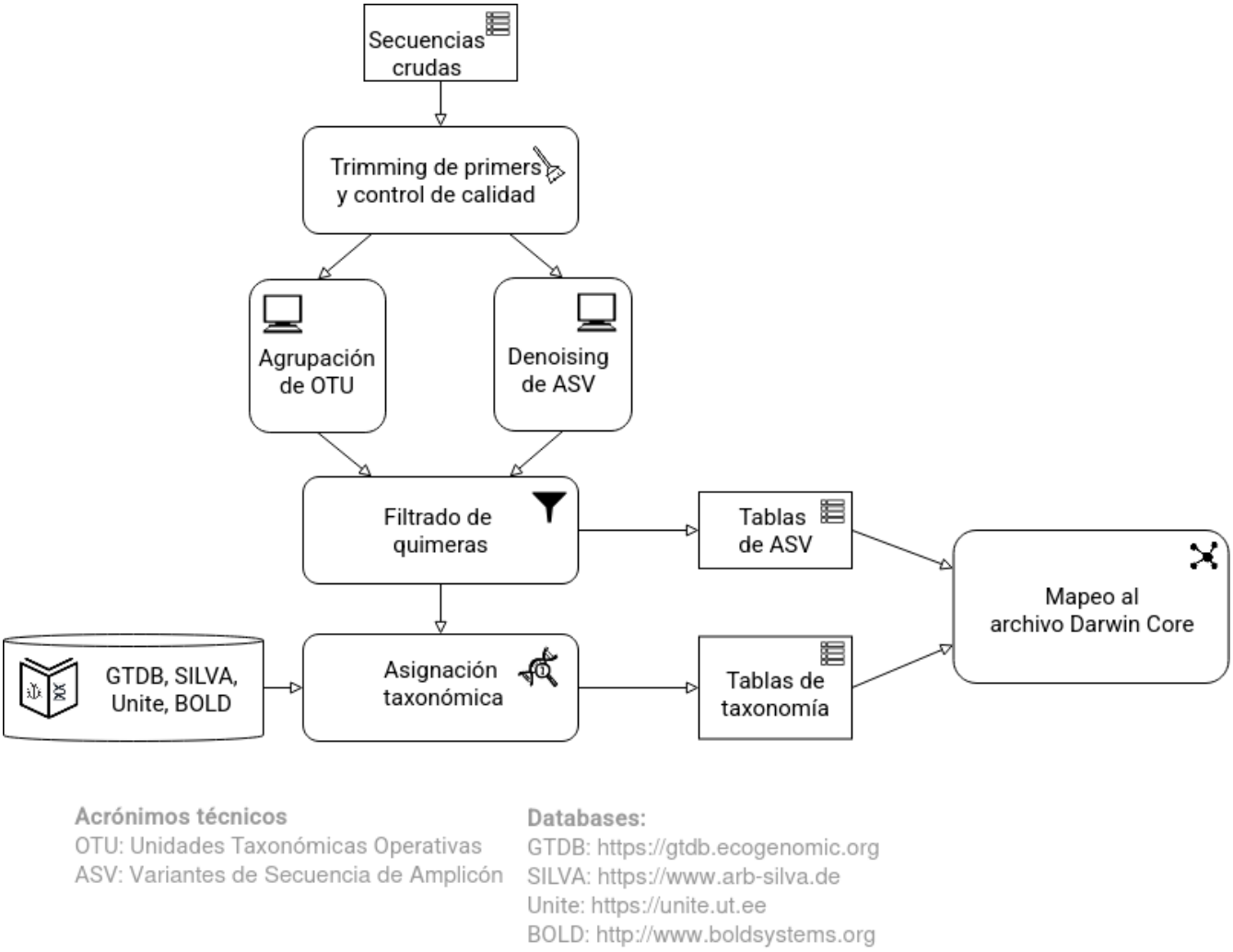

Figura 4. Esquema del procesamiento bioinformático del metabarcoding.

Por lo general, las secuencias de ADN son pre-procesadas eliminando las secuencias primarias y, dependiendo del método de secuenciación utilizado, bases de baja calidad, generalmente hacia los extremos de la secuencia 5' y 3'. Se eliminan las secuencias que no cumplen los requisitos de longitud, calidad general, presencia de primers, etiquetas, etc.

Las secuencias pre-procesadas pueden ser asignadas a un taxón al compararlas con las bases de datos de referencia. Cuando las bases de datos de referencia están incompletas, la clasificación de secuencias puede hacerse sin identificaciones taxonómicas, o agrupando secuencias en unidades taxonómicas operativas basadas en su similitud (OTUs; Blaxter et al. 2005) o mediante el denoising de los datos, p.e. detección y exclusión explícita de secuencias de errores PCR/secuenciación para producir variantes de secuencia de amplicón (ASV; también referida como cero radio OTU (zOTU)). Los intentos de denoising para corregir errores que se han introducido en el PCR y/o pasos de secuenciación, tal que las secuencias denoised son el conjunto de secuencias biológicamente reales únicas presentes en la mezcla de la secuencia original. En caso de secuencias de extremos emparejados, las secuencias de adelante e inversa pueden ser denoised por separado antes de fusionarse o fusionarse antes de ser denoised. Las ASV en el conjunto resultante pueden diferir por tan solo una base que es indicativa de variación de secuencia inter o intraspectiva. Operacionalmente, las ASV pueden ser consideradas como OTUs sin radio definido y, mientras que los denoising algoritmos son normalmente muy buenos, no eliminan por completo los problemas de las secuencias de sobredivisión o repartición.

El PCR utilizado para generar la biblioteca de secuenciación puede resultar en la generación de secuencias artefactuales en forma de quimeras; una secuencia única que se origina a partir de múltiples secuencias parentales. Estas secuencias pueden ser detectadas bioinformáticamente y

eliminarse, y esto se hace típicamente después del denoising.

Finalmente, las secuencias preprocesadas, OTU o ASV, son clasificadas taxonómicamente al compararlas con una base de datos de secuencias anotadas (a menudo referidas como bibliotecas de referencia, véase [taxonomy-of-sequences]). Como en los pasos anteriores, varios métodos alternativos están disponibles. La mayoría de estos se basan en alinear de las secuencias de metabarcoding con las secuencias de referencia o en contar los k-mers compartidos (secuencias cortas exactas).

Existen varias herramientas y algoritmos de código abierto para el procesamiento bioinformático de datos de metabarcoding (QIIME2 (Bolyen et al. 2019), DADA2 (Callahan et al. 2016), SWARM (Mahé et al. 2014), USEARCH (Edgar 2010), Mothur (Schloss et al. 2009), LULU (Frøslev et al. 2017), PROTAX (Somervuo et al. 2016), VSEARCH (Rognes et al. 2016)). Dada la existencia de muchos flujos de trabajo populares y bien utilizados, a continuación hacemos algunas recomendaciones sobre el análisis de datos para enviarlos a plataformas de datos sobre biodiversidad. Esto no sugiere que estos sean los mejores métodos o los más apropiados para todos los fines, sino que es un intento de fomentar la presentación de datos relativamente estandarizados que puedan compararse fácilmente a través de las plataformas. Si es posible, debería utilizarse un flujo de trabajo bien documentado y mantenido (por ejemplo, nf-core/ampliseq pipeline). Los metadatos deben incluir los detalles del flujo de trabajo y las versiones, ya sea en los pasos de los métodos en los metadatos, o como referencia en el campo SOP en la extensión de datos derivados de ADN (ver el mapeo en Table 4). Los datos de secuencia deben depositarse en un archivo de nucleótidos apropiado NCBI's SRA: Leinonen et al. 2011) or EMBL's ENA (Amid et al. 2020 y los datos enviados a la plataforma de biodiversidad deben incluir el biosample ID obtenido del archivo (ver mapeo de datos en [data-mapping]). Hacer uso de estos identificadores reducirá las posibilidades de duplicación y asegurará que los datos de secuencia sean fácilmente alcanzables en caso de que surjan oportunidades de reanálisis a medida que mejoren las bilbiotecas de referencia y las herramientas bioinformáticas. El producto final principal de estos pipelines es típicamente un archivo de recuentos de OTUs o ASVs individuales en cada muestra junto con la taxonomía asignada a estos. Lo anterior se genera en formato tabular o en formato BIOM (McDonald et.al 2012). Las secuencias OTU o ASV también se proporcionan a menudo en formato FASTA ((Pearson & Lipman 1988).

1.6. Taxonomía de secuencias

La anotación taxonómica de secuencias es un paso crítico en el procesamiento de los conjuntos de datos de biodiversidad molecular, ya que los nombres científicos son clave para acceder y comunicar información sobre los organismos observados. La exactitud y precisión de tal anotación de secuencia dependerá de la disponibilidad de bases de datos de referencia fiables y bibliotecas en todas las ramas del árbol de la vida, que a su vez requerirá esfuerzos conjuntos de taxónomos y ecólogos moleculares. Las bases de datos de secuencia pública siempre deben ser utilizadas a sabiendas del hecho de que sufren de diverdad deficiencias relacionadas, por ejemplo, con la confiabilidad taxonómica y la falta de vocabularios de metadatos estandarizados (Hofstetter et al. 2019; Durkin et al. 2020).

Las especies, tal como las describen los taxónomos, son fundamentales para la biología y, por lo tanto, los intentos de caracterizar la biodiversidad, pueden utilizar los productos finales de la investigación taxonómica. Sin embargo, a diferencia de los datos de la secuencia de ADN, las salidas taxonómicas no siempre son fácilmente susceptibles a la interpretación algorítmica directa o computacional: la taxonomía clásica es un proceso impulsado por el hombre que inclye pasos manuales de delimitación de taxones, descripción y nombramiento, culminando en una publicación formal de acuerdo con los Códigos internacionales de Nomenclatura. Como se discutió en capítulos anteriores, los estudios basados en secuencias de ADN son muy eficaces para detectar especies difíciles de observar y, a menudo, identificarán la presencia de organismos que actualmente están fuera del conocimiento taxonómico tradicional de Linneo. Si bien estas directrices no abordan la

publicación de listas de especies alternativas derivadas de datos de secuencia, la desconexión entre la taxonomía tradicional y los esfuerzos del ADN ambiental es indeseable. Por ello ofrecemos las siguientes recomendaciones a los lectores de esta guía.

Dado que la taxonomía es fundamental para el descubrimiento de datos sobre biodiversidad, es altamente recomendable que con cualquier esfuerzo de secuenciación de ADN ambiental se busque incluir experiencia taxonómica relevante en su estudio. De manera similar, será beneficioso si los estudios de secuenciación de ADN ambiental puedan asignar una parte de su presupuesto a la generación y publicación de secuencias de referencia a partir de especímenes tipo no secuenciados previamente u otro material de referencia importante del herbario, museo o colección biológica local. Los taxónomos también pueden contribuir a este objetivo incluyendo siempre secuencias de ADN relevantes con cada descripción de una nueva especie (Miralles et al. 2020) y centrándose en las muchas entidades biológicas novedosas desveladas por los esfuerzos del ADN ambiental (por ejemplo, Tedersoo et al. 2017).

La mayoría de las plataformas de datos sobre biodiversidad actuales están organizadas en torno a listas de nombres e índices taxonómicos tradicionales. Dado que las ocurrencias derivadas de secuencias de ADN se están convirtiendo rápidamente en una fuente importante de datos sobre biodiversidad, y como la taxonomía y nomenclatura oficiales para dichos datos van retrasadas, se recomienda que los proveedores de datos y las plataformas continúen explorando e incluyendo representaciones más flexibles de la taxonomía en sus árboles taxonómicos. Estas nuevas representaciones incluyen bases de datos de referencia molecular (por ejemplo, GTDB, BOLD, UNITE) que reconocen los datos de secuencia como material de referencia para organismos no clasificados previamente. Además, sugerimos que otras bases de datos moleculares de uso común (por ejemplo, PR2, RDP, SILVA) desarrollen identificadores estables para taxones y pongan a disposición secuencias de referencia para esos taxones, para permitir su uso como referencias taxonómicas.

A diferencia de la taxonomía clásica, que es un proceso altamente manual, el agrupamiento de secuencias de ADN en conceptos taxonómicos se basa en el análisis algorítmico de similitud y otras señales (como filogenia y probabilidad), así como en cierta edición humana. Las OTUs resultantes varian en estabilidad, presencia de secuencias de referencia, material físico, alineaciones y valores de corte, así como en identificadores de OTU como DOIs (Nilsson et al. 2019). Aún mas importante, estas OTUs varían en escala, desde bibliotecas específicas de estudios o proyectos locales hasta bases de datos globales que permiten una comparación más amplia entre estudios. A diferencia de la centralización y codificación de los taxones Linneanos que se describen formalmente en publicaciones de investigación, las OTUs se distribuyen en múltiples bibliotecas digitales de referencia en constante evolución, que difieren en enfoque taxonómico, genes de códigos de barras y otros factores. Al asociar secuencias estándar con especímenes de referencia identificados, BOLD y UNITE están estableciendo una capa de mapeo esencial para vincular ASVs y OTUs con la taxonomía linneana. La taxonomía principal de GBIF incluye identificadores para las Hipótesis de Especies UNITE (SHs) así como los Números de Índice de Código de Barras (BINs), lo que permite indexar datos de ocurrencia de especies anotados taxonómicamente a nivel de OTU, principalmente para hongos y animales (https://www.gbif.org/news/2LrgV5t3ZuGeU2WIymSEuk/adding-sequence-based-identifiers-to-backbone-taxonomy-reveals-dark-taxa-fungi [Secretariado de GBIF 2018^], https://data-blog.gbif.org/post/gbif-backbone-taxonomy [Grosjean 2019^]).

