Référentiel des séquences génétiques des espèces de France : note pour la mise en place d'un nouvel outil national

Aurélie Lacoeuilhe, Gaël P.J. Denys, Pascal Dupont, Anne-Sophie Archambeau, Bertrand Bed'hom, Thomas Bouix, Julien Brisset, Maxime Cammas, Agnès Dettaï, François Dusoulier, et al.

Référentiel des séquences génétiques des espèces de France : *note pour la mise en place d'un nouvel outil national*

Aurélie Lacoeuilhe, Gaël Denys, Pascal Dupont, Anne-Sophie Archambeau, Bertand Bed'Hom, Thomas Bouix, Julien Brisset, Maxime Cammas, Agnès Dettaï, François Dusoulier, Claire Gachon, Olivier Gargominy, Myriam Gaudeul, Vincent Haÿ, Katia Herard, Yvan Le Bras, Line Le Gall, Noëlie Maurel, Valentin de Mazancourt, Marion Mennesson, Sophie Pamerlon, Rémy Poncet, Nicolas Puillandre, Rodolphe Rougerie, Sarah Samadi, Lucas Sire, Chloé Vinet, Laurent Poncet, Julien Touroult

Février 2025

PATRINAT

Centre d'expertise et de données sur le patrimoine naturel

Un service commun
de l'Office français de la biodiversité,
du Muséum national d'Histoire naturelle,
du Centre national de la recherche scientifique
et de l'Institut pour la recherche et le développement

Responsables de l'étude : Aurélie Lacoeuilhe, Rodolphe Rougerie, Laurent Poncet, Julien Touroult

Contributeurs et experts mobilisés : Gaël Denys, Pascal Dupont, Anne-Sophie Archambeau, Bertrand Bed'Hom, Thomas Bouix, Julien Brisset, Maxime Cammas, Agnès Dettaï, François Dusoulier, Claire Gachon, Olivier Gargominy, Myriam Gaudeul, Vincent Haÿ, Katia Herard, Aurélie Lacoeuihe, Yvan Le Bras, Line Le Gall, Noëlie Maurel, Valentin de Mazancourt, Marion Mennesson, Sophie Pamerlon, Laurent Poncet, Rémy Poncet, Nicolas Puillandre, Rodolphe Rougerie, Sarah Samadi, Lucas Sire, Julien Touroult, Chloé Vinet

Référence du rapport conseillée : Lacoeuilhe A., Denys G., Dupont P., Archambeau A.-S., Bed'Hom B., Bouix T., Brisset J., Cammas M., Dettaï A., Dusoulier F., Gachon C., Gargominy O., Gaudeul M., Haÿ V., Herard K., Le Bras Y., Le Gall L., Maurel N., de Mazancourt V., Mennesson M., Pamerlon S., Poncet R., Puillandre N., Rougerie R., Samadi S., Sire L., Vinet C., Poncet L. & Touroult J., 2025. *Référentiel des séquences génétiques des espèces de France : note pour la mise en place d'un nouvel outil national*. PatriNat (OFB-MNHN-CNRS-IRD), 32 pp. ⟨hal-04931482⟩

PatriNat

Centre d'expertise et de données sur la nature

Cette unité scientifique de l'OFB, du MNHN, du CNRS et de l'IRD regroupe des experts de la biodiversité et de la donnée ainsi que des coordinateurs de programmes nationaux.

Ce centre d'expertise et de données coordonne des programmes d'inventaire et de suivi des écosystèmes, des espèces et des aires protégées, contribue à répertorier les zones clefs pour la conservation de la nature, et produit des référentiels scientifiques et des standards de données. Ces programmes associent de nombreux partenaires nationaux et régionaux et fédèrent les citoyens à travers des observatoires de sciences participatives.

En tant que centre de données, PatriNat développe des systèmes d'information permettant d'organiser, diffuser et faire parler les données pour les politiques publiques (SIB, SINP) en coopération avec les infrastructures de recherche. L'ensemble des informations (de la donnée brute à la donnée de synthèse) est rendu publique dans les portails du service public d'information sur la biodiversité. Des outils et services numériques complètent une offre pour les acteurs de la conservation de la nature.

En tant que centre d'expertise, PatriNat réalise des études fondées sur les informations fiables et pertinentes (publications, données, experts…) pour accompagner les politiques de biodiversité : indicateurs, chiffres clés, Listes rouges, revues systématiques, rapportages, avis scientifiques CITES… PatriNat développe des méthodes et transfère des technologies innovantes pour accompagner les acteurs de la transition écologique.

En savoir plus : www.patrinat.fr

Direction : Laurent PONCET et Julien TOUROULT

Le portail des indicateurs et des informations sur des politiques de biodiversité	Le portail dédié aux espèces, aux habitats, aux espaces naturels et au patrimoine géologique	Le portail des indicateurs, des enjeux et des initiatives sur la biodiversité en Outre-mer
naturefrance.fr	www.inpn.fr	biodiversite-outre-mer.fr

SOMMAIRE

Liste des tableaux et figure

Résumé exécutif

Le référentiel dont il est question dans ce document consiste en un nouvel outil national mettant à disposition des séquences génétiques de qualité maîtrisée pour les espèces présentes dans les territoires français. L'objectif principal de cet outil est d'assurer la robustesse et la fiabilité des liens entre les séquences intégrées au référentiel et les identifications taxonomiques des spécimens associés. Il permettra ainsi d'améliorer la qualité des données de biodiversité (ex. occurrences, écologie des espèces, etc.) produites en employant des techniques moléculaires, aussi bien dans le contexte de travaux de recherche que dans le cadre des politiques publiques. La création de ce référentiel répond à la fois à un enjeu de mutualisation des données entre acteurs, et de qualité des inventaires, des suivis, des évaluations et des recherches, pour, *in fine,* l'amélioration de la prise en compte de la diversité des espèces et de leur conservation.

Pourquoi un référentiel des séquences génétiques des espèces présentes en France ?

- Durant les 20 dernières années, l'utilisation de séquences génétiques pour l'identification des espèces a fait ses preuves, notamment grâce aux approches dites de codes-barres ADN (« DNA barcoding ») qui se sont multipliées dans les territoires et au niveau mondial sur une diversité de taxons (principalement eucaryotes : animaux, plantes, champignons).

- Les approches d'identification moléculaire pour les inventaires de biodiversité telles que l'ADN environnemental (ADNe) et le metabarcoding ADN sont en plein essor, notamment grâce au développement rapide durant les 15 dernières années des méthodes de séquençage à haut-débit.

- Les bases de séquences génétiques internationales disponibles en libre-accès (ex. GenBank, BOLD) présentent une couverture taxonomique, géographique et moléculaire encore fragmentaire et peu représentative de la biodiversité des territoires français.

- Les séquences disponibles dans le domaine public ne sont pas fréquemment associées à des informations précises permettant d'en apprécier et d'en contrôler la qualité. L'existence de nombreuses erreurs d'assignations taxonomiques est une assertion commune au sein de la communauté scientifique qui n'a cependant été ni évaluée ni mesurée.

- L'assignation taxonomique de séquences moléculaires est plus fiable lorsqu'elle est réalisée avec un référentiel de séquences portant sur un périmètre géographique défini.

Objectifs du référentiel

- Améliorer et objectiver la fiabilité des assignations taxonomiques dérivées des analyses d'échantillons environnementaux ou de spécimens par des techniques moléculaires.
- Améliorer la qualité des données d'occurrence provenant d'analyses moléculaires portant sur les espèces présentes dans les territoires français, en faciliter l'accès, l'interopérabilité et leur réutilisation.
- Mutualiser les efforts de construction, de maintenance et d'alimentation des bases de séquences de référence, en proposant un référentiel unique, public et partagé entre les acteurs utilisant des techniques moléculaires.

Le référentiel c'est quoi ? Points clés et périmètre

- Un référentiel national, conçu en lien avec les standards internationaux relatifs aux données de biodiversité, dont les données moléculaires, mettant à disposition des utilisateurs des séquences de référence validées et fiables pour l'identification taxonomique en appui aux politiques publiques et à la recherche, pour des applications telles que les analyses de l'ADNe ou d'autres approches fondées sur des techniques moléculaires (ex. barcoding, metabarcoding).

- Un nouvel outil public, conçu en lien avec les systèmes d'information nationaux et internationaux sur les données relatives à la biodiversité, publié dans le cadre du système d'information sur la biodiversité (SIB)

via ses portails d'information (INPN/NatureFrance) et orienté vers les utilisateurs (chercheurs, gestionnaires, bureaux d'études, etc.) pour :
    - Améliorer la qualité, la fiabilité et le partage des données génétiques ayant fait l'objet d'une assignation taxonomique experte ;
    - Accompagner l'utilisation des données génétiques pour l'inventaire, le suivi et la surveillance de la biodiversité.

- Un référentiel associé à TAXREF, le référentiel national qui liste les organismes vivants, ayant vécu ou à rechercher dans les territoires français et qui constitue la base nomenclaturale du Système d'information de l'inventaire du patrimoine naturel (SINP) pour les données d'observation et de suivi des espèces.

- Un référentiel qui concerne les espèces d'Eucaryotes (dont les unicellulaires), en Hexagone, Corse et Outre-mer, dans tous les milieux (continentaux et marins). Le référentiel de séquences cible les espèces natives ou introduites, présentes, ayant été présentes ou à rechercher dans les territoires français. Les groupes prioritaires ciblés sont ceux concernés par des politiques publiques et des programmes de recherche utilisant l'ADN et l'ADNe, notamment les poissons continentaux et marins, les mammifères marins, les élasmobranches, les mammifères terrestres et amphibies, les amphibiens et bivalves d'eau douce, les insectes aquatiques, les insectes pollinisateurs ou saproxyliques, la faune et la fonge du sol, les espèces exotiques envahissantes (EEE), les espèces non indigènes (ENI) (présentes ou émergentes).

- Un référentiel qui s'appuie sur une stratégie de développement dédiée, intégrant et qualifiant des séquences préexistantes, et visant à représenter la variabilité intraspécifique, notamment géographique, en mettant à disposition des séquences provenant de différentes localités de l'aire de distribution des espèces.

1. Introduction

Alors que les compétences naturalistes et taxonomiques expertes se raréfient, et que pour certaines spécialités il devient pratiquement impossible de trouver des experts capables d'identifier morphologiquement les espèces (ex. Pentinsaari *et al.,* 2020 ; Engel *et al.,* 2021 ; Stroud *et al.,* 2022), l'essor récent des approches moléculaires d'identification des espèces (ex. DNA barcoding (Hebert *et al*., 2003)) et de leurs applications (ex. metabarcoding, ADN environnemental) représente une opportunité de pallier à ce « déficit taxonomique » (Hebert *et al*., 2005 ; Engel *et al.,* 2021). Les sources de données sur la biodiversité sont désormais multiples : expertise, imagerie satellitaire et aéroportée, approches moléculaires, *etc*. La production de données d'observation et de suivi des espèces basée sur l'analyse de données moléculaires est en très forte augmentation depuis quelques années (Makiola *et al.,* 2020), et les cas d'usage se multiplient (Dejean *et al.,* 2012 ; Taberlet, *et al.,* 2012 ; Valentini *et al.,* 2016 ; Hongsanan, *et al.,* 2018 ; Kelly *et al.,* 2019 ; Agostinetto *et al.,* 2022 ; Arjona *et al.,* 2022 ; Waterhouse, *et al.,* 2022).

