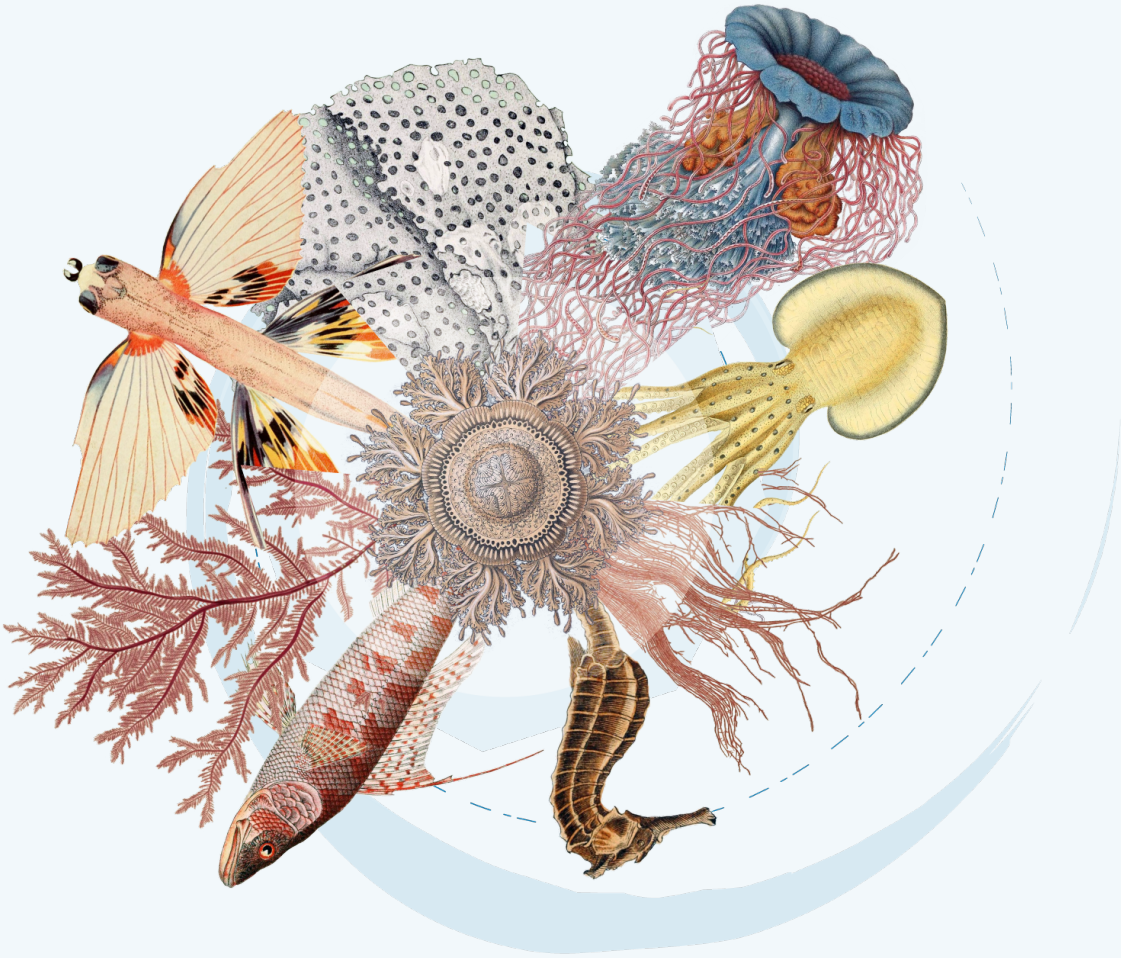

Publier des données dérivées de l'ADN sur les plateformes de données sur la biodiversité

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

Version 1.3.0, 7 June 2023

GBIF Global Biodiversity Information Facility

OBIS OCEAN BIODIVERSITY INFORMATION SYSTEM

Table des matières

Colophon

Citation suggérée

Andersson AF, Bissett A, Finstad AG, Fossøy F, Grosjean M, Hope M, Jeppesen TS, Kõljalg U, Lundin D, Nilsson RN, Prager M, Svenningsen C & Schigel D (2021) Publishing DNA-derived data through biodiversity data platforms. v1.0 Copenhagen: GBIF Secretariat. https://doi.org/10.35035/doc-vf1a-nr22.

Auteurs

- Kessy Kõljalg, 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, Department of Natural History, Centre for Biodiversity Dymamics, Norwegian University of Science and Technology, Trondheim, Norway
- 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, Norway
- Marie Grosjean, mgrosjean@gbif.org, Global Biodiversity Information Facility, Universitetsparken 15, 2100 København Ø, Denmark
- Michael Hope, Michael.Hope@ga.gov.au, Atlas of Living Australia, CSIRO National Collections & Marine Infrastructure, GPO Box 1700, Canberra ACT 2601, Australia.
- Thomas S. Jeppesen, tsjeppesen@gbif.org, Global Biodiversity Information Facility, Universitetsparken 15, 2100 København Ø, Denmark
- Urmas Kõljalg, urmas.koljalg@ut.ee, Natural History Museum and Botanical Garden, University of Tartu, 46 Vanemuise Street, 51003 Tartu, Estonia.
- Daniel Lundin, daniel.lundin@lnu.se, Centre for Ecology and Evolution in Microbial model Systems - EEMiS, Linnaeus University, SE-39182 Kalmar, Sweden
- R. Henrik Nilsson, henrik.nilsson@bioenv.gu.se, University of Gothenburg, Department of Biological and Environmental Sciences, Box 461, 405 30 Göteborg, Sweden
- Maria Prager, maria.prager@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 System, Jacobsenstraat 1, 8400 Oostende, Belgium
- Dmitry Schigel, dschigel@gbif.org, Global Biodiversity Information Facility, Universitetsparken 15, 2100 København Ø, Denmark
- Saara Suominen, s.suominen@unesco.org, Ocean Biodiversity Information System, Jacobsenstraat 1, 8400 Oostende, Belgium
- Cecilie Svenningsen, csvenningsen@gbif.org, Global Biodiversity Information Facility, Universitetsparken 15, 2100 København Ø, Denmark
- Tobias Guldberg Frøslev, tfroeslev@gbif.org, Global Biodiversity Information Facility,

Universitetsparken 15, 2100 København Ø, Denmark

Contributeurs

De précieuses discussions avec les membres des réseaux ELIXIR, iBOL, GGBN, GLOMICON et OBIS ont contribué à la compilation de ce projet. Nous sommes particulièrement reconnaissants pour les contributions et les encouragements d'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 qui ont contribué au processus de révision de la communauté GBIF.

Licence

Le document *Publier des données dérivées de l'ADN sur les plateformes de données sur la biodiversité* est sous licence Creative Commons Attribution-ShareAlike 4.0 Unported License.

URI permanent

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

Contrôle du document

Version 1.3.0 publiée sur 7 June 2023.

Cette version ajoute un paragraphe à propos des jeux de données marines et l' Ocean Biodiversity Information System (OBIS), ainsi que quelques modifications mineures du texte.

Résumé

Lorsque des informations génétiques sont utilisées pour décrire ou classer un taxon, la plupart des utilisateurs prévoient son utilisation dans le contexte de l'écologie moléculaire ou de la recherche phylogénétique. Il est important de se rendre compte qu'une séquence avec des coordonnées et une date/heure est une occurrence de biodiversité précieuse, qui est utile dans un contexte beaucoup plus large que son objectif initial. Pour réaliser ce potentiel, les données dérivées de l'ADN doivent être visibles sur les plateformes de données sur la biodiversité. Ce guide vous enseignera les principes et les approches afin d'exposer les « séquences avec dates et coordonnées » dans le contexte plus large des données sur la biodiversité. Le guide couvre les choix de schémas et de termes particuliers, les pièges communs et les bonnes pratiques, sans toutefois entrer dans les détails spécifiques à une plateforme. Il aidera toute personne intéressée par une meilleure exposition des données dérivées de l'ADN sur des plateformes générales de données sur la biodiversité, y compris les portails nationaux sur la biodiversité.

Préface

Les travaux sur ce guide ont commencé suite aux discussions à la conférence biodiversity_next de 2019 qui consolidaient les contributions de diverses sources, telles que:

- Rapport final du projet Plateforme Scientifique sur l'Avenir de l'Environnement

- ALA blog post enregistrements eDNA maintenant disponibles sur ALA
- ADN environnemental (eDNA) dans ALA
- Modèle de données eDNA ALA
- Critère norvégien pour le dépôt des échantillons et des données eDNA, y compris les spécimens étiquetés
- Données sur la biodiversité moléculaire dans SBDI, Suède
- Ressources du GBIF (Comment) puis-je publier des données moléculaires dérivées de séquences/ADN dans GBIF?
- Données moléculaires dans GBIF
- Guide rapide de publication des données et guides détaillés de publication des données du GBIF
- Comment publier des données dans GBIF, ainsi que l'aperçu des champs de l'extension DwC.
- Groupe d'intérêt sur la biodiversité génomique

1. Introduction

1.1. Justification

Les 20 dernières années ont permis de mieux comprendre l'immense pouvoir des méthodes moléculaires pour documenter la diversité de la vie sur terre. Des substrats apparemment sans vie et banals, tels que le sol et l'eau de mer, s'avèrent regorger de vie, même si ce n'est pas d'une manière que l'observateur occasionnel pourrait immédiatement apprécier. Des études basées sur l'ADN ont montré que des groupes d'organismes tels que les champignons, les insectes, les oomycètes, les bactéries et les archaea sont partout, même si nous ne pouvons souvent pas les observer physiquement (Debroas et al. 2017). Les avantages des méthodes moléculaires ne se limitent pas au monde microscopique : il existe de nombreux organismes, tels que certaines espèces de poissons, qui peuvent, au moins théoriquement, être observés physiquement, mais pour lesquels il est très coûteux, laborieux et peut-être invasif de le faire (Boussarie et al. 2018). Dans de telles situations, les données ADN nous permettent d'enregistrer la présence (et la présence passée) de ces organismes de manière non invasive et avec un minimum d'effort. Ces développements signifient qu'il n'est pas toujours nécessaire d'avoir des manifestations tangibles et physiques de tous les organismes présents sur un site donné pour les enregistrer. Tous les organismes, qu'ils soient ou non physiquement observables, peuvent être importants pour comprendre la biodiversité, l'écologie et la conservation biologique.

Les données dérivées de l'ADN nous permettent d'enregistrer des taxons peu visibles ou autrement inobservables qui passent sous le radar de protocoles reconnus pour le travail de terrain, les checklists, les dépôts dans les collections de sciences naturelles, etc. La maturité actuelle des méthodes d'analyse de l'ADN permet d'enregistrer la présence de ces organismes à un niveau de détail qui dépasse celui des observations macroscopiques des organismes en général. Toutefois, compte tenu du fait que les méthodes basées sur l'ADN s'accompagnent de leurs propres problèmes et biais, il est important de profiter de cette occasion pour définir et convenir de la manière dont nous devrions enregistrer et signaler la présence d'un organisme dans un substrat ou une localité donnée au moyen de données moléculaires. Cela permettra d'éviter les inefficacités importantes qui ont été signalées dans d'autres domaines, où l'absence de normes et d'orientations a conduit à des jeux de données très hétérogènes et largement incomparables (Berry et al. 2021 ; Leebens-Mack et al. 2006 ; Yilmaz et al. 2011 ; Nilsson et al. 2012 ; Shea et al. 2023). En outre, une documentation claire du traitement informatique, depuis la lecture de séquences brutes jusqu'à l'observation des espèces déduites, permettra de procéder à une nouvelle analyse lorsque des méthodes améliorées apparaîtront.

Les données d'occurrence des espèces dérivées de l'ADN devraient être aussi normalisées et reproductibles que possible, que les espèces détectées aient ou non des noms scientifiques formels. Dans certains cas, ces relevés d'occurrences indiqueront des propriétés géographiques et écologiques des espèces décrites précédemment inconnues, enrichissant ainsi notre corpus de connaissances sur ces taxons. Dans d'autres cas, les données peuvent nous permettre de fusionner et de visualiser des informations sur les espèces actuellement non décrites, ce qui peut éventuellement accélérer leur description formelle. La capacité de collecter des données utilisables, même pour les espèces sans nom, ajoute de manière significative aux nombreuses façons dont le GBIF et d'autres plateformes de données sur la biodiversité indexent le monde vivant et rendent ces connaissances disponibles à tous et à des fins diverses, y compris la conservation de la biodiversité. Selon des estimations récentes, au moins 85 % de toutes les espèces existantes ne sont pas décrites (Mora et al. 2011; Tedesco et al. 2014). Les standards de données existants ont été conçus pour la minorité de taxons décrits. Les bonnes pratiques pour traiter les données dérivées de l'ADN aideront à caractériser les occurrences de tous les organismes, qu'ils soient décrits ou non.

Ce guide explique comment les données d'occurrence dérivées de l'ADN doivent être rapportées pour être intégrées dans GBIF et dans d'autres plateformes de données sur la biodiversité. Il n'exprime aucune opinion sur la question de l'accès et du partage des bénéfices pour l'information sur les séquences numériques, qui a fait l'objet de discussions approfondies dans le cadre de Convention sur la diversité biologique (CBD). Toutefois, il convient de noter que les codes-barres génétiques et les métacodes-barres sont généralement des gènes ou des fragments d'ADN non codant, qui ne se prêtent pas à l'exploitation commerciale. Comme l'archivage des séquences via Collaboration internationale de base de données sur la séquence des nucléotides (INDSC) est une norme répandue dans la recherche basée sur le séquençage de l'ADN, la publication de données d'occurrence issues de séquences n'implique pas la publication de nouvelles séquences. Dans la plupart des cas, celles-ci ont déjà été placées dans un répositoire génétique public. Ce guide aborde donc la valeur ajoutée possible de la dérivation des données spatio-temporelles d'occurrence et des noms basés sur l'ADN plutôt que la valeur de l'information génétique elle-même. En plus de traiter les données dérivées des séquences ADN, ce guide contient également des suggestions pour la publication de données sur les occurrences d'espèces dérivées d'analyses qPCR ou (d)PCR.

Signaler les occurrences dérivées de l'ADN de manière ouverte et reproductible apporte de nombreux avantages : notamment, cela accroit la citabilité, met en évidence les taxons concernés dans le contexte de la conservation biologique et contribue aux connaissances taxonomiques et écologiques. De plus, ça fournit également un mécanisme pour stocker les occurrences d'espèces non décrites. Quand ce taxon, qui n'est pas encore décrit, est enfin lié à un nouveau nom linnéen, tous les enregistrements d'occurrences qui lui sont liés seront immédiatement disponibles. Chacun de ces avantages justifie fortement l'adoption par les professionnels des pratiques décrites dans ce guide, qui les aideront à mettre en évidence une part importante de la biodiversité existante, à accélérer sa découverte et à l'intégrer dans la conservation biologique et l'élaboration des politiques.

1.2. Public cible

Ce guide a été élaboré à l'intention de plusieurs publics cibles : les étudiants qui planifient une première étude basée sur l'ADN, les chercheurs qui possèdent d'anciennes séquences et d'anciens tableaux d'abondance qu'ils souhaitent faire revivre ou préserver, les spécialistes des données sur la biodiversité qui s'initient aux occurrences dérivées de l'ADN, et les bioinformaticiens qui sont familiers avec les séquences ADN, mais qui ne connaissent pas les plateformes de données sur la biodiversité. Le guide ne s'adresse pas directement aux utilisateurs des données moléculaires dans les plateformes de données sur la biodiversité, mais ces utilisateurs pourront trouver un intérêt particulier à la section 1.7 en Résultats sur la sortie des données. L'intention des auteurs est de conseiller et d'instruire sur la publication des données et des attributs associés aux séquences génétiques par le biais de plateformes générales de données sur la biodiversité.

Le diagramme décrit les étapes de traitement nécessaires à la publication de données de biodiversité moléculaire dans des référentiels tels que le GBIF et les plateformes nationales de données sur la biodiversité, y compris celles construites sur la plateforme ALA. Ce guide se concentre principalement sur les étapes qui suivent l'acquisition des séquences brutes FASTQ issues de l'étape de séquençage. En se familiarisant avec le diagramme — et en notant toute étape qui semble familière ou peu claire — les utilisateurs seront en mesure de voguer dans les contenus de ce guide.