Los algoritmos para la anotación taxonómica del ADN ambiental típicamente asignarán cada secuencia única al grupo taxonómico más cercano en un conjunto de referencia, basado en algunos criterios de parentesco y confianza. Para grupos de organismos poco conocidos, como los procariotas, insectos y hongos, la anotación puede ser un nombre provisional no Linneano (basado en clústeres) para un taxón (es decir, el ID/número del SH o BIN relevante), y este taxón puede representar una especie o incluso una unidad taxonómica por encima del nivel de especie. Ninguna base de datos de referencia contiene todas las especies de un grupo dado debido a las muchas especies desconocidas, no identificadas y no descritas en la tierra. La ignorancia frecuente de este

hecho ha sido la fuente de numerosas identificaciones taxonómicas erróneas durante los últimos 30 años.

Durante la importación en la plataforma de biodiversidad (por ejemplo, GBIF u OBIS), la resolución taxonómica para estas ocurrencias basadas en ADN puede reducirse aún más, ya que los nombres/IDs obtenidos al comparar con la base de datos de referencia (por ejemplo, UNITE, BOLD) pueden no estar incluidos totalmente en el índice taxonómico de esa plataforma en el momento de la publicación. Sin embargo, la inclusión de la secuencia subyacente de OTU o ASV para cada registro permitirá a los futuros usuarios identificar potencialmente la secuencia a un nivel de granularidad mayor, especialmente a medida que las bibliotecas de referencia mejoren con el tiempo. Por lo tanto, también recomendamos publicar todas las secuencias en un estudio, incluso aquellas que actualmente están completamente sin clasificar, ya que es posible que se puedan identificar con bases de datos de referencia mejoradas. En los casos en los que la secuencia subyacente no pueda incluirse como parte de los datos enviados, abogamos por la deposición de un nombre (científico o de marcador de posición) del taxón (por ejemplo, el BOLD BIN o UNITE SH) junto con la suma de verificación MD5 de la secuencia como un ID de taxón único (ver [data-mapping]). La suma de verificación MD5 es un algoritmo de hash unidireccional comúnmente utilizado para verificar la integridad de archivos. En este caso proporciona una representación única y repetible de la secuencia original que, sin embargo, no permitiría recuperar la secuencia en sí. Esto puede ser necesario en casos en los que exista sensibilidad en torno al acceso. La suma de verificación MD5 permite consultas eficientes para determinar si se ha recuperado la misma secuencia exacta en otros esfuerzos de eDNA, pero no reemplaza por completo la secuencia, ya que los MD5 no habilitan análisis adicionales. Dos secuencias que difieren en incluso una sola base obtendrán dos sumas de verificación MD5 completamente diferentes, de modo que las búsquedas de similitud de secuencias al estilo BLAST no funcionarán.

1.7. Salidas

El propósito de exponer datos derivados del ADN a través de plataformas de biodiversidad es permitir la reutilización de estos datos en combinación con otros tipos de datos sobre biodiversidad. Es muy importante tener en cuenta esta reutilización al preparar sus datos para la publicación. Idealmente, los metadatos y los datos deben contar una historia completa de tal manera que los nuevos usuarios no informados puedan utilizar esta evidencia sin ninguna consulta o correspondencia adicional. Las plataformas de datos sobre biodiversidad proporcionan funcionalidades de búsqueda, filtrado, navegación, visualizaciones, acceso a datos y citación de datos. Para los datos de metabarcoding, alentamos a los usuarios a configurar filtros para la abundancia mínima absoluta y relativa de lecturas para hacer un filtrado adecuado de los datos. Las ocurrencias individuales o cualquier ocurrencia con un recuento absoluto de lecturas por debajo de algún valor seleccionado pueden ser filtradas estableciendo una abundancia mínima de lecturas por OTU o ASV (usando el campo organismQuantity). Las ocurrencias con una abundancia relativa de lecturas por debajo de un umbral seleccionado pueden ser filtradas estableciendo un valor mínimo de cantidad relativa del organismo, que se calcula a partir de las lecturas detectadas (organismQuantity) y el total de lecturas en la muestra correspondiente (sampleSizeValue) ([mapping-metabarcoding-edna-and-barcoding-data]). Los usuarios a menudo pueden elegir formatos de salida de datos (por ejemplo, DwC-A, CSV) y luego procesar, limpiar y transformar los datos en la forma y formato necesarios para los análisis.

En GBIF.org o a través de la API de GBIF, los usuarios registrados pueden buscar, filtrar y descargar datos de biodiversidad en los siguientes tres formatos:

- **Simple**: Un formato sencillo delimitado por tabulaciones que incluye solo la versión interpretada por GBIF de los datos, como resultado del proceso de indexación. Este formato es adecuado para pruebas rápidas y para importar directamente a hojas de cálculo.
- **Archivo Darwin Core**: Formato más completo que incluye tanto los datos interpretados como la

versión original literal proporcionada por el publicador (antes de la indexación e interpretación por parte de GBIF). Debido a que incluye todos los metadatos y las indicaciones de problemas, este formato aporta una vista más detallada del conjunto de datos descargado.

- **Lista de especies**: Un formato de tabla sencillo que incluye solo una lista interpretada de nombres únicos de especies de un conjunto de datos o de un resultado de una consulta.

Independientemente del formato seleccionado, cada descarga de usuario de GBIF recibe un enlace reutilizable a la consulta y una cita de datos que incluye un DOI. Este sistema de citación basado en DOI proporciona los medios para reconocer y acreditar el uso de conjuntos de datos y sus creadores, mejorando tanto la credibilidad como la transparencia de los hallazgos basados en los datos. Es esencial seguir las recomendaciones de citación de datos y utilizar DOIs, ya que una buena cultura de citación de datos no solo es la norma académica, sino también un mecanismo poderoso para acreditar y, por lo tanto, incentivar a los publicadores de datos.

2. Empaquetado y mapeo de datos

Este capítulo se enfoca en detalles prácticos necesarios para convertir los datos exportados en un conjunto de datos que pueda ser indexado en una plataforma de datos sobre biodiversidad. [categorization-of-your-data] le ayudará a entender cuál es el esquema de mapeo óptimo para sus datos. [data-mapping] aporta una descripción de estos mapeos en detalle.

Esta guía combina los estándares para la publicación de datos generales sobre biodiversidad con datos genéticos de biodiversidad derivados de ADN (Figura 5). En esta “sección práctica”, se proporcionan recomendaciones para el mapeo de diferentes tipos de datos derivados de ADN.

Las rutas de empaquetado y publicación de datos varían de una plataforma a otra y se describen en la documentación general. Uno de los métodos ampliamente utilizados para empaquetar archivos de datos es DwC-A, donde las tablas de datos están organizadas en un esquema estelar, con registros (filas) en archivos de extensión periférica que apuntan a un solo registro en el archivo central o "core" (Figura 5). Los diferentes tipos de archivos principales (por ejemplo, registros y eventos de muestreo) corresponden a diferentes clases de conjuntos de datos. Aunque los conjuntos de datos derivados de ADN a menudo se basan en la naturaleza del evento, p.ej. cientos o incluso miles de registros de secuencias cuantificadas pueden derivarse de un solo evento de muestreo y por lo tanto, compartir la mayoría de los atributos del metadato, la recomendación actual es publicar datos como "core" de registros (Categoría I o II) con la extensión de datos derivados de ADN. Este enfoque compensa las limitaciones del esquema estelar del DwC, que no permite que un dato a nivel de registro que se encuentra en archivos de extensión (como secuencias de código de barras procesadas) apunte a registros en un archivo de "core" de eventos. Sin embargo, recomendamos incluir un eventID para cada registro, para indicar la asociación entre los registros derivados del mismo evento de muestreo.

[dwca structure.es] | *img/print/dwca-structure.es.png*

Figura 5. Ampliación de DwC-A / IPT de la figura 3 del capítulo 1.2. La elección de la entidad central o "core" es principalmente una cuestión de adecuar los datos al mecanismo de importación de los mismos (ingestión) de las plataformas de datos sobre biodiversidad. La mayoría de los datos podrían ser formulados como "core" de Registros, Evento o Taxón, pero, dado que solo el "core" puede tener extensiones, esto afectará la elección. Por ejemplo, no es posible incluir la secuencias de ADN como extensión de un registro si los datos se empaquetan utilizando el "core" de Evento.

2.1. Categorización de los datos

Para el propósito de esta guía, categorizamos los datos en cinco categorías, enlazados por un campo ID clave (*eventID*), equivalente a los estándares para datos generales de biodiversidad, y se incluyen campos relevantes para los datos derivados de ADN (see [data-mapping]). Estas cinco categorías

buscan reflejar los enfoques moleculares más comúnes utilizados para la caracterización de biodiversidad y son I) Registros biológicos derivados de ADN, II) Registros biológicos enriquecidos, III) Detección de especies objetivo (qPCR/ddPCR), IV) Referencia de nombres científicos y V) Conjuntos de solo metadatos. Examine el árbol de decisión y proceda a la sección correcta.

Tabla 1. Árbol de decisión para la categorización de datos derivados de ADN.

ⓘ *¿Sus datos son derivados de (meta)barcoding o basados en qPCR?*			
(Meta)barcoding ↓ ⓘ *¿Los datos consisten en materia genético digitalizado, o secuencias, asociadas con un momento y lugar?*			qPCR ↓ **Category III** Detección de especies objetivo
Sí ↓ ⓘ *¿El material genético es la **únicA** evidencia del organismo dado o comunidad?*		No ↓ ⓘ *¿El conjunto de datos es una lista de nombres derivados de ADN?*	
Sí ↓ **Category I** Registros biológicos derivados de ADN	No ↓ **Category II** Registros biológicos enriquecidos	Sí ↓ **Category IV** Referencias de nombre	No ↓ **Category V** Conjunto de solo metadatos

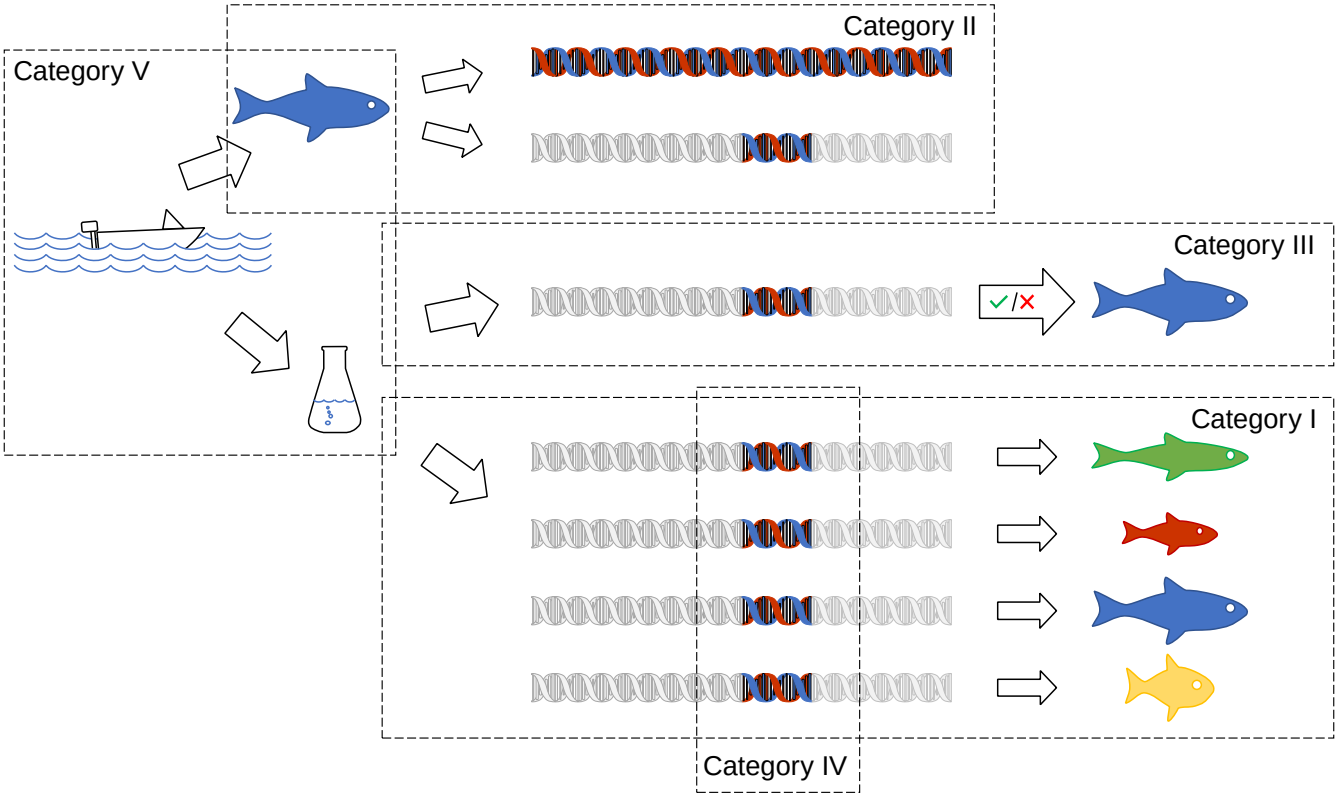

Figura 6. Representación visual de las categorías I-V.

2.1.1. Categoría I: Registros biológicos derivados de ADN

Esta categoría hace referencia a datos en los que una secuencia de ADN o la detección a través de

PCR son la única evidencia de la presencia de un organismo o comunidad específica. En otras palabras, los datos no pueden ser rastreados hasta un espécimen observable. Este es el caso de muchos estudios de metagenómica, metabarcoding y eADN.