Les techniques moléculaires constituent ainsi une approche complémentaire de plus en plus usitée, notamment pour des applications taxonomiques (*ex.* codes-barres ADN ou « DNA barcoding » (voir glossaire) pour l'identification et la délimitation des espèces, la description d'espèces nouvelles)) ou pour l'analyse de communautés et l'évaluation et la gestion de la biodiversité et des services écosystémiques via l'assignation taxonomique de séquence de spécimens ou d'ADNe à l'aide d'une base de séquences connues (ex. utilisation de l'ADN environnemental [ADNe] ; Rourke *et al.*, 2022 ; Heuertz *et al.*, 2023).

L'identification moléculaire permet une standardisation du processus et dans certains cas une augmentation de la résolution taxonomique (c'est-à-dire le fait d'assigner l'échantillon à un rang taxonomique plus précis dans le cas de taxons difficilement différenciables et identifiables morphologiquement). Des approches à haut-débit (Porter & Hajibabaei, 2018 ; Buchner *et al*., 2021), aussi bien dans le traitement en amont des échantillons que dans les méthodes de séquençage, rendent l'identification moléculaire généralement beaucoup plus rapide et moins onéreuse que le recours à l'examen morphologique des spécimens (mais voir par ex. Behrens-Chapuis *et al*., 2021). Dans tous les cas, la qualité des résultats obtenus dépend cependant de l'existence de bases de référence moléculaires complètes et fiables pour les groupes taxonomiques visés. À défaut, l'assignation taxonomique risque d'être incertaine, voire fausse (voir l'Annexe 1 pour quelques éléments techniques supplémentaires). Les erreurs d'assignation taxonomique peuvent entraîner une cascade de problèmes, elles ont en particulier un impact dans le cadre des évaluations environnementales (Bortolus, 2008), de la gestion de la biodiversité (Dudgeon *et al.,* 2006 ; Browett *et al.,* 2020), de la santé publique et des personnes (Fuchs *et al.,* 2011 ; Mezzasalma *et al.,* 2017 ; Giusti *et al.,* 2021 ; Nithaniyal *et al.,* 2021), ou de la production industrielle (Bottero & Dalmasso, 2011 ; Faunce, 2011 ; Collins *et al.,* 2012 ; Faunce *et al.,* 2015).

Les données de biodiversité issues des méthodes moléculaires sont acquises grâce à des prélèvements faits directement dans l'environnement (ex. échantillons d'eau, de sol, d'air, *etc*.), ou à partir d'un organisme (ex. feuille, appendice, morceau de nageoire, *etc*.) ou de multiples organismes simultanément (ex. insectes collectés par piégeage). Les séquences (un ou plusieurs loci) d'acides nucléiques extraits (ADN ou ARN) sont ensuite comparées et mises en correspondance avec des séquences de référence portant sur la ou les mêmes régions génomiques, et ainsi rattachées à un taxon. Toutefois, des travaux récents montrent d'une part, que de nombreuses séquences ne peuvent être assignées à une espèce en raison de l'absence de séquence génétique associée à l'espèce en question, ou de leur non-disponibilité dans le domaine public (Zafeiropoulos *et al.,* 2021), et d'autre part, que la disponibilité d'une séquence mise en correspondance avec un taxon ne préjuge pas de la justesse de l'information taxonomique associée (Gold *et al.,* 2021 ; Hleap *et al.,* 2021). Ces travaux mettent en exergue le besoin de consolidation de la chaine de production des séquences de référence, ainsi que de la validation et de la diffusion des données d'occurrence d'espèces produites par des méthodes moléculaires. Le cycle de la donnée fondé sur l'assignation moléculaire bénéficierait notamment d'un référentiel de séquence validé qui permettrait l'accès à des séquences de référence fiables et liées à des spécimens identifiés à l'espèce selon l'état des connaissances taxonomiques (à date), et révisables. Il ressort

également de ces travaux que la qualité des données ainsi obtenues est améliorée quand la base de référence utilisée correspond au périmètre géographique de l'étude, c'est-à-dire une base régionale, associée à une base de référence plus générale et accessible (ex. voir 4.1) (Bourret *et al.,* 2023).

La première partie de cette note explicite les besoins, les enjeux et les objectifs relatifs au référentiel de séquences. La seconde présente l'environnement scientifique et technique du projet, et la troisième propose un état des lieux des différentes ressources existantes sur lesquelles l'élaboration du référentiel pourra s'appuyer. Les principaux moyens techniques nécessaires à la réalisation d'un référentiel permettant l'accès à des données qualifiées de séquences génétiques sont ensuite exposés. Ces moyens concernent l'environnement technique du référentiel avec :

1. l'utilisation de standards pour la gestion des données ;
2. une définition des périmètres taxonomiques et génétiques couverts par le référentiel ;
3. les critères de qualité des données constitutives du référentiel (relatifs à leur acquisition et à leur gestion notamment).

Pour chaque type de besoin des propositions sont faites en s'appuyant, lorsque cela est possible, sur l'existant. La dernière partie est consacrée à la proposition d'une organisation (structurelle et fonctionnelle) pour l'élaboration du référentiel de séquences.

2. Besoins, enjeux et objectifs d'un référentiel de séquences génétiques

Lorsque les méthodes moléculaires sont utilisées pour l'identification des espèces, la précision et la qualité des assignations taxonomiques reposent respectivement sur l'adéquation des marqueurs génétiques ou génomiques représentés dans les bases de référence utilisées, et sur la fiabilité des assignations taxonomiques des spécimens dont sont issus les séquences. L'accès à des séquences de référence pertinentes et fiables, associées à un référentiel taxonomique à jour, est donc un point essentiel pour améliorer la fiabilité des données d'observation et de suivi issues des techniques moléculaires. La création d'un référentiel de séquences génétiques, qui relève du système d'information sur la biodiversité (SIB), et qui associe les principales parties prenantes – dont les gestionnaires des collections naturalistes du Muséum national d'Histoire naturelle et d'autres collections – constitue de fait un enjeu important parallèlement à l'augmentation de l'utilisation des techniques moléculaires. Ce référentiel doit concerner l'ensemble des taxons présents sur les territoires français ou pour lesquels la France doit réaliser une veille (ex. espèce exotique envahissante (EEE), espèce non indigène (ENI)).

Un référentiel des séquences génétiques utilisées pour l'identification des espèces répond à l'enjeu de qualité des données produites à partir de l'analyse moléculaire d'échantillons de tissus ou environnementaux, aussi bien dans le contexte de travaux de recherche que dans le cadre des politiques publiques relatives à la biodiversité. Les finalités sont de :

- Améliorer et objectiver la **fiabilité des assignations taxonomiques** dérivées des analyses d'échantillons environnementaux ou de spécimens par des techniques moléculaires ;
- Améliorer la **qualité des données d'occurrence provenant d'analyses moléculaires** portant sur les espèces présentes dans les territoires français, en **faciliter l'accès, l'interopérabilité et leur réutilisation** ;
- **Mutualiser les efforts de construction, de maintenance et d'alimentation des bases de séquences de référence**, en proposant un référentiel public et partagé.

Le référentiel de séquences vise ainsi à **rendre accessibles gratuitement des séquences génétiques, vérifiées et fiables** (répondant à certaines exigences de qualité, voir parties 4 et 5), **pour les espèces de France (Hexagone, Corse et Outre-mer**) **en s'inscrivant dans le paysage existant** (grâce à une interopérabilité entre les systèmes d'information (SI) et outils).

3. Contexte français des infrastructures numériques et outils relatifs aux données sur la biodiversité

L'élaboration d'un référentiel de séquences doit être entreprise en ayant au préalable une bonne connaissance de l'environnement scientifique et technique actuel. Une première analyse non exhaustive fait ressortir au moins quatre structurations importantes.

3.1. DiSSCo

DiSSCo[1] (**Distributed System of Scientific Collections**) est une infrastructure de recherche européenne pour les collections relatives aux sciences naturelles. Elle vise à harmoniser les pratiques en matière de gestion numérique des spécimens afin que les données soient facilement trouvables, accessibles, interopérables et réutilisables (principes FAIR[2], Wilkinson *et al.,* 2016). L'un des enjeux fondamentaux de l'infrastructure est de générer un écosystème cohérent de données permettant de tendre vers la notion de « spécimen digital » ou « spécimen étendu ».

L'infrastructure de recherche **Récolnat** (**Réseau national des collections naturalistes**)[3] pilotée par le MNHN[4] est le point nodal de DiSSCo pour la France. Dans le cadre de DiSSCo, plusieurs groupes travaillent sur la problématique des standards de données en lien avec le **TDWG** (Taxonomic Databases Working Group historiquement, devenu le **Biodiversity Information Standards**)[5]. Ce qui est important à prendre en compte dans l'approche de DiSSCo pour le référentiel dont il est question ici, c'est le concept de **"spécimen étendu" ou "spécimen digital"** qui concerne l'ensemble des objets numériques associés au **"spécimen physique"** (Walton *et al.,* 2020), dont les séquences génétiques produites à partir du spécimen physique font partie. La standardisation de ces données a fait l'objet de travaux spécifiques (Droege *et al*. 2016 ; Corales & Astrin, 2023).

3.2. MOLECULAIRE et JACIM

Le MNHN dispose déjà d'un outil de gestion interne des séquences, la base de référence appelée MOLECULAIRE qui permet de faire le lien entre des séquences nucléotidiques et leurs spécimens[6] enregistrés dans les collections du MNHN (ou ailleurs). Cette base de données a initialement été développée dans le cadre du projet **MARBOL (MARine Barcode Of Life**) indépendamment de DISSCo mais pourrait être mobilisable dans ce contexte. Elle est utilisée pour assurer la traçabilité et la qualité des données produites à partir de spécimens conservés dans certaines collections du MNHN afin d'en garantir la robustesse. Elle permet notamment la gestion de collections de tissus, d'extraits d'ADN et des données de laboratoire associées aux séquences produites (voir Annexe 2 pour plus d'informations).

JACIM est une application développée pour l'informatisation et la gestion des collections naturalistes du MNHN, dont ARTHROTHER (arthropodes terrestres), COLLVERS (vers parasites), GICIM (ichtyologie ;

[1] https://www.dissco.eu/fr-fr/
[2] *Findable* / Facilement trouvable ; Accessible ; Interoperable ; *Reusable* / Réutilisable
[3] https://www.recolnat.org/fr/gis-recolnat
[4] https://www.mnhn.fr/fr/dissco
[5] Normes internationales d'information sur la biodiversité, https://www.tdwg.org/
[6] L'existence d'un spécimen en collection est un élément clé pour une base de séquences de référence, notamment pour revenir sur l'identification morphologique si le résultat issu de l'analyse moléculaire n'est pas concordant.