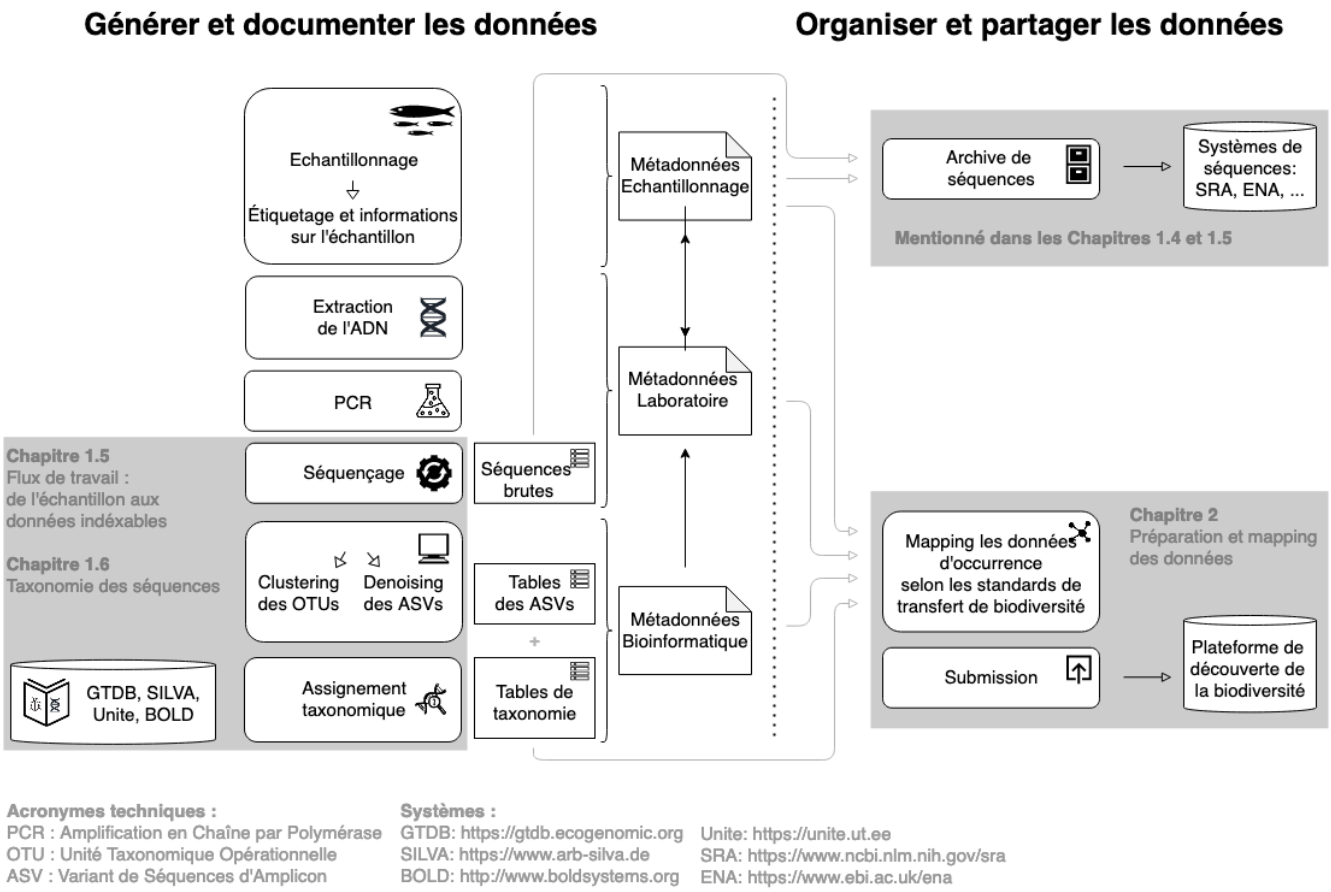

Figure 1. Flux de travail global pour les données sur la biodiversité dérivées de séquences ADN, tel que décrit dans ce guide.

Les auteurs se sont efforcés de rendre les informations de ce guide utiles à chacun des publics décrits ci-dessus, mais des lectures plus approfondies (par exemple GBIF quick guide to data publishing) peuvent être nécessaires dans certains cas.

1.3. Introduction aux données d'occurrence dérivées de l'ADN

Les données d'occurrence biologique dérivées de l'ADN comprennent des informations dérivées de l'ADN d'organismes individuels, mais aussi de l'ADN environnemental (eDNA, c'est-à-dire l'ADN extrait d'échantillons environnementaux, (Thomsen & Willerslev 2015) et d'échantillons en vrac comprenant de nombreux individus (par exemple, des échantillons de plancton ou des échantillons de pièges Malaise constitués de plusieurs individus de nombreuses espèces). Actuellement, le plus grand volume de données d'occurrence dérivées de l'ADN provient de l'ADN environnemental. Étant donné que les méthodes d'analyse et les produits finaux sont largement similaires pour toutes les sources d'échantillons, la discussion ci-dessous se concentrera sur l'ADN environnemental ([catégorie-i] et [catégorie-ii]), tout en notant que les grandes lignes sont applicables aux autres sources. Les études utilisent souvent le séquençage ciblé de marqueurs génétiques informatifs sur le plan taxonomique et phylogénétique, mais peuvent également utiliser, par exemple, des approches basées sur la qPCR qui n'aboutissent pas directement à des données de séquence d'ADN ([catégorie-iii] et [mapping-ddpcr-

qpcr-data]). Ce guide peut sembler lourd en termes liés à l'ADN ; si c'est le cas, veuillez consulter le Glossaire.

1.3.1. L'ADN environnemental comme source de données d'occurrence

Le terme ADN environnemental est utilisé depuis 1987, lorsqu'il a été utilisé pour la première fois pour décrire l'ADN de microbes dans des échantillons de sédiments (https://doi.org/10.1016/0167-7012(87)90025-x [Ogram et al. 1987^]). L'eDNA est maintenant utilisé plus largement pour décrire un mélange complexe d'ADN provenant de différents organismes (https://doi.org/10.1093/oso/9780198767220.001.0001 [Taberlet et al. 2018^] et https://doi.org/10.1111/j.1365-294X.2012.05542.x [2012^]). Ainsi, l'eDNA comprend tout l'ADN extrait d'un échantillon environnemental spécifique, indépendamment du substrat et des espèces qu'il contient. Il peut être extrait d'un large éventail de sources, y compris les cellules de la peau et des cheveux, la salive, le sol, les fèces et les organismes vivants ou récemment morts (Pietramellara et al. 2009). Souvent, l'ADN environnemental représente suffisamment tous les organismes d'un échantillon donné. Dans la pratique, cependant, la présence d'ADN dans l'échantillon environnemental dépend de la sélection de l'habitat, de la taille du corps, de la morphologie et du niveau d'activité de l'organisme. En outre, les méthodes d'échantillonnage utilisées pour capturer l'ADN (Taberlet et al. 2018) et le stade de dégradation de celui-ci peuvent influer sa détection.

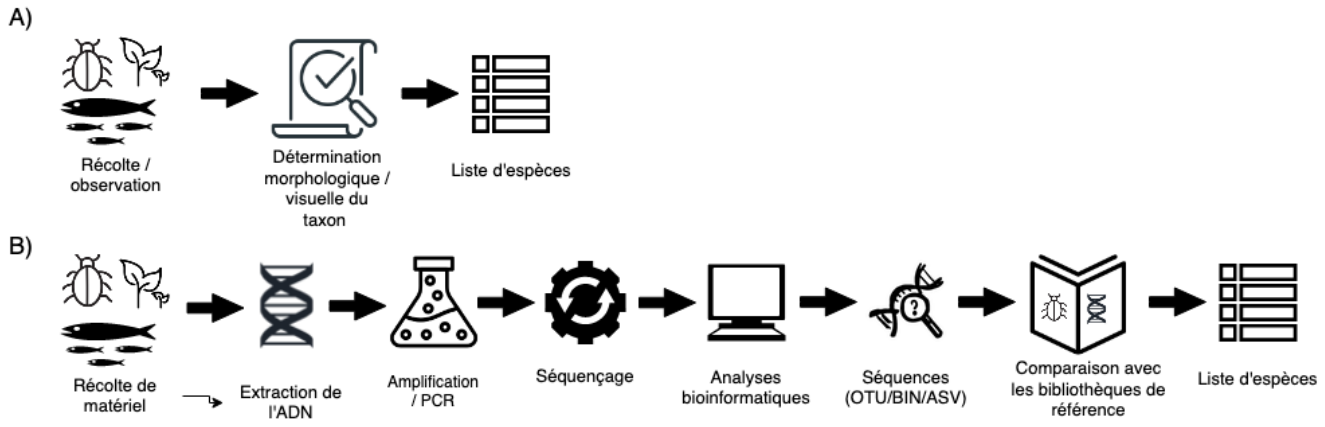

Figure 2. Représentation simplifiée des processus d'échantillonnage comparant la récolte de données par A) les méthodes traditionnelles d'échantillonnage utilisées en écologie/biodiversité, et B) les études basées sur l'eDNA, illustrées ici par le métabarcoding. Pour l'eDNA, la plupart des étapes jusqu'au séquençage impliquent des réplications techniques ou biologiques, qui permettent d'identifier les contaminations et les faux positifs ainsi que les faux négatifs, résultant en une structure hiérarchique des données et des métadonnées. Cependant, les études comprendront souvent les deux types d'échantillonnage. Par exemple, si la "Bibliothèque de référence" utilisée en B) ne contient pas toutes les espèces pertinentes d'un groupe donné d'organismes, il sera nécessaire de revenir à A). Il se peut également que la "résolution par rapport à la bibliothèque de référence" produise des résultats inattendus ou improbables, auquel cas d'autres études utilisant la méthodologie traditionnelle seront nécessaires pour déterminer si les espèces identifiées par l'analyse bioinformatique peuvent être vérifiées.

L'eDNA est donc un type d'échantillon, et non une méthode, qui inclut l'ADN dérivé de tout échantillon environnemental plutôt que de la capture et séquençage d'un individu ciblé. Ces types d'échantillons comprennent l'eau, le sol, les sédiments et l'air, mais aussi les échantillons de contenu intestinal et les tissus (végétaux/animaux) où l'ADN de l'hôte n'est pas la cible (Taberlet et al. 2018). Un certain nombre de méthodes analytiques existent pour étudier l'eDNA. Elles peuvent être divisées en deux grandes catégories : 1) celles qui visent à détecter un organisme spécifique et 2) celles qui décrivent un assemblage ou une communauté de nombreux organismes. Les différentes méthodes d'analyse génèrent différents types et volumes de données. Le plus souvent, les concentrations d'ADN sont faibles et des réplications techniques et biologiques doivent être utilisées afin de valider la présence/absence des espèces.

Plusieurs études montrent que, pour les échantillons d'eau, les analyses basées sur l'eDNA peuvent avoir une meilleure probabilité de trouver des espèces rares et difficiles à étudier que les méthodes conventionnelles (Thomsen et al. 2012; Biggs et al. 2015; Valentini et al. 2016; Bessey et al. 2020). Il peut en être de même pour d'autres environnements, où des traces d'ADN peuvent encore être détectées, bien que l'organisme réel n'y soit plus présent. Par conséquent, l'eDNA peut convenir à la surveillance des espèces rares figurant sur les listes rouges, ainsi que des espèces exotiques indésirables, qui sont souvent présentes en faibles densités, rendant la détection avec les méthodes conventionnelles particulièrement difficile. Les méthodes d'analyse de l'eDNA permettent de détecter des organismes cryptiques, notamment ceux de petite taille, qui ne peuvent pas être détectés à l'œil nu (bactéries et champignons, par exemple). En outre, l'eDNA peut également être utilisé pour l'observation de nombreuses espèces simultanément, et peut décrire des communautés biologiques entières ou des composants majeurs de celles-ci (Ekrem & Majaneva 2019).

Certaines études montrent une relation entre la quantité d'ADN d'une espèce donnée dans un échantillon environnemental et la biomasse de l'espèce dans l'environnement. On peut donc éventuellement considérer que l'ADN environnemental permet une estimation semi-quantitative (cible indirecte) de la biomasse des organismes, tant à partir d'échantillons environnementaux que d'échantillons en vrac (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). Cependant, d'autres études montrent peu de corrélation entre la quantité d'ADN environnemental et la densité de population estimée (Knudsen et al. 2019). La PCR, la quantification, la préparation et d'autres biais sont fréquemment débattus. Par exemple, la mue, la reproduction et la mort massive peuvent contribuer à augmenter les niveaux d'ADN environnemental des crustacés dans l'eau, tandis que la turbidité et la mauvaise qualité de l'eau réduisent la quantité d'ADN environnemental détectable (Strand et al. 2019). Par conséquent, nous encourageons les éditeurs de données à fournir à la fois le nombre de reads pour chaque OTU ou ASV par échantillon, ainsi que le nombre total de reads par échantillon, car il s'agit d'informations nécessaires pour que les utilisateurs puissent tirer leurs propres conclusions sur la présence/absence et l'abondance (relative).

1.3.2. Métabarcoding : données dérivées de séquences

La génération de données dérivées de séquences ADN augmente rapidement en raison du développement du du métabarcoding de l'ADN. Cette méthode utilise des amorces universelles pour générer des milliers, voire des millions, de courtes séquences d'ADN pour un groupe donné d'organismes à l'aide du séquençage à haut débit (HTS, alt. next-generation sequencing (NGS)). En comparant chaque séquence d'ADN à une base de données de référence telle que GenBank (Benson et al. 2006), BOLD (Ratnasingham et al. 2007) ou UNITE (Nilsson et al. 2019), chaque séquence peut être attribuée à une espèce ou à une identité taxonomique de rang supérieur. Le DNA-metabarcoding est utilisé pour des échantillons provenant d'environnements terrestres et aquatiques, y compris l'eau, le sol, l'air, les sédiments, les biofilms, le plancton, les échantillons en vrac et les fèces, identifiant simultanément des centaines d'espèces (Ruppert et al. 2019).

L'identification et la classification des organismes à partir de données de séquences et d'études basées sur des marqueurs ADN dépendent de l'accès à une bibliothèque de référence de séquences provenant de spécimens morphologiquement identifiés, qui sont comparées aux séquences nouvellement générées. L'efficacité de la classification dépend de l'exhaustivité (couverture) et de la fiabilité des bibliothèques de référence, ainsi que des outils utilisés pour effectuer la classification. Il s'agit là d'éléments en constante évolution, ce qui rend indispensable l'application d'une expertise taxonomique et la prudence dans l'interprétation des résultats ([taxonomie des séquences]). La disponibilité de tous les variants de séquences d'amplicons vérifiés (Callahan et al. 2017) permet une réinterprétation précise des données, des analyses génétiques des populations intraspécifiques (Sigsgaard et al. 2019) et est susceptible d'accroître la précision de l'identification, et c'est pourquoi nous recommandons de partager les données ASV (non regroupées).

1.3.3. Métagénomique : données dérivées de séquences ADN

Les données de biodiversité dérivées de séquences ADN peuvent également être générées en utilisant des méthodes métagénomiques sans amplification, par lesquelles tout l'ADN d'un échantillon est ciblé pour le séquençage (Tyson & Hugenholtz 2005), plutôt que des amplicons ou des codes-barres spécifiques, comme décrit ci-dessus. Les données de biodiversité dérivées de séquences ADN obtenues à partir du séquençage métagénomique peuvent se présenter sous la forme de correspondances de séquences avec des bases de données de gènes annotés (comme ci-dessus) ou en tant que génomes assemblés (MAGs) (presque) complets. Alors que les méthodes de métabarcoding dominent toujours en termes d'informations sur la biodiversité dérivée de séquences ADN, les données métagénomiques prennent de plus en plus d'importance, comme en témoigne le nombre croissant de MAGS et leur utilité pour améliorer la phylogénie et la taxonomie (Parks et al. 2020); la discussion sur les méthodes associées à l'analyse du métagénome, qui évoluent actuellement de manière très rapide, dépasse le cadre du présent document. Ce document utilise le métabarcoding comme modèle de discussion autour des concepts et des méthodes de publication des données sur la biodiversité dérivées de séquences ADN, et bien que les voies bioinformatiques soient différentes pour les données métagénomiques, le résultat final (une séquence, souvent sous la forme d'un contig/assemblage) est conforme aux concepts suggérés pour les données de métabarcoding (c'est-à-dire que les métadonnées du flux de travail spécifiques à l'échantillon, à la récolte de l'échantillon, à la génération de données et au traitement doivent être saisies).

1.3.4. qPCR / (d)dPCR : données d'occurrence

Pour la détection ciblée d'espèces dans les échantillons d'eDNA, la plupart des analyses utilisent des amorces spécifiques aux espèces, et la qPCR (amplification en chaîne quantitative par polymérase) ou la dPCR (amplification en chaîne numérique par polymérase). Ces méthodes ne génèrent pas de séquences ADN, et les données d'occurrence dépendent entièrement de la spécificité des amorces/essais. Par conséquent, il y a des recommandations strictes pour valider ces tests et des exigences pour la publication des données (Bustin et al. 2009, Huggett et al. 2013), ainsi que pour la préparation de ces tests pour la surveillance de routine (Thalinger et al. 2020). L'analyse d'échantillons d'eDNA utilisant la qPCR nécessite peu de ressources et peut être réalisée dans la plupart des laboratoires d'analyse de l'ADN. Le premier exemple d'utilisation d'échantillons d'eau contenant de l'eDNA a utilisé la qPCR pour détecter la grenouille américaine envahissante (*Rana catesbeiana*) (Ficetola et al. 2008), et les analyses par qPCR de l'eDNA d'échantillons d'eau sont régulièrement utilisées pour détecter des espèces ciblées de poissons, d'amphibiens, de mollusques, de crustacés et autres, ainsi que leurs parasites (Hernandez et al. 2020, Wacker et al. 2019, Fossøy et al. 2019, Wittwer et al. 2019). Les détections d'eDNA à l'aide de la qPCR génèrent donc d'importantes données isolées d'occurrence des espèces.