Ejemplos de conjuntos de datos de registros biológicos derivados de ADN

- Holman L E, Bohmann K (2025). Eukaryotic metabarcoding (18S rRNA V9 region) of environmental DNA from an archived marine sediment record, Skagerrak, North Sea, spanning 8000 years. Globe Institute. Occurrence dataset https://doi.org/10.15468/2ve69k accessed via GBIF.org on 2025-02-27.
- Flanders Research Institute for Agriculture, Fisheries and Food (ILVO): BulkDNA macrobenthos from sandextraction sites in the Belgian part of the North Sea https://doi.org/10.15468/djwzhu accessed via GBIF.org on 2025-02-27.
- Okrasińska A, Pawłowska J (2025). Metabarcoding of fungi in post-industrial soils. Institute of Evolutionary Biology, University of Warsaw. Occurrence dataset https://doi.org/10.15468/e4maz4 accessed via GBIF.org on 2025-02-28.

To specifically format and share metabarcoding datasets, we suggest the user friendly Metabarcoding Data Toolkit (MDT) – and consult the dedicated guide material: Metabarcoding Data Toolkit – user guide. For more general guidance on how to format and share these dna-derived data, see [mapping-metabarcoding-edna-and-barcoding-data]. General guidelines for Darwin Core occurrence datasets are also available through the DwC-A template for occurrence datasets and Data quality requirements for occurrences.

2.1.2. Categoría II: Registros biológicos enriquecidos

Si algún material genético está, o puede estar, asociado con una observación o un espécimen, clasificaremos este tipo de datos como “registros biológicos enriquecidos”. En este contexto, las secuencias no son la única evidencia para un registro biológico. Siempre se puede rastrear la información a un espécimen catalogado o a un organismo observado. Esta categoría incluye, por ejemplo, conjuntos de datos de barcoding (códigos de barras) y algunos conjuntos de datos de metabarcoding de ADN con material de referencia. Para obtener más orientación sobre barcoding, remitase al siguiente enlace Centro de Genómica de la Biodiversidad, Universidad de Guelph (2021).

Ejemplos de conjuntos de datos de registros biológicos enriquecidos

- The International Barcode of Life Consortium (2016) International Barcode of Life project (iBOL). Occurrence dataset https://doi.org/10.15468/inygc6 accessed via GBIF.org on 2020-04-16.
- Takamura K (2019) Chironomid Specimen records in the Chironomid DNA Barcode Database. Version 1.9. National Institute of Genetics, ROIS. Occurrence dataset https://doi.org/10.15468/hxhow5 accessed via GBIF.org on 2020-04-16.
- Bessey C, Jarman SN, Stat M, Rohner CA, Bunce M, Koziol A, Power M, Rambahiniarison JM, Ponzo A, Richardson AJ & Berry O (2019) DNA metabarcoding assays reveal a diverse prey assemblage for Mobula rays in the Bohol Sea, Philippines. Ecology and Evolution 9 (5) 2459-2474. https://doi.org/10.1002/ece3.4858, (Atlas of Living Australia website at https://collections.ala.org.au/public/show/dr11663. Accessed 24 June 2020)

Para obtener orientación sobre cómo dar formato y compartir estos conjuntos de datos, consulta [mapping-metabarcoding-edna-and-barcoding-data]. También están disponibles las directrices generales para los conjuntos de datos de ocurrencias de Darwin Core a través de plantilla DwC-A para conjuntos de datos de ocurrencias y recomendaciones de calidad de datos para ocurrencias.

2.1.3. Categoría III: Detección de especies objetivo (qPCR/ddPCR)

Esta categoría hace referencia a los datos en los que se utiliza un ensayo específico (qPCR/ddPCR) para detectar la presencia (o ausencia) de una secuencia de ADN específica para el organismo objetivo en una muestra ambiental. En este caso, el registro biológico puede que ni siquiera contenga datos de la secuencia, ya que es el proceso como tal el que determina el registro biológico. Con los análisis qPCR/ddPCR para la detección de especies específicas, muchos estudios también reportan la ausencia de esa especie en particular para una muestra determinada. La ausencia de datos depende en gran medida del límite de detección del ensayo específico, así como de los protocolos de campo y laboratorio. En cuanto a los datos de ADN-metabarcoding hay un problema de falsos negativos y falsos positivos, y es importante que se reporte suficiente información para evaluar los registros.

Ejemplos de conjuntos de datos de especies objetivo

- Strzelecki, Joanna; Feng, Ming; Berry, Olly; Zhong, Liejun; Keesing, John; Fairclough, David; Pearce, Alan; Slawinski, Dirk; Mortimer, Nick. Location and transport of early life stages of Western Australian Dhufish Glaucosoma hebraicum. Floreat, WA: Fisheries Research and Development Corporation; 2013. http://hdl.handle.net/102.100.100/97533 (Atlas of Living Australia website at https://collections.ala.org.au/public/show/dr8131. Accessed 22 July 2020)

Para obtener orientación sobre cómo dar formato y compartir estos conjuntos de datos, consulta [mapping-ddpcr-qpcr-data]. También están disponibles las directrices generales para los conjuntos de datos de ocurrencias de Darwin Core a través de plantilla DwC-A para conjuntos de datos de ocurrencias y recomendaciones de calidad de datos para ocurrencias.

2.1.4. Categoría IV: Referencias de nombres

Esta categoría corresponde a nombres derivados del ADN, obtenidos mediante agrupación o denoising (modelos basados en corrección de errores), como las Unidades Taxonómicas Operativas (OTU) no Linneanas estables, las Variantes de Secuencia de Amplicón (ASV) y los Números de Índice de Código de Barras (BIN). En otras palabras, se refiere a cualquier mención de taxones o nombres provisionales definidos fuera de la taxonomía Linneana. Numerosos proyectos generan bibliotecas locales de OTUs específicas para proyectos o estudios, y aunque es técnicamente posible publicarlas como listas de verificación, tienen un valor limitado o nulo para la vinculación o interpretación de datos. Por lo tanto, no alentamos su publicación a través de plataformas de datos de biodiversidad. Sin embargo, la inclusión de las OTUs ampliamente adoptadas, estables, globales y digitalmente referenciables en las estructuras taxonómicas Linneanas es de vital importancia para indexar la biodiversidad “oscura” sin nombre. GBIF ha acumulado experiencia en la integración de estas grandes bibliotecas de referencia globales de OTUs en la estructura taxonómica principal de GBIF, lo que permite mostrar las OTUs bajo el taxón padre más cercano que tiene un nombre científico (Figura 7).

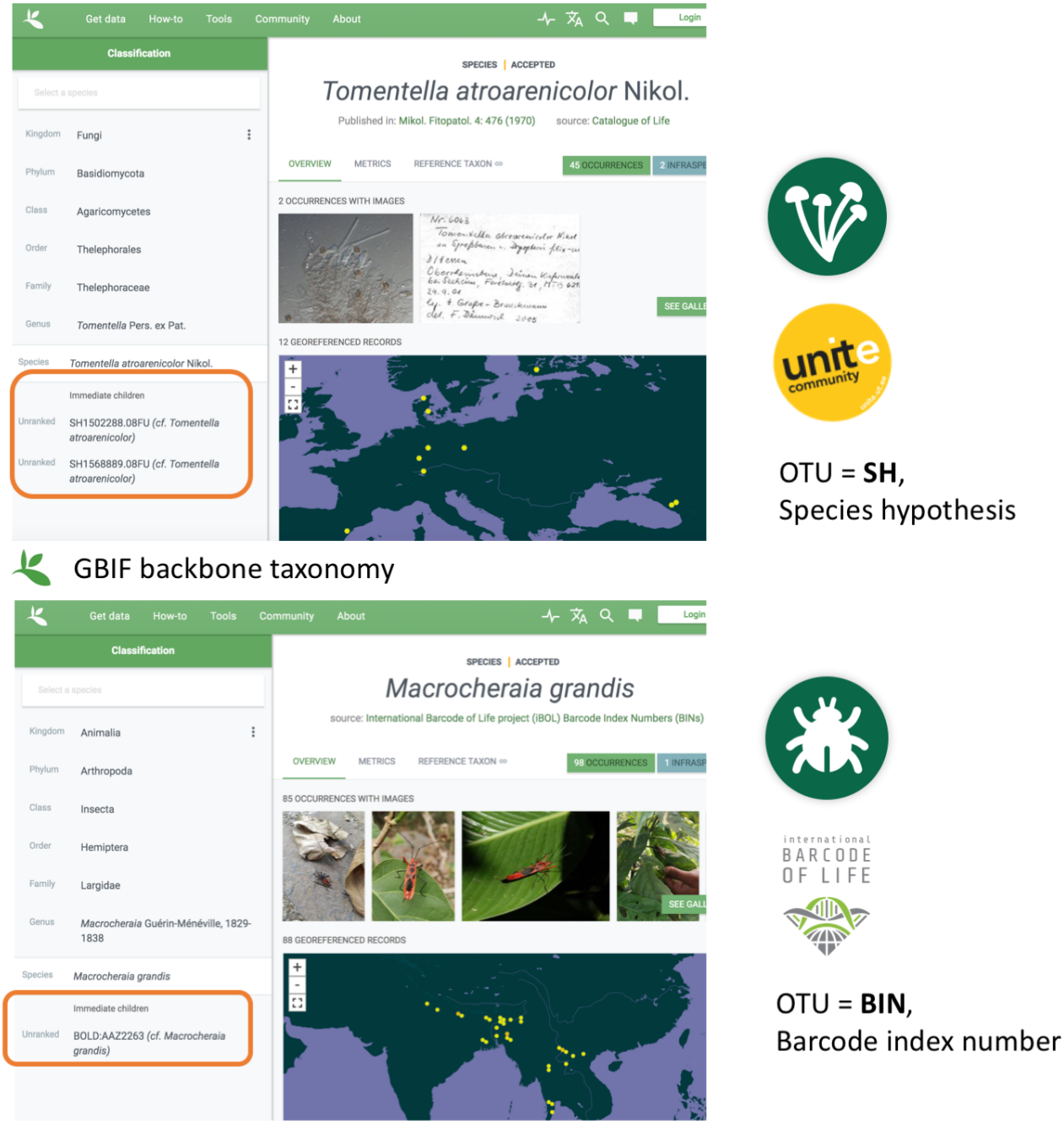

Figura 7. Los OTUs (SHs) de UNITE (principalmente hongos, arriba) y de BOLD (BINs) (principalmente artrópodos, abajo) se muestran en la taxonomía principal de GBIF bajo sus taxones parentales correspondientes que tienen nombres científicos. Múltiples ocurrencias observadas individualmente de biodiversidad críptica se vuelven descubribles junto con evidencia no genética a través de un único punto de acceso.

Ejemplos de listas de verificación de referencias de nombres

- The International Barcode of Life Consortium (2016). International Barcode of Life project (iBOL) Barcode Index Numbers (BINs). Checklist dataset https://doi.org/10.15468/wvfqoi accessed via GBIF.org on 2020-04-16.
- PlutoF (2019). UNITE - Unified system for the DNA based fungal species linked to the classification. Version 1.2. Checklist dataset https://doi.org/10.15468/mkpcy3 accessed via GBIF.org on 2020-04-16.

Este guía no proporciona recomendaciones de mapeo para listas de verificación globales de OTUs / bibliotecas de referencia (Categoría IV), y se desaconseja publicar bibliotecas de OTUs referenciables

(proyecto o estudio específico) como listas de verificación. Para obtener orientación sobre cómo dar formato y compartir listas de verificación de OTUs, consulte las siguientes directrices generales de Darwin Core en plantilla DwC-A para listas de verificación, recomendaciones de calidad de datos para listas de verificación y directrices generales para listas de verificación MIxS. Si necesita asesoramiento sobre cómo mapear bibliotecas de referencia globales de OTUs para su inclusión en la estructura taxonómica principal de GBIF, comunicarse con centro de ayuda de GBIF.

2.1.5. Categoría V: Conjuntos de solo metadatos

Los metadatos son datos acerca de los datos y son una descripción del conjunto de datos en términos generales, con información como los autores, la afiliación de los autores, el objetivo original de investigación del conjunto de datos, los DOIs asociados, la cobertura taxonómica, la cobertura temporal y la cobertura geográfica. La información asociada a los métodos de laboratorio y métodos generales de secuenciación están incluídos en esta categoría. Esta categoría incluye conjuntos de datos o colecciones que no pueden estar disponibles en línea en este momento, por ejemplo trabajos sin digitalizar.

Ejemplos de conjuntos de datos de solo metadatos

- Collins E, Sweetlove M (2019). Arctic Ocean microbial metagenomes sampled aboard CGC Healy during the 2015 GEOTRACES Arctic research cruise. SCAR - Microbial Antarctic Resource System. Metadata dataset https://doi.org/10.15468/iljmun accessed via GBIF.org on 2020-04-16.
- Cary S C (2015). New Zealand Terrestrial Biocomplexity Survey. SCAR - Microbial Antarctic Resource System. Metadata dataset https://doi.org/10.15468/xnzrhq accessed via GBIF.org on 2020-04-16.