Pruvost *et al.*, 2023), INVMAR (invertébrés marins), MYCOBASE (fonge), SONNERAT (herbier) et ZAC (mammifères et anatomie comparée). Un projet de refonte de JACIM a démarré en 2023.

Le référentiel de séquences s'inscrira donc dans la logique mise en place au travers de ces outils du MNHN, mais également à l'échelle nationale et à l'échelle de l'Europe.

3.3. SIB

Le **Système d'Information sur la Biodiversité (SIB)**[7] est un des systèmes d'information de l'État. Il est porté par l'OFB et coordonné par PatriNat (OFB-MNHN-CNRS-IRD). C'est un dispositif qui vise à fédérer l'ensemble des données issues des politiques publiques portant sur la biodiversité – chaque politique publique étant identifiée par un système d'information qui lui est propre, appelé "système d'information métier" (ex. le système d'information de l'inventaire du patrimoine naturel, SINP). Le SIB cherche à donner un cadre de cohérence commun pour ces SI métiers pour rendre les données plus interopérables et réutilisables par tous. Pour cela le SIB a mis en place un centre d'administration du référentiel (CARET) afin de mettre à disposition des référentiels interopérables. Par exemple, le référentiel TAXREF (Gargominy *et al.*, 2022) concernant les noms scientifiques des taxons présents en France, produit dans le cadre du SINP et diffusé sur l'INPN, est un des référentiels promus dans le cadre du SIB. Le **référentiel de séquences de référence s'inscrira dans ce cadre**.

3.4. PNDB

Le **Pôle National de Données de Biodiversité (PNDB)**[8] est une infrastructure de recherche portée par le MNHN. Il a pour objectif de mettre à disposition des communautés de recherche un ensemble cohérent d'outils et de services pour la description, l'accès, la validation, l'analyse et la réutilisation des données de biodiversité, dans le but de prendre en compte la biodiversité sur le temps long (depuis les origines de la vie), à tous les niveaux d'organisation biologique (de la molécule à l'anthropo-écosystème). Dans la structuration mise en place, le PNDB a établi des liens en termes de services avec le SIB. **Le référentiel de séquences pourrait ainsi à la fois être utilisé dans le cadre du PNDB pour accompagner la communauté de la recherche et être alimenté avec les séquences produites par les acteurs fédérés au sein du PNDB.** N'ayant pas vocation à produire un référentiel, l'organisme gestionnaire du PNDB pourra accompagner le projet d'élaboration du référentiel de séquences (processus d'alimentation, utilisation et valorisation) et assurer les liens avec les communautés de recherche.

3.5. IFB

L'**institut français de bioinformatique (IFB)**[9] est une infrastructure de recherche, point nodal de l'infrastructure européenne de bioinformatique **ELIXIR**[10]. L'IFB permet aux scientifiques travaillant dans le domaine des sciences de la vie d'accéder à des ressources et services multiples telles que des bases de données, des outils logiciels, du matériel de formation, du stockage cloud et des clusters de calcul. L'objectif est de coordonner ces ressources afin de partager les compétences et d'aller vers des pratiques plus efficientes. Dans ce cadre, un certain nombre de **recommandations ont été produites pour améliorer les interconnexions entre le domaine de la génomique** et les autres domaines thématiques relatifs à la biodiversité (Waterhouse *et al.,* 2022). **Le référentiel de séquences bénéficiera ainsi des bonnes pratiques mises en œuvre dans le cadre de l'IFB.**

---

[7] https://naturefrance.fr/systeme-information-biodiversite
[8] https://www.pndb.fr/
[9] https://www.france-bioinformatique.fr/
[10] https://elixir-europe.org/ Elixir est une infrastructure européenne des sciences de la vie, réunissant des scientifiques de 23 pays et plus de 250 instituts de recherche, qui permet aux chercheurs d'accéder aux données des sciences de la vie et de les analyser.

4. Dispositifs gérant et/ou donnant accès à des séquences de référence

4.1. Principales bases de données internationales de séquences moléculaires

La majorité des revues scientifiques demandent que les données moléculaires utilisées dans les publications soient rendues publiques, souvent sans imposer de plateforme en particulier. Certaines demandent par exemple un enregistrement préalable avec un numéro d'accession, au niveau d'un institut national américain : le NCBI (National Center for Biotechnology Information), qui centralise les séquences au niveau de la base de données GenBank (Benson *et al.,* 2006). Mais d'autres options existent aussi, les séquences peuvent être déposées sur une plateforme d'un institut européen : EMBL-EBI, sur des sites comme *Dryad* ou *figshare*, ou sur des serveurs privés, l'important étant que l'information soit accessible. L'INSDC (International Nucleotide Sequence Database Collaboration)[11] est une initiative de longue date entre EMBL-EBI, NCBI et du National Institute of Genetics (NIG), et qui assure un échange régulier de données de couverture mondiale entre les 3 bases suivantes :

- L'**European Nucleotide Archive** (**ENA**)[12] est portée par l'**Institut Européen de Bioinformatique du Laboratoire européen de biologie moléculaire** (**EMBL-EBI)** qui est une organisation de recherche intergouvernementale financée par plus de 20 États membres et située sur le Wellcome Genome Campus près de Cambridge au Royaume-Uni. En décembre 2023, l'ENA contenait 20 000 milliards de bases de nucléotides agrégées dans plus de 3,2 milliards de séquences[13].
- **GenBank** est géré par le **NCBI.** La version de décembre 2023 contenait 2 570 milliards de bases pour 249 millions de séquences[14]. Toutefois dans la mesure où le NCBI a séparé schématiquement les données de GenBank et les données de WGS (Whole Genome Sequencing) qui représente 24 652 milliards de bases pour 2,86 milliards de séquences, le total représente donc 3,112 milliards de séquences pour 27 222 milliards de bases[15]. Il peut y être déposé tous types de données moléculaires (du génome complet aux zones non codantes). Par ailleurs, les séquences doivent faire au moins **200** paires de bases (sans les amorces) pour être déposées dans GenBank.
- La **DNA Data Bank of Japan** (**DDBJ**)[16] fait partie du National Institute of Genetics (NIG) de la Research Organization of Information and Systems[17] à Mishima au Japon.

Ces trois bases sont ainsi pour partie des miroirs les unes des autres et contiennent aussi des informations qui leur sont propres. Si GenBank semble aujourd'hui la base la plus complète disponible, certaines séquences n'y sont pas disponibles mais peuvent l'être dans d'autres bases. De plus, aucune de ces trois bases, à la différence de BOLD, UNITE ou SILVA par exemple (voir plus bas), n'a pour objectif de consolider l'assignation taxonomique des séquences obtenues à partir d'organismes ou d'échantillons environnementaux – ce sont des entrepôts de séquences, qui bien que comportant des outils intégrés de comparaison de séquences, ne prétendent pas fournir et servir directement à l'expertise taxonomique.

---

[11] https://www.insdc.org/
[12] https://www.ebi.ac.uk/ena/browser/home
[13] https://www.ebi.ac.uk/ena/browser/about/statistics
[14] https://www.ncbi.nlm.nih.gov/genbank/statistics/
[15] ENA et GenBank contiennent presque le même nombre de séquences dans la mesure où les bases sont en bonne partie synchronisées mais chacune a aussi des bases spécifiques. Par ailleurs les données de type SRA (Sequence Read Archive) ne sont visiblement pas comptées côté NCBI. Pour l'ENA en 2023, cela représente 28,7 millions de runs pour 54,1 Po de données et pour GenBank, il y avait 9 millions de runs pour 12 Po de données en 2019.
[16] https://www.ddbj.nig.ac.jp/index-e.html
[17] https://www.nig.ac.jp/nig/

De plus, les bases en accès libres présentent l'avantage de centraliser beaucoup de données et d'être accessibles, mais l'inconvénient de contenir certaines erreurs dans les séquences (ex. des séquences qui ne sont pas associées à un référentiel taxonomique à jour ou qui ne sont pas associées à un spécimen) (Benson *et al.,* 2018 ; Renner *et al.*, 2024 ; van den Burg & Vieites, 2023). Il n'y a pas systématiquement d'étape de validation des métadonnées pour diffuser des séquences, et le contrôle qualité et la mise à jour des données et des métadonnées sont très limités (Benson *et al.,* 2018), comme cela a été montré pour les données génétiques de champignons (Nilsson *et al.*, 2006), d'oiseaux (van den Burg & Vieites 2023) ou de poissons téléostéens (Hinsinger *et al.,* 2015).

Le Barcode of Life Datasystems (**BOLD** ; boldsystems.org ; Ratnasingham & Hebert, 2007) est une plateforme d'acquisition, de stockage, d'analyse et de publication de données génétiques constituant une base de référence moléculaire à vocation d'identification des espèces. BOLD a été développé et est maintenu dans le cadre des projets portés par le consortium **iBOL** (international Barcode of Life, ibol.org). Créé en 2018, iBOL porte aujourd'hui le projet **BIOSCAN** visant à poursuivre le développement des bases de référence ainsi que l'adoption de standards et l'usage des outils génétiques dans la mise en place d'un système mondial de surveillance de la biodiversité. BOLD a été conçu et est maintenu par une équipe dédiée à l'Université de Guelph au Canada. Un miroir européen est en cours de lancement dans le cadre d'une collaboration entre iBOL et le projet Biodiversity Genomics Europe (BGE) qui porte le volet européen de BIOSCAN (BIOSCAN Europe ; bioscaneurope.org). Au 17 janvier 2024, BOLD héberge plus de 20 millions de codes-barres ADN représentant environ 259 000 espèces animales, 72 000 espèces végétales et 25 000 espèces fongiques ; 16 millions de ces codes-barres ADN sont accessibles publiquement, représentant plus d'un millions d'espèces selon l'utilisation d'un registre automatique d'assignation des séquences à des groupements (clusters) génétiques (Barcode Index Number (BINs) ; Ratnasingham & Hebert, 2013), implémenté dans BOLD afin de représenter un proxy des espèces indépendamment de leur identification taxonomique. Cette base de référence est très largement centrée sur les codes-barres ADN standards (DNA barcoding) (voir glossaire). Il est cependant possible d'y intégrer d'autres marqueurs génétiques. Le moteur d'identification intégré à BOLD peut rechercher des séquences faisant au minimum 100 paires de bases.

Une autre plateforme de stockage et d'analyse baptisée mBrave (mbrave.net a également été développée sous le pilotage du consortium iBOL) représente un volet metagénomique de BOLD, permettant l'intégration, la visualisation, le stockage, l'analyse et la publication de données issues des nouvelles technologies de séquençage. Cette plateforme permet d'utiliser des bases de références pour des analyses de metabarcoding en lien direct avec les données hébergées dans BOLD.