1.4. Introduction à la publication de données sur la biodiversité

La publication des données sur la biodiversité consiste en grande partie à rendre les données d'occurrence des espèces identifiables, accessibles, interopérables et réutilisables, conformément aux principes FAIR (Wilkinson et al. 2016). Les plateformes de données sur la biodiversité aident à exposer et à découvrir les données de séquences ADN, en tant que registres d'occurrence de la biodiversité en parallèle avec d'autres types de données sur la biodiversité, tels que les spécimens de collections de musées, les observations issues de la science citoyenne et les études classiques de terrain. La structure, la gestion et le stockage de chaque source originale de données varient en fonction des besoins de chaque communauté. Les plateformes de données sur la biodiversité favorisent la découverte, l'accès et la réutilisation des données en rendant ces ensembles de données compatibles entre eux, et en palliant aux incohérences taxonomiques, spatiales et autres dans les données disponibles sur la biodiversité. Les points d'accès uniques qui mettent les données à

disposition favorisent la recherche, la gestion et la politique à grande échelle. La compatibilité entre les jeux de données est obtenue grâce au processus de normalisation.

Un certain nombre de standards de données sont utilisés pour les données générales sur la biodiversité (https://www.gbif.org/standards) et un ensemble distinct de standards sont utilisés pour les données sur les séquences génétiques (voir MIxS et GGBN). Ce guide reflète les efforts en cours pour améliorer la compatibilité entre les standards relatifs aux données générales sur la biodiversité et aux données génétiques. Les standards mettent souvent en évidence les sous-ensembles de champs les plus importants ou les plus fréquemment applicables. Ces sous-ensembles peuvent être appelés "cores". Le format préféré pour la publication des données dans les réseaux GBIF et ALA est actuellement le Darwin Core Archive (DwC-A) qui utilise le standard de données Darwin Core (DwC). En pratique, il s'agit d'un dossier compressé (un fichier zip) contenant des fichiers de données, dans un format texte standard délimité par des virgules ou des tabulations, un fichier de métadonnées (eml.xml) qui décrit la ressource de données, et un métafichier (meta.xml) qui spécifie la structure des fichiers et des champs de données inclus dans l'archive. La préparation normalisée garantit que les données peuvent circuler entre les systèmes en utilisant des protocoles d'échange de données spécifiques. La Section 2 de ce guide fournit des recommandations pour le mapping des fichiers de données, tandis que des lignes directrices et des outils pour la construction des fichiers xml peuvent être trouvés ici : TDWG, GBIF, et ALA.

Un élément central du processus de normalisation est le mapping des informations, qui est nécessaire pour transformer la structure originale des informations (colonnes) d'un export de données source en une structure standardisée des informations. La normalisation peut également affecter les informations contenues dans chaque enregistrement, par exemple en recalculant les coordonnées selon un système commun, en réorganisant les éléments de date ou en faisant correspondre le contenu des champs à un ensemble standard de valeurs, souvent appelé vocabulaire. Le processus de normalisation offre également la possibilité d'améliorer la qualité des données, par exemple en comblant les omissions, en corrigeant les fautes de frappe et les espaces inutiles et en gérant l'utilisation incohérente des informations. De telles améliorations rehaussent la qualité des données et augmentent leur aptitude à la réutilisation, mais quoi qu'il en soit, des données publiées dans n'importe quel état sont meilleures que des données qui restent non publiées et inaccessibles. La normalisation est généralement appliquée à une copie ou à un export des données source, laissant l'original intact.

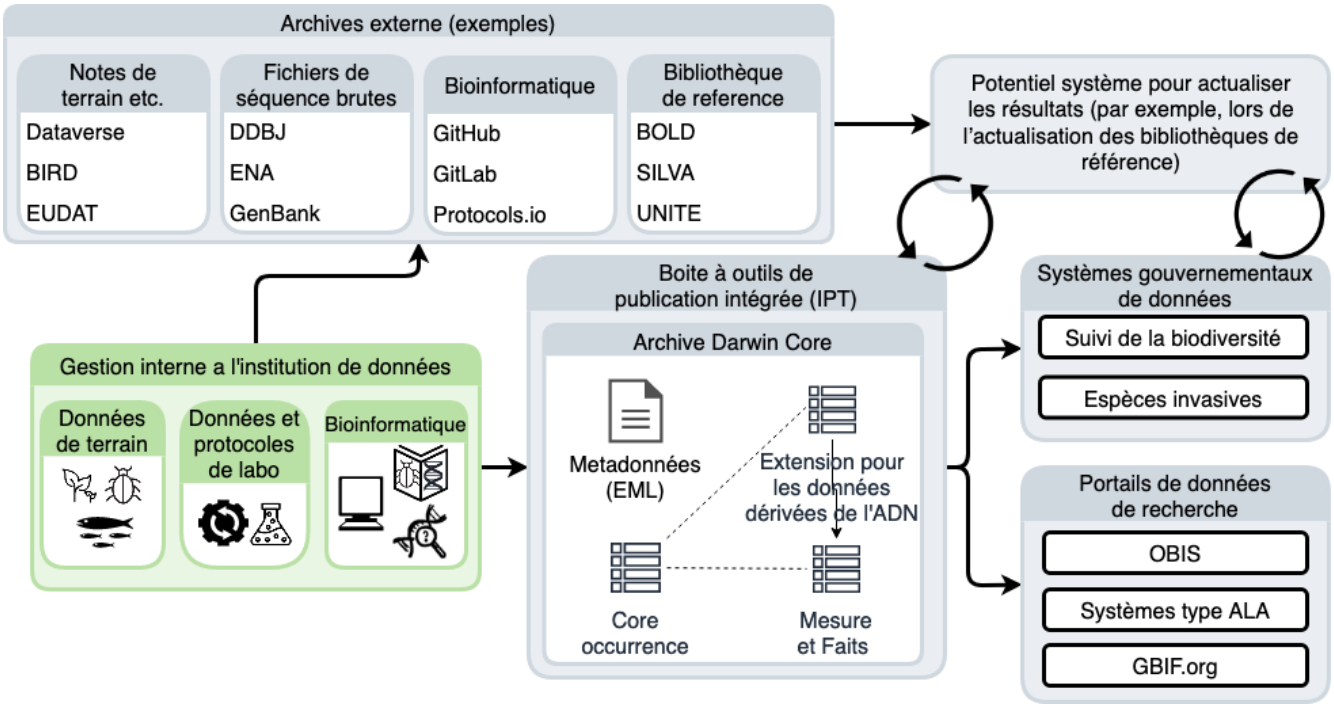

Figure 3. Esquisse d'une plateforme pour le signalement et la publication de séquences ADN et des métadonnées associées (boîte verte) basée sur les systèmes existants et les normes de données (boîtes grises). Un système envisagé pour la mise à jour régulière (basé sur la lecture automatique des données) des résultats (boîte blanche) peut lire et mettre à jour les informations soit de l'Archive Darwin Core, soit de divers systèmes d'administration. Le transfert de données entre les différents éléments (flèches noires) nécessite différents degrés de transformation et d'harmonisation des informations et peut inclure une évaluation automatique ou humaine de la qualité.

Une fois qu'un jeu de données a été soumis aux processus de normalisation et d'amélioration de la qualité des données, il doit être placé à un endroit accessible en ligne et associé à des métadonnées pertinentes. Les métadonnées – données ou informations sur le jeu de données – comprennent des paramètres clés qui décrivent le jeu de données et améliorent encore son accessibilité et sa réutilisation. Les métadonnées devraient inclure d'autres éléments importants tels que les auteurs, les identifiants d'objets numériques (DOI), les affiliations institutionnelles et d'autres informations sur la provenance des données, ainsi que des informations sur les procédures et méthodes liés au traitement du jeu de données. Nous encourageons à ce qu'une description des détails et des versions du flux de travail, y compris des contrôles de qualité, soit fournie dans la section méthodes du fichier EML.

Les jeux de données et les métadonnées associées sont indexés par chaque portail de données : ce processus permet aux utilisateurs d'interroger, de filtrer et de traiter les données à travers les API et les portails web. Contrairement aux publications scientifiques, les jeux de données peuvent être des produits dynamiques qui passent par de multiples versions, avec un nombre évolutif d'enregistrements et de métadonnées remplaçables sous le même titre et le même DOI.

Il convient de noter que les détenteurs de données de séquences génétiques sont censés les soumettre et les déposer dans des archives de données de séquences brutes tels que le SRA, EMBL's ENA ou le DDBJ. Le sujet à propos de l'archivage des séquences n'est pas abordé ici, mais à titre d'exemple, Penev et al. (2017) donnent un aperçu général de l'importance de la soumission des données et des directives en lien avec la publication scientifique. Les plateformes de données sur la biodiversité telles que l'ALA, le GBIF et la plupart des portails nationaux sur la biodiversité ne sont pas des archives pour les reads de séquences brutes et les fichiers associés. Nous soulignons toutefois l'importance de maintenir des liens entre ces données primaires et les occurrences dérivées dans la Section 2.

1.5. Flux de travail : de l'échantillon aux données indexables

Les données de métabarcoding peuvent être produites à partir de différentes plateformes de séquençage (Illumina, PacBio, Oxford Nanopore, Ion Torrent, etc.), qui s'appuient sur différents principes pour la lecture et la génération de données qui se distinguent en ce qui concerne la longueur des séquences et le profil d'erreur, que les séquences soient simples ou à double-sens, etc. Actuellement, la plateforme Illumina à lecture courte est la plus largement adoptée et, en tant que telle, est à la base des descriptions ici. Cependant, le traitement bioinformatique des données suit les mêmes principes généraux (contrôle de la qualité, suppression du bruit - denoising, classification) indépendamment de la technologie de séquençage utilisée (Hugerth et al. 2017, Figure 2).

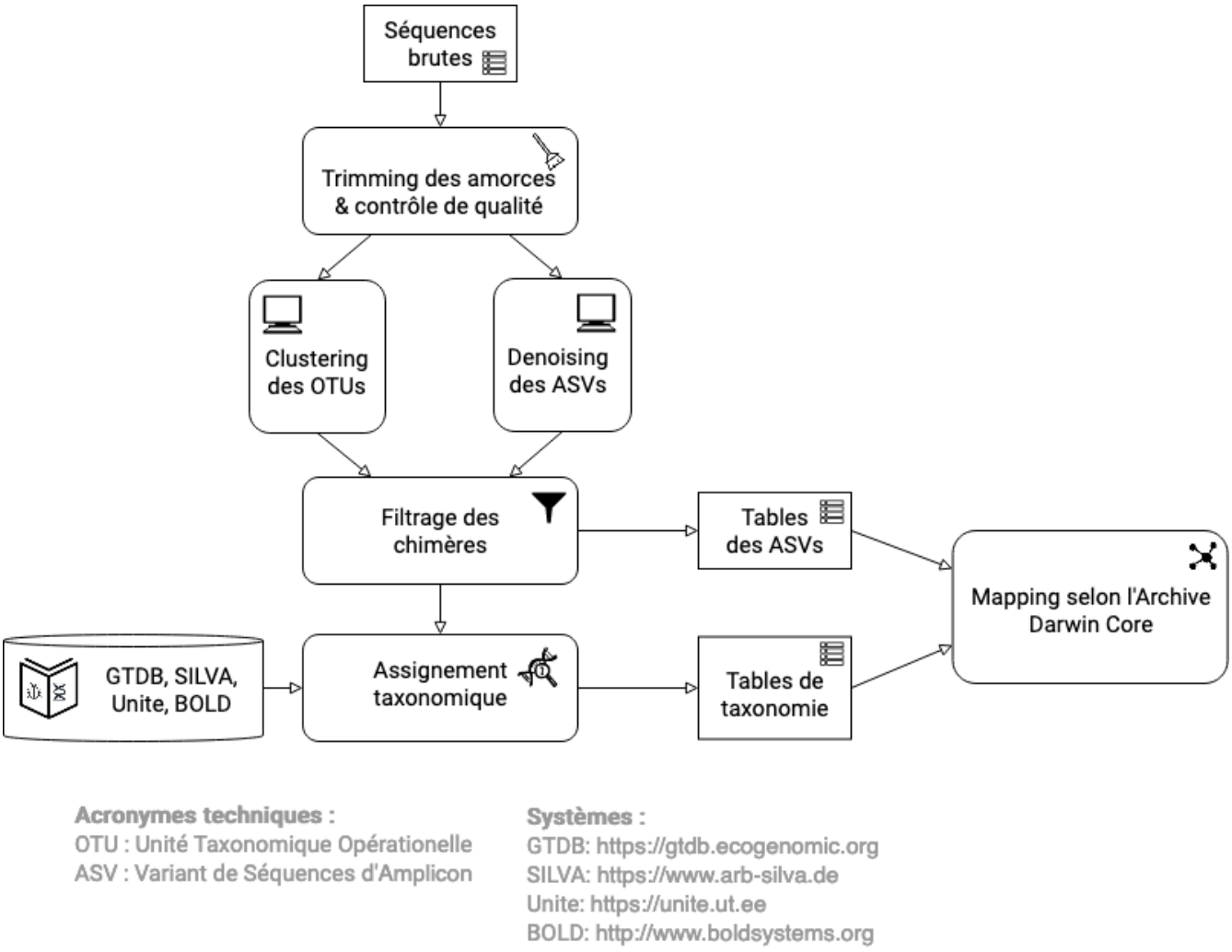

Figure 4. Aperçu du traitement bioinformatique des données de métabarcoding.

Généralement, les séquences ADN sont d'abord prétraitées en supprimant les séquences des amorces (triming) et, en fonction de la méthode de séquençage utilisée, les bases de faible qualité, généralement vers les extrémités 5' et 3' de la séquence. Les séquences qui ne satisfont pas aux exigences de longueur, de qualité globale, de présence d'amorces, d'étiquettes, etc. sont enlevées.

Les séquences prétraitées peuvent ensuite être attribuées à un taxon en les comparant à des bases de données de référence. Lorsque les bases de données de référence sont incomplètes, la classification des séquences peut se faire sans identification taxonomique, soit en regroupant les séquences en unités taxonomiques opérationnelles sur la base de leur similarité (OTU ; Blaxter et al. 2005), soit en faisant un denoising des données, c'est-à-dire en détectant et en excluant explicitement les séquences comportant des erreurs de PCR/séquençage pour produire des variants de séquence d'amplicon (ASV ; également appelés OTU à rayon zéro (zOTU)). Le denoising tente de

corriger les erreurs qui ont été introduites dans les étapes de PCR et/ou de séquençage, de sorte que les séquences représentent l'ensemble des séquences uniques biologiquement réelles présentes dans l'ensemble de séquences d'origine. Dans le cas de séquençage double-sens, le denoising des séquences directe et inverse peut être fait séparément avant d'être assemblées ou les séquences peuvent être assemblées avant le denoising. Les ASV de l'ensemble résultant peuvent différer d'aussi peu qu'une base, ce qui indique une variation de séquence inter- ou intraspécifique. D'un point de vue opérationnel, les ASV peuvent être considérés comme des OTU sans rayon défini et, bien que les algorithmes de denoising soient généralement très bons, ils n'éliminent pas entièrement les problèmes liés au fractionnement ou agrégation excessif des séquences.

La PCR utilisée pour générer la librairie de séquençage peut entraîner la génération de séquences artefactuelles sous la forme de chimères, c'est-à-dire une séquence unique issue de plusieurs séquences parentales. Ces séquences peuvent être détectées de manière bioinformatique et supprimées, ce qui est généralement fait après le denoising.

Enfin, les séquences prétraitées, OTUs ou ASVs, sont classées taxonomiquement en les comparant à une base de données de séquences annotées (souvent appelées bibliothèques de références, voir § 1.6). Comme pour les étapes précédentes, plusieurs méthodes alternatives sont disponibles. La plupart d'entre elles sont basés soit sur l'alignement des séquences de métabarcoding sur les séquences de référence, soit sur le nombre de k-mers partagées (séquences courtes exactes).

Plusieurs outils et algorithmes open source existent pour le traitement bioinformatique des données de métabarcoding (QIIME2 (Bolyen et al. 2019), DADA2 (Callahan et al. 2016), SWARM https://doi.org/10.7717/peerj.593 [(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)). Étant donnée l'existence de nombreux flux de travail populaires et bien utilisés, nous formulons ci-dessous quelques recommandations sur l'analyse des données en vue de leur soumission aux plateformes de données sur la biodiversité. Il ne s'agit pas de suggérer que ce sont les meilleures méthodes ou qu'elles sont les plus appropriées à toute fin, mais d'encourager la soumission de données relativement standardisées qui peuvent être facilement comparées par le biais des plateformes. Si possible, il convient d'utiliser un flux de travail bien documenté et mis à jour (par exemple, nf-core/ampliseq pipeline). Les métadonnées doivent inclure les détails et les versions du flux de travail, soit dans les étapes de la méthode des métadonnées, soit en tant que référence dans le champ SOP de l'extension des données dérivées de l'ADN (voir la correspondance dans le tableau 4). Les données de séquence doivent être déposées dans une archive de nucléotides appropriée (NCBI's SRA : Leinonen et al. 2011) ou EMBL's ENA (Amid et al. 2020)) et les données soumises à la plateforme de biodiversité doivent inclure l'ID de l'échantillon biologique obtenu à partir de l'archive (voir le mapping des données dans [data-mapping]). L'utilisation de ces identifiants d'échantillons réduira les risques de duplication et garantira que les données de séquences ADN soient facilement accessibles si des opportunités de réanalyse se présentent, à mesure que les bibliothèques de référence et les outils bioinformatiques s'améliorent. Le principal produit final de ces pipelines est généralement un fichier contenant le nombre d'OTUs ou d'ASVs individuels dans chaque échantillon, ainsi que la taxonomie qui leur a été attribuée. Ce fichier est généré soit sous forme de tableau, soit sous forme de BIOM (McDonald et.al 2012). Les séquences d'OTU ou d'ASV sont également souvent fournies au format FASTA (Pearson & Lipman 1988).