Las recomendaciones de mapeo para conjuntos de datos derivados del ADN que contienen solo metadatos (Categoría V) son las mismas que para cualquier otro conjunto de datos que contenga solo metadatos, y esta guía no proporciona recomendaciones específicas de mapeo para los metadatos. Por favor, siga las recomendaciones generales de los portales de datos de biodiversidad, prestando atención a metadatos requeridos y recomendados. Las descripciones de los campos, laboratorio y pasos de bioinformática deben ser lo más detalladas posible. Describir los métodos como pasos de método en los metadatos EML permite que se muestren en la página de inicio del conjunto de datos en GBIF (https://www.gbif.org/es/dataset/3b8c5ed8-b6c2-4264-ac52-a9d772d69e9f#methodology Frøslev T, Ejrnæs R (2018). Conjunto de datos de hongos eDNA BIOWIDE. Servicio Danés de Información sobre Biodiversidad. Conjunto de datos de ocurrencias https://doi.org/10.15468/nesbvx accedido a través de GBIF.org el 2021-07-06). Sin embargo, si ya existe una descripción de método estructurada y posiblemente más detallada publicada en algún lugar (por ejemplo, en protocols.io o en colección de protocolos NEON), es sencillo proporcionar un enlace a través del campo MIxS SOP (ver [mapping-metabarcoding-edna-and-barcoding-data]).

2.2. Mapeo de datos

Mientras que los archivos core guardan información general de un registro referente al "Qué, donde y cuándo", los archivos de extensión son utilizados para describir las especificaciones de un cierto tipo de observación. Proponemos utilizar la extensión datos derivados de ADN para complementar los registros biológicos derivados tanto de barcoding, metabarcoding (eDNA) o qPCR/ddPCR. La extensión datos derivados de ADN se basa en el estándar mínimo de información desarrollado por el Consorcio de Estándares Genómicos (GSC por sus siglas en inglés) y aplicado por ENA para el envío de metadatos de muestra de eDNA, por ejemplo. Nosotros seguimos y hemos contribuido a los lineamientos propuestos por el grupo de trabajo de Interoperatibilidad Sostenible DwC-MIxS alojado por TDWG. Para mejorar la indexación y búsqueda, hemos optado por separar algunos elementos de MIxS, por ejemplo separando los nombres y secuencias del primer forward y el primer reverse.

Además, algunos elemenos del estándar GGBN y elementos de la guía para datos de qPCR y ddPCR MIQE (información mínima para la publicación de PCR cuantitativa en tiempo real), han sido incluídos para hacerlo aplicable a un amplio rando de datos derivados de ADN.

Como primer paso en la preparación de sus datos para su publicación, debe asegurarse de que los nombres de sus campos / encabezados de sus columnas siguen el https://dwc. dwg.org/terms/[Estándar Darwin Core]. En muchos casos esto es sencillo, como renombrar su campo `lat` o `latitude` a `decimalLatitude`. Sin embargo, el estándar Darwin Core es bastante flexible y algunos términos se utilizan de diferentes maneras, dependiendo del tipo de datos. Un ejemplo de esto son los elementos `organismQuantity` y `organismQuantityType`, que podrían utilizarse para describir el número de individuos, porcentaje de biomasa o una valor en la escala Braun-Blanquet, así como el número de lecturas de una ASV dentro de una muestra. Por lo tanto, aquí proporcionamos tablas de elementos obligatorios y recomendados con descripciones y ejemplos (Table 1, Table 2, Table 3 and Table 4). La recomendación de utilizar el core de Registros biológicos para los datos derivados de ADN, surge del fuerte deseo de compartir la secuencia para ayudar a calificar la determinación. Elementos y extensiones adicionales (como la extensión http://rs.gbif. rg/extension/obis/extended_measurement_or_fact.xml[Medidas o Hechos Extendida (eMoF)]) son aplicables - tanto al core de Registro biológicos como al core de eventos. Cuando una secuencia es derivada de un organismo (por ejemplo, parásitos, contenidos estomacales, epibiontes, etc.). la observación puede estar vinculada a la observación del organismo anfitrión. Esto se puede lograr utilizando la extensión (https://dwc. dwg.org/terms/#resourcerelationship[Relación del Recurso]) de Darwin Core (por ejemplo, https://www.gbif. rg/species/143610775/verbatim). Tal vez la recomendación más importante sea utilizar identificadores únicos globales (cuando estén disponibles) y otros identificadores permanentes para tantos elementos y parámetros como sea posible (en todos los elementos ID de las siguientes tablas).

2.2.1. Mapeando datos de metabarcoding (eDNA) y barcoding

NB: To format and share metabarcoding datasets, we suggest the user friendly Metabarcoding Data Toolkit (MDT) which has a specific guide (Metabarcoding Data Toolkit – user guide) that includes dedicated versions of the tables in this section.

Esta sección proporciona recomendaciones de mapeo para las Categorías I y II.

Tabla 2. Elementos recomendados del core de Registros biológicos para datos de Metabarcoding

Elemento	Ejemplos	Descripción	Obligatoriedad
basisOfRecord	MaterialSample	Denota el origen o evidencia específica de la que se deriva el organismo - un subtipo de dcterms:type. Para los registros biológicos derivados de ADN, (ver la Category I y la Category III) utilice MaterialSample. Para registros biológicos enriquecidos, utilice PreservedSpecimen o LivingSpecimen según sea apropiado.	Obligatorio
occurrenceID	urn:catalog:UWBM:Bird:89776	Un identificador único del registro biológico, permitiendo que el mismo registro sea reconocido a través de diferentes versiones de un conjunto de datos, así como a través de las descargas y usos del mismo. Puede ser un identificador único global o un identificador específico para el conjunto de datos.	Obligatorio
eventID	urn:uuid:a964765b-22c4-439a-jkgt-2	Un identificador único para el Evento de muestreo, que ocurre en un lugar y tiempo determinado. Puede ser un identificador único global o un identificador específico para el conjunto de datos.	Altamente recomendado
eventDate	2020-01-05	La fecha durante la cual se produjo el evento de observación. La práctica recomendada es utilizar una fecha documentada en el esquema de codificación ISO 8601-1:2019. Para más información, revisar https://dwc.tdwg.org/terms/#dwc:eventDate	Obligatorio
recordedBy	"Oliver P. Pearson | Anita K. Pearson"	Una lista (concatenada y separada) de los nombres de las personas, grupos u organizaciones responsables de la colecta u observación del espécimen. La práctica recomendada es sepearar la valores con una barra vertical (' | '). Incluir información sobre el observador mejora la reproducibilidad científica (Groom et al. 2020).	Altamente recomendado
organismQuantity	33	Número de lecturas de este OTU o ASV en la muestra.	Altamente recomendado
organismQuantityType	DNA sequence reads	Siempre debe ser “DNA sequence reads”	Altamente recomendado

Elemento	Ejemplos	Descripción	Obligatoriedad
sampleSizeValue	1233890	Número total de lecturas en la muestra. Esto es importante dado que permite calcular la abundancia relativa de cada OTU o ASV dentro de la muestra. Este número preferiblemente debe ser calculado despues de un procesamiento universal (control de calidad, ASV denoising, remoción de quimeras, etc.), pero antes de la remoción manual/selectiva de OTUs o ASV no objetivos del conjunto de datos. La rarefacción (remuestreo parar igualar la profundidad de la secuenciación a través de muestras) no es necesaria o aconsejada.	Altamente recomendado
sampleSizeUnit	DNA sequence reads	Siempre debe ser “DNA sequence reads”	Altamente recomendado
materialSampleID	https://www.ncbi.nlm.nih.gov/biosample/15224856 https://www.ebi.ac.uk/ena/browser/view/SAMEA3724543 urn:uuid:a964805b-33c2-439a-beaa-6379ebbfcd03	Un identificador para muestras de material (no hace referencia a muestras digitales sino físicas, como exicados o tejidos). Use el ID BioSample si uno fue obtenido de un archivo de nucleótidos. En ausencia de un identificador único global persistente, puede construir uno usando una combinación de identificadores en el registro, de tal forma que el materialSampleID sea globalmente único.	Altamente recomendado
samplingProtocol	Trampa de luz UV	El nombre, la descripción o la referencia del método o protocolo de muestreo usado para realizar el muestreo. https://dwc.tdwg.org/terms/#dwc:samplingProtocol	Recomendado
associatedSequences	https://www.ncbi.nlm.nih.gov/nuccore/MK405371	Una lista (en una fila continua y separada por una barra vertical “|”) de los identificadores (publicación, identificador único global, URI) de la información de la secuencia genética asociada al registro biológico. Puede ser utilizada para relacionar las lecturas de códigos de barras sin procesar o secuencias de genomas asociados, disponibles en un repositorio público.	Recomendado

Elemento	Ejemplos	Descripción	Obligatoriedad
identificationRemarks	Confianza en la anotación RDP (para el taxon especificado más bajo): 0.96, para la base de datos de referencia: GTDB	Especificación del proceso de identificación taxonómico, idealmente incluyendo datos del algoritmo aplicado y la base de datos de referencia, así como del nivel de confianza en el resultado de la identificación. Specification of taxonomic identification process, ideally including data on applied algorithm and reference database, as well as on level of confidence in the resulting identification.	Recomendado
identificationReferences	https://www.ebi.ac.uk/metagenomics/pipelines/4.1 https://github.com/terrimporter/CO1Classifier	Una lista (en una fila continua y separada por una barra vertical "|") de las referencias (publicación, identificador único global, URI) usadas en la identificación.	Recomendado
decimalLatitude	60.545207	La latitud geográfica (en grados decimales, utilizando el sistema de referencia espacial provisto en geodeticDatum) del centro geográfico de una ubicación. Los valores positivos se encuentran al norte del ecuador, los valores negativos están al sur del mismo. Los valores admitidos se encuentran entre -90 y 90.	Altamente recomendado
decimalLongitude	24.174556	La longitud geográfica (en grados decimales, mediante el sistema de referencia espacial provisto en geodeticDatum) del centro geográfico de una ubicación. Los valores positivos se encuentran al este del meridiano de Greenwich, los valores negativos se encuentran al oeste de la misma. Los valores admitidos se encuentran entre -180 y 180.	Altamente recomendado
taxonID	ASV:7bdb57487bee022ba30c03c3e7ca50e1	Para datos de eDNA, es recomendado utilizar un hash MD5 de la secuencia precedida por "ASV:". Más información [taxonomy-of-sequences].	Altamente recomendado, si el elemento DNA_sequence no está presente
scientificName	*Gadus morhua* L. 1758, BOLD:ACF1143	Nombre científico del taxón conocido más cercano (especie o superior) o un identificador para un OTU de BOLD (BIN) o UNITE (SH)	Obligatorio
kingdom	Animalia	Clasificación superior	Altamente recomendado

Elemento	Ejemplos	Descripción	Obligatoriedad
`phylum`	Chordata	Clasificación superior	Recomendado
`class`	Actinopterygii	Clasificación superior	Recomendado
`order`	Gadiformes	Clasificación superior	Recomendado
`family`	Gadidae	Clasificación superior	Recomendado
`genus`	*Gadus*	Clasificación superior	Recomendado

Tabla 3. Elementos recomendados de la extensión datos derivados de ADN (una selección) para datos de metabarcoding

Elemento	Ejemplos	Descripción	Obligatoriedad
DNA_sequence	TCTATCCTCAATTATAGGTCATAATTCACCATCAGTAGATTTAGGAATTTTCTCTATTCATATTGCAGGTGTATCATCAATTATAGGATCAATTAATTTTATTGTAACAATTTTAAATATACATACAAAAACTCATTCATTAAACTTTTTACCATTATTTTCATGATCAGTTCTAGTTACAGCAATTCTCCTTTTATTATCATTA	La secuencia de ADN (ASV). La interpretación taxonómica de la secuencia depende de la tecnología y la librería de referencia disponible en el momento de la publicación. Por eso, el manejo taxonómico más objetiva es la secuencia que puede ser reinterpretada en el futuro.	Altamente recomendado
sop	https://www.protocols.io/view/emp-its-illumina-amplicon-protocol-pa7dihn	Los procedimientos operativos estándar utilizados en el montaje y/o anotación de los genomas, metagenomas o secuencias ambientales. + Una referencia a un protocolo bien documentado, por ejemplo usar protocols.io	Recomendado
target_gene	16S rRNA, 18S rRNA, ITS	El gen objetivo o nombre del marcador para estudios basados en marcadores.	Altamente recomendado
target_subfragment	V6, V9, ITS2	Nombre del subfragmento de un gen o marcador importante, por ejemplo para identificar regiones especiales en genes marcadores como la región hipervariable V6 del gen rRNA 16S	Altamente recomendado
pcr_primer_forward	GGACTACHVGGGTWTCTAAT	La secuencia del primer directo utilizado para el proceso de amplificación del gen, locus o subfragmento	Altamente recomendado
pcr_primer_reverse	GGACTACHVGGGTWTCTAAT	La secuencia del primer inverso utilizado para el proceso de amplificación del gen, locus o subfragmento	Altamente recomendado
pcr_primer_name_forward	jgLCO1490	El nombre del primer directo utilizado	Altamente recomendado

Elemento	Ejemplos	Descripción	Obligatoriedad
pcr_primer_name_reverse	jgHCO2198	El nombre del primer inverso utilizado	Altamente recomendado
pcr_primer_reference	https://doi.org/10.1186/1742-9994-10-34	Referencia para los primers	Altamente recomendado
env_broad_scale	forest biome [ENVO:01000174]	**Equivalente al env_biome en MIxS v4** En este elemento, se reporta de cuál ecosistema superior provienen los especímenes o muestras. Los sistemas identificados deben tener una granularidad espacial gruesa, para proveer de información general del contexto medioambiental donde la muestra fue realizada (e.g. ¿estaba en el desierto o en el bosque lluvioso?). Se recomienda utilizar subclases de ENVO pertenecientes a la clase bioma: http://purl.obolibrary.org/obo/ENVO_00000428	Recomendado
env_local_scale	litter layer [ENVO:01000338]	**Equivalente al env_feature en MIxS v4** En este elemento, se reporta la entidad o entidades que están en la localidad cercana a los especímenes o muestras, y cree que tienen influencias causales significativas sobre la muestra o especímen. Por favor utilice términos que estén presentes en ENVO y tengan una granularidad espacial menor a la documentada en env_broad_scale	Recomendado
env_medium	soil[ENVO:00001998]	**Equivalente al env_material en MIxS v4** En este elemento, se reporta cuál material o materiales (separados por una barra vertical “	”) estaban en la vecindad inmediata de los especímenes o muestras antes del muestreo, utiliando una o más subclases de ENVO pertenecientes a la clase material medioambiental: http://purl.obolibrary.org/obo/ENVO_00010483

Elemento	Ejemplos	Descripción	Obligatoriedad
Recomendado	`lib_layout`	Pareada	**Equivalente a lib_const_meth en MIxS v4** Especifica si se espera una configuración de lecturas individual, pareada o de otro tipo
Recomendado	`seq_meth`	Illumina HiSeq 1500	El método o plataforma de secuenciación utilizada
Altamente recomendado	`otu_class_appr`	dada2; 1.14.0; ASV	El algoritmo y el nivel de agrupamiento (si es relevante) utilizado para la definición de OTUs o ASVs
Altamente recomendado	`otu_seq_comp_appr`	blastn;2.6.0+;e-value cutoff: 0.001	La herramienta y los umbrales utilizado para asignar nombres a los OTUs o ASVs al nivel de especie "species-level"

Elemento	Ejemplos	Descripción	Obligatoriedad
Altamente recomendado	`otu_db`	Genbank nr;221, UNITE;8.2	La base de datos de referencia (es decir, las secuencias que no fueron generadas como parte del estudio) utilizadas para asignar la taxonomía a los OTUs o ASVs

2.2.2. Mapeando datos ddPCR / qPCR

Esta sección proporciona recomendaciones de mapeo para https://academic.oup.com/view-large/199871507 [Categoría III].