En France, **le réseau FrBOL**, visant à fédérer les acteurs français académiques et non académiques (à l'instar, par exemple, de UKBOL au Royaume-Uni ou GBOL en Allemagne) impliqués dans la génération ou l'utilisation des codes-barres ADN, a été initié dès 2019 et lancé officiellement en 2023. Il est piloté par le MNHN dans le cadre du projet BIOSCAN-FrBOL, financé par l'intermédiaire d'une convention de coopération OFB/MNHN.

UNITE est une base de données de séquences d'ADNr (ADN qui code pour de l'ARN ribosomique) pour l'identification des espèces fongiques et qui contient un peu plus de 800 000 séquences du marqueur nucléaire ITS. Cette plateforme fournit un moyen de délimiter, identifier, communiquer et travailler avec des hypothèses d'espèces (Species Hypotheses = SH) basées sur l'ADN. Toutes les séquences ITS fongiques présentes dans les bases de données internationales de séquences (GenBank, ENA, DDBJ) sont regroupées. Toutes les SH reçoivent un nom unique et stable sous la forme d'un DOI (Digital Object Identifier) dans GBIF. Les SH sont reliées à un nom de taxon et à sa classification dans la mesure du possible (phylum, classe, ordre, *etc.*) en prenant en compte les identifications pour toutes les séquences de la SH. Ces séquences sont publiées (unite.ut.ee/repository.php) pour être utilisées par la communauté scientifique dans, par exemple, des recherches de similarité de séquences locales et des processus d'analyse des données issues des

technologies de séquençage à haut débit. Le système et les données sont mis à jour automatiquement à mesure que le nombre de séquences ITS fongiques publiques augmente[18].

Cette liste n'est pas exhaustive, il existe en effet d'autres bases de séquences spécialisées, comme par exemple **SILVA,** une des bases de référence d'ADNr 16S et 18S les plus riches qui est mise à jour très fréquemment (Quast *et al.*, 2013) pour les eucaryotes, les bactéries et les archées notamment. Il existe aussi **PR2** pour les protistes (18S) (Guillou *et al.,* 2013) et pour le phytoplancton, **µgreen-db** (23S) (Djemiel *et al.,* 2020) et **PhytoREF** (16S) (Decelle *et al.,* 2015), **Diat.barcode** pour les Diatomées (Rimet *et al.,* 2019) et **Phytool** (16S, 23S et 18S) (Canino *et al.,* 2021) ou encore **DictDB**, une base dédiée à l'étude des flores digestives d'insectes (Mikaelyan *et al.*, 2015). On peut aussi citer **FishBase** (fishbase.mnhn.fr) pour les poissons et aussi **R-SYST**, un réseau national regroupant une douzaine d'équipes de recherche[19] qui a pour objectif de développer un outil accessible à tous et composé de bases de données regroupant tous les organismes étudiés par l'INRAe (insectes ravageurs, abeilles, parasitoïdes, champignons, *etc*.) *etc*.

Enfin, certaines bases comme GenBank et BOLD proposent non seulement une base de données, accessible via un portail d'accès, mais aussi un outil de comparaison de séquences avec toutes les séquences présentes en base pour une éventuelle identification moléculaire et des API qui permettent de faire des requêtes dynamiques à partir de scripts *via* des serveurs de calcul.

4.2. Implications pour le référentiel de séquences des espèces de France

Les bases internationales présentées au 4.1 sont alimentées par de nombreux projets générant des séquences au niveau mondial. Les données contenues dans ces entrepôts de séquences moléculaires pourront pour partie alimenter le référentiel de séquences des espèces de France (selon le niveau de qualité des séquences, et la possibilité d'administrer, de valider et de réviser la taxonomie associée, ou des métadonnées associées [ex. méthodologie de séquençage, lien avec un spécimen en collection, *etc*.]). Il faut aussi noter que certaines de ces bases de données moléculaires possèdent des outils ou même des équipes dédiées à la validation et la révision des données et des métadonnées.

Elles renferment des séquences issues à la fois des technologies de **séquençage de nouvelle génération (NGS)**[20] **et de séquençage à haut débit**, mais aussi des séquences issues de la technologie **Sanger**[21] ; toutes sont considérées potentiellement pertinentes pour rejoindre le référentiel national, mais l'appréciation de leur qualité et de leur fiabilité nécessitera de faire appel à des critères différents (présence d'électrophérogrammes pour les séquences Sanger, information sur la profondeur de lecture, le degré d'hétéroplasmie pour les séquences NGS, etc.)**.**

Du fait de la masse considérable de séquences déjà rassemblées dans les bases présentées précédemment, dont de nombreuses représentent des spécimens provenant de France, ou des espèces présentes en France, l**e référentiel national s'appuiera en priorité sur les bases de référence existantes. Il cherchera à représenter la diversité intraspécifique en incluant plusieurs séquences par espèce, provenant si possible de différentes localités de l'aire de distribution géographique des espèces. Pour les nombreuses espèces qui sont aussi distribuées hors de France, il s'efforcera de compléter les séquences disponibles par ailleurs pour d'autres pays par d'autres issues de l'intérieur de nos frontières.** Il est ainsi proposé, à titre indicatif, de veiller à rassembler dans la mesure du possible au moins 3 séquences par marqueur et par espèce, dont une *a minima* en France, afin de renforcer la précision et la fiabilité du référentiel.

---

[18] UNITE Community, Abarenkov K (2023). UNITE - Unified system for the DNA based fungal species linked to the classification. PlutoF. Checklist dataset https://doi.org/10.15468/mkpcy3 accessed via GBIF.org on 2023-01-26.

[19] https://www6.inrae.fr/r-syst/Qui-sommes-nous

[20] voir glossaire

[21] voir glossaire

5. Identification et proposition de moyens techniques

5.1. Contributeurs et public cible du référentiel

Le référentiel de séquences impliquera principalement :

- Les équipes de l'OFB et du MNHN impliquées dans l'exploitation métier et la gestion technique du Système d'information sur la Biodiversité (PatriNat), notamment celles en charge du référentiel taxinomique TAXREF et de la base de données de séquences génétiques du MNHN (ISYEB – DGD REVE).
- Les équipes impliquées dans la gestion métier et technique des spécimens de collection (MNHN, Recolnat).
- Les équipes produisant des séquences moléculaires dans un cadre de recherche ou d'expertise ou à partir de spécimens en collection.
- Des experts responsables de la sélection et de la validation des séquences de référence de chaque groupe biologique.
- Les futurs utilisateurs du référentiel de la sphère publique ou privée (gestionnaires, chercheurs, bureaux d'études...).

5.2. Standards de données

Le référentiel de séquences intègrera deux grands ensembles de données :

1. Des données concernant les **séquences génétiques** (composition en nucléotides, longueur, indice de qualité, nombre d'ambiguïtés, *etc.*), leur gestion (identifiant de la séquence, lien url vers la séquence, *etc.*) et les métadonnées associées (méthode de séquençage, amorces, producteur/structure ayant fourni la séquence, référence bibliographique associée le cas échéant, *etc.*) ;
2. Des données concernant les **spécimens dont sont issus les séquences** (identification, origine géographique, sexe, stade de développement, etc.), leur gestion (lieux de dépôt, numéro d'inventaire, etc.) et les métadonnées associées (source, date et méthode de l'identification, statut de type, etc.).

Comme tout référentiel, le référentiel de séquences reposera sur des standards de données adaptés à ces ensembles de données. Il pourra notamment s'appuyer sur les quatre standards suivants et s'assurer de leur interopérabilité :

1. Le standard du Global Genome Biodiversity Network (GGBN) qui permettrait de gérer l'ensemble des types de données précédemment cités (Droege *et al*., 2016) (GGBN Data Standard v1 - GGBN Wiki).
2. le standard du "Minimum information about any sequence" (MIxS) (Yilmaz *et al.,* 2011) (genomicsstandardsconsortium.github.io/mixs).
3. le standard du "Access to Biological Collection Data" (ABCD) (Holetschek *et al.,* 2012) (tdwg.org/standards/abcd).
4. le standard "Collection Descriptions" (Woodburn *et al.,* 2021) (Collection Descriptions - TDWG).

5.3. Périmètres associés aux données de séquences génétiques

5.3.1. Périmètre taxinomique

Il est proposé de limiter le périmètre taxinomique :

- au cadre de la **France (Hexagone, Corse et Outre-mer**), c'est-à-dire aux espèces présentes ou ayant été présentes sur le territoire national (natives et introduites) et aux espèces non présentes en France mais pour lesquelles la France met en place un dispositif spécifique de surveillance. C'est le cas notamment de certaines espèces exotiques envahissantes, ou risquant d'être introduites (EEE, ENI), voire concernées par des importations commerciales.

Étant donné qu'une identification taxinomique n'est robuste que si la base de référence contient une bonne représentation des diversités intra- et interspécifiques, il est proposé d'inclure dans ce référentiel des spécimens d'espèces françaises collectés dans des populations hors de France, voire des espèces proches même si elles ne sont pas présentes en France, afin de pouvoir vérifier qu'il s'agisse bien de l'espèce connue sur le territoire français, et pas d'une autre espèce dont la présence en France n'était pas encore documentée.

- aux **Eucaryotes** (dont les unicellulaires) (les bactéries et archées disposant déjà d'un référentiel organisé). **TAXREF** (Gargominy *et al.,* 2022) est le référentiel taxinomique de l'**Inventaire National du Patrimoine naturel (INPN,** https://inpn.mnhn.fr/**) (SINP/SIB)** pour les taxons de France. Chaque taxon est identifié par un code unique (le CD_NOM). La gestion de ce référentiel est assurée par PatriNat (OFB/MNHN) et prend en compte régulièrement les évolutions taxinomiques et nomenclaturales. Les séquences du référentiel seront associées à une espèce valide « à date ». Il s'agira donc de prévoir un système **d'archivage et de versionnage** du référentiel qui sera mis à jour en fonction de l'évolution des connaissances scientifiques (Weigand *et al.*, 2019).

5.3.2. Périmètre génétique

On entend par périmètre génétique, les marqueurs utilisés pour réaliser l'identification taxinomique.

En fonction des groupes taxinomiques, des applications et des méthodes employées, les marqueurs utilisés ne sont pas les mêmes. En effet, les utilisateurs, selon leurs objectifs, visent généralement un compromis entre le niveau de résolution souhaité (c'est-à-dire le niveau d'identification atteint : genre, espèce, sous-espèce) et leurs capacités à obtenir des données génétiques dans un temps donné et pour un coût contrôlé. L'utilisation de l'ADNe par exemple permet d'échantillonner facilement, sans capture d'organismes, mais impose d'analyser des séquences d'ADN courtes du fait de la dégradation de l'ADN ciblé dans l'environnement ; un référentiel pertinent pour ce type d'application devra donc inclure des séquences de référence incluant ces fragments d'ADN courts. Les approches par barcoding ou metabarcoding visent quant à elles des fragments d'ADN plus longs, ce qui facilite une meilleure résolution de l'identification si le référentiel contient également les séquences correspondantes. Ainsi, chez les animaux les marqueurs les plus utilisés sont les marqueurs mitochondriaux tels que le COI-5P ou Cytb pour des approches de type barcoding ou metabarcoding, mais plutôt les 12S, 16S ou 18S lors de l'utilisation de fragments plus courts pour des analyses d'ADNe par exemple. Pour les plantes, le génome chloroplastique est plus souvent utilisé que les génomes mitochondrial ou nucléaire. Le tableau suivant illustre les marqueurs les plus fréquemment analysés en fonction des grands groupes taxinomiques (Tableau 1).