1.6. Taxonomie des séquences

L'annotation taxonomique des séquences est une étape critique dans le traitement des jeux de données de biodiversité moléculaire, car les noms scientifiques sont essentiels pour accéder et communiquer des informations sur les organismes observés. La précision et l'exactitude de cette annotation de séquences dépendront de la disponibilité de bases de données de référence et de bibliothèques fiables à travers toutes les branches de l'arbre de vie, qui, à son tour, nécessitera des efforts conjoints des taxonomistes et des écologistes moléculaires. Les bases de données de

séquences publiques devraient toujours être utilisées en prenant conscience du fait qu'elles souffrent de diverses lacunes, par ex. la fiabilité taxonomique et le manque de vocabulaires normalisés de métadonnées (Hofstetter et al. 2019; Durkin et al. 2020).

Les espèces, telles que décrites par les taxonomistes, sont primordiales en biologie et les tentatives de caractérisation de la biodiversité peuvent donc utiliser les résultats de la recherche taxonomique. Cependant, contrairement aux données de séquences d'ADN, les résultats taxonomiques ne sont pas toujours exploitables par des algorithmes directs ou des interprétations informatiques : la taxonomie classique est un processus dirigé par l'homme qui comprend les étapes manuelles de la délimitation des taxons, de description et de désignation, aboutissant à une publication formelle conforme aux Codes internationaux de Nomenclature. Comme discuté dans les chapitres précédents, les analyses basées sur les séquences d'ADN sont très efficaces pour détecter les espèces difficiles à observer et identifieront souvent la présence d'organismes actuellement en dehors des connaissances taxonomiques linnéennes traditionnelles. Bien que ces lignes directrices n'abordent pas la publication de listes d'espèces alternatives dérivées de données de séquences, la déconnexion entre la taxonomie traditionnelle et les efforts basés sur le eDNA n'est pas souhaitable. En conséquence, nous proposons aux lecteurs de ce guide les recommandations suivantes.

La taxonomie étant au cœur de la découverte de données sur la biodiversité, il est fortement recommandé que les efforts de séquençage eDNA cherche à inclure l'expertise taxonomique pertinente pour l'étude en question. Il serait également bénéfique que les projets de séquençage d'eDNA puissent allouer une partie de leur budget à la génération et à la publication de séquences de référence à partir de spécimens types non séquencés ou d'autres éléments de référence importants à partir des herbiers, musées ou collections biologiques locales. Les taxonomistes peuvent également contribuer à cet objectif en incluant toujours des séquences d'ADN pertinentes avec chaque nouvelle description d'espèce (Miralles et al. 2020) et en ciblant les nombreuses entités biologiques découvertes par les efforts d'eDNA (par exemple Tedersoo et al. 2017).

La plupart des plateformes actuelles de données sur la biodiversité sont organisées autour de listes de noms et d'index taxonomiques traditionnels. Étant donné que les occurrences dérivées de séquences ADN deviennent rapidement une source importante de données sur la biodiversité, et comme la taxonomie et la nomenclature officielles pour ce type de données manquent, il est recommandé que les fournisseurs et les plateformes de données continuent d'explorer et d'inclure des représentations plus souples de la taxonomie dans leur squelette taxonomique. Ces nouvelles représentations comprennent des bases de référence de données moléculaires (par exemple, GTDB, BOLD, UNITE) qui reconnaissent les données de séquences comme matériel de référence pour les organismes non classifiés précédemment. En outre, nous suggérons que d'autres bases de données moléculaires couramment utilisées (par ex. PR2, RDP, SILVA) développent des identifiants stables pour les taxa et rendent disponibles les séquences de référence pour ces taxa, afin de permettre leur utilisation comme références taxonomiques.

Contrairement à la taxonomie classique, qui est un processus fortement manuel, le regroupement des séquences d'ADN en concepts taxonomiques repose sur l'analyse algorithmique de la similarité et d'autres signaux (tels que la phylogénie et la probabilité), ainsi que sur une certaine édition humaine. Les OTU qui en résultent varient en termes de stabilité, de présence de séquences de référence et de matériel physique, d'alignements et de valeurs de cut-off, ainsi que d'identifiants d'OTU tels que les DOI (Nilsson et al. 2019). Plus important encore, elles varient en termes d'échelle, des bibliothèques locales spécifiques à une étude ou à un projet aux bases de données mondiales qui permettent une comparaison plus large entre les études. Contrairement à la centralisation et à la codification des taxons linnéens qui sont formellement décrits dans les publications de recherche, les OTU sont répartis dans de multiples bibliothèques de référence numériques évolutives qui diffèrent par leur focus taxonomique, leurs gènes de code-barres et d'autres facteurs. En associant des séquences standard à des spécimens de référence identifiés, BOLD et UNITE établissent une couche de correspondance essentielle pour relier les ASV et les OTU à la taxonomie linnéenne. La taxonomie de

base du GBIF comprend des identifiants pour les hypothèses d'espèces UNITE (SH) ainsi que des numéros d'index de code-barres (BIN) qui permettent d'indexer les données d'occurrence d'espèces annotées taxonomiquement au niveau de l'OTU, principalement pour les champignons et les animaux (GBIF secretariat 2018, Grosjean 2019).

Les algorithmes d'annotation taxonomique de l'eDNA attribuent généralement chaque séquence unique au groupe taxonomique le plus proche dans un ensemble de référence, sur la base de certains critères de parenté et de confiance. Pour les groupes d'organismes mal connus, tels que les procaryotes, les insectes et les champignons, l'annotation pour un taxon (basé sur un cluster) peut être un nom de réserve non linnéen (c'est-à-dire l'ID/le numéro du SH ou du BIN concerné), et ce taxon peut représenter une espèce ou même une unité taxonomique supérieure au niveau de l'espèce. Aucune base de données de référence ne contient toutes les espèces d'un groupe donné en raison du grand nombre d'espèces inconnues, non identifiées et non décrites sur terre. L'ignorance fréquente de ce fait a été la source de nombreuses erreurs d'identification taxonomique au cours des 30 dernières années.

Lors de l'importation dans la plateforme de biodiversité (par exemple GBIF ou OBIS), la résolution taxonomique de ces occurrences dérivées de l'ADN peut être encore plus réduite, étant donné que les noms/ID obtenus par comparaison avec la base de données de référence (par exemple UNITE, BOLD) peuvent ne pas être tous inclus dans l'index taxonomique de la plateforme utilisée au moment de la publication. Toutefois, l'inclusion de la séquence OTU ou ASV sous-jacente pour chaque enregistrement permettra aux futurs utilisateurs d'identifier potentiellement la séquence à un niveau de granularité plus élevé, notamment parce que les bibliothèques de référence s'améliorent au fil du temps. C'est pourquoi nous recommandons également de publier toutes les séquences d'une étude - y compris celles qui sont actuellement entièrement non classifiées - car il pourrait être possible de les identifier grâce à l'amélioration des bases de données de référence. Dans les cas où la séquence sous-jacente ne peut pas être incluse dans les données soumises, nous préconisons le dépôt d'un nom (scientifique ou de réserve) du taxon (par exemple BOLD BIN ou UNITE SH) plus une somme de contrôle MD5 de la séquence en tant qu'identifiant unique du taxon (voir [data-mapping]). Les sommes de contrôle MD5 sont des algorithmes de hachage unidirectionnels couramment utilisés pour vérifier l'intégrité des fichiers. Dans ce cas, elles fournissent une représentation unique et reproductible de la séquence originale qui ne permet cependant pas de récupérer la séquence elle-même. Cela peut être nécessaire dans les cas où l'accès est sensible. Les sommes de contrôle MD5 permettent une interrogation efficace pour déterminer si la même séquence exacte a été récupérée dans d'autres efforts d'eDNA, mais il ne s'agit pas d'un remplacement complet de la séquence, car les MD5 ne permettent pas d'effectuer d'autres analyses. Deux séquences différant d'une seule base obtiendront deux sommes de contrôle MD5 complètement différentes, de sorte que les recherches de similarité de séquence de type BLAST ne fonctionneront pas.

1.7. Résultats

L'objectif d'exposer des données dérivées de l'ADN par l'intermédiaire des plateformes de biodiversité est de permettre la réutilisation de ces données en combinaison avec d'autres types de données sur la biodiversité. Il est très important de garder cette réutilisation à l'esprit lorsque vous préparez vos données pour la publication. Idéalement, les métadonnées et les données devraient raconter une histoire complète de telle sorte que de nouveaux utilisateurs non informés puissent utiliser ces évidences sans aucune consultation ou correspondance supplémentaire. Les plateformes de données sur la biodiversité offrent des fonctionnalités de recherche, de filtrage, de navigation, de visualisation, d'accès et de citation des données. Pour les données de métabarcoding, nous encourageons les utilisateurs à configurer les filtres pour l'abondance minimale absolue et relative des reads afin d'effectuer un filtrage approprié des données. En définissant une abondance minimale de reads par OTU ou ASV (à l'aide du champ organismQuantity), les singletons ou toute occurrence dont le nombre absolu de reads est inférieur à une certaine valeur peuvent être filtrés. En définissant une valeur minimale de la quantité relative d'organismes, calculée à partir des reads détectés (organismQuantity)

et des reads totales dans l'échantillon correspondant (sampleSizeValue) ([mapping-metabarcoding-edna-and-barcoding-data]), les occurrences dont l'abondance relative des reads est inférieure à un seuil sélectionné peuvent être éliminées. L'utilisateur peut souvent choisir les formats de sortie des données (par exemple DwC-A, CSV) et ensuite traiter, nettoyer et transformer les données dans la forme et le format requis pour ses analyses.

Sur GBIF.org ou via l'API GBIF, les utilisateurs enregistrés peuvent chercher, filtrer et télécharger des données sur la biodiversité dans les trois formats suivants :

- **Simple** : un format simple, délimité par des tabulations qui n'inclut que la version interprétée par GBIF des données, à la suite du processus d'indexation. Ceci est approprié pour les tests rapides et l'importation directe dans les feuilles de calcul.
- **Archive Darwin Core** : format plus riche qui inclut à la fois les données interprétées et la version originale verbatim fournie par l'éditeur (avant l'indexation et l'interprétation par GBIF). Puisqu'il inclut toutes les métadonnées et les indicateurs de problème, ce format fournit une vue plus riche du jeu de données téléchargé.
- **Liste d'espèces** : un format de table simple qui ne comprend qu'une liste interprétée de noms d'espèces uniques à partir d'un jeu de données ou d'un résultat de requête.

Quel que soit le format sélectionné, chaque téléchargement d'utilisateur du GBIF reçoit un lien réutilisable vers la requête et une citation des données incluant un DOI. Ce système de référence basé sur le DOI fournit un moyen permettant de reconnaître et de créditer les utilisations des jeux de données et des fournisseurs de données, améliorant à la fois la crédibilité et la transparence des résultats basés sur ces données. Il est essentiel de suivre les recommandations de citation de données et d'utiliser les DOIs, une bonne culture de citation de données n'étant pas seulement la norme académique, mais aussi un mécanisme puissant pour créditer, reconnaitre et, par conséquent, encourager les éditeurs de données.

2. Préparation et mapping des données

Ce chapitre se concentre sur les détails pratiques pour transformer votre export de données en un jeu de données indexé par une plateforme de données sur la biodiversité. § 2.1 vous aidera à décider quel est le schéma de mapping optimal pour les données à votre disposition. § 2.2 décrit ces mappings en détail.

Ce guide combine les standards pour la publication de données générales sur la biodiversité avec les données génétiques sur la biodiversité dérivées de l'ADN (Figure 5). Cette section "comment faire" fournit des recommandations de mapping pour différents types de données dérivées de l'ADN.

Les modes de préparation et de publication des données varient d'une plateforme à l'autre et sont décrits dans la documentation générale. Actuellement, une des manières les plus répandues de préparer les fichiers de données est DwC-A, où les tableaux de données sont disposés selon un schéma en étoile, et les enregistrements (lignes) des fichiers d'extension périphériques pointent vers un seul enregistrement du fichier central (Figure 5). Les différents types de fichiers centraux (par exemple, occurrence et événement d'échantillonnage) correspondent à différentes classes de jeux de données. Bien que les jeux de données dérivées de l'ADN soient souvent de nature événementielle, c'est-à-dire que des centaines, voire des milliers d'occurrences de séquences ADN peuvent provenir d'un même événement d'échantillonnage et donc partager la plupart des attributs des métadonnées, la recommandation actuelle est de publier les données en tant qu'Occurrences (catégorie I ou II) avec l'extension pour les données dérivées de l'ADN. Cette approche compense les limites du schéma en étoile du DwC, qui ne permet pas aux données relatives aux occurrences dans les fichiers d'extension (telles que les séquences de codes-barres traitées) de pointer vers les enregistrements d'un fichier événement. Nous recommandons cependant d'inclure un eventID pour chaque enregistrement

central, afin d'indiquer l'association entre les occurrences dérivées du même événement d'échantillonnage.

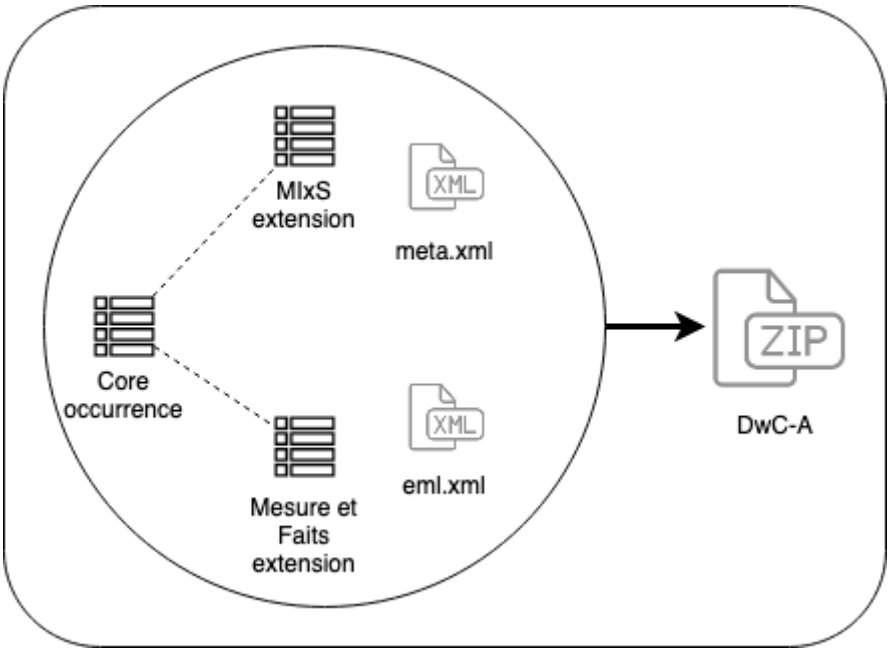

Figure 5. Zoom sur le DwC-A / IPT de la figure 3 du chapitre 1.2. Le choix de l'entité centrale est principalement une question d'adaptation des données au mécanisme d'importation de données (ingestion) des plateformes de données sur la biodiversité. La plupart des données pourraient être formulées en tant qu'Occurrence, Evénement ou de Taxon, mais comme seul le core peut avoir des extensions, cela affectera forcément le choix. Par exemple, ce n'est pas possible d'ajouter des séquences ADN aux occurrences si les données sont présentées sous forme d'événements.

2.1. Catégorisation des données

Pour la finalité de ce guide, nous classons les données en cinq catégories, reliées par un champ d'identification clé (*eventID*), qui correspondent aux standards applicables aux données générales sur la biodiversité, en incluant des champs pertinents pour les données dérivées de l'ADN (voir § 2.2, "Mapping des données"). Ces cinq catégories représentent les approches moléculaires les plus couramment utilisées pour la caractérisation de la biodiversité et sont les suivantes : I) occurrences dérivées de l'ADN, II) occurrences enrichies, III) détection ciblée d'espèces, IV) références de noms et V) métadonnées. Examinez l'arbre de décision ci-dessous et allez directement à la section qui correspond à vos données.

Tableau 1. Arbre de décision pour la catégorisation des données dérivées de l'ADN.