Tabla 4. Campos recomendados para el core de registros biológicos para datos ddPCR/qPCR

Elemento	Ejemplos	Descripción	Obligatoriedad
basisOfRecord	MaterialSample	Denota el origen o evidencia específica de la que se deriva el registro- un subtipo de dcterms:type. Para registros derivados de ADN (ver Category I y Category III), use MaterialSample.	Obligatorio
occurrenceStatus	Present, Absent	Estado que da cuenta de la presencia o ausencia de un taxón en una ubicación.	Obligatorio
eventID	urn:uuid:a964765b-22c4-439a-jkgt-2	Un identificador único para la información asociada con el evento (algo que ocurre en un lugar y tiempo determinado). Puede ser un identificador único global o un identificador específico para el conjunto de datos.	Altamente recomendado
eventDate	2020-01-05	La fecha durante la cual se produjo el evento de observación. LaDate when the event was recorded. Debe estar documentada en el esquema de codificación ISO 8601-1:2019. Para más información, revise https://dwc.tdwg.org/terms/#dwc:eventDate	Obligatorio
recordedBy	"Oliver P. Pearson | Anita K. Pearson"	Una lista (concatenada y separada) de los nombres de las personas, grupos u organizaciones responsables de la colecta u observación del espécimen. La práctica recomendada es sepearar la valores con una barra vertical (' | '). Incluir información sobre el observador mejora la reproducibilidad científica (Groom et al. 2020).	Altamente recomendado
organismQuantity	50	Numero positivo de droplets/chambers en la muestra	Altamente recomendado para ddPCR, dPCR
organismQuantityType	ddPCR droplets dPCR chambers	El tipo de partición	Altamente recomendado para ddPCR, dPCR
sampleSizeValue	20000	El número de particiones aceptadas (n), e.g. representa el número aceptado de droplets en ddPCR o chambers en dPCR.	Altamente recomendado para ddPCR, dPCR
sampleSizeUnit	ddPCR droplets dPCR chambers	El tipo de partición, debería ser igual al valor presente en organismQuantityType	Altamente recomendado para ddPCR, dPCR

Elemento	Ejemplos	Descripción	Obligatoriedad
materialSampleID	https://www.ncbi.nlm.nih.gov/biosample/15224856 urn:uuid:a964805b-33c2-439a-beaa-6379ebbfcd03	Un identificador para muestras de material (no hace referencia a muestras digitales sino físicas, como exicados o tejidos). Use el ID BioSample si uno fue obtenido de un archivo de nucleótidos. En ausencia de un identificador único global persistente, puede construir uno usando una combinación de identificadores en el registro, de tal forma que el materialSampleID sea globalmente único.	Altamente recomendado
samplingProtocol	Trampa de luz UV	El nombre, la descripción o la referencia del método o protocolo de muestreo usado durante el evento de muestreo. https://dwc.tdwg.org/terms/#dwc:samplingProtocol	Recomendado
decimalLatitude	60.545207	La latitud geográfica (en grados decimales, utilizando el sistema de referencia espacial provisto en geodeticDatum) del centro geográfico de una ubicación. Los valores positivos se encuentran al norte del ecuador, los valores negativos están al sur del mismo. Los valores admitidos se encuentran entre -90 y 90.	Altamente recomendado
decimalLongitude	24.174556	La longitud geográfica (en grados decimales, mediante el sistema de referencia espacial provisto en geodeticDatum) del centro geográfico de una ubicación. Los valores positivos se encuentran al este del meridiano de Greenwich, los valores negativos se encuentran al oeste de la misma. Los valores admitidos se encuentran entre -180 y 180.	Altamente recomendado
scientificName	*Gadus morhua* L. 1758, BOLD:ACF1143	Nombre científico del taxón conocido más cercano (especie o superior) o un identificador para un OTU de BOLD o UNITE	Obligatorio
kingdom	Animalia	Taxonomía superior	Altamente recomendado
phylum	Chordata	Taxonomía superior	Recomendado
class	Actinopterygii	Taxonomía superior	Recomendado
order	Gadiformes	Taxonomía superior	Recomendado
family	Gadidae	Taxonomía superior	Recomendado
genus	*Gadus*	Taxonomía superior	Recomendado

Tabla 5. Campos recomendados de la extensión de datos derivados de ADN https://rs.gbif.org/extension/gbif/1.0/dna_derived_data_2021-07-05.xml (una selección) para datos de ddPCR/qPCR

Elemento	Ejemplos	Descripción	Obligatoriedad
sop	https://www.protocols.io/view/protocol-for-dna-extraction-and-quantitative-pcr-d-vwie7ce https://doi.org/10.17504/protocols.io.vwie7ce	Los procedimientos operativos estándar utilizados en el montaje y/o anotación de los genomas, metagenomas o secuencias ambientales. Una referencia a un protocolo bien documentado, por ejemplo usar protocols.io	Altamente recomendado
annealingTemp	60	La temperatura de reacción durante la cual se realizó la fase de anillado de la PCR.	Obligatorio si se presentó fase de anillado
annealingTempUnit	Grados Centígrados		Altamente recomendado
pcr_cond	desnaturalización inicial:94_3;anillado:50_1;elongación:72_1.5;elongación final:72_10;35	Descripción de las condiciones de reacción y componentes de la PCR siguiendo la estructura "desnaturalización inicial:94degC_1.5min; anillado=…"	Altamente recomendado
probeReporter	FAM	Tipo de fluoróforo (reportado) utilizado. La sonda se hibrida dentro del ADN objetivo amplificado. La actividad de la Polimerasa degrada la sonda que se hibrido a la plantilla y la sonda libera los fluoróforos desde y rompe las proximidades del quencher, permitiendo la fluorescencia del fluoróforo.	Altamente recomendado
probeQuencher	NFQ-MGB	Tipo de quencher utilizado. La molécula de quencher disminuye la fluorescencia emitida por el fluoróforo cuando es excitada por la fuente de luz del ciclo, siempre y cuando el fluoróforo y el quencher estén cerca, el quencher inhibe cualquier secuencia de fluorescencia.	Altamente recomendado
ampliconSize	83	Longitúd del amplicón en pares de bases	Altamente recomendado

Elemento	Ejemplos	Descripción	Obligatoriedad
thresholdQuantificationCycle	0.3	Umbral para el cambio en fluorescencia entre señales de ciclos	qPCR: Altamente recomendado
baselineValue	15	El número de ciclos cuando la señal de fluorescencia del objetivo de amplificación está por debajo de la fluorescencia de fondo que no es originada por el objetivo real de amplificación.	qPCR: Altamente recomendado
quantificationCycle	37.9450950622558	El número de ciclos requerido para que la señal de fluorescencia cruce el valor de umbral para la línea base. Ciclo de cuantificación (Cq), umbral de ciclo (Ct), punto de cruce (Cp), y punto de partida (TOP) se refieren al mismo valor del instrumento de medida en tiempo real. Usar el umbral de ciclo (Cq), es preferible de acuerdo al estándar de datos [RDML (Real-Time PCR Data Markup Language)](http://www.rdml.org)	
automaticThresholdQuantificationCycle	no	Especifica si el umbral fue fijado, ya sea por el instrumento o manualmente	
automaticBaselineValue	no	Especifica si la línea base fue fijada, ya sea por el instrumento o manualmente	
contaminationAssessment	no	Especifica si se realizó evaluación de contaminación para ADN o ARN	
estimatedNumberOfCopies	10300	Número de moléculas objetivo por µl. El número de copias promedio por partición (?) puede ser calculado usando el número de particiones (n) y el número estimado de copias en el volumen total de todas las particiones (m) con la fórmula ?=m/n.	
amplificationReactionVolume	22	Volumen de reacción de la PCR	
amplificationReactionVolumeUnit	µl	Unidad utilizada para el volumen de reacción de la PCR. Muchos de los instrumentos requiere la preparación de una cantidad inicial de muestra más grande que el volumen finalmente analizado.	
pcr_analysis_software	BIO-RAD QuantaSoft	El programa utilizado para analizar las ejecuciones de la d(d)PCR.	

Elemento	Ejemplos	Descripción	Obligatoriedad
experimentalVariance		Se recomienda realizar multiples réplicas para evaluar la varianza total experimental. Cuando experimentos simples de dPCR son realizados, un estimador mínimo de la varianza debido solamente al error de conteo debe ser calculado a partir de la distribución binomial (o una equivalente).	
target_gene	16S rRNA, 18S rRNA, nif, amoA, rpo	El gen objetivo o nombre del marcador para estudios basados en marcadores.	Altamente recomendado
target_subfragment	V6, V9, ITS	Nombre del subfragmento de un gen o marcador importante, por ejemplo para identificar regiones especiales en genes marcadores como la región hipervariable V6 del gen rRNA 16S	Altamente recomendado
pcr_primer_forward	GGACTACHVGGGTWTCTAAT	La secuencia del primer directo utilizado para el proceso de amplificación del gen, locus o subfragmento.	Altamente recomendado
pcr_primer_reverse	GGACTACHVGGGTWTCTAAT	La secuencia del primer inverso utilizado para el proceso de amplificación del gen, locus o subfragmento.	Altamente recomendado
pcr_primer_name_forward	jgLCO1490	El nombre del primer directo utilizado	Altamente recomendado
pcr_primer_name_reverse	jgHCO2198	El nombre del primer inverso utilizado	Altamente recomendado
pcr_primer_reference	https://doi.org/10.1186/1742-9994-10-34	Referencia para los primers	Altamente recomendado
env_broad_scale	forest biome [ENVO:01000174]	**Equivalente al env_biome en MIxS v4** En este elemento, se reporta de cuál ecosistema superior provienen los especímenes o muestras. Los sistemas identificados deben tener una granularidad espacial gruesa, para proveer de información general del contexto medioambiental donde la muestra fue realizada (e.g. ¿estaba en el desierto o en el bosque lluvioso?). Se recomienda utilizar subclases de ENVO pertenecientes a la clase bioma: http://purl.obolibrary.org/obo/ENVO_00000428	Recomendado

Elemento	Ejemplos	Descripción	Obligatoriedad
env_local_scale	litter layer [ENVO:01000338]	**Equivalente al env_feature en MIxS v4** En este elemento, se reporta la entidad o entidades que están en la localidad cercana a los especímenes o muestras, y cree que tienen influencias causales significativas sobre la muestra o especímen. Por favor utilice términos que estén presentes en ENVO y tengan una granularidad espacial menor a la documentada en env_broad_scale.	Recomendado
env_medium	soil [ENVO:00001998]	**Equivalente al env_material en MIxS v4** En este elemento, se reporta cuál material o materiales (separados por una barra vertical “	”) estaban en la vecindad inmediata de los especímenes o muestras antes del muestreo, utiliando una o más subclases de ENVO pertenecientes a la clase material medioambiental: http://purl.obolibrary.org/obo/ENVO_00010483
Recomendado	concentration	67.5	Concentración del ADN (peso ng/volumen µl), ver también http://terms.tdwg.org/wiki/ggbn:concentration

Elemento	Ejemplos	Descripción	Obligatoriedad
Recomendado	`concentrationUnit`	ng/µl	Unidad utilizada para la medida de concentración, ver tamvbién http://terms.tdwg.org/wiki/ggbn:concentrationUnit
Recomendado	`methodDeterminationConcentrationAndRatios`	Nanodrop, Qubit	Descripción del método utilizado para medir la concentración, ver también http://terms.tdwg.org/wiki/ggbn:methodDeterminationConcentrationAndRatios

Elemento	Ejemplos	Descripción	Obligatoriedad
Recomendado	`ratioOfAbsorbance260_230`	1.89	Relación de absorbancia a 260 nm y 230 nm para evaluar la pureza del ADN. (Generalmente una medida secundaria, indicando principalmente EDTA, carbohidratos, fenol), (solo para muestras de ADN). Ver también http://terms.tdwg.org/wiki/ggbn:ratioOfAbsorbance260_230