Groupes tax**inomiques**	**Marqueurs mitochondriaux ou chloroplastiques utilisés**	**Marqueurs nucléaires utilisés**
Animaux	Cytb, COI-5P, 12S, 16S, d-loop	18S, 28S
Plantes	rbcL, matK	ITS1, ITS2
Fonge	COI-5P	ITS1, ITS2
Chromistes et protozoaires	rbcL, COI-5P	ITS, 18S, 28S

Tableau 1 : Exemple de marqueurs les plus utilisés en fonction des grands groupes taxinomiques.

L'adoption par le référentiel de séquences de standards de données inclura idéalement la gestion d'une liste ouverte de marqueurs génétiques s'appuyant sur une nomenclature dédiée et adaptée aux problématiques taxinomiques et/ou méthodologiques (ADNe, barcoding et metabarcoding), en tenant compte des marqueurs les plus fréquemment utilisés pour les inventaires et les suivis (Tableau 1). Nous soulignons notamment la pertinence pour le référentiel des génomes complets d'organelles (mitochondries et chloroplastes) qui incluent de nombreux marqueurs parmi les plus utilisés.

5.3.3. Critères de qualité du référentiel et des données

La conception du référentiel de séquences génétiques requiert de définir un modèle conceptuel de données et un ou des standards de données qui permettent de véhiculer et diffuser en libre-accès les données moléculaires et les informations associées en conformité avec les principes FAIR[22]. A cette fin, il s'agit de définir les informations minimales et optionnelles à inclure dans le référentiel, leurs sources, leurs relations entre-elles, et leur encodage. Ainsi, lors de la phase d'élaboration du modèle conceptuel de données, plusieurs éléments clés devront être définis, tels que ceux portant sur l'utilisation d'un référentiel taxonomique à jour (Weigand et al., 2019 ; Somervuo et al., 2017), la formalisation de règles de standardisation des données et des métadonnées liées aux séquences, et la gestion des relations entre les séquences et les spécimens conservés en collection. Le modèle conceptuel de données devra notamment inclure des informations rendant compte de la qualité intrinsèque des données (DISIC, 2013), et intégrer les six principes suivants : unicité, complétude, exactitude, conformité, intégrité et cohérence.

Concernant la qualité des données qualifiées de « séquences de référence », elle pourra être transcrite en s'appuyant sur un dictionnaire de données prédéfinit, tel que celui présenté dans le Tableau 2.

Nomenclature GenSeq	**Matériel source**
GenSeq-1	Types primaires : holotype, lectotype, néotype, syntype
GenSeq-2	Types secondaires : paratype, paralectotype, *etc*.
GenSeq-3	Topotypes : spécimens ou colonies non-types en collection et provenant de la localité type
GenSeq-4	Spécimens enregistrés ou colonies cultivées en collection
GenSeq-5	Photo seule
GenSeq-6	Pas d'information permettant de revenir sur le spécimen séquencé.

Tableau 2 : Exemple de la nomenclature GenSeq des séquences moléculaires d'après Chakrabarty et al. (2013).

[22] *Findable* / Facilement trouvable ; Accessible ; Interoperable ; *Reusable* / Réutilisable

5.4. Métadonnées associées aux données de séquences

Les données de séquences sont associées à un ensemble de métadonnées spécifiques relatives :

1. **au spécimen qui a permis de produire la séquence** : son identifiant unique pouvant être un numéro d'inventaire en collection qui permet de faire un lien avec une base des collections du MNHN ou d'autres collections, et le statut du spécimen (ex. spécimen type). La donnée d'observation du spécimen (lieu et date d'observation, identité du collecteur et déterminateur, organisme gestionnaire et lieu de stockage, *etc.*) est associée à l'identifiant unique de la donnée (ex. identifiant SINP ou identifiant GBIF).
2. **à l'identification taxinomique**, nécessitant une expertise (ex. identification morphologique du spécimen, ou assignation taxinomique moléculaire), et accompagnée de certaines métadonnées importantes : identité de l'expert.e, date de l'identification, méthode d'identification, etc.)
3. **au protocole ayant permis l'obtention de la séquence**. Il est important de disposer d'informations concernant chacune des étapes de laboratoire (extraction, amplification et séquençage) car elles peuvent influencer le résultat obtenu (marqueur, amorces, méthodes de séquençage, producteur/structure ressource, la référence bibliographique associée le cas échéant, *etc.*). Ces informations sont généralement disponibles dans les articles scientifiques ayant généré ces séquences, dans les bases de données de séquences, ou via les systèmes de gestion de l'information de laboratoire (cahiers de laboratoire, cahiers numériques, LIMS (« Laboratory Information Management Systems »).
4. à la **base de référence dans laquelle la séquence est déposée** (numéro d'accession unique, date version de la base de référence, etc.).

L'ensemble de ces informations confère à la donnée de séquence un gage de qualité notamment sur l'identification ou la possibilité de revenir sur le spécimen pour réviser l'assignation taxonomique, et son accessibilité en libre accès.

Le référentiel pourra s'appuyer sur les recommandations du **Genomic Standards Consortium** (GSC ; gensc.org) (voir aussi concernant les métadonnées : Field *et al.,* 2008 ; Yilmaz *et al.,* 2011).

6. Proposition d'une préfiguration pour le référentiel

6.1. Choix du type de référentiel

Le référentiel consistera en un **système de diffusion secondaire** et de **qualification de séquences** issues de bases de données moléculaires existantes. Au regard de l'existant (voir 4.1.), nous préconisons d'établir un **référentiel de consolidation**[23] qui s'appuierait sur des applications sources (Figure 1). Le périmètre fonctionnel du référentiel concerne des séquences génétiques caractérisant des taxons eucaryotes de France (Hexagone, Corse et territoires d'Outre-Mer) ou pertinents pour le pays. Par "application source", on entend tout système mettant à disposition des données pouvant alimenter le référentiel. Le choix d'héberger au sein du référentiel la totalité ou une partie des données et métadonnées liées aux séquences devra être l'objet d'une réflexion menée lors de la phase de conception de celui-ci, avec l'aide des organes de gouvernance (voir 6.2). Il sera notamment important de considérer les questions liées à la gestion des **droits des séquences** provenant des différentes plateformes, ou encore liées à la réglementation en vigueur (ex. réglementations en vigueur en matière de prélèvement, transport ou analyse des échantillons des organismes, notamment, la réglementation relative au bien-être animal, les statuts de protection des espèces et l'accord de Nagoya concernant les règles de diffusion, en particulier dans les territoires d'Outre-mer qui restent souverains de leurs données biologiques, même si celle-ci s'applique aux producteurs primaires des séquences.

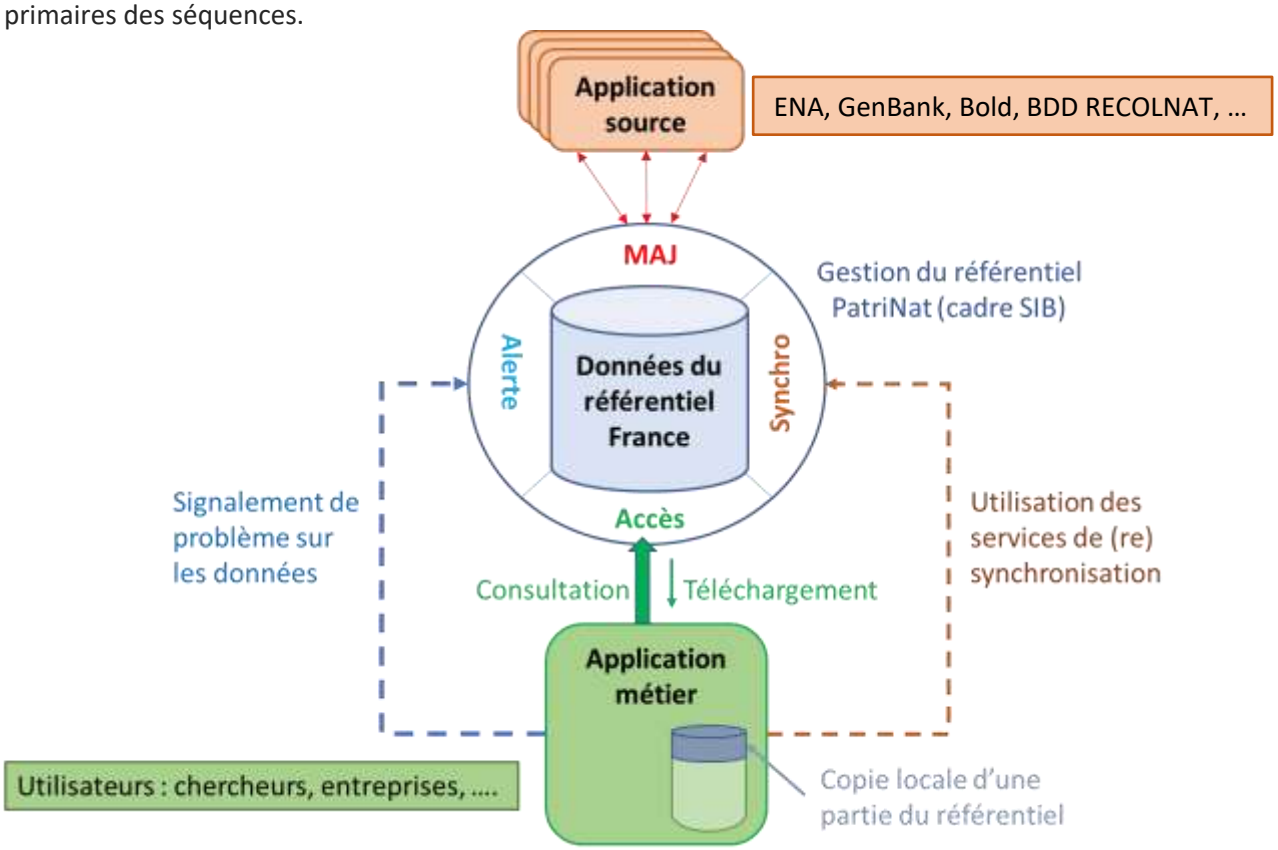

Figure 1 : Schéma conceptuel proposé pour le référentiel

[23] D'après Direction interministérielle des systèmes d'information et de communication. Cadre Commun d'Architecture des Référentiels de données Complément n°2 au Cadre Commun d'Urbanisation du Système d'Information de l'État version 1.0 du 18/12/2013. https://ged.ofb.fr/share/s/7q_XnMpfQceg4kFzgPHpPA

6.2. Gouvernance du référentiel

Le référentiel de séquence est intégré au Système d'Information sur la Biodiversité (SIB), sa gouvernance s'appuie sur trois instances, décrites dans le Tableau 3.