<table>
<tr><td colspan="3">? Est-ce que vos données sont basées sur le (méta)barcoding ou sur la qPCR ?</td></tr>
<tr><td colspan="2">(Méta)barcoding
↓
? Est-ce que vos données consistent de matériel génétique numérisé ou de séquences ADN, associés à un endroit et à une date ?</td><td>qPCR
↓
Catégorie III
Détection ciblée d'espèces</td></tr>
<tr><td>Oui
↓
? Est-ce que le matériel génétique est la seule évidence d'un organisme ou d'une communauté ?</td><td>Non
↓
? Est-ce que le jeu de données consiste en une liste de noms basés sur l'ADN ?</td><td></td></tr>
</table>

Oui ↓ **Catégorie I** Occurrences dérivées de l'ADN	Non ↓ **Catégorie II** Occurrences enrichies	Oui ↓ **Catégorie IV** Références de noms	Non ↓ **Catégorie V** Métadonnées	

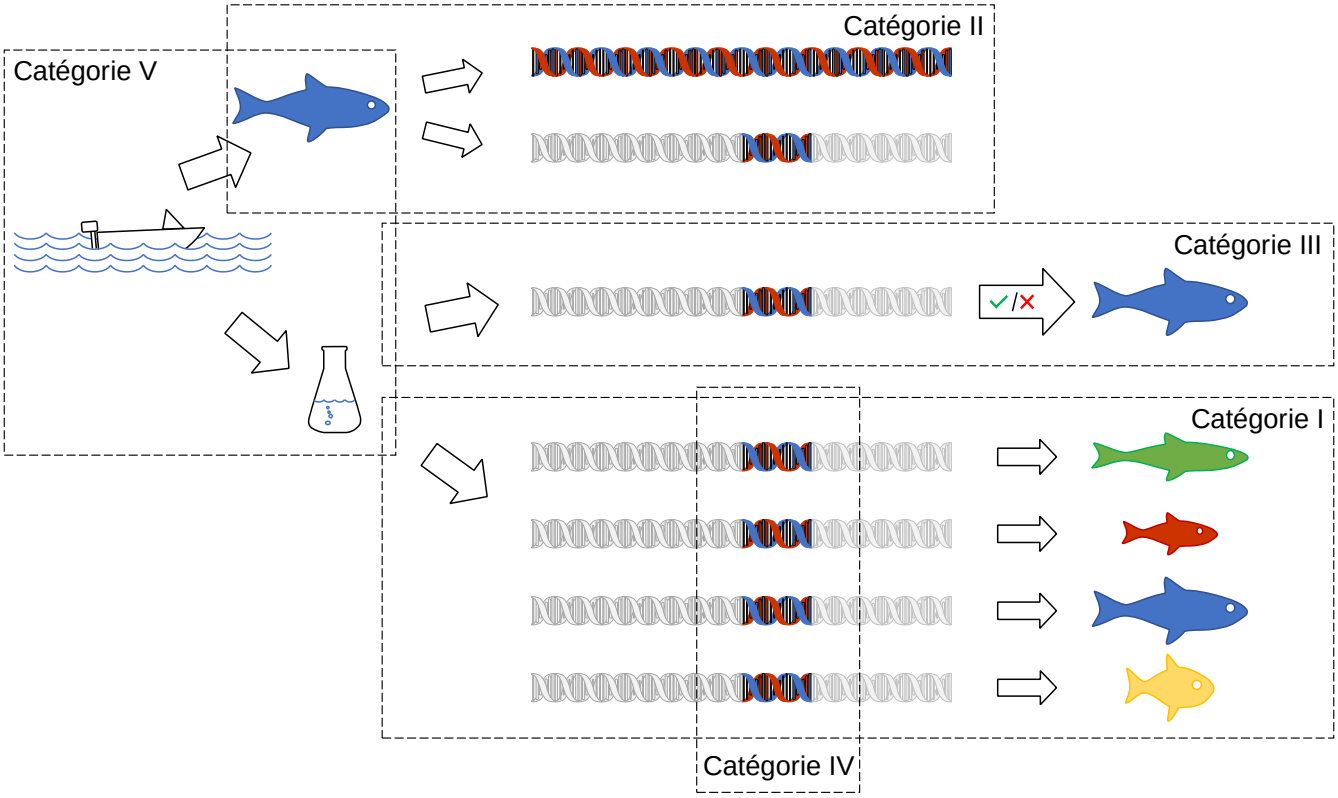

Figure 6. Représentation visuelle des catégories I-V.

2.1.1. Catégorie I : occurrences dérivées de l'ADN

Cette catégorie concerne les données pour lesquelles une séquence ADN est la seule preuve de la présence d'un organisme ou d'une communauté donnée. En d'autres termes, les données ne peuvent pas être rattachées à un spécimen observable. C'est le cas de nombreuses études de métagénomique, de métabarcoding et d'eDNA.

Exemples de jeux de données d'occurrences dérivées de l'ADN

- MGnify (2019) Impact of rainforest transformation on phylogenetic and functional diversity of soil prokaryotic communities in Sumatra (Indonésie). Jeu de données d'événements d'échantillonnage https://doi.org/10.15468/osp7hi accédé sur GBIF.org le 2020-04-16.
- MGnify (2020) Métagénomes marins du projet bioGEOTRACES. Jeu de données sur les événements d'échantillonnage https://doi.org/10.15468/oifcho accessible via GBIF.org le 2020-04-16.
- Bessey C, Jarman SN, Berry O et al. (2020) Maximizing fish detection with eDNA metabarcoding. Environmental DNA: 1–12. https://doi. rg/10.1002/edn3.74 (site internet Atlas of Living Australia sur https://collections.ala.org.au/public/show/dr14581. Accessible le 24 juin 2020)

Pour des conseils sur la façon de formater et de partager les jeux de données, voir § 2.2.1. Les directives générales pour les jeux de données d'occurrence Darwin Core sont également disponibles via DwC-A template for occurrence datasets et Data quality requirements for occurrences.

2.1.2. Catégorie II: Occurrences enrichies

Si du matériel génétique est, ou peut être, associé à une observation ou à un spécimen, nous qualifierons ce type de données d'"occurrences enrichies". Dans ce contexte, les séquences ne sont pas la seule preuve d'occurrences. On peut toujours remonter l'information jusqu'à un spécimen ou un organisme observé. Cette catégorie comprend les jeux de données de codes-barres ADN et certains jeux de données de métabarcoding de l'ADN avec du matériel de référence, par exemple. Pour plus de conseils sur les codes-barres ADN, suivez le Centre for Biodiversity Genomics, University of Guelph (2021).

Exemples de jeux de données d'occurrences enrichies

- The International Barcode of Life Consortium (2016) International Barcode of Life project (iBOL). Jeu de données sur les occurrences https://doi.org/10.15468/inygc6 accessible via GBIF.org le 2020-04-16.
- Takamura K (2019) Chironomid Specimen records in the Chironomid DNA Barcode Database. Version 1.9. National Institute of Genetics, ROIS. Jeu de données sur les occurrences https://doi.org/10.15468/hxhow5 accessible via GBIF.org le 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, (Site internet de l'Atlas of Living Australia à https://collections.ala.org.au/public/show/dr11663. Accessible le 24 juin 2020).

Pour des conseils sur la façon de formater et de partager les jeux de données, voir § 2.2.1. Les directives générales pour les jeux de données d'occurrence Darwin Core sont également disponibles via DwC-A template for occurrence datasets et Data quality requirements for occurrences.

2.1.3. Catégorie III: Détection ciblée d'espèces (qPCR / (d)dPCR)

Cette catégorie concerne les données pour lesquelles un test spécifique (qPCR / (d)dPCR) est utilisé pour détecter la présence (ou l'absence) d'une séquence ADN spécifique à l'organisme cible dans un échantillon environnemental. Dans ce cas, l'enregistrement de l'occurrence peut même ne pas contenir de données de séquence, car c'est le processus lui-même qui détermine l'occurrence. Avec les analyses qPCR / (d)dPCR pour la détection ciblée d'espèces, de nombreuses études rapportent également l'absence de l'espèce spécifique pour un échantillon donné. Les données d'absence dépendent fortement de la limite de détection de l'analyse spécifique, ainsi que des protocoles de terrain et de laboratoire. Comme pour les données de métabarcoding, il existe un problème de faux négatifs et de faux positifs, et il est important que des informations suffisantes soient rapportées pour évaluer les enregistrements.

Exemples de jeux de données d'occurrences d'espèces ciblées

- 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 site at https://collections.ala.org.au/public/show/dr8131. Accessible le 22 juillet 2020).

Pour des conseils sur la manière de formater et de partager ces jeux de données, voir [mapping-des-données-ddpcr-qpcr]. Les directives générales pour les jeux de données sur les occurrences Darwin Core sont également disponibles via DwC-A template for occurrence datasets and Data quality requirements for occurrences.

2.1.4. Catégorie IV: Références de noms

Cette catégorie correspond aux noms dérivés de l'ADN, issus du clustering ou du denoising (modèles basés sur la correction d'erreur), tels que les unités taxonomiques opérationnelles (OTU) stables et non linnéennes, les variants de séquences d'amplicons (ASV) et les index de code-barres (BIN) - en d'autres termes, toute référence à des taxons ou à des noms provisoires définis en dehors de la taxonomie linnéenne. De nombreux projets produisent des bibliothèques d'OTU spécifiques à un projet ou à une étude et, bien qu'il soit techniquement possible de les publier sous forme de checklists, elles n'ont qu'une valeur limitée, voire nulle, pour la mise en relation ou l'interprétation des données. Cependant, l'inclusion des OTUs largement adoptées, stables, globales et numériquement référençables dans les classifications taxonomiques linnéennes est d'une importance cruciale pour l'indexation de la biodiversité "obscure" sans nom. Le GBIF a accumulé de l'expérience dans l'intégration de ces grandes bibliothèques de référence mondiales d'OTUs dans la base taxonomique du GBIF, qui permet l'affichage des OTUs sous le taxon parent le plus proche qui a un nom scientifique (Figure 7).

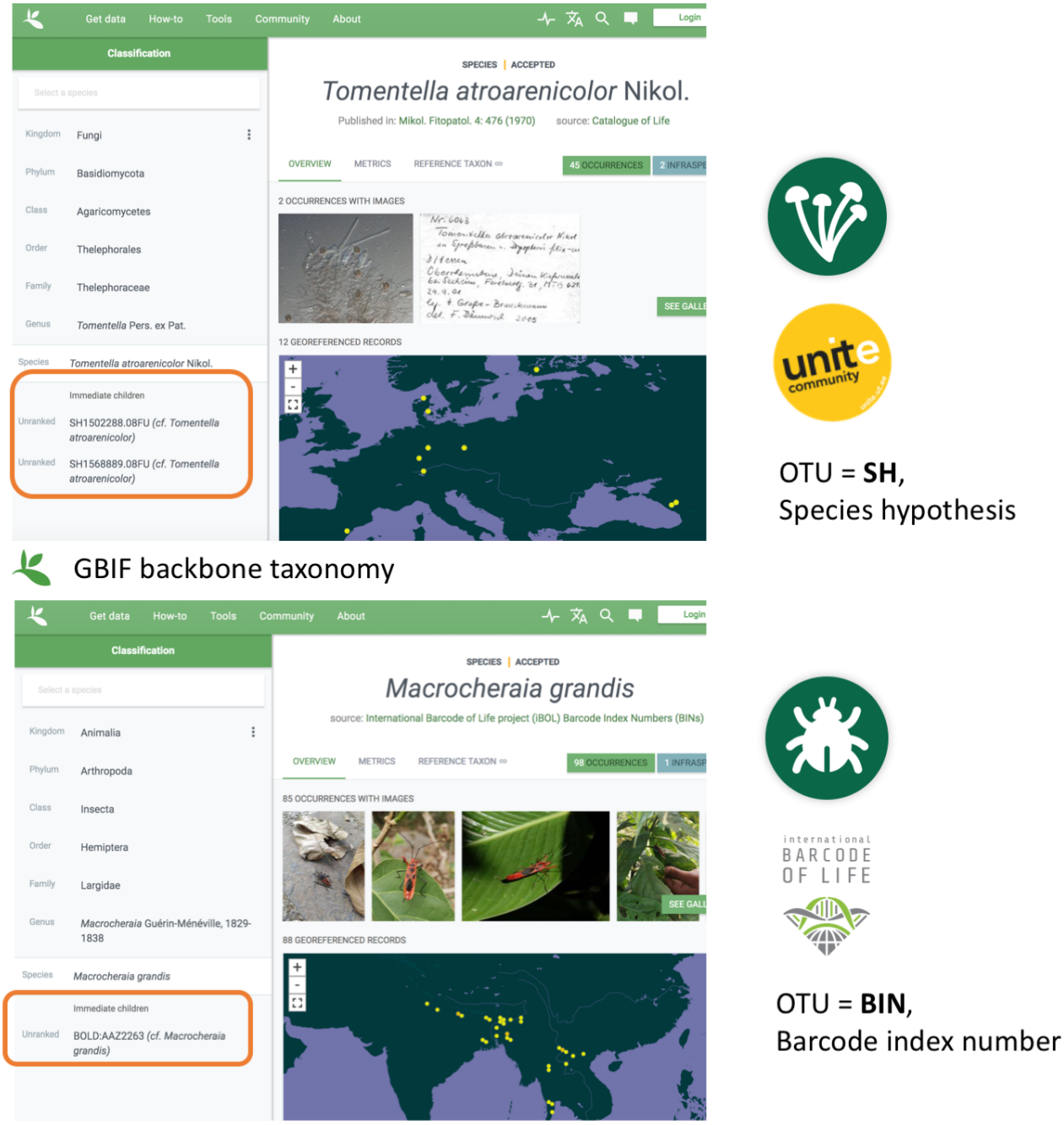

Figure 7. Les OTUs (SHs) de UNITE (principalement des champignons, ci-dessus) et de BOLD (BINs) (principalement des arthropodes, ci-dessous) sont affichés dans la taxonomie de base du GBIF sous leurs taxons parents correspondants qui ont des noms scientifiques. Les multiples occurrences de biodiversité cryptique observées individuellement peuvent être découvertes en même temps que les évidences non génétiques par le biais d'un point d'accès unique.

Exemples de checklists de références de noms

- The International Barcode of Life Consortium (2016) International Barcode of Life project (iBOL). Jeu de données sur les occurrences https://doi.org/10.15468/wvfqoi accessible via GBIF.org le 2020-04-16.
- PlutoF (2019) - Système unifié pour les espèces fongiques basées sur l'ADN liées à la classification. Version 1.2. Jeu de données de la checklist https://doi.org/10.15468/mkpcy3 accédé via GBIF.org le 2020-04-16.

Ce guide ne fournit pas de recommandations de mapping pour les checklists globales d'OTUs / bibliothèques de référence (Catégorie IV), et il est déconseillé de publier des bibliothèques OTU

référençables (spécifiques à un projet ou à une étude) sous forme de checklists. Pour obtenir des conseils sur la façon de formater et de partager les checklists d'OTUs, consultez les directives de Darwin Core suivantes sur DwC-A template for checklists, Data quality requirements for checklists et General guidelines for MIxS checklists. Pour obtenir des conseils sur la façon de mapper les bibliothèques de référence mondiales d'OTUs pour les inclure dans le squelette taxonomique du GBIF, contactez le GBIF help desk.

2.1.5. Catégorie V : jeux de métadonnées uniquement

Les métadonnées sont des données sur les données et consistent en une description du jeu de données en termes généraux, tels que les auteurs, les affiliations des auteurs, l'objectif original de la recherche liée au jeu de données, les DOI, la portée taxonomique, la portée temporelle, et la portée géographique. Les informations concernant les méthodes de laboratoire et les méthodes générales de séquençage sont incluses dans cette catégorie. Cette catégorie comprend des jeux de données ou des collections qui ne peuvent pas être mis en ligne pour le moment, comme par exemple les travaux non numérisés.

Exemples de jeux de métadonnées uniquement

- 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. Jeu de métadonnées https://doi.org/10.15468/iljmun accessible via GBIF.org on 2020-04-16.
- Cary S C (2015). New Zealand Terrestrial Biocomplexity Survey. SCAR - Microbial Antarctic Resource System. Jeu de métadonnées https://doi.org/10.15468/xnzrhq accessible via GBIF.org on 2020-04-16.

Les recommandations de mapping pour les jeux de données dérivées de l'ADN (Catégorie V) ne comportant que des métadonnées sont les mêmes que pour tous les autres jeux de données qui incluent que des métadonnées, et ce guide ne fournit pas de recommandations de mapping plus spécifiques. Veuillez suivre les recommandations générales des portails de données sur la biodiversité, en prêtant attention à required and recommended metadata. Les descriptions des étapes de terrain, de laboratoire et de bioinformatique doivent être aussi détaillées que possible. La description de vos méthodes en tant qu'étapes de méthode dans les métadonnées EML permet de les afficher sur la page d'accueil du GBIF (https://www.gbif.org/dataset/3b8c5ed8-b6c2-4264-ac52-a9d772d69e9f#methodology Frøslev T, Ejrnæs R (2018). BIOWIDE eDNA Fungi dataset. Danish Biodiversity Information Facility. Occurrence dataset https://doi.org/10.15468/nesbvx accessed via GBIF.org on 2021-07-06). Cependant, si une description structurée et éventuellement plus détaillée de la méthode est déjà publiée quelque part (par exemple, sur protocols.io ou dans NEON protocols collection), c'est facile de fournir un lien via le champ MIxS SOP (voir [mapping du métabarcoding eDNA et des données de codes-barres ADN]).