Elemento	Ejemplos	Descripción	Obligatoriedad
Recomendado	`ratioOfAbsorbance260_280`	1.91	Relación de absorbancia a 280 nm y 230 nm para evaluar la pureza del ADN. (Generalmente una medida secundaria, indicando principalmente EDTA, carbohidratos, fenol), (solo para muestras de ADN). Ver también http://terms.tdwg.org/wiki/ggbn:ratioOfAbsorbance260_280
Recomendado	`samp_collect_device`	biopsia, botella Niskin, Ahoyador	El método o dispositivo utilizado para colectar la muestra

Elemento	Ejemplos	Descripción	Obligatoriedad
Recomendado	samp_mat_process	Filtrado de agua marina, muestras guardadas en etanol	Cualquier proceso aplicado a la muestra durante o después de extraerla del medio ambiente. Este elemento acepta OBI, para buscar los término OBI (v 2018-02-12) por favor vea http://purl.bioontology.org/ontology/OBI
Recomendado	samp_size	5 litros	La cantidad o tamaño de la muestra (volumen, masa o área) que fue colectada
Recomendado	size_frac	0-0.22 micrometros	El tamañao del poro de filtrado utilizado en la preparación de la muestra
Recomendado	pcr_primer_lod	51	La habilidad del ensayo para detectar el objetivo a bajos niveles
Altamente recomendado	pcr_primer_loq	184	La habilidad del ensayo para cuantificar el número de copias a bajos niveles

2.3. Conjuntos de datos marinos y el Sistema de Información sobre la Biodiversidad Oceánica (OBIS)

Cuando se trabaja con conjuntos de datos originarios del entorno marino, se recomienda que la información se publique también en el Ocean Biodiversity Information System (OBIS) además de GBIF. OBIS es una base de datos mundial sobre biodiversidad, que se especializa en proporcionar datos fiables y accesibles relacionados con la vida marina y forma parte del IOC-UNESCO. Al igual que GBIF, y ALA, OBIS utiliza el formato DwC-A para la indexación y publicación de datos. Al publicar conjuntos de datos marinos a través de OBIS además de otras bases de datos de biodiversidad, los datos pueden llegar a un público más amplio y a diversos grupos que trabajan en el campo de la biodiversidad marina, ya que los conjuntos de datos de OBIS se utilizan a menudo para los procesos de las Naciones Unidas. Con el foco en los conjuntos de datos marinos, los estrictos controles de calidad de los datos aumentan la fiabilidad de los datos y dan lugar a pequeñas diferencias en la información que se necesita para publicar en OBIS en comparación con GBIF.

Para asegurar una nomenclatura taxonómica consistente OBIS utiliza el Registro Mundial de Especies Marinas (WoRMS) como única fuente taxonómica. Este es el caso también de las ocurrencias derivadas de datos genéticos; un nombre científico vinculado al identificador de un nombre científico de la base de datos de WoRMS es información altamente recomendada para su publicación. Si no se proporciona un identificador de nombre científico, OBIS tratará de hacer coincidir el nombre científico con WoRMS durante el proceso de ingesta de la base de datos, pero esto debe evitarse siempre que sea posible. Los nombres científicos no listados en WoRMS son aceptables, y serán enviados a WoRMS para su revisión y posible inclusión en el registro. Se recomienda que las secuencias no clasificadas sean documentadas como "incertae sedis", con el `scientificNameID` urn:lsid:marinespecies.rg:taxname:12 de WoRMS. Esto garantizará una interpretación correcta tanto por GBIF como por OBIS. Adicionalmente, se recomienda que los identificadores de secuencia de las bases de datos de referencia usadas (p.ej. los Barcode Index Numbers: los BINs de BOLD) se agregan en el campo `taxonConceptID` de la tabla principal de la ocurrencia. De esta manera, OBIS conservará su columna vertebral taxonómica basada en WoRMS, al tiempo que permitirá enlazar las bases de datos de secuencias de referencia. Los nombres de bases de datos de referencia que no son nombres estrictamente científicos, pueden ser añadidos como `verbatimIdentification`. La clasificación automática de nombres de especies se puede hacer a menudo a través del servicio de coincidencia de taxones de WoRMS y paquetes de R como worrms y taxize. En el futuro, OBIS planea buscar y actualizar periódicamente las asignaciones taxonómicas de secuencias enviadas a medida que las bases de datos de referencia se desarrollen con el tiempo, por lo que es altamente recomendable documentar la información de la secuencia genética vinculada a cada ocurrencia.

Otro elemento requerido para la publicación de datos a través de OBIS son las coordenadas geográficas. OBIS realiza controles de calidad adicionales relacionados a los datos marinos, por ejemplo que las coordenadas para especies estrictamente marinas no estén en tierra y que el valor de profundidad reportado esté en un rango razonable. Finalmente, se debe mencionar que adicionalmente OBIS suporta el uso de la extensión de Medidas o Hechos Extendida (eMoF). Esta extensión permite conectar datos abióticos ambientales y medidas o hechos asociadas al protocolo de muestreo, con los eventos de muestreo o los registros biológicos, así como a medidas bióticas asociadas a los registros, de una forma flexible y estandarizada. OBIS tiene un ejemplo de conjunto de datos de metabarcoding (eDNA) con scripts para realizar la transformación de los datos, que está disponible en https://github.com/iobis/dataset-edna.

Tabla 6. Requisitos y recomendaciones de OBIS para registrar ocurrencias basadas en ADN. La tabla resalta diferencias importantes en los valores de campo y requisitos en comparación con la publicación en GBIF. Aquí se ejemplifica con una detección de ADN de la ballena azul (Balaenoptera musculus).

Elemento	Valor/ejemplo (OBIS)	Descripción	Obligatoriedad
scientificName	Balaenoptera musculus	Nombre científico, preferiblemente como aparece en la base de datos WoRMS. Esto difiere de GBIF, donde se recomienda utilizar el nombre del taxón derivado del enfoque de clasificación utilizado.	Obligatorio
scientificNameID	urn:lsid:marinespecies.org:taxname:137090	El ID del nombre científico de "Balaenoptera musculus" según la base de datos de WoRMS.	Altamente recomendado
taxonConceptID	NCBI:txid9771	El ID de NCBI vinculado a Balaenoptera musculus en la base de datos taxonómica de NCBI. También puede ser un BIN-ID si BOLD fue utilizado para la identificación, u otro ID de una base de datos diferente.	Recomendado
verbatimIdentification	Balaenoptera musculus	El nombre correspondiente al ID de NCBI (Balaenoptera musculus) (o otro ID). No es necesario que corresponda al valor documentado en scientificName.	Recomendado

Tabla 7. Requisitos y recomendaciones de OBIS para registrar secuencias que no pueden clasificarse como un nombre científico en ningún nivel taxonómico.

Elemento	Valor/ejemplo (OBIS)	Descripción	Obligatoriedad
scientificName	incertae sedis	El nombre científico para secuencias desconocidas recomendado por OBIS. Use este nombre cuando la secuencia/taxonomía es desconocida. Esta recomendación difiere de la de GBIF, donde se recomienda utilizar el nombre del taxón tal como fue recuperado del clasificador, incluso cuando no es estrictamente un nombre científico.	Obligatorio
scientificNameID	urn:lsid:marinespecies.org:taxname:12	El ID del nombre científico de "incertae sedis" según la base de datos de WoRMS utilizado para secuencias desconocidas por recomendación de OBIS. Use este ID cuando la secuencia/taxonomía es desconocida.	Altamente recomendado
taxonConceptID	NCBI:txid1899546	El identificador (ID) es una base de datos taxonómica externa, por ejemplo un repositorio de secuenias de referencia.	Recomendado
verbatimIdentification	Eucariota fototrópico	El nombre del taxón en una base de datos externa, correspondiente al ID del concepto del taxón.	Recomendado

3. Perspectivas a futuro

El interés actual por exponer los datos derivados de ADN a través de las plataformas de datos sobre biodiversidad es muy alto, y probable la demanda siga en crecimiento. Nuestro objetivo es que las recomendaciones de mapeo proporcionadas aquí sigan siendo válidas y evolucionen lentamente, incluso en el caso de que el empaquetamiento y la indexación por parte de las plataformas de datos sobre biodiversidad puedan desarrollarse más rápidamente. Los autores conocen, pero aún no han consultado BOLD Handbook, BIOM format and http://edamontology.org/page.

Sugerimos que las plataformas de datos como ALA y GBIF trabajen para adoptar formatos de datos que soporten datos relacionales y jerárquicos más complejos. Algunos ejemplos podrían ser los Frictionless Data Format y de dominio más específico Biological Observation Matrix (formato BIOM). Este último, utilizado por varias herramientas bioinformáticas (QIIME2, Mothur, USEARCH etc.), por lo tanto, podría ayudar a los editores a omitir la conversión de datos al formato DwC-A. Un formato de datos más flexible que el actual esquema en estrella de DwC es crucial para permitir eventos de muestreo jerárquicos y muestras de materiales; así como para adjuntar datos de secuencias a los registros biológicos individuales dentro de un evento de muestreo.

Las plataformas de datos de la biodiversidad también tendrán que permitir a los investigadores incluir o excluir fácilmente los datos de ocurrencia derivados del ADN en los resultados de consultas. Los formatos de datos sugeridos anteriormente podrían abrir oportunidades para una clasificación más rica de los tipos de evidencia en las que se basa un registro biológico específico. Sin embargo, por el momento falta un valor apropiado en el vocabulario del BasisOfRecord para este tipo de datos. Sugerimos, como solución pragmática inmediata, que el BasisOfRecord se extienda con un valor como “DNA”, “DNA-derived”, o similar. Como se ha descrito anteriormente, los datos derivados de ADN pueden provenir de muestras bien documentadas u organismos individuales, estar respaldado por material físico preservado o no, resultar de secuencias genéticas u otros métodos de detección de ADN, como qPCR. Las plataformas de datos de la biodiversidad y TDWG deberían proporcionar los medios para diferenciar entre estos tipos de datos y sus orígenes.

También recomendamos que las plataformas de datos indexen las secuencias actuales, o al menos una suma de verificación MD5 de estos, para facilitar las búsquedas de ASVs a través de los conjuntos de datos. Si se proporcionan ASVs, los MD5s deben ser generados por las plataformas de descubrimiento de la biodiversidad; si no se proporcionan los ASVs, los MD5 deben ser obligatorios.

Como se menciona en [taxonomy-of-sequences] y § 2.1.4, alentamos a las plataformas de datos sobre biodiversidad a continuar trabajando en implementar bases de datos taxonómicas moleculares relevantes en sus árboles taxonómicos.

Aplicación de otros métodos y tecnologías, como Oxford Nanopore, PacBio y shotgun sequencing, probablemente desencadenarán la necesidad de ajustes a esta guía para integrar nuevos datos específicos y campos de metadatos.

Glosario

Atlas de la Vida de Australia (ALA)

ALA es una plataforma web que reúne datos de biodiversidad de Australia de múltiples fuentes, haciéndolos accesibles y reutilizables para cualquier persona (ver https://www.ala.org.au/about-ala/). La plataforma de infraestructura abierta desarrollada por ALA también es utilizada por varios otros países para sus propias plataformas nacionales de datos de biodiversidad (ver https://living-atlases.gbif.org/).

Variante de Secuencia de Amplicón (ASV)

Secuencia única de ADN derivado de técnicas de secuenciación masiva y denoising, se asume que representa una variante de secuencia biológica real. Vea también Operational Taxonomic Unit (OTU) y (Callahan et al. 2017).

Interfaz de programación de aplicaciones (API)

Conjunto de protocolos y herramientas para la interacción y transmisión de datos entre diferentes aplicaciones informáticas.

Números de índice de códigos de barras (BIN)

Operational Taxonomic Units (OTUs) a nivel de especie (species-level) derivados de un proceso de agrupamiento del gen Citocromo oxidasa I (COI) en animales. Cada BIN es asignado con un identificador único global y está disponible para búsqueda en la base de datos Barcode of Life Data System (BOLD).

Sistema de datos de código de barras de vida (BOLD)

BOLD es la base de datos de referencia mantenida por el Centro de Genómica de la biodiversidad en Guelph, respaldado por el Consorcio Internacional de códigos de barras (IBOL) por sus siglas en inglés. Alberga datos de códigos de barras para especímenes de referencia y secuencias para especies de Eucariota, particularmente COI para animales, y mantiene el sistema Barcode Index Number (BIN; Ratnasingham & Hebert 2013), que son identificadores para OTUs aproximadamente al rango de especie, basado en agrupamientos de secuencias muy similares.

Plataforma de datos de biodiversidad

Recurso general en línea para descubrir y acceder a datos de biodiversidad derivados de diversas fuentes, como colecciones de historia natural, ciencia ciudadana, proyectos de ecología y monitoreo, y secuencias genéticas. Puede ser global (GBIF) o nacional (ALA).

Clusterizar

En la clasificación taxonómica, el proceso de agrupar organismos juntos de acuerdo a algún criterio de similaridad. Ver Operational Taxonomic Unit.

ADN comunitario (masivo)

ADN de muestras masivas (por ejemplo, muestras de plancton o muestras de trampas de Malaise compuestas por varios individuos de muchas especies). A los efectos de esta guía, las muestras de ADN en masa se incluyen en el concepto de eADN.

Archivo Darwin Core (DwC-A)

Formato de archivo comprimido (ZIP) para el intercambio de datos de biodiversidad de acuerdo al Darwin Core (DwC) standard. Esencialmente es un grupo de archivos CSV autocontenidos e interconectados y un documento XML que describe los archivos incluidos y las columnas de datos, y las relaciones mutuas.