Comités	Rôles	Composition
Comité de pilotage	• statue sur les stratégies et démarches d'acquisition de données, de validation et de diffusion des données et des services associés au référentiel ; • arbitre les propositions du comité technique.	Des représentants de l'OFB (DSUED + Direction de la Recherche et Appui Scientifique (DRAS)), du MNHN (DGD-REVE, direction des collections, départements scientifiques, UAR DOHNEE, DINSI, DIREC) et de PatriNat (Membres du CODIR et chef.fe de programme ADNe), un.e représentant.e du comité d'utilisateurs, un.e représentant.e du/des comité.s technique.s et un.e/des responsables scientifiques. Il pourra être étendu à des structures comme l'Institut Français de Bioinformatique (IFB).
Comité d'utilisateurs	• identifie les besoins ; • conseille sur les réponses techniques en matière d'utilisation (API, *etc.*).	Acteurs dépendants de l'utilisation du référentiel, appartenant à la sphère publique ou privée (représentants de collectifs de chercheurs, des labos, des départements d'universités / autres établissements de recherche, acteurs privés, *etc.*).
Un ou plusieurs comités techniques	• propose des dispositifs applicatifs autour de l'acquisition, la gestion (dont le stockage) et la diffusion des données (développements informatiques, API, serveurs, *etc.*) ainsi qu'autour des services (gestion des DOI, *etc.*).	Inclus notamment les organismes et les acteurs producteurs de séquences de référence

Tableau 3 : Comités envisagés pour la gouvernance du référentiel national de séquences génétiques

7. BIBLIOGRAPHIE

Agostinetto, G., Brusati, A., Sandionigi, A., Chahed, A., Parladori, E., Balech, B., Bruno, A., Pescini, D. & Casiraghi, M. 2022. ExTaxsI: an exploration tool of biodiversity molecular data. *GigaScience*, 11 : giab092.

Arjona, Y., Arribas, P., Salces-castellano, A., López, H., Emerson, B. & Andújar, C. 2022. Metabarcoding for biodiversity inventory blind spots: A test case using the beetle fauna of an insular cloud forest. *Molecular Ecology*, 32(23): 6130-6146.

Behrens-Chapuis, S., Herder, F. & Geiger, M. F. 2021. Adding DNA barcoding to stream monitoring protocols – What's the additional value and congruence between morphological and molecular identification approaches?. *PLoS ONE*, 16(1): e0244598.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. 2006. GenBank. *Nucleic Acids Research*, 34(supl.1): D16-D20.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., & Sayers, E. W. 2018. GenBank. Nucleic acids research, 46(D1), D41-D47.

Bortolus, A. 2008. Error Cascades in the Biological Sciences: The Unwanted Consequences of Using Bad Taxonomy in Ecology. *AMBIO: A Journal of the Human Environment*, 37(2): 114-118.

Bottero, M. T., & Dalmasso, A. 2011. Animal species identification in food products: evolution of biomolecular methods. *The Veterinary Journal*, *190*(1): 34-38.

Bourret, A., Nozères, C., Parent, E. & Parent, G. 2023. Maximizing the reliability and the number of species assignments in metabarcoding studies using a curated regional library and a public repository. *Metabarcodingmics*, 7: 37-49.

Browett, S. S., O'Meara, D. B., & McDevitt, A. D. 2020. Genetic tools in the management of invasive mammals: recent trends and future perspectives. *Mammal Review*, 50(2): 200-210.

Buchner, D., Macher, T.H., Beermann, A.J., Werner, M.T., Leese, F. 2021. Standardized high-throughput biomonitoring using DNA metabarcoding: Strategies for the adoption of automated liquid handlers. *Environ Sci Ecotechnol,* 8**,** 100122.

Canino A., Bouchez A., Laplace-Treyture C., Domaizon I. & Rimet F., 2021. Phytool, a ShinyApp to homogenise taxonomy of freshwater microalgae from DNA barcodes and microscopic observations. Metabarcoding and Metagenomics 5, e74096.

Chakrabarty, P., Warren, M., Page, L. & Baldwin, C. 2013. GenSeq: An updated nomenclature and ranking for genetic sequences from type and non-type sources. *ZooKeys*, 346: 29-41.

Collins, R., Armstrong, K., Meier, R., Yi, Y., Brown, S., Cruickshank, R., Keeling, S. & Johnston, C. 2012. Barcoding and Border Biosecurity: Identifying Cyprinid Fishes in the Aquarium Trade. *PLoS ONE*, 7(1): e28381.

Corales, C. & Astrin, J.J. 2023. *Biodiversity Biobanking – a Handbook on Protocols and Practices*. PENSOFT, Sofia, Bulgaria, 271 pp.

Decelle J, Romac S, Stern RF, Bendif el M, Zingone A, Audic S, Guiry MD, Guillou L, Tessier D, Le Gall F, Gourvil P, Dos Santos AL, Probert I, Vaulot D, de Vargas C, Christen R. 2015. PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. Mol Ecol Resour. 2015 Nov;15(6):1435-45.

Dejean, T., Valentini, A., Miquel, C., Taberlet, P., Bellemain, E., & Miaud, C. 2012. Improved detection of an alien invasive species through environmental DNA barcoding: the example of the American bullfrog *Lithobates catesbeianus*. Journal of applied ecology, 49(4): 953-959.

Direction interministérielle des systèmes d'information et de communication (DISIC), 2013. Cadre Commun d'Architecture des Référentiels de données - Complément n°2 au Cadre Commun d'Urbanisation du Système d'Information de l'État version 1.0 du 18/12/2013. 48p.

Djemiel, C., Dequiedt, S., Karimi, B., Cottin, A., Girier, T., El Djoudi, Y., Maron, P.A. and Ranjard, L., 2020. µgreen-db: a reference database for the 23S rRNA gene of eukaryotic plastids and cyanobacteria. Scientific Reports, 10(1), 5410

Droege, G., Barker, K., Seberg, O., Coddington, J., Benson, E., Berendsohn, W., Bunk, B., Butler, C., Cawsey, E., Deck, J., Döring, M., Flemons, P., Gemeinholzer, B., Güntsch, A., Hollowell, T., Kelbert, P., Kostadinov, I., Kottmann, R., Lawlor, R., Lyal, C., Mackenzie-dodds, J., Meyer, C., Mulcahy, D., Nussbeck, S., O'tuama, É., Orrell, T., Petersen, G., Robertson, T., Söhngen, C., Whitacre, J., Wieczorek, J., Yilmaz, P., Zetzsche, H., Zhang, Y. & Zhou, X. 2016. The Global Genome Biodiversity Network (GGBN) Data Standard specification. *Database*, 2016: baw125.

Dudgeon, D., Arthington, A., Gessner, M., Kawabata, Z., Knowler, D., Lévêque, C., Naiman, R., Prieur-richard, A., Soto, D., Stiassny, M. & Sullivan, C. 2006. Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological Reviews*, 81(2): 163-182.

Engel, M., Ceríaco, L., Daniel, G., Dellapé, P., Löbl, I., Marinov, M., Reis, R., Young, M., Dubois, A., Agarwal, I., Lehmann a., P., Alvarado, M., Alvarez, N., Andreone, F., Araujo-vieira, K., Ascher, J., Baêta, D., Baldo, D., Bandeira, S., Barden, P., Barrasso, D., Bendifallah, L., Bockmann, F., Böhme, W., Borkent, A., Brandão, C., Busack, S., Bybee, S., Channing, A., Chatzimanolis, S., Christenhusz, M., Crisci, J., D'elía, G., Da costa, L., Davis, S., De lucena, C., Deuve, T., Fernandes elizalde, S., Faivovich, J., Farooq, H., Ferguson, A., Gippoliti, S., Gonçalves, F., Gonzalez, V., Greenbaum, E., Hinojosa-díaz, I., Ineich, I., Jiang, J., Kahono, S., Kury, A., Lucinda, P., Lynch, J., Malécot, V., Marques, M., Marris, J., Mckellar, R., Mendes, L., Nihei, S., Nishikawa, K., Ohler, A., Orrico, V., Ota, H., Paiva, J., Parrinha, D., Pauwels, O., Pereyra, M., Pestana, L., Pinheiro, P., Prendini, L., Prokop, J., Rasmussen, C., Rödel, M., Rodrigues, M., Rodríguez, S., Salatnaya, H., Sampaio, Í., Sánchez-garcía, A., Shebl, M., Santos, B., Solórzano-kraemer, M., Sousa, A., Stoev, P., Teta, P., Trape, J., Dos santos, C., Vasudevan, K., Vink, C., Vogel, G., Wagner, P., Wappler, T., Ware, J., Wedmann, S. & Zacharie, C. 2021. The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches. *Zoological Journal of the Linnean Society*, 193(2): 381-387.

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M., Angiuoli, S., Ashburner, M., Axelrod, N., Baldauf, S., Ballard, S., Boore, J., Cochrane, G., Cole, J., Dawyndt, P., De vos, P., Depamphilis, C., Edwards, R., Faruque, N., Feldman, R., Gilbert, J., Gilna, P., Glöckner, F., Goldstein, P., Guralnick, R., Haft, D., Hancock, D., Hermjakob, H., Hertz-fowler, C., Hugenholtz, P., Joint, I., Kagan, L., Kane, M., Kennedy, J., Kowalchuk, G., Kottmann, R., Kolker, E., Kravitz, S., Kyrpides, N., Leebens-mack, J., Lewis, S., Li, K., Lister, A., Lord, P., Maltsev, N., Markowitz, V., Martiny, J., Methe, B., Mizrachi, I., Moxon, R., Nelson, K., Parkhill, J., Proctor, L., White, O., Sansone, S., Spiers, A., Stevens, R., Swift, P., Taylor, C., Tateno, Y., Tett, A., Turner, S., Ussery, D., Vaughan, B., Ward, N., Whetzel, T., San gil, I., Wilson, G. & Wipat, A. 2008. The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*, 26(5): 541-547.

Faunce, C. H. 2011. A comparison between industry and observer catch compositions within the Gulf of Alaska rockfish fishery. *ICES Journal of Marine Science*, 68(8): 1769-1777.

Faunce, C. H., Cahalan, J., Bonney, J., & Swanson, R. 2015. Can observer sampling validate industry catch reports from trawl fisheries?. *Fisheries Research*, 172: 34-43.

Fuchs, J., Rauber-Lüthy, C., Kupferschmidt, H., Kupper, J., Kullak-Ublick, G. A., & Ceschi, A. 2011. Acute plant poisoning: analysis of clinical features and circumstances of exposure. *Clinical toxicology*, *49*(7), 671-680.