2.2. Mapping des données

Alors que les fichiers de base stockent des données omniprésentes sur le "quoi, où et quand" d'un enregistrement, les fichiers d'extension sont utilisés pour décrire les spécificités d'un certain type d'observation. Nous proposons d'utiliser DNA derived data extension pour compléter les données d'occurrence dérivées soit du barcoding, du métabarcoding (eDNA) ou de la qPCR / (d)dPCR. L'extension des données dérivées de l'ADN s'appuie sur Minimum information standards développés par le Genomic Standards Consortium (GSC) et appliqués par ENA pour submission of eDNA sample metadata, par exemple. Nous suivons et avons contribué aux directives proposées par Sustainable DwC-MIxS interoperability task group under TDWG. Afin d'améliorer l'indexation et la recherche, nous avons choisi de séparer certains termes MIxS, par exemple en partageant les séquences et les noms des amorces forward et reverse. De plus, afin de rendre le système applicable à un large éventail de

données, nous avons inclus certains champs des standards GGBN, et des champs de MIQE (informations minimales pour la publication de la PCR quantitative en temps réel) pour les données qPCR et (d)dPCR.

La première étape de la préparation de vos données pour la publication consiste à assurer que les noms des champs et les en-têtes des colonnes soient conformes à https://dwc.tdwg.org/terms/ [Darwin Core data standard]. Dans de nombreux cas, c'est assez simple, comme par exemple renommer votre champ `lat` ou `latitude` en `decimalLatitude`. Cependant, le Darwin Core Standard est assez flexible et certains termes sont utilisés de différentes manières, en fonction du type de données. Les champs `organismQuantity` et `organismQuantityType` en sont un exemple. Ils peuvent être utilisés pour décrire le nombre d'individus, le pourcentage de biomasse ou un score sur l'échelle de Braun-Blanquet, ainsi que le nombre de reads d'un ASV dans un échantillon. C'est pourquoi nous fournissons ici des tableaux des champs obligatoires et recommandés avec des descriptions et des exemples (Table 1, Table 2, Table 3 et Table 4). La recommandation d'utiliser le core Occurrence pour les données dérivées de l'ADN découle de la volonté de partager la séquence pour aider à qualifier la détermination. Des champs supplémentaires et des extensions (tels que extended Measurement or Fact (eMoF)) sont applicables - à la fois au core Occurrences et au core Evénements. Lorsqu'une séquence est dérivée d'un organisme (par exemple un parasite, le contenu d'un intestin, un épibionte, etc.), l'observation peut être liée à l'observation de l'organisme hôte. Ceci peut être réalisé en utilisant l'extension (Resource Relation extension) de Darwin Core (par exemple https://www.gbif.org/species/143610775/verbatim). La recommandation la plus importante est peut-être d'utiliser des identifiants uniques au niveau mondial (lorsqu'ils sont disponibles) et d'autres identifiants permanents pour le plus grand nombre possible de champs de données et de paramètres (dans tous les champs ID des tableaux ci-dessous).

2.2.1. Mapping du métabarcoding (eDNA) et des données de codes-barres ADN

Cette section fournit des recommandations de mapping pour les catégories I et II.

Tableau 2. Champs recommandés pour Occurrence core pour les données de métabarcoding

Nom du champ	Exemples	Description	Exigence
basisOfRecord	MaterialSample	Nature spécifique de l'enregistrement de données - un sous-type de dcterms:type. Pour les occurrences dérivées de l'ADN (voir catégorie I et catégorie III), utiliser MaterialSample. Pour les occurrences enrichies, utiliser PreservedSpecimen ou LivingSpecimen selon le cas.	Obligatoire
occurrenceID	urn:catalog:UWBM:Bird:89776	Identifiant unique pour l'occurrence, qui permet de la reconnaître dans les différentes versions du jeu de données, ainsi que lors du téléchargement et de l'utilisation des données. Il peut s'agir d'un identifiant global unique ou d'un identifiant spécifique au jeu de données.	Obligatoire
eventID	urn:uuid:a964765b-22c4-439a-jkgt-2	Identifiant pour l'ensemble des informations associées à un événement (quelque chose qui se produit à un endroit et à un moment donnés). Il peut s'agir d'un identifiant global unique ou d'un identifiant spécifique au jeu de données.	Fortement recommandé
eventDate	2020-01-05	Date à laquelle l'événement a été enregistré. La meilleure pratique recommandée est d'utiliser une date conforme à la norme ISO 8601-1:2019. Pour plus d'informations, consultez le site https://dwc.tdwg.org/terms/#dwc:eventDate	Obligatoire
recordedBy	"Oliver P. Pearson | Anita K. Pearson"	Liste (concaténée et séparée) de noms de personnes, de groupes ou d'organisations responsables de l'enregistrement de l'occurrence originale. La meilleure pratique recommandée est de séparer les valeurs par une barre verticale (' | '). L'inclusion d'informations sur l'observateur améliore la reproductibilité scientifique (Groom et al. 2020).	Fortement recommandé
organismQuantity	33	Nombre de reads de cet OTU ou ASV dans l'échantillon.	Fortement recommandé
organismQuantityType	DNA sequence reads	Le terme devrait toujours être "DNA sequence reads"	Fortement recommandé

Nom du champ	Exemples	Description	Exigence
sampleSizeValue	1233890	Nombre total de reads dans l'échantillon. Ce nombre est important car il permet de calculer l'abondance relative de chaque OTU ou ASV dans l'échantillon. Ce nombre devrait de préférence être calculé après le traitement universel (contrôle de la qualité, denoising des ASV, élimination des chimères, etc.), mais avant l'élimination manuelle/sélective, par exemple, des OTU ou ASV non ciblés de l'ensemble de données. La raréfaction (rééchantillonnage pour uniformiser la profondeur de séquençage des échantillons) n'est ni nécessaire ni conseillée.	Fortement recommandé
sampleSizeUnit	DNA sequence reads	Le terme devrait toujours être "DNA sequence reads"	Fortement recommandé
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	Identifiant pour le MaterialSample (par opposition à un enregistrement numérique particulier de l'échantillon). Utiliser l'ID de l'échantillon biologique (biosample) s'il a été obtenu à partir d'une archive de nucléotides. En l'absence d'un identifiant global unique persistant, en construire un à partir d'une combinaison d'identifiants dans l'enregistrement qui rendra l'identifiant materialSampleID globalement unique.	Fortement recommandé
samplingProtocol	Piège à lumière UV	Nom, référence ou description de la méthode ou du protocole utilisé lors d'un événement d'échantillonnage. https://dwc.tdwg.org/terms/#dwc:samplingProtocol	Recommandé
associatedSequences	https://www.ncbi.nlm.nih.gov/nuccore/MK405371	Liste (concaténée et séparée) d'identifiants (publication, identifiant global unique, URI) des séquences génétiques associées à l'occurrence. Pourrait être utilisée pour établir un lien avec des reads bruts de codes-barres archivées et/ou des séquences génomiques associées, par exemple dans un répositoire public.	Recommandé

Nom du champ	Exemples	Description	Exigence
identificationRemarks	Confiance de l'annotation RDP (au taxon spécifié le plus bas) : 0,96, par rapport à la base de données de référence : GTDB	Spécification du processus d'identification taxonomique, comprenant idéalement des données sur l'algorithme appliqué et la base de données de référence, ainsi que sur le niveau de confiance dans l'identification résultante.	Recommandé
identificationReferences	https://www.ebi.ac.uk/metagenomics/pipelines/4.1 https://github.com/terrimporter/CO1Classifier	Liste (concaténée et séparée) des références (publication, identifiant global unique, URI) utilisées dans l'identification. La meilleure pratique recommandée est de séparer les valeurs d'une liste par un espace vertical (|).	Recommandé
decimalLatitude	60.545207	Latitude géographique (en degrés décimaux, en utilisant le système de référence spatiale indiqué dans geodeticDatum) du centre géographique d'un lieu. Les valeurs positives correspondent au nord de l'équateur, les valeurs négatives au sud de l'équateur. Les valeurs légales sont comprises entre -90 et 90 inclus.	Fortement recommandé
decimalLongitude	24.174556	Longitude géographique (en degrés décimaux, en utilisant le système de référence spatiale indiqué dans geodeticDatum) du centre géographique d'un lieu. Les valeurs positives correspondent à l'est du méridien de Greenwich, les valeurs négatives à l'ouest. Les valeurs légales sont comprises entre -180 et 180 inclus.	Fortement recommandé
taxonID	ASV:7bdb57487bee022ba30c03c3e7ca50e1	Pour les données d'eDNA, il est recommandé d'utiliser un hachage MD5 de la séquence et de le faire précéder de "ASV :". Voir également [taxonomie des séquences].	Fortement recommandé, si la séquence ADN n'est pas fournie
scientificName	*Gadus morhua* L. 1758, BOLD:ACF1143	Nom scientifique du taxon connu le plus proche (espèce ou supérieur) ou un identifiant OTU de BOLD (BIN) ou UNITE (SH)	Obligatoire
kingdom	Animalia	Taxonomie supérieure	Fortement recommandé
phylum	Chordata	Taxonomie supérieure	Recommandé

Nom du champ	Exemples	Description	Exigence
term : dwc[class]	Actinopterygii	Taxonomie supérieure	Recommandé
`order`	Gadiformes	Taxonomie supérieure	Recommandé
`family`	Gadidae	Taxonomie supérieure	Recommandé
`genus`	*Gadus*	Higher taxonomy	Recommended

Tableau 3. Champs recommandés de l'extension de données dérivées de l'ADN (une sélection) pour les données de métabarcoding

Nom du champ	Exemples	Description	Exigence
DNA_sequence	TCTATCCTCAATTATAGGATAATTCACCATCAGTAGATTTAGGAATTTTCTATTCATGCAGGTATATCATCAATTAGATTAATTAATTTGTAACAATTTTAATACAAACTCATTAAACTTTACCATTTTCATGATCAGTTAGTTACCAATTCTCCTTATTATCATTA	La séquence d'ADN (ASV). L'interprétation taxonomique de la séquence dépend de la technologie et de la bibliothèque de référence disponibles au moment de la publication. Par conséquent, l'approche taxonomique la plus objective est la séquence qui peut être réinterprétée à l'avenir.	Fortement recommandé
terme:mixs[sop]	https://www.protocols.io/view/emp-its-illumina-amplicon-protocol-pa7dihn	Les procédures opérationnelles standard utilisées dans l'assemblage et/ou l'annotation des génomes, des métagénomes ou des séquences environnementales. Référence à un protocole bien documenté, par exemple en utilisant protocols.io.	Recommandé
target_gene	ARNr 16S, ARNr 18S, ITS	Nom du gène ou du marqueur ciblé pour les études basées sur les marqueurs.	Fortement recommandé
target_subfragment	V6, V9, ITS2	Nom du sous-fragment d'un gène ou d'un marqueur. Important pour, par exemple, identifier des régions spéciales sur les gènes marqueurs, comme la région hypervariable V6 du gène de l'ARNr 16S.	Fortement recommandé
terme:mixs[pcr_primer_forward]	GGACTACHVGGGTWTCTAAT	Amorce PCR directe utilisée pour amplifier la séquence du gène, du locus ou du sous-fragment ciblé.	Fortement recommandé
terme:mixs[pcr_primer_reverse]	GGACTACHVGGGTWTCTAAT	Amorce PCR inverse utilisée pour amplifier la séquence du gène, locus ou sous-fragment ciblé.	Fortement recommandé
terme:mixs[pcr_nom_de_l'amorce_forward]	jgLCO1490	Nom de l'amorce PCR directe	Fortement recommandé

Nom du champ	Exemples	Description	Exigence
pcr_primer_name_reverse	jgHCO2198	jgHCO2198	Nom de l'amorce PCR inverse
Fortement recommandé	pcr_primer_reference	https://doi.org/10.1186/1742-9994-10-34	Référence pour les amorces
Fortement recommandé	env_broad_scale	biome forestier [ENVO:01000174] **Equivalent à env_biome dans MIxS v4** Dans ce champ, indiquez de quel système environnemental majeur provient votre échantillon ou spécimen. Les systèmes identifiés doivent avoir un grain spatial grossier, afin de fournir le contexte environnemental général de l'endroit où l'échantillonnage a été effectué (par exemple, étiez-vous dans le désert ou dans une forêt tropicale ?) Nous recommandons d'utiliser des sous-classes de la classe de biome de l'OEV : http://purl.obolibrary.org/obo/ENVO_00000428	Recommandé

Nom du champ	Exemples	Description	Exigence
env_local_scale	couche de litière [ENVO:01000338] **Equivalent de env_feature dans MIxS v4** Dans ce champ, indiquez l'entité ou les entités qui se trouvent dans le voisinage local de votre échantillon ou spécimen et qui, selon vous, ont des influences causales significatives sur votre échantillon ou spécimen. Veuillez utiliser des termes présents dans ENVO et dont le grain spatial est plus petit que celui de votre entrée pour env_broad_scale.	Recommandé	env_medium

Nom du champ	Exemples	Description	Exigence
sol [ENVO:00001998] **Equivalent de env_material dans MIxS v4** Dans ce champ, indiquez quel(s) matériau(x) environnemental(aux) (valeurs séparées par une barre verticale ' | ') entourait(nt) immédiatement votre échantillon ou spécimen avant l'échantillonnage, en utilisant une ou plusieurs sous-classes de la classe de matériau environnemental de l'OEV : http://purl.obolibrary.org/obo/ENVO_00010483	Recommandé	terme:mixs[lib_layout]	Paired **Equivalent à lib_const_meth dans MIxS v4** Spécifier si l'on doit s'attendre à des reads simples, paired ou à une autre configuration.
Recommandé	`seq_meth`	Illumina HiSeq 1500	Méthode/plateforme de séquençage utilisée

Nom du champ	Exemples	Description	Exigence
Fortement recommandé	otu_class_appr	"dada2 ; 1.14.0 ; ASV"	Approche/algorithme et niveau de clustering (le cas échéant) lors de la définition des OTUs ou des ASVs
Fortement recommandé	otu_seq_comp_appr	"blastn;2.6.0+;e-value cutoff : 0.001"	Outil et seuils utilisés pour attribuer des noms "au niveau de l'espèce" aux OTUs ou ASVs
Fortement recommandé	otu_db	"Genbank nr;221", "UNITE;8.2"	Base de données de référence (i.e. séquences non générées dans le cadre de l'étude actuelle) utilisée pour attribuer une taxonomie aux OTUs ou ASVs.

2.2.2. Mapping des données qPCR / (d)dPCR

Cette section fournit des recommandations de mapping pour la Category III.