Estándar Darwin Core (DwC)

Estándar para compartir y publicar datos sobre biodiversidad, desarrollado por la organización de información de estándares para biodiversidad (TDWG). En principio, es un grupo de elementos usados para describir distintos tipos de observaciones de biodiversidad, como eventos de muestreo, registros biológicos y listas de chequeo. Los elementos del Darwin Core actuales están descritos en la Guía de Referencia Rápida.

Vocabulario Controlado

Grupo de términos o conceptos preferidos, con significados específicos e interrelaciones bien definidas, facilitando el intercambio y reutilización de datos.

ddPCR (Reacción en cadena de polimerasa digital en gota)

PCR digital en gota. Método para medir la cantidad absoluta de ADN (número de copias) de un marcador en una muestra. Ver también qPCR.

Denoising

En metabarcoding, el método para separar las secuencias biológicas reales (vea ASVs) de secuencias espurias causadas por la amplificación de la PCR y el error de secuenciación.

Identificador Único Digital (DOI)

Referencia de larga duración utilizada para identificar (y localizar) de forma única un Objeto Digital, como un conjunto de datos o una publicación científica.

Código de barras de ADN y metabarcoding (secuencia de amplicón)

Uso de fragmentos cortos y estandarizados de ADN para identificar organismos individuales a través de secuenciación. El metabarcoding combina el uso de códigos de baras con secuenciación de alto rendimiento, utilizando primers universales para amplificar y secuenciar grupos generales de organimos en muestras de ADN ambiental (eDNA).

Marcador de AND

Un fragmento de ADN utilizado como marcador para alguna propiedad (e.g. afiliación taxonómica). Puede, pero no tiene que ser un gen o parte de un gen.

Base de datos para ADN de metabarcoding

Base de datos que contiene secuencias de ADN (códigos de barras de ADN) de organismos previamente estudiados o recuperados. Las secuencias de referencia son idealmente generadas a partir de individuos de especies descritas y bien estudiadas (utilizando el espécimen tipo como fuente ideal) o de un nivel taxonómico superior (e.g. género, familia), pero también pueden provenir de esfuerzos de secuenciación de ADN ambiental. Es recomendado no confiar ciegamente en las "secuencias de referencia".

Sonda de ADN

Un fragmento corto de ADN sintético de hebra simple con un etiquetado fluorescente que se une a la región seleccionada del ADN objetivo (marcador) durante la PCR. Incrementa la especificidad y puede ser utilizado como adición a los primers en qPCR y ddPCR para detectar y cuantificar un marcador genérico.

Instituto Europeo de Bioinformática (EMBL-EBI)

Organización intergubernamental para la investigación bioinformática y de servicios, parte del Laboratorio Europeo de Biología Molecular (EMBL), proporcionando lecturas de secuencias (sin procesar) y datos de ensamblaje a través de the European Nucleotide Archive (ENA).

ADN ambiental (eDNA)

ADN de una muestra ambiental, e.g. suelo, agua, aire o un organismo hospedero. Una definición utilizada frecuentemente es que el ADN ambiental es el material genético (ADN) obtenido de muestras ambientales sin ninguna fuente biológica de evidencia obvia (Thomsen and Willerslev 2015).

Repositorio Europeo de Nucleótidos (ENA)

Repositorio Europeo para secuencias de nucleotidos, abarcando secuencias sin procesar, secuencias con informacion ensamblada y anotaciones funcionales. Incluye la Sequence Read Archive (SRA) y es mantenido por el Instituo Europeo de Bioinformática (EMBL-EBI), como parte del the International Nucleotide Sequence Database Collaboration (INSDC).

FASTQ

Archivo de texto estándar para guardar información molecular de secuencias y los controles de calidad asociados, derivados de High-throughput sequencing (HTS). Para cada posición de secuencia, se utilizan caracteres individuales tipo ASCII que representan un llamado base (nucleotido identificado) y un puntaje, respectivamente.

Sistema Global de Información sobre Biodiversidad (GBIF)

Red internacional e infraestructura de datos, enfocada principalmente en movilizar y proveer acceso abierto a datos globales de biodiversidad.

Red Global de Biodiversidad Genómica (GGBN)

Red internacional de instituciones interesadas en el eficiente intercambio y uso de muestras genéticas de biodiversidad y sus metadatos asociados, por ejemplo promoviendo el estándar de datos GGBN, que es compatible con el Darwin Core.

Sistema de Posicionamiento Global (GPS)

Sistema de navegación satelital operado por la Fuerza Espacial de los Estados Unidos de América.

Secuenciación de alto rendimiento (HTS)

Diferentes tecnologías utilizadas para la secuenciación masiva en paralelo, produciendo millones de lecturas de secuencias de ADN a partir de preparaciones de bibliotecas de material genético, en vez de tener como objetivo amplicones simples como es tradicional en la secuenciación Sanger. También es llamada Secuenciación de Nueva Generación (NGS).

Ingestión

Proceso de importar datos de fuentes heterogéneas, como bases de datos locales, archivos de texto u hojas de cálculo a un sistema de destino común, como una biodiversity data platform en línea, para su almacenamiento y posterior análisis. Normalmente incluye pasos de extracción, transformación (limpieza) y carga (ETL).

Indexación

Organización de la información de acuerdo a un esquema o estructura específica, logrando que los datos sean más fáciles de acceder y presentar.

Colaboración Internacional de Base de datos de Secuencias de Nucleótidos (INSDC)

Esfuerzo conjunto del Banco de ADN de Japón (DDBJ), EMBL y NCBI para proveer acceso público global a los datos de secuenciación de nucleótidos e información asociada.

Metagenómica

Secuenciación sin PCR de fragmentos genéticos aleatorios en una muestra mixta.

Estándar de Información mínima para cualquier (x) secuencia (MIxS)

Familia de estándares (listas de chequeo) para metadatos de secuencias, desarrollado por el Consorcio de Estándares Genéticos (GSC).

Unidad molecular taxonómica operativa (mOTU)

Ver Operational Taxonomic Unit (OTU).

Centro Nacional para la Información Biotecnológica (NCBI)

División de la Librería Nacional de Medicina (NLM) de Estados Unidos de América que alberga importantes recursos bioinformáticos, como la base de datos de secuencias de AND GenBank, y el Sequence Read Archive (SRA) para datos de secuencias de alto rendimiento.

Secuenciación de nueva generación (NGS)

Ver High-throughput sequencing (HTS).

Registro biológico

La existencia de un organismo (sensu http://rs.tdwg.org/dwc/terms/Organism) en un lugar particular en un tiempo particular.

Unidad taxonómica operativa (OTU)

Agrupamiento de organismos basados en la similitud de marcadores de secuencia(s) de ADN, utilizados para clasificación taxonómica. Incluye por ejemplo Species Hypothesis en UNITE, y Barcode Index Numbers en el Sistema de datos de código de barras de vida (BOLD). Los Amplicon Sequence Variants (ASVs) pueden ser considerados análogos a los zero radius OTUs (zOTUs).

Reacción en Cadena de la Polimerasa (PCR)

Técnica para la amplificación rápida y la detección de fragmentos específicos de secuencias de ADN objetivo (o de ARN). Las regiones amplificadas son determinadas por el par de PCR primers usados en la reacción.

Pipeline

En bioinformática, es un grupo de algoritmos o herramientas aplicadas en un flujo de trabajo preestablecido, por ejemplo para datos de High-throughput sequencing (HTS).

Primers (PCR primers)

Fragmentos cortos sintéticos de hebra simple de ADN que se emparejan con la región objetivo de ADN (marcador) seleccionada para iniciar la replicación durante una PCR. Un par de primers es necesario para que la enzima polimerasa amplifique el marcador seleccionado.

qPCR (Reacción en cadena de polimerasa cuantitativa)

PCR cuantitativa. Método que mide la cantidad relativa de ADN de un marcador en una muestra. Vea también ddPCR.

Muestra

Material (agua, suelo, contenido estomacal, etc.) obtenido para análisis.

Alineamiento de secuencias

Proceso bioinformático para comparar y reorganizar dos o más secuencias de moléculas (ADN, ARN o proteína) para detectar similitudes causadas por relaciones evolutivas.

Hipótesis de especie (SH)

Operational Taxonomic Unit (OTU) a nivel de especie (species-level) como está definido en la base de datos de UNITE y en el entorno de gestión de secuencias para hongos.

Espécimen

Un individuo, animal, planta, hongos, etc. Usado como un ejemplo de su especie o un espécimen tipo para estudio científico o exhibición.

Archivador de lecturas de secuencias (SRA)

Repositorio público de datos de secuenciación de alto rendimiento (NGS), con instancias operadas por the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EMBL-EBI) y el Banco de ADN de Japón (DDBJ). Incluye tanto resultados se secuencias sin procesar (sin denoising) y sequence alignments. Uno de los tres componentes de the European Nucleotide Archive (ENA), y previamente conocido como Archivador de secuencias cortas.

Secuenciación por captura de objetivo

Secuenciación de fragmentos de ADN asilados con sondas de hibridación.

UNITE

UNITE es una ambiente web para el manejo de secuencias, centrado en la región del ribosoma nuclear de los eucariotas ITS. Todas las secuencias públicas son agrupadas en hipótesis de especies (SH), a las que se les asigna un DOI único. Un servicio de comparación de SH produce varios elementos de información, incluyendo que especies están presentes en las muestras de eDNA, si estas especies son potencialmente nuevas especies no descritas, otros estudios en los cuales hayan sido recuperadas, si alas especies son invasoras en la región y si están amenazadas. Los DOIs está conectados con el árbol taxonómico de PlutoF platform y GBIF, de esa manera están acompañados por el nombre de un taxón cuando está disponible. Los datos usados en UNITE están alojados y manejados rn PlutoF. Los datos son representados a través de un rango de estándares, principalmente Darwin Core, MIxS, y DMP Common Standard; con un suporte parcial disponible para EML, MCL, y GGBN. PlutoF exporta datos principalmente a través de los formatos CSV y FASTA. PlutoF también puede ser utilizado para publicar los datos en GBIF (usando el formato DwC) y para preparar los archivos para someterlos a GenBank. También es posible descargar las listas de especies a partir de sus datos y descargar su proyecto como un documento JSON con los datos del proyecto en una estructura jerárquica.

Radio cero de otu (zOTU)

Ver ASV.

Referencias

- Amid C, Alako BT, Balavenkataraman Kadhirvelu V, Burdett T, Burgin J, Fan J, Harrison PW, Holt S, Hussein A, Ivanov E & Jayathilaka S (2020) The European Nucleotide Archive in 2019. Nucleic acids research 48(D1): D70-D76. https://doi.org/10.1093/nar/gkz1063
- Andersen K, Bird KL, Rasmussen M, Haile J, Breuning-Madsen H, Kjaer KH, Orlando L, Gilbert MTP and Willerslev E (2012) Meta-Barcoding of 'Dirt' DNA from Soil Reflects Vertebrate Biodiversity. Molecular Ecology 21(8): 1966–79. https://doi.org/10.1111/j.1365-294X.2011.05261.x
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2006) GenBank, Nucleic Acids Research, 34(1): D16–D20, https://doi.org/10.1093/nar/gkj157
- Berry O, Jarman S, Bissett A, Hope M, Paeper C, Bessey C, Schwartz MK, Hale J & Bunce M (2021) Making environmental DNA (eDNA) biodiversity records globally accessible. Environmental DNA, 3(4), 699-705. https://doi.org/10.1002/edn3.173
- Bessey C, Jarman SN, Berry O et al. (2020) Maximizing fish detection with eDNA metabarcoding. Environmental DNA: 1–12. https://doi.org/10.1002/edn3.74
- Biggs J, Ewald N, Valentini A, Gaboriaud C, Dejean T, Griffiths RA, Foster J, et al. (2015) Using eDNA to Develop a National Citizen Science-Based Monitoring Programme for the Great Crested Newt (*Triturus cristatus*). Biological Conservation 183: 19–28. https://doi.org/10.1016/j.biocon.2014.11.029
- Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R & Abebe E (2005) Defining operational taxonomic units using DNA barcode data. Philosophical Transactions of the Royal Society B: Biological Sciences 360(1462): 1935-1943. https://doi.org/10.1098/rstb.2005.1725
- Bolyen E, Rideout JR, Dillon MR et al. (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37: 852–857. https://doi.org/10.1038/s41587-019-0209-9
- Boussarie G, Bakker J, Wangensteen OS, Mariani S, Bonnin L, Juhel JB, Kiszka JJ, Kulbicki M, Manel S, Robbins WD & Vigliola L (2018) Environmental DNA illuminates the dark diversity of sharks. Science Advances 4(5): eaap9661. https://doi.org/10.1126/sciadv.aap9661
- Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, … & Wittwer CT (2009). The MIQE Guidelines: *M*inimum *I*nformation for Publication of *Q*uantitative Real-Time PCR *E*xperiments. https://doi.org/10.1373/clinchem.2008.112797
- Callahan B, McMurdie P & Holmes S (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal 11: 2639–2643. https://doi.org/10.1038/ismej.2017.119
- Callahan B, McMurdie P, Rosen M et al. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13, 581–583. https://doi.org/10.1038/nmeth.3869
- Centre for Biodiversity Genomics, University of Guelph (2021) The Global Taxonomy Initiative 2020: A Step-by-Step Guide for DNA Barcoding. Technical Series No. 94. Secretariat of the Convention on Biological Diversity, Montreal, 66 pp. https://www.cbd.int/doc/publications/cbd-ts-94-en.pdf
- Convention on Biological Diversity (2020) Report of the ad hoc Technical Expert Group on Digital Sequence Information On Genetic Resources, 17-20 March 2020. Montreal, Canada. https://www.cbd.int/doc/c/ba60/7272/3260b5e396821d42bc21035a/dsi-ahteg-2020-01-07-en.pdf
- Debroas D, Domaizon I, Humbert JF, Jardillier L, Lepère C, Oudart A & Taïb N (2017) Overview of freshwater microbial eukaryotes diversity: a first analysis of publicly available metabarcoding data. FEMS Microbiology Ecology 93(4): fix023. https://doi.org/10.1093/femsec/fix023