Gargominy, O., Tercerie, S., Régnier, C., Dupont, P., Daszkiewicz, P., Antonetti, P., Léotard, G., Ramage, T., Idczak, L., Vandel, E., Petitteville, M., Leblond, S., Boullet, V., Denys, G., De Massary, J.C., Dusoulier, F., Lévêque, A., Jourdan, H., Touroult, J., Rome, Q., Le Divelec, R., Simian, G., Savouré-Soubelet, A., Page, N., Barbut, J., Canard, A., Haffner, P., Meyer, C., Van Es, J., Poncet, R., Demerges, D., Mehran, B., Horellou, A., Ah-Peng, C., Bernard, J.-F., Bounias-Delacour, A., Caesar, M., Comolet-Tirman, J., Courtecuisse, R., Delfosse, E., Dewynter, M., Hugonnot, V., Lavocat Bernard, E., Lebouvier, M., Lebreton, E., Malécot, V., Moreau, P.A., Moulin, N., Muller, S., Noblecourt, T., Noël, P., Pellens, R., Thouvenot, L., Tison, J.M., Robbert Gradstein, S., Rodrigues, C., Rouhan, G. & Véron, S. 2022. *TAXREF v16.0, référentiel taxonomique pour la France*. PatriNat (OFB-CNRS-MNHN), Muséum national d'Histoire naturelle, Paris. Archive de téléchargement contenant 8 fichiers. https://inpn.mnhn.fr/telechargement/referentielEspece/taxref/16.0/menu

Giusti, A., Ricci, E., Gasperetti, L., Galgani, M., Polidori, L., Verdigi, F., Narducci, R. & Armani, A. 2021. Building of an Internal Transcribed Spacer (ITS) Gene Dataset to Support the Italian Health Service in Mushroom Identification. *Foods*, 10(6): 1193.

Gold, Z., Curd, E., Goodwin, K., Choi, E., Frable, B., Thompson, A., Walker, H., Burton, R., Kacev, D., Martz, L. & Barber, P. 2021. Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem. *Molecular Ecology Resources*, 21(7): 2546-2564.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C. *et al*. 2013. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Research 41:D597–D604

Hebert, P.D.N., Cywinska, A., Ball, S.L., Dewaard, J.R., 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences,* 270**,** 313-321.

Hebert, P.D.N., Gregory, T.R. 2005. The promise of DNA barcoding for taxonomy. *Systematic Biology,* 54(5)**,** 852-859.

Hleap, J., Littlefair, J., Steinke, D., Hebert, P. & Cristescu, M. 2021. Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. *Molecular Ecology Resources*, 21(7): 2190-2203.

Heuertz, M., Carvalho, S., Galindo, J., Rinkevich, B., Robakowski, P., Aavik, T., Altinok, I., Barth, J., Cotrim, H., Goessen, R., González-martínez, S., Grebenc, T., Hoban, S., Kopatz, A., Mcmahon, B., Porth, I., Raeymaekers, J., Träger, S.,

Valdecantos, A., Vella, A., Vernesi, C. & Garnier-Géré, P. 2023. The application gap: Genomics for biodiversity and ecosystem service management. *Biological Conservation*, 278: 109883.

Hinsinger, D., Debruyne, R., Thomas, M., Denys, G., Mennesson, M., Utge, J. & Dettai, A. 2015. Fishing for barcodes in the Torrent: from COI to complete mitogenomes on NGS platforms. *DNA Barcodes*, 3(1): 170-186.

Holetschek, J., Dröge, G., Güntsch, A. & Berendsohn, W. 2012. The ABCD of primary biodiversity data access. *Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology*, 146(4): 771-779.

Hongsanan, S., Jeewon, R., Purahong, W., Xie, N., Liu, J., Jayawardena, R., Ekanayaka, A., Dissanayake, A., Raspé, O., Hyde, K., Stadler, M. & Peršoh, D. 2018. Can we use environmental DNA as holotypes?. *Fungal Diversity*, 92(1): 1-30.

Kelly, R. P., Shelton, A. O., & Gallego, R., 2019. Understanding PCR processes to draw meaningful conclusions from environmental DNA studies. *Scientific reports*, *9*(1), 1-14.

Makiola, A., Compson, Z., Baird, D., Barnes, M., Boerlijst, S., Bouchez, A., Brennan, G., Bush, A., Canard, E., Cordier, T., Creer, S., Curry, R., David, P., Dumbrell, A., Gravel, D., Hajibabaei, M., Hayden, B., Van der hoorn, B., Jarne, P., Jones, J., Karimi, B., Keck, F., Kelly, M., Knot, I., Krol, L., Massol, F., Monk, W., Murphy, J., Pawlowski, J., Poisot, T., Porter, T., Randall, K., Ransome, E., Ravigné, V., Raybould, A., Robin, S., Schrama, M., Schatz, B., Tamaddoni-nezhad, A., Trimbos, K., Vacher, C., Vasselon, V., Wood, S., Woodward, G. & Bohan, D. 2020. Key Questions for Next-Generation Biomonitoring. *Frontiers in Environmental Science*, 7: 197.

Mezzasalma, V., Ganopoulos, I., Galimberti, A., Cornara, L., Ferri, E., & Labra, M. 2017. Poisonous or non-poisonous plants? DNA-based tools and applications for accurate identification. *International journal of legal medicine*, *131*, 1-19.

Mikaelyan, A., Dietrich, C., Köhler, T., Poulsen, M., Sillam-Dussès, D. & Brune, A. 2015. Diet is the primary determinant of bacterial community structure in the guts of higher termites. *Molecular Ecology*, 24(20): 5284-5295.

Nilsson, R., Ryberg, M., Kristiansson, E., Abarenkov, K., Larsson, K. & Kõljalg, U. 2006. Taxonomic Reliability of DNA Sequences in Public Sequence Databases: A Fungal Perspective. *PLoS ONE*, 1(1): e59.

Nithaniyal, S., Majumder, S., Umapathy, S., & Parani, M. 2021. Forensic application of DNA barcoding in the identification of commonly occurring poisonous plants. *Journal of forensic and legal medicine*, 78: 102126.

Pentinsaari, M., Ratnasingham, S., Miller, S. & Hebert, P. 2020. BOLD and GenBank revisited – Do identification errors arise in the lab or in the sequence libraries?. *PLOS ONE*, 15(4): e0231814.

Porter, T.M., Hajibabaei, M. 2018. Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Mol Ecol,* 27(2)**,** 313-338.

Pruvost, P., Causse, R. & Bailly, N. 2023. Historical review of the computerization of the MNHN Fish Collection and its collaboration with FishBase. *Cybium*, 47(3): 225-248.

Quast C, Pruesse E, Yilmaz P, *et al*. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41(D1):D590–D596. doi: 10.1093/nar/gks1219

Ratnasingham, S. & Hebert, P. 2007. BARCODING: BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, 7(3): 355-364.

Ratnasingham, S. & Hebert, P. 2013. A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLoS ONE*, 8(7): e66213.

Renner, S. S., Scherz, M. D., Schoch, C. L., Gottschling, M., & Vences, M. 2024. Improving the gold standard in NCBI GenBank and related databases: DNA sequences from type specimens and type strains. Systematic Biology, 73(2), 486-494., https://doi.org/10.1093/sysbio/syad068

Rimet F., Gusev E., Kahlert M., Kelly M., Kulikovskiy M., Maltsev Y., Mann D., Pfannkuchen M., Trobajo R., Vasselon V., Zimmermann J., Bouchez A., 2019. Diat.barcode, an open-access curated barcode library for diatoms. Scientific Reports.

Rourke, M., Fowler, A., Hughes, J., Broadhurst, M., Dibattista, J., Fielder, S., Wilkes Walburn, J. & Furlan, E. 2022. Environmental DNA (eDNA) as a tool for assessing fish biomass: A review of approaches and future considerations for resource surveys. *Environmental DNA*, 4(1): 9-33.

Somervuo, P., Yu, D., Xu, C., Ji, Y., Hultman, J., Wirta, H. & Ovaskainen, O. 2017. Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. *Methods in Ecology and Evolution*, 8(4): 398-407.

Stroud, S., Fennell, M., Mitchley, J., Lydon, S., Peacock, J., & Bacon, K. L. (2022). The botanical education extinction and the fall of plant awareness. *Ecology and Evolution*, 12(7), e9019.

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

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G., Geniez, P., Pont, D., Argillier, C., Baudoin, J., Peroux, T., Crivelli, A., Olivier, A., Acqueberge, M., Le brun, M., Møller, P., Willerslev, E. & Dejean, T. 2016. Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, 25(4): 929-942.

van den Burg, M. P., & Vieites, D. R. 2023. Bird genetic databases need improved curation and error reporting to NCBI. *Ibis*, *165*(2), 472-481. https://onlinelibrary.wiley.com/doi/full/10.1111/ibi.13143

Walton, S., Livermore, L., Bánki, O., Cubey, R., Drinkwater, R., Englund, M., Goble, C., Groom, Q., Kermorvant, C., Rey, I., Santos, C., Scott, B., Williams, A. & Wu, Z. 2020. Landscape Analysis for the Specimen Data Refinery. *Research Ideas and Outcomes*, 6: e57602.

Waterhouse, R., Adam-blondon, A., Agosti, D., Baldrian, P., Balech, B., Corre, E., Davey, R., Lantz, H., Pesole, G., Quast, C., Glöckner, F., Raes, N., Sandionigi, A., Santamaria, M., Addink, W., Vohradsky, J., Nunes-jorge, A., Willassen, N. & Lanfear, J. 2022. Recommendations for connecting molecular sequence and biodiversity research infrastructures through ELIXIR. *F1000Research*, 10: 1238.

Weigand, H., Beermann, A., Čiampor, F., Costa, F., Csabai, Z., Duarte, S., Geiger, M., Grabowski, M., Rimet, F., Rulik, B., Strand, M., Szucsich, N., Weigand, A., Willassen, E., Wyler, S., Bouchez, A., Borja, A., Čiamporová-zaťovičová, Z., Ferreira, S., Dijkstra, K., Eisendle, U., Freyhof, J., Gadawski, P., Graf, W., Haegerbaeumer, A., Van der hoorn, B., Japoshvili, B., Keresztes, L., Keskin, E., Leese, F., Macher, J., Mamos, T., Paz, G., Pešić, V., Pfannkuchen, D., Pfannkuchen, M., Price, B., Rinkevich, B., Teixeira, M., Várbíró, G. & Ekrem, T. 2019. DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work *Science of The Total Environment*, 678: 499-524.

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., Da Silva Santos, L., Bourne, P., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers, R., Gonzalez-Beltran, A., Gray, A., Groth, P., Goble, C., Grethe, J., Heringa, J., 't Hoen, P., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S., Martone, M., Mons, A., Packer, A., Persson, B., Rocca-Serra, P., Roos, M., Van Schaik, R., Sansone, S., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M., Thompson, M., Van der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018.

Woodburn, M., Droege, G., Grant, S., Groom, Q., Jones, J., Trekels, M., Vincent, S. & Webbink, K. 2021. A Data Standard for Dynamic Collection Descriptions. *Biodiversity Information Science and Standards*, 5: e73902.