Tableau 4. Champs recommandés pour l'Occurrence core pour les données qPCR / (d)dPCR

Nom du champ	Exemples	Description	Exigence
basisOfRecord	MaterialSample	Nature spécifique de l'enregistrement de données - un sous-type de dcterms:type. Pour les occurrences dérivées de l'ADN (voir catégorie I et catégorie III), utiliser MaterialSample.	Obligatoire
occurrenceStatus	Présence, absence	Déclaration sur la présence ou l'absence d'un taxon à un endroit donné.	Obligatoire
eventID	urn:uuid:a964765b-22c4-439a-jkgt-2	Identifiant pour l'ensemble des informations associées à un événement (quelque chose qui se produit à un endroit et à un moment donnés). Il peut s'agir d'un identifiant global unique ou d'un identifiant spécifique au jeu de données.	Fortement recommandé
eventDate	2020-01-05	Date à laquelle l'événement a été enregistré. La meilleure pratique recommandée est d'utiliser une date conforme à la norme ISO 8601-1:2019. Pour plus d'informations, consultez le site https://dwc.tdwg.org/terms/#dwc:eventDate	Obligatoire
recordedBy	"Oliver P. Pearson | Anita K. Pearson"	Liste (concaténée et séparée) de noms de personnes, de groupes ou d'organisations responsables de l'enregistrement de l'occurrence originale. La meilleure pratique recommandée est de séparer les valeurs par une barre verticale (' | '). L'inclusion d'informations sur l'observateur améliore la reproductibilité scientifique (Groom et al. 2020).	Fortement recommandé
organismQuantity	50	Nombre de gouttelettes/chambres positives dans l'échantillon	Fortement recommandé pour la méthode ddPCR et dPCR
organismQuantityType	gouttelettes ddPCR + chambres dPCR	Type de partition	Fortement recommandé pour la méthode ddPCR et dPCR

Nom du champ	Exemples	Description	Exigence
sampleSizeValue	20000	Le nombre de partitions acceptées (n), c'est-à-dire les gouttelettes acceptées en ddPCR ou les chambres en dPCR.	Il est fortement recommandé pour la ddPCR, la dPCR et la dPCR.
terme:dwc[sampleSizeUnit]	gouttelettes ddPCR + chambres dPCR	Type de partition doit être égal à la valeur de organismQuantityType.	Fortement recommandé pour la méthode ddPCR et dPCR
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	Identifiant pour le MaterialSample (par opposition à un enregistrement numérique particulier de l'échantillon). Utiliser l'ID de l'échantillon biologique (biosample) s'il a été obtenu à partir d'une archive de nucléotides. En l'absence d'un identifiant global unique persistant, en construire un à partir d'une combinaison d'identifiants dans l'enregistrement qui rendra l'identifiant materialSampleID globalement unique.	Fortement recommandé
samplingProtocol	Piège à lumière UV	Nom, référence ou description de la méthode ou du protocole utilisé lors d'un événement d'échantillonnage. https://dwc.tdwg.org/terms/#dwc:samplingProtocol	Recommandé
decimalLatitude	60.545207	Latitude géographique (en degrés décimaux, en utilisant le système de référence spatiale indiqué dans geodeticDatum) du centre géographique d'un lieu. Les valeurs positives correspondent au nord de l'équateur, les valeurs négatives au sud de l'équateur. Les valeurs légales sont comprises entre -90 et 90 inclus.	Fortement recommandé
decimalLongitude	24.174556	Longitude géographique (en degrés décimaux, en utilisant le système de référence spatiale indiqué dans geodeticDatum) du centre géographique d'un lieu. Les valeurs positives correspondent à l'est du méridien de Greenwich, les valeurs négatives à l'ouest. Les valeurs légales sont comprises entre -180 et 180 inclus.	Fortement recommandé

Nom du champ	Exemples	Description	Exigence
scientificName	*Gadus morhua* L. 1758, BOLD:ACF1143	Nom scientifique du taxon connu le plus proche (espèce ou supérieur) ou un identifiant OTU de BOLD (BIN) ou UNITE (SH)	Obligatoire
kingdom	Animalia	Taxonomie supérieure	Fortement recommandé
phylum	Chordata	Taxonomie supérieure	Recommandé
term : dwc[class]	Actinopterygii	Taxonomie supérieure	Recommandé
order	Gadiformes	Taxonomie supérieure	Recommandé
family	Gadidae	Taxonomie supérieure	Recommandé
genus	*Gadus*	Taxonomie supérieure	Recommandé

Tableau 5. Champs recommandés de l'extension de données dérivées de l'ADN DNA derived data extension (une sélection) pour les données qPCR / (d)dPCR

Nom de champ	Exemples	Description	Recommandation
sop	https://www.protocols.io/view/protocol-for-dna-extraction-and-quantitative-pcr-d-vwie7ce https://doi.org/10.17504/protocols.io.vwie7ce	Protocole opérationnel standard utilisé lors de l'assemblage et/ou l'annotation de génomes, métagénomes et séquences environnementales. Référence vers un protocole bien documenté, e.g. en utilisant protocols.io	Fortement recommandé
annealingTemp	60	La température de réaction pendant la phase d'hybridation de la PCR.	Requis si annealingTemp a été spécifié
annealingTempUnit	Degrees Celsius		Fortement recommandé
pcr_cond	initial denaturation:94_3; annealing:50_1;elongation:72_1.5;final elongation:72_10;35	Description des conditions de réactions et des composantes de la PCR, sous la forme de "dénaturation initiale :94degC_1.5min; hybridation=…"	Fortement recommandé
probeReporter	FAM	Type de fluorophore (rapporteur) utilisé. La sonde s'hybride avec l'ADN cible amplifié. L'activité de la polymérase dégrade la sonde s'étant hybridée au modèle, et la sonde relâche le fluorophore et brise la proximité avec le "quencher", permettant la fluorescence du fluorophore.	Fortement recommandé
probeQuencher	NFQ-MGB	Type de "quencher" utilisé. La molécule "quencher" absorbe la fluorescence émise par le fluorophore lorsque excité par la source lumineuse du thermocycleur tant que le fluorophore et le quencher sont à proximité l'un de l'autre, l'absorption inhibe tout signal de fluorescence.	Fortement recommandé
ampliconSize	83	La longueur de l'amplicon en paires de bases.	Fortement recommandé

Nom de champ	Exemples	Description	Recommandation
thresholdQuantificationCycle	0.3	Seuil pour le changement de signal de fluorescence entre les cycles.	qPCR : Fortement recommandé
baselineValue	15	Le nombre de cycles pendant lesquels le signal de fluorescence est inférieur à la fluorescence de fond ne provenant pas de la véritable cible d'amplification.	qPCR : Fortement recommandé
quantificationCycle	37.9450950622558	Le nombre de cycles nécessaires pour que le signal de fluorescence dépasse une valeur seuil supérieure à la valeur de base. Le cycle de quantification (Cq), le cycle seuil (Ct), le point de franchissement (Cp) et le point de décollage (TOP) font référence à la même valeur provenant de l'instrument en temps-réel. L'utilisation du cycle de quantification (Cq) est préférable selon le standard de données RDML (language Markup PCR en temps-réel)	
automaticThresholdQuantificationCycle	no	Information indiquant si le seuil a été fixé par l'instrument ou manuellement.	
automaticBaselineValue	no	Information indiquant si la valeur de base a été fixée par l'instrument ou manuellement.	
contaminationAssessment	no	Information indiquant si l'évaluation de la contamination ADN ou ARN a été faite ou non.	
estimatedNumberOfCopies	10300	Nombre de molécules cibles par µl. La moyenne des copies par partition (?) peut être calculée en utilisant le nombre de partitions (n) et le nombre estimé de copies dans le volume total de toutes les partitions (m) à l'aide de la formule ?=m/n.	
amplificationReactionVolume	22	Volume de la réaction PCR.	
amplificationReactionVolumeUnit	µl	Unité utilisée pour le volume de la réaction PCR. De nombreux instruments nécessitent la préparation d'un volume initial d'échantillon plus grand que ce qui sera réellement analysé.	
pcr_analysis_software	BIO-RAD QuantaSoft	Le programme utilisé pour analyser les runs de d(d)PCR.	

Nom de champ	Exemples	Description	Recommandation
experimentalVariance		Il est encouragé d'obtenir de multiples réplicats biologiques afin d'évaluer la variation expérimentale totale. Lorsqu'un seul essai de dPCR est réalisé, une estimation minimale de la variance due au dénombrement des erreurs seulement doit être calculée à partir de la distribution binomiale (ou un équivalent valable).	
target_gene	16S rRNA, 18S rRNA, nif, amoA, rpo	Gène ciblé ou nom du marqueur pour les études basées sur les marqueurs.	Fortement recommandé
target_subfragment	V6, V9, ITS	Nom du sous-fragment d'un gène ou d'un marqueur. Important pour e.g. identifier des régions spécifiques sur des marqueurs de gènes tel que la région hypervariable V6 du gène 16S rARN.	Fortement recommandé
pcr_primer_forward	GGACTACHVGGGTWTCTAAT	Amorce PCR sens ayant été utilisée lors de l'amplification de la séquence du gène cible, du locus ou du sous-fragment.	Fortement recommandé
pcr_primer_reverse	GGACTACHVGGGTWTCTAAT	Amorce PCR anti-sens ayant été utilisée lors de l'amplification de la séquence du gène cible, du locus ou du sous-fragment.	Fortement recommandé
pcr_primer_name_forward	jgLCO1490	Nom de l'amorce PCR sens.	Fortement recommandé
pcr_primer_name_reverse	jgHCO2198	Nom de l'amorce PCR anti-sens.	Fortement recommandé
pcr_primer_reference	https://doi.org/10.1186/1742-9994-10-34	Références des amorces.	Fortement recommandé
env_broad_scale	forest biome [ENVO:01000174]	**Équivalent de env_biome dans MIxS v4** Dans ce champ, indiquez de quel système environnemental majeur provient votre échantillon ou votre spécimen. Les systèmes identifiés devraient avoir une granularité spatiale grossière, afin de fournir un contexte environnemental général par rapport au lieu d'échantillonnage (e.g. étiez-vous dans un désert ou une forêt tropicale ?). Nour recommandons l'utilisation des sous-classes des classes de biomes de ENVO : http://purl.obolibrary.org/obo/ENVO_00000428	Recommandé

Nom de champ	Exemples	Description	Recommandation
env_local_scale	litter layer [ENVO:01000338]	**Équivalent de env_feature dans MIxS v4** Dans ce champs, mentionner la ou les entité(s) présente(s) dans le voisinage proche de votre échantillon ou spécimen et qui pourrai(en)t avoir une importante influence causale sur ceux-ci. Veuillez utiliser les termes présents dans ENVO ayant une granularité spatial plus fine que ceux utilisés pour env_broad_scale.	Recommandé
env_medium	soil [ENVO:00001998]	**Équivalent de env_material dans MIxS v4** Dans ce champ, mentionner tout le matériel environnemental (séparer les valeurs dans une liste par un espace barre verticale (|)) entourant directement votre échantillon ou votre spécimen avant l'échantillonnage, en utilisant une ou plusieurs sous-classes des classes ENVO pour le matériel environnemental : http://purl.obolibrary.org/obo/ENVO_00010483	Recommandé
concentration	67.5	Concentration d'ADN (poids ng/volume µl). Voir aussi : http://terms.tdwg.org/wiki/ggbn:concentration	Recommandé
concentrationUnit	ng/µl	Unité utilisée pour mesurer la concentration. Voir aussi : http://terms.tdwg.org/wiki/ggbn:concentrationUnit	Recommandé
methodDeterminationConcentrationAndRatios	Nanodrop, Qubit	Description de la méthode utilisée pour mesurer la concentration. Voir aussi : http://terms.tdwg.org/wiki/ggbn:methodDeterminationConcentrationAndRatios	Recommandé
ratioOfAbsorbance260_230	1.89	Ratio de l'absorbance à 260 nm et 230 nm évaluant la pureté de l'ADN (mesure secondaire principalement, indiquant surtout l'EDTA, les carbohydrates et phenol), (échantillons d'ADN seulement). Voir aussi : http://terms.tdwg.org/wiki/ggbn:ratioOfAbsorbance260_230	Recommandé
ratioOfAbsorbance260_280	1.91	Ratio de l'absorbance à 280 nm et 230 nm évaluant la pureté de l'ADN (mesure secondaire principalement, indiquant surtout l'EDTA, les carbohydrates et phenol), (échantillons d'ADN seulement). Voir aussi : http://terms.tdwg.org/wiki/ggbn:ratioOfAbsorbance260_280	Recommandé
samp_collect_device	biopsy, niskin bottle, push core	La méthode ou l'appareil utilisé pour récolter l'échantillon.	Recommandé
samp_mat_process	filtering of seawater, storing samples in ethanol	Tout traitement appliqué à l'échantillon pendant ou après la collecte de celui-ci dans l'environnement. Ce champ accepte OBI. Pour parcourir les termes OBI (v 2018-02-12) veuillez consulter : http://purl.bioontology.org/ontology/OBI	Recommandé

Nom de champ	Exemples	Description	Recommandation
samp_size	5 litre	Quantité ou taille de l'échantillon (volume, masse ou aire) qui a été collecté.	Recommandé
size_frac	0-0.22 micrometer	Taille des pores de filtrations utilisés lors de la préparation des échantillons.	Recommandé
pcr_primer_lod	51	La capacité du test PCR à détecter la cible avec de faibles niveaux.	Fortement recommandé
pcr_primer_loq	184	La capacité du test PCR à quantifier le nombre de copies à de faibles niveaux.	Fortement recommandé

2.3. Jeux de données marines et système d'information sur la biodiversité des océans (OBIS)

Lorsqu'on travaille avec des jeux de données provenant de l'environnement marin, il est recommandé de publier les informations dans Ocean Biodiversity Information System (OBIS) en plus de GBIF. L'OBIS est une base de données mondiale sur la biodiversité, spécialisée dans la mise à disposition de données fiables et accessibles sur la vie marine, qui fait partie de la COI-UNESCO. Comme GBIF et ALA, OBIS utilise le format DwC-A pour l'indexation et la publication des données. En publiant des jeux de données marines par l'intermédiaire d'OBIS, en plus d'autres bases de données sur la biodiversité, les données peuvent atteindre un public plus large et divers groupes travaillant dans le domaine de la biodiversité marine, car les jeux de données d'OBIS sont souvent utilisés dans le cadre de processus des Nations unies. Avec un accent sur les jeux de données marines, des contrôles rigoureux de qualité augmentent la fiabilité des données et conduisent à de petites différences dans les informations requises pour la publication dans l'OBIS par rapport à GBIF.

Afin d'assurer une nomenclature taxonomique cohérente, OBIS utilise le World Register of Marine Species (WoRMS) comme seule base de référence taxonomique. C'est également le cas pour les occurrences dérivées de données génétiques ; un nom scientifique lié à un ID de nom scientifique de WoRMS est une information fortement recommandée pour la publication. Si un ID de nom scientifique n'est pas fourni, OBIS essaiera de faire correspondre le nom scientifique avec WoRMS pendant l'indexation, mais cela devrait être évité dans la mesure du possible. Les noms scientifiques non répertoriés dans le WoRMS sont acceptables et seront soumis à WoRMS pour examen et éventuelle inclusion dans le registre. Il est recommandé de classer les séquences entièrement non classées comme "incertae sedis", avec le WoRMS `scientificNameID` urn:lsid:marinespecies.org:taxname:12. Cela garantira une interprétation correcte par GBIF et OBIS. En outre, il est recommandé d'ajouter les identifiants de séquence des bases de données de référence utilisées (par exemple les numéros d'index de code-barres : BINs de BOLD) dans le champ `taxonConceptID` du core occurrences. De cette manière, OBIS conservera son système taxonomique basée sur WoRMS, tout en permettant l'établissement de liens avec des bases de données de séquences de référence disparates. Les noms provenant de bases de données de référence qui ne sont pas strictement des noms scientifiques peuvent être ajoutés en tant que `verbatimIdentification`. La classification automatique des noms d'espèces peut souvent être réalisée grâce aux services de correspondance de taxons de WoRMS et des packages R, tels que worrms et taxize. A l'avenir, OBIS prévoit de rechercher et de mettre à jour périodiquement les assignations taxonomiques des séquences soumises, au fur et à mesure que les bases de données de référence se développent, donc l'enregistrement des informations de séquences ADN liées à chaque occurrence est fortement recommandé.

Les coordonnées géographiques sont un autre champ obligatoire dans les données soumises à l'OBIS. L'OBIS effectue des contrôles de qualité supplémentaires pour les données marines, pour assurer, par exemple, que les coordonnées des espèces strictement marines ne se trouvent pas sur la terre ferme et que la profondeur indiquée se situe dans une fourchette raisonnable. Enfin, il convient de mentionner que l'OBIS permet également l'utilisation de extended Measurement or Fact (eMoF). Cette extension permet de relier les données environnementales et les informations d'échantillonnage aux événements d'échantillonnage ou aux occurrences, ainsi que les mesures biologiques aux occurrences d'une manière souple et normalisée. OBIS dispose d'un exemple de jeu de données de métabarcoding d'eDNA avec des scripts pour le formatage des données disponibles à l'adresse https://github.com/iobis/dataset-edna.

Tableau 6. Exigences et recommandations d'OBIS pour l'enregistrement d'occurrences dérivées de l'ADN. Le tableau met en évidence les différences importantes dans les valeurs des champs et les exigences par rapport à la publication dans GBIF. L'exemple suivant illustre la détection de l'ADN du rorqual bleu (Balaenoptera musculus).

Nom du champ	**Valeur/exemple (OBIS)**	**Description**	**Exigence**
Balaenoptera musculus - `scientificName`	Balaenoptera musculus	Nom scientifique, de préférence tel qu'il figure dans la base de données du WoRMS. Ceci diffère de GBIF, où il est recommandé d'utiliser le nom du taxon dérivé de l'approche de classification utilisée.	Obligatoire
`scientificNameID`	urn:lsid:marinespecies.org:taxname:137090	ID du nom scientifique de "Balaenoptera musculus" selon la base de données WoRMS.	Fortement recommandé
`taxonConceptID`	NCBI:txid9771	ID NCBI lié à Balaenoptera musculus dans la base de données taxonomique NCBI. Il peut également s'agir d'un BIN-ID si BOLD a été utilisé pour l'identification, ou d'un autre ID provenant d'une base de données différente.	Recommandé
`verbatimIdentification`	Balaenoptera musculus	Nom correspondant à l'identifiant NCBI (Balaenoptera musculus) (ou autre identifiant). Il ne correspond pas nécessairement à la valeur du nom scientifique.	Recommandé Traduit avec www.DeepL.com/Translator (version gratuite)

Tableau 7. Exigences et recommandations de l'OBIS pour l'enregistrement des séquences qui ne peuvent pas être classées sous un nom scientifique à aucun niveau taxonomique.