- Doi H, Fukaya K, Oka SI, Sato K, Kondoh M & Miya M (2019) Evaluation of Detection Probabilities at the Water-Filtering and Initial PCR Steps in Environmental DNA Metabarcoding Using a Multispecies Site Occupancy Model. Scientific Reports 9(1): 3581. https://doi.org/10.1038/s41598-019-40233-1
- Durkin L, Jansson T, Sanchez M, Khomich M, Ryberg M, Kristiansson E, Nilsson RH (2020) When mycologists describe new species, not all relevant information is provided (clearly enough). MycoKeys 72: 109-128. https://doi.org/10.3897/mycokeys.72.56691
- Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26(19): 2460–2461, https://doi.org/10.1093/bioinformatics/btq461
- Ekrem T & Majaneva M (2019) DNA-Metastrekkoding Til Undersøkelser Av Invertebrater I Ferskvann. NTNU Vitenskapsmuseet Naturhistorisk Notat. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2612638.
- Elbrecht V & Leese F (2015) Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. PLoS ONE 10(7): e0130324. https://doi.org/10.1371/journal.pone.0130324
- Ficetola GF, Miaud C, Pompanon F, & Taberlet P (2008). Species detection using environmental DNA from water samples. Biology letters, 4(4), 423-425. https://doi.org/10.1098/rsbl.2008.0118
- Fossøy F, Brandsegg H, Sivertsgård R, Pettersen O, Sandercock BK, Solem Ø, Hindar K & Tor AM (2019) Monitoring Presence and Abundance of Two Gyrodactylid Ectoparasites and Their Salmonid Hosts Using Environmental DNA. Environmental DNA. https://doi.org/10.1002/edn3.45.
- Frøslev TG, Kjøller R, Bruun HH et al. (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nat Commun 8, 1188 . https://doi.org/10.1038/s41467-017-01312-x
- Groom Q, Güntsch A, Huybrechts P, Kearney N, Leachman S, Nicolson N, Page RDM, Shorthouse DP, Thessen, AE, Haston E. People are essential to linking biodiversity data. 2020. Database 2020:baaa072 https://doi.org/10.1093/database/baaa072.
- Hernandez C, Bougas B, Perreault-Payette A, Simard A, Côté G, & Bernatchez L (2020). 60 specific eDNA qPCR assays to detect invasive, threatened, and exploited freshwater vertebrates and invertebrates in Eastern Canada. Environmental DNA, 2(3): 373-386. https://doi.org/10.1002/edn3.89
- Hofstetter, V, Buyck, B, Eyssartier, G, Schnee S, Gindro K (2019) The unbearable lightness of sequenced-based identification. Fungal Diversity 96, 243–284. https://doi.org/10.1007/s13225-019-00428-3
- Huggett JF, Foy CA, Benes V, Emslie K, Garson JA, Haynes R, … & Bustin SA (2013). The Digital MIQE Guidelines: M inimum I nformation for Publication of Q uantitative Digital PCR E xperiments. Clinical chemistry, 59(6), 892-902. https://doi.org/10.1373/clinchem.2013.206375
- Hugerth LW, Andersson AF (2017) Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. Frontiers in Microbiology 8: 1561. https://doi.org/10.3389/fmicb.2017.01561
- Knudsen SW, Ebert RB, Hesselsøe M, Kuntke F, Hassingboe J, Mortensen PB, Thomsen PF et al (2019) Species-Specific Detection and Quantification of Environmental DNA from Marine Fishes in the Baltic Sea. Journal of Experimental Marine Biology and Ecology 510: 31–45. https://doi.org/10.1016/j.jembe.2018.09.004
- Lacoursière-Roussel A, Rosabal M & Bernatchez L (2016) Estimating Fish Abundance and Biomass from eDNA Concentrations: Variability among Capture Methods and Environmental Conditions. Molecular Ecology Resources 16(6): 1401–14. https://doi.org/10.1111/1755-0998.12522

- Leebens-Mack J, Vision T, Brenner E, Bowers JE, Cannon S, Clement MJ, Cunningham CW, DePamphilis C, DeSalle R, Doyle JJ & Eisen JA (2006) Taking the first steps towards a standard for reporting on phylogenies: Minimum Information About a Phylogenetic Analysis (MIAPA). Omics: a journal of integrative biology 10(2): 231-237. https://doi.org/10.1089/omi.2006.10.231
- Leinonen R, Sugawara H, Shumway M & International Nucleotide Sequence Database Collaboration (2011) The sequence read archive. Nucleic Acids Research 39(suppl_1): D19-D21. https://doi.org/10.1093/nar/gkq1019
- Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. (2014) Swarm: robust and fast clustering method for amplicon-based studies. PeerJ 2:e593 https://doi.org/10.7717/peerj.593
- McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, … & Caporaso JG (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. Gigascience, 1(1), 2047-217X. https://doi.org/10.1186/2047-217X-1-7
- Miralles A, Bruy T, Wolcott K, Scherz MD, Begerow D, Beszteri B, Bonkowski M, Felden J, Gemeinholzer B, Glaw F & Glöckner FO (2020) Repositories for Taxonomic Data: Where We Are and What is Missing. Systematic Biology: syaa026. https://doi.org/10.1093/sysbio/syaa026
- Mora C, Tittensor DP, Adl S, Simpson AG & Worm B (2011) How many species are there on Earth and in the ocean? PLoS Biology 9(8): e1001127. https://doi.org/10.1371/journal.pbio.1001127
- Nilsson RH, Tedersoo L, Abarenkov K, Ryberg M, Kristiansson E, Hartmann M, Schoch CL, Nylander JA, Bergsten J, Porter TM & Jumpponen A (2012) Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. MycoKeys 4: 37-63. https://doi.org/10.3897/mycokeys.4.3606
- Nilsson RH, Larsson KH, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L, Saar I, Kõljalg U, Abarenkov K (2019) The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Research, Volume 47, Issue D1, D259–D264. https://doi.org/10.1093/nar/gky1022
- Ogram A, Sayler GS, Barkay T (1987) The Extraction and Purification of Microbial DNA from Sediments. Journal of Microbiological Methods. https://doi.org/10.1016/0167-7012(87)90025-x.
- Ovaskainen O, Schigel D, Ali-Kovero H et al. (2013) Combining high-throughput sequencing with fruit body surveys reveals contrasting life-history strategies in fungi. The ISME Journal 7: 1696–1709. https://doi.org/10.1038/ismej.2013.61
- Parks, DH, Chuvochina, M, Chaumeil, P, Rinke C, Mussig AJ, Hugenholtz P (2020) A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol 38, 1079–1086. https://doi.org/10.1038/s41587-020-0501-8
- Pearson, WR & Lipman DJ (1988) Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences 85(8): 2444-2448. https://dx.doi.org/10.1073%2Fpnas.85.8.2444
- Penev P, Mietchen D, Chavan VS, Hagedorn G, Smith VS, Shotton D, Tuama ÉÓ, Senderov V, Georgiev T, Stoev P, Groom QJ, Remsen D, Edmunds SC (2017) Strategies and guidelines for scholarly publishing of biodiversity data. Research ideas and outcomes 3: e12431, https://doi.org/10.3897/rio.3.e12431
- Pietramellara G, Ascher J, Borgogni F, Ceccherini MT, Guerri G & Nannipieri P (2009) Extracellular DNA in Soil and Sediment: Fate and Ecological Relevance. Biology and Fertility of Soils 45: 219-235. https://doi.org/10.1007/s00374-008-0345-8.
- Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System. Molecular Ecology Notes, 7: 355-364. https://doi.org/10.1111/j.1471-8286.2007.01678.x
- Ratnasingham S, Hebert PDN (2013). A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. PloS one, 8(7), e66213. https://doi.org/10.1371/journal.pone.0066213

- Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4, e2584. https://doi.org/10.7717/peerj.2584
- Ruppert KM, Kline RJ, Rahman MS (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. Global Ecology and Conservation, 17, e00547. https://doi.org/10.1016/j.gecco.2019.e00547
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, … & Weber CF (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology, 75(23), 7537-7541. https://doi.org/10.1128/AEM.01541-09
- Shea MM, Kuppermann J, Rogers MP, Smith DS, Edwards P & Boehm AB (2023) Systematic review of marine environmental DNA metabarcoding studies: toward best practices for data usability and accessibility. PeerJ, 11, p.e14993. https://doi.org/10.7717/peerj.14993
- Sigsgaard EE, Jensen MR, Winkelmann IE, Møller PR, Hansen MM, Thomsen PF (2020). Population -level inferences from environmental DNA—Current status and future perspectives. Evolutionary Applications, 13(2), 245-262. https://doi.org/10.1111/eva.12882
- Somervuo P, Koskela S, Pennanen J, Nilsson RH, Ovaskainen O (2016) Unbiased probabilistic taxonomic classification for DNA barcoding. Bioinformatics 32(19):2920–2927, https://doi.org/10.1093/bioinformatics/btw346
- Strand DA, Johnsen SI, Rusch JC, Agersnap S, Larsen WB, Knudsen SW, Møller PR & Vrålstad T (2019) Monitoring a Norwegian Freshwater Crayfish Tragedy: eDNA Snapshots of Invasion, Infection and Extinction. Journal of Applied Ecology 56(7): 1661-1673. https://doi.org/10.1111/1365-2664.13404.
- Taberlet P, Bonin A, Coissac E & Zinger L (2018) Environmental DNA: For Biodiversity Research and Monitoring. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oso/9780198767220.001.0001
- Taberlet P, Coissac E, Hajibabaei M & Rieseberg LH (2012) Environmental DNA. Molecular Ecology 21(8): 1789–93. https://doi.org/10.1111/j.1365-294X.2012.05542.x
- Takahara T, Minamoto T, Yamanaka H, Doi H & Kawabata Z (2012) Estimation of Fish Biomass Using Environmental DNA. PLoS ONE 7(4): e35868. https://doi.org/10.1371/journal.pone.0035868
- Tedersoo, L, Bahram M, Puusepp R, Nilsson RH & James TY (2017) Novel soil-inhabiting clades fill gaps in the fungal tree of life. Microbiome 5: 42. https://doi.org/10.1186/s40168-017-0259-5
- Tedesco PA, Bigorne R, Bogan AE, Giam X, Jézéquel C & Hugueny B (2014) Estimating how many undescribed species have gone extinct. Conservation Biology 28(5): 1360-1370. https://doi.org/10.1111/cobi.12285
- Thalinger B, Deiner K, Harper LR, Rees HC, Blackman RC, Sint D, … & Bruce K (2021). A validation scale to determine the readiness of environmental DNA assays for routine species monitoring. Environmental DNA. https://doi.org/10.1101/2020.04.27.063990
- Thomsen PF, Kielgast JOS, Iversen LL, Wiuf C, Rasmussen M, Gilbert MTP Orlando L & Willerslev E (2012) Monitoring Endangered Freshwater Biodiversity Using Environmental DNA. Molecular Ecology 21(11): 2565–73. https://doi.org/10.1111/j.1365-294X.2011.05418.x
- Thomsen PF, Møller PR, Sigsgaard EE, Knudsen SW, Jørgensen OA & Willerslev E (2016) Environmental DNA from Seawater Samples Correlate with Trawl Catches of Subarctic, Deepwater Fishes. PLoS ONE 11(11): e0165252. https://doi.org/10.1371/journal.pone.0165252
- Thomsen PF & Willerslev E (2015) Environmental DNA – An Emerging Tool in Conservation for Monitoring Past and Present Biodiversity. Biological Conservation 183: 4–18. https://doi.org/10.1016/j.biocon.2014.11.019

- Tyson, GW & Hugenholtz, P (2005). Environmental shotgun sequencing. Encyclopedia of genetics, genomics, proteomics, and bioinformatics. Edited by Lynn B. Jorde. West Sussex, UK: John Wiley & Sons.1386-1391. https://doi.org/10.1002/047001153X.g205313
- Valentini A, Taberlet P, Miaud C, Civade R, Herder J, Thomsen PF, Bellemain E et al. (2016) Next-Generation Monitoring of Aquatic Biodiversity Using Environmental DNA Metabarcoding. Molecular Ecology 25(4): 929–42. https://doi.org/10.1111/mec.13428
- Wacker S, Fossøy F, Larsen BM, Brandsegg H, Sivertsgård R, & Karlsson S (2019). Downstream transport and seasonal variation in freshwater pearl mussel (Margaritifera margaritifera) eDNA concentration. Environmental DNA, 1(1), 64-73. https://doi.org/10.1002/edn3.10
- Wilkinson M, Dumontier M, Aalbersberg I et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18
- Wittwer C, Stoll S, Strand D, Vrålstad T, Nowak C, & Thines M (2018). eDNA-based crayfish plague monitoring is superior to conventional trap-based assessments in year-round detection probability. Hydrobiologia, 807(1), 87-97. https://doi.org/10.1007/s10750-017-3408-8
- Yates MC, Fraser DJ & Derry AM (2019) Meta-analysis Supports Further Refinement of eDNA for Monitoring Aquatic Species-specific Abundance in Nature. Environmental DNA. https://doi.org/10.1002/edn3.7.
- Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G & Vaughan R (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology 29(5): 415. https://doi.org/10.1038/nbt.1823