Nom du champ	**Valeur/exemple (OBIS)**	**Description**	**Exigence**
terme:dwc[nomscientifique]	incertae sedis	Nom scientifique des séquences inconnues recommandé par OBIS. Utilisez ce nom lorsque la séquence/taxonomie est inconnue. Ceci diffère de GBIF, où il est recommandé d'utiliser le nom du taxon dérivé de l'approche de classification utilisée, même s'il ne s'agit pas d'un nom scientifique à proprement parler.	Obligatoire
`scientificNameID`	urn:lsid:marinespecies.org:taxname:12	ID du nom scientifique de "incertae sedis" selon la base de données WoRMS pour les séquences inconnues recommandées par OBIS. Utilisez cet identifiant lorsque la séquence/taxonomie est inconnue.	Fortement recommandé

Nom du champ	Valeur/exemple (OBIS)	Description	Exigence
`taxonConceptID`	NCBI:txid1899546	ID dans une base de données taxonomique externe, comme une base de données de référence de séquences, par exemple.	Recommandé
Les termes : `verbatimIdentification`	eucaryote phototrophe	Nom du taxon dans une base de données externe, correspondant à l'ID du concept de taxon.	Recommandé

3. Perspectives futures

L'intérêt actuel à exposer les données dérivées de l'ADN par le biais des plateformes de données sur la biodiversité est énorme, et probablement la demande va encore augmenter. Notre objectif est que les recommandations de mapping fournies ici restent valables et évoluent lentement, même si la préparation et l'indexation par les plateformes de données sur la biodiversité se développent plus rapidement. Les auteurs ont connaissance de BOLD Handbook, BIOM format et http://edamontology.org/page, mais ne les ont pas encore consultés.

Nous proposons que les plateformes de données telles que l'ALA et le GBIF s'efforcent d'adopter des formats de données qui prennent en compte des données relationnelles et hiérarchiques plus complexes. Des exemples pourraient être le Frictionless Data Format et le format plus spécifique au domaine Biological Observation Matrix (BIOM) format. Ce dernier est utilisé par plusieurs outils bioinformatiques (QIIME2, Mothur, USEARCH etc.) et pourrait donc aider les éditeurs à sauter une étape dans la conversion des données au format DwC-A. Un format de données plus flexible que le schéma actuel DwC en étoile est essentiel pour permettre des événements d'échantillonnage hiérarchiques et des échantillons de matériaux, ainsi que pour lier des données de séquence ADN à des occurrences individuelles dans un même événement d'échantillonnage.

Les plateformes de données sur la biodiversité devront également permettre aux utilisateurs d'inclure ou d'exclure facilement les données d'occurrence dérivées de l'ADN dans les résultats de leurs recherches. Les formats de données suggérés ci-dessus pourraient ouvrir la voie à une classification plus riche des types d'évidence sur lesquelles repose un enregistrement d'occurrence spécifique. Toutefois, pour l'instant, il manque une valeur appropriée dans le vocabulaire BasisOfRecord pour ces types de données. Nous suggérons, à titre de solution pragmatique immédiate, d'ajouter à la base de données BasisOfRecord une valeur telle que "ADN", "dérivé de l'ADN" ou similaire. Comme décrit ci-dessus, les données dérivées de l'ADN peuvent provenir d'un échantillonnage bien documenté ou d'organismes individuels, peuvent être soutenues par du matériel physique préservé ou non, ou peuvent résulter d'un séquençage génétique ou d'autres méthodes de détection de l'ADN, comme la qPCR. Les plateformes de données sur la biodiversité et le TDWG devraient fournir les moyens de différencier ces types de données et leurs origines.

Nous recommandons également que les plateformes de données indexent les vraies séquences ADN, ou au moins un checksum MD5 de celles-ci, afin de faciliter les recherches d'ASVs dans les jeux de données. Si les ASVs sont fournis, les MD5 devraient être générés par les plateformes de données sur la biodiversité ; si les ASVs ne sont pas fournis, les MD5 doivent être obligatoires.

Comme mentionné dans § 1.6 et § 2.1.4, nous encourageons les plateformes de données sur la biodiversité à poursuivre leur travail d'adoption de bases de données de référence en taxonomie moléculaire dans leurs structures taxonomiques centrales.

L'application plus large d'autres méthodes et technologies, telles que Oxford Nanopore, PacBio et le séquençage shotgun, rendra probablement nécessaire l'adaptation du présent guide à de nouveaux champs de données et de métadonnées spécifiques.

Glossaire

Atlas du Vivant d'Australie (ALA)

ALA est une plateforme web qui rassemble des données sur la biodiversité australienne provenant de sources multiples, ce qui la rend accessible et réutilisable pour tout le monde (voir https://www.ala.org.au/about-ala/). La plateforme d'infrastructure ouverte développée par ALA est également utilisée par plusieurs autres pays pour leur propre plateforme nationale de données sur la biodiversité (voir https://living-atlases.gbif.org/).

Variant de Séquences d'Amplicons (ASV)

Séquence ADN individuelle produite par séquençage à haut débit d'amplicons ou par denoising, et supposée représenter une variante de séquence biologiquement réelle. Voir également <otu,Operational Taxonomic Unit (OTU)> et (Callahan et al. 2017).

Interface d'Applications de Programmation (API)

Ensemble de protocoles et d'outils pour l'interaction et la transmission de données entre différentes applications informatiques.

Numéros d'Index de Codes-barres (BINs)

Unité taxonomique opérationnelle au niveau de l'espèce Operational Taxonomic Units (OTUs) dérivée du clustering de séquences du gène cytochrome c oxydase I (COI) chez les animaux. Chaque BIN se voit assigner un identifiant unique au niveau mondial et est disponible dans la base de données consultable Barcode of Life Data System (BOLD).

Système Barcode of Life Data (BOLD)

BOLD est la base de données de référence maintenue par le Centre for Biodiversity Genomics à Guelph au nom du Consortium international Barcode of Life (IBOL). Elle héberge des données sur les spécimens et les séquences de référence des codes-barres ADN pour les espèces d'eukaryotes, en particulier le COI pour les animaux, et maintient le système des numéros d'index de codes-barres (BIN; Ratnasingham & Hebert 2013), identifiants pour les OTUs des rangs approximatifs d'espèces, basés sur des groupes de séquences étroitement similaires.

Plateforme de données sur la biodiversité

Ressource générale en ligne permettant de découvrir et d'accéder aux données sur la biodiversité provenant de diverses sources, comme les collections d'histoire naturelle, les projets de science citoyenne, d'écologie et de surveillance, et de séquençage génétique. Peut être mondial (GBIF) ou national (ALA).

Clustering

Dans la classification taxonomique, le processus de regroupement des séquences ADN selon certains critères de similarité. Voir Operational Taxonomic Unit.

ADN de la communauté (échantillons mixtes)

ADN provenant d'échantillons mixtes (par exemple, des échantillons de plancton ou des échantillons de pièges de Malaise composés de plusieurs individus provenant de nombreuses espèces). Dans le but de ce guide, des échantillons d'ADN mixte sont inclus dans le concept d'eDNA.

Archive Darwin Core (DwC-A)

Format de fichier compressé (ZIP) pour l'échange de données sur la biodiversité compilées conformément au standard Darwin Core (DwC) standard. Essentiellement un ensemble autonome de fichiers CSV interconnectés et un document XML décrivant les fichiers et les colonnes de données, ainsi que leurs relations mutuelles.

Standard Darwin Core (DwC)

Norme de partage et de publication des données sur la biodiversité, provenant de la communauté des Normes d'Information sur la Biodiversité (Biodiversity Information Standards, TDWG). En principe, un ensemble de termes utilisés pour décrire les différentes entités d'observations de la biodiversité, telles que les événements d'échantillonnage, les occurrences et les taxa. Les termes actuels du Darwin Core sont décrits dans le Guide de référence rapide.

Vocabulaire des données

Ensemble privilégié de termes ou de concepts avec des significations et des relations spécifiques et bien définies, facilitant l'échange et la réutilisation des données.

(d)dPCR (Polymerase Chain Reaction numérique par goutelette)

Droplet digital PCR. Méthode permettant de mesurer la quantité absolue d'ADN (nombre de copies) d'un marqueur dans un échantillon. Voir également qPCR.

Suppression du bruit - Denoising

En métabarcoding, méthode de séparation des véritables séquences biologiques (voir ASVs) des variantes de séquences parasites causées par l'amplification par PCR et l'erreur de séquençage.

Identifiant d'Objet Numérique (DOI)

Référence pérenne utilisée pour identifier (et localiser) de façon unique des objets d'information numérique, comme un jeu de données sur la biodiversité ou une publication scientifique.

Barcoding ADN et métabarcoding (séquence amplicon)

Utilisation de fragments d'ADN courts et standardisés pour identifier des organismes individuels par séquençage. Le métabarcoding combine l'utilisation des codes-barres ADN avec le séquençage à haut débit, en utilisant des amorces universelles pour amplifier et séquencer de grands groupes d'organismes dans des échantillons d'ADN environnemental.

Marqueur ADN

Fragment d'ADN utilisé pour différentier certaines propriétés (par exemple, l'affiliation taxonomique). Peut être un gène ou une partie d'un gène, mais ce n'est pas obligatoire.

Base de données de métabarcoding ADN

Base de données contenant des séquences ADN (codes-barres ADN) d'organismes précédemment récupérés ou étudiés. Les séquences de référence ont été idéalement générées à partir d'individus d'espèces décrites et bien étudiées - le spécimen type servant de référence idéale - ou d'un niveau taxonomique supérieur (par exemple, le genre, la famille), mais elles peuvent également provenir d'efforts de séquençage d'eDNA. Il est recommandé de ne pas se fier aveuglément aux « séquences de référence ».

Sonde ADN

Fragment d'ADN court, simple brin synthétisé avec un marquage fluorescent qui se lie à une région sélectionnée de l'ADN cible (marqueur) pendant la PCR. Augmente la spécificité et peut être utilisée en plus des amorces dans qPCR et ddPCR pour détecter et quantifier un marqueur génétique.

Institut de Bioinformatique Européen (EMBL-EBI)

Organisation intergouvernementale pour la recherche et les services en bioinformatique, faisant partie du Laboratoire Européen de biologie moléculaire (EMBL), fournissant par exemple des séquences (brutes) et des données d'assemblage via Archives européennes de nucléotides (ENA).

ADN environnemental (eDNA)

ADN provenant d'un échantillon environnemental, par exemple le sol, l'eau, l'air ou l'organisme hôte. Une définition souvent utilisée est que l'ADN environnemental est le matériel génétique (ADN) obtenu à partir d'échantillons environnementaux sans preuve évidente de matériel biologique source.

Archives Européennes de Nucléotides (ENA)

Dépôt européen de séquences de nucléotides, couvrant les données de séquençage brutes, les

informations d'assemblage de séquences et les annotations fonctionnelles. Inclut le Sequence Read Archive (SRA), et est maintenu par l'Institut Européen de Bioinformatique (EMBL-EBI), dans le cadre de la International Nucleotide Sequence Database Collaboration (INSDC).

FASTQ

Standard textuel pour le stockage des séquences moléculaires et des mesures de qualité associées dérivant de High-throughput sequencing (HTS). Pour chaque position de séquence, des caractères ASCII uniques sont utilisés pour représenter l'appel de base (nucléotide identifié) et le score, respectivement.

Système Mondial d'Information sur la Biodiversité (GBIF)

Réseau international et infrastructure de recherche, principalement axé sur la mobilisation et la publication en accès libre de données mondiales sur la biodiversité.

Réseau Mondial sur la Biodiversité Génomique (GGBN)

Réseau international d'institutions soucieuses du partage et de l'utilisation efficaces d'échantillons génomiques sur la biodiversité et des métadonnées associées, faisant par exemple la promotion du standard de données GGBN compatible avec Darwin Core.

Système de Positionnement Global (GPS)

Système de navigation par satellite exploité par la Force spatiale américaine.

Séquençage à Haut Débit (HTS)

Différentes technologies pour le séquençage massivement parallèle, produisant des millions de séquences ADN à partir de la préparation de la librairie de matériel génétique, plutôt que de cibler des amplicons simples comme dans le séquençage Sanger traditionnel. Également appelé séquençage nouvelle génération (NGS).

Ingestion

Processus d'importation de données à partir de sources hétérogènes, telles que des bases de données locales, des fichiers texte ou des feuilles de calcul, vers un système de destination commun, tel qu'une plateforme de données de biodiversité en ligne, pour stockage et analyse ultérieure. Inclut généralement les étapes d'extraction, de transformation (nettoyage) et de chargement (ETL).

Indexation

Organisation des informations selon un schéma ou une structure spécifique, facilitant l'accès et la présentation des données.

Système International Collaboratif des Séquences de Nucléotides (INSDC)

Effort conjoint de la base de données d'ADN du Japon (DDBJ), EMBL et NCBI pour fournir un accès public global aux données de séquences de nucléotides et aux informations associées.

Métagénomique

Séquençage sans PCR de fragments génomiques aléatoires dans un échantillon mixte.

Standard de l'Information Minimale sur toute (x) Séquence (MIxS)

Famille de normes (listes de contrôle) pour les métadonnées de séquence, élaborées par le Consortium des Normes Généomiques (GSC).

Unité Taxonomique Opérationnelle moléculaire (mOTU)

Voir Operational Taxonomic Unit (OTU).

Centre National d'Information sur la Biotechnologie (NCBI)

Division de la Bibliothèque Nationale de Médecine des États-Unis (NLM) hébergeant d'importantes ressources bioinformatiques, comme la base de données GenBank des séquences ADN, et le Sequence Read Archive (SRA) des données de séquences à haut débit.

Séquençage Nouvelle Génération (NGS)

Voir High-throughput sequencing (HTS).

Occurrence

Une existence d'un organisme (sensu http://rs.tdwg.org/dwc/terms/Organism) à un endroit particulier à un moment donné.

Unité Taxonomique Opérationnelle (OTU)

Regroupement d'organismes basé sur la similarité d'une ou de plusieurs séquences de marqueurs d'ADN spécifiques, utilisé pour la classification taxonomique. Comprend, par exemple, Species Hypothesis dans UNITE, et Barcode Index Numbers dans le Barcode of Life Data System (BOLD). Amplicon Sequence Variants (ASVs) peuvent être considérés comme analogues à zero radius OTUs (zOTUs).

Amplification en Chaîne par Polymérase (PCR)

Technique pour l'amplification rapide et la détection de fragments spécifiques de séquences ADN (ou ARN). Les régions amplifiées sont déterminées par la paire de PCR primers utilisée dans la réaction.

Pipeline

En bioinformatique, ensemble d'algorithmes ou d'outils appliqués dans un flux de travail prédéfini pour traiter, par exemple, des données High-throughput sequencing (HTS).

Amorces (PCR primers)

Fragments d'ADN courts, simple brin qui se lient à une région sélectionnée de l'ADN cible (marqueur) pour initier la réplication pendant PCR. Une paire d'amorces est généralement utilisée pour que l'enzyme polymérase amplifie le marqueur sélectionné.

qPCR (Amplification quantitative en Chaîne par Polymérase)

PCR quantitative. Méthode qui mesure la quantité relative d'ADN d'un marqueur dans un échantillon. Voir aussi ddPCR.

Échantillon

Matériel (eau, sol, contenu intestinal, etc.) obtenu pour analyse.

Alignement des séquences

Processus bioinformatique de comparaison et de disposition de deux ou plusieurs séquences moléculaires (ADN, ARN ou protéine) pour détecter les similitudes causées par exemple par une parenté évolutionnaire.

Hypothèse d'Espèce (SH)

Niveau d'espèce Operational Taxonomic Unit (OTU) tel que défini dans la base de données UNITE et l'environnement de gestion des séquences, pour les champignons.

Spécimen

Un animal, plante, champignon, etc utilisé comme exemple de son espèce ou de son type pour l'étude ou exposition scientifique.

Archive de Reads de Séquences (SRA)

Dépôt public de données de séquençage à haut débit (NGS) avec des instances opérées par the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EMBL-EBI), et la banque de données ADN du Japon (DDBJ). Inclut à la fois les séquences brutes (sans application du denoising) et sequence alignments. L'une des trois composantes de the European Nucleotide Archive (ENA), et précédemment connue sous le nom d'archive de lectures courtes.

Séquençage de cibles capturées

Séquençage de fragments d'ADN isolés par des sondes d'hybridation.

UNITE

UNITE est un environnement basé sur le web de gestion des séquences, centré sur la région nucléaire ITS du ribosome eucaryote. Toutes les séquences publiques sont regroupées en hypothèses d'espèces (SH), auxquelles sont attribuées des DOI uniques. Un service de mise en correspondance des SH produit divers éléments d'information, notamment les espèces présentes dans les échantillons d'eDNA, si ces espèces sont potentiellement de nouvelles espèces non décrites, les autres études dans lesquelles elles ont été récupérées, si les espèces sont étrangères à une région et si elles sont menacées. Les DOI sont connectés au backbone taxonomique de la plateforme PlutoF et de GBIF, de sorte qu'ils sont accompagnés d'un nom de taxon lorsqu'il est disponible. Les données utilisées dans UNITE sont hébergées et gérées dans PlutoF. Les données sont représentées par une série de standards, principalement Darwin Core, MIxS, et DMP Common Standard ; un support partiel est disponible pour EML, MCL, et GGBN. PlutoF exporte les données principalement via les formats CSV et FASTA. PlutoF peut également être utilisé pour publier des données dans GBIF (en utilisant le format DwC) et pour préparer des fichiers de soumission GenBank. Il est en outre possible de télécharger des listes d'espèces à partir de vos données et de télécharger votre projet sous forme de document JSON avec les données du projet dans une structure hiérarchique.

OTU Zero rayon (zOTU)

Voir ASV.

Références

- 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 Experiments. 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