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J., Amaral-zettler, L., Gilbert, J., Karsch-mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., Birren, B., Blaser, M., Bonazzi, V., Booth, T., Bork, P., Bushman, F., Buttigieg, P., Chain, P., Charlson, E., Costello, E., Huot-creasy, H., Dawyndt, P., Desantis, T., Fierer, N., Fuhrman, J., Gallery, R., Gevers, D., Gibbs, R., Gil, I., Gonzalez, A., Gordon, J., Guralnick, R., Hankeln, W., Highlander, S., Hugenholtz, P., Jansson, J., Kau, A., Kelley, S., Kennedy, J., Knights, D., Koren, O., Kuczynski, J., Kyrpides, N., Larsen, R., Lauber, C., Legg, T., Ley, R., Lozupone, C., Ludwig, W., Lyons, D., Maguire, E., Methé, B., Meyer, F., Muegge, B., Nakielny, S., Nelson, K., Nemergut, D., Neufeld, J., Newbold, L., Oliver, A., Pace, N., Palanisamy, G., Peplies, J., Petrosino, J., Proctor, L., Pruesse, E., Quast, C., Raes, J., Ratnasingham, S., Ravel, J., Relman, D., Assunta-sansone, S., Schloss, P., Schriml, L., Sinha, R., Smith, M., Sodergren, E., Spor, A., Stombaugh, J., Tiedje, J., Ward, D., Weinstock, G., Wendel, D., White, O., Whiteley, A., Wilke, A., Wortman, J., Yatsunenko, T. & Glöckner, F. 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, 29(5): 415-420.

Zafeiropoulos, H., Gargan, L., Hintikka, S., Pavloudi, C. & Carlsson, J. 2021. The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data. Metabarcoding *and Metagenomics*, 5: 163-174.

8. GLOSSAIRE

COI : gène mitochindrial codant pour la sous-unité 1 de l'enzyme cyctochrome oxydase ; marqueur dont la partie 5' (env. 650 pb) est utilisée comme code-barres ADN standard chez les animaux.

DNA barcode / code-barres ADN : un fragment du génome (marqueur) de quelques centaines de pb utilisé pour l'identification moléculaire des espèces. Le code-barres fréquemment utilisé en barcoding pour les animaux est un fragment du gène COI mitochondrial (ou *cox1*) ; pour les plantes c'est un fragment du gène du plaste ribulose 1,5-bisphosphate carboxylase (rbcL) combiné à un fragment du gène nucléaire de la maturase (matK), et complété parfois par un 3ème marqueur (trnH, ITS par ex.), alors que le code-barres des champignons est l'espaceur interne transcrit nucléaire (ITS) de l'ADN ribosomique. Le consortium iBOL (https://ibol.org/) regroupe une communauté internationale de scientifiques qui est à l'origine des propositions de marqueurs standards ; il dirige et coordonne également le développement de la base de données de référence de séquence correspondante pour ces marqueurs (BOLD).

D**NA barcoding** : l'identification des spécimens à l'aide de fragments d'ADN standardisés. La procédure idéale de codage à barres d'ADN commence par des spécimens de référence bien conservés déposés dans des collections d'histoire naturelle et se termine par une séquence unique déposée dans une bibliothèque de référence publique d'identificateurs d'espèces qui pourraient être utilisés pour attribuer des séquences inconnues à des espèces connues.

ITS : Espaceur interne transcrit (*Internal transcribed spacer)* qui est une région de l'ADN ribosomique non codante et hautement polymorphe.

Metabarcoding : une méthode rapide d'identification à haut débit basée sur le séquençage massif parallèle d'un ou plusieurs marqueurs communs à plusieurs espèces ou unités évolutives (parfois de rang supra-spécifique, notamment si le marqueur n'est pas suffisamment résolutif dans certains groupes) à partir d'un substrat complexe pouvant prendre la forme d'un échantillon environnemental (sol, eau, air, contenu digestif par exemple)) ou d'un ensemble d'individus ou de fragments d'individus (collecte massive de spécimens par piégeage par exemple). L'approche metabarcoding appliquée aux communautés microbiennes peut également être appliquée à la méiofaune ou même à la mégafaune.

NGS : Séquençage de Nouvelle Génération. Ensemble des méthodes de séquençages post-Sanger. Il existe deux technologies principales : le séquençage par synthèse (MGI, Illumina, PacBio, Ion Torrent) et le séquençage par passage à travers un pore (ONT), qui produisent des séquences courtes (< 1 kb, MGI, Illumina, Ion Torrent) ou longues (> 10 kb, PacBio, ONT). Ces technologies massivement parallèlisées produisent généralement de très grandes quantités de séquences, qui nécessitent un traitement bio-informatique spécifique.

Sanger : méthode classique, à faible débit, de séquençage par terminaison à l'aide d'une ADN polymérase, de di-désoxyribonucléotides (ddNTP) marqués avec un fluorochrome, et d'une séparation des brins par électrophorèse.

Taxinomie, taxonomie : la science de la découverte, de la description, de la classification et de la dénomination des organismes.

Voucher : spécimen "témoin" rattaché à une donnée de séquence et enregistré en collection avec un numéro d'inventaire.

9. ANNEXES

Annexe 1 : Éléments techniques

A. Outils et méthodes d'identification

Un référentiel moléculaire disponible pour un large public permettrait de répondre à plusieurs types d'expertises moléculaires (Figure B, C, D).

- **PCR assay**

Les premières identifications moléculaires ont été développées sur la biocénose bactérienne et elles consistaient à amplifier un marqueur à l'aide d'amorces spécifiques aux espèces en question. Ainsi, le succès de l'amplification par PCR (réaction de polymérisation en chaine) est l'attestation de l'identification. Cette technique d'identification moléculaire sans séquençage dite "PCR assay" (Figure B) est largement utilisée pour des détections de fraudes alimentaires, des dépistages de maladies infectieuses (ex: COVID 19) ainsi que des identifications d'espèces dans le milieu naturel pour ses faibles coûts (ex. Dalmasso *et al.,* 2004 ; Borland & Kading, 2021 ; Gupta *et al.,* 2021). Elle représente un intérêt pour des identifications en routine.

- **Barcoding**

L'analyse des séquences d'ADN a révolutionné la systématique (ex. Avise, 1994). Désormais, les données moléculaires permettent d'obtenir des séquences caractéristiques d'espèce. Le Barcoding (Figure C) consiste à identifier un individu en comparant sa séquence avec d'autres présentes dans une base de référence moléculaire (ex : GenBank). Parmi les programmes internationaux les plus connus, il y a ceux initiés en 2010 par le consortium iBOL (international Barcode of Life) et s'appuyant sur la base de données de référence communautaire BOLD[1] (Ratnasingham & Hebert, 2007) qui regroupe plusieurs millions de séquences des codes-barres ADN standards pour les animaux, végétaux et champignons.

- **Metabarcoding**

Avec l'arrivée des séquenceurs à haut débit et ceux de nouvelle génération (*Next-Generation Sequencing*, NGS), le débit de séquençage est largement augmenté et les coûts diminués. Il est désormais possible de séquencer plus facilement des marqueurs de longueurs nettement supérieures à 1000 paires de bases mais également de séquencer plusieurs spécimens en un seul séquençage (ex. Dettai *et al.,* 2012). Le metabarcoding permet à partir d'un échantillon contenant le matériel génétique de plusieurs espèces d'obtenir une liste des espèces présentes dans cet échantillon (Voir Figure) (ex. Taberlet *et al.,* 2012). Cette méthode est bien moins coûteuse que le barcoding qui traite les spécimens un par un. Néanmoins elle peut nécessiter plus de temps de main d'œuvre et des compétences en traitements bioinformatiques (Coissac *et al.,* 2012 ; Cristescu, 2014). Mais elle permet une approche sans *a priori* et une plus grande couverture d'exploration de la biodiversité (Taberlet *et al.,* 2012 ; Adamowicz *et al.,* 2019).

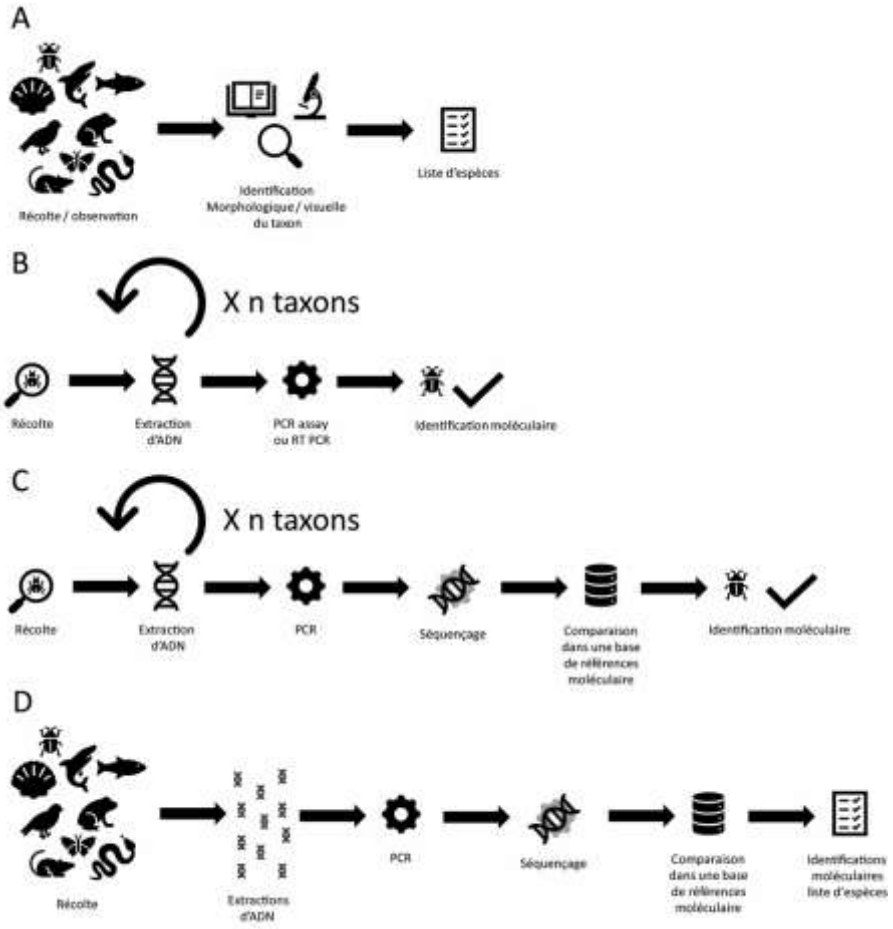

Figure : A) : échantillonnage et détermination des espèces classiques ; B) : échantillonnage et détermination par PCR assay ; C) échantillonnage et détermination par barcoding ; D) échantillonnage et détermination par metabarcoding.

Ces trois méthodes d'identification moléculaire ne sont pertinentes que si l'on dispose d'une base de référence moléculaire complète et de qualité (Taberlet *et al.,* 2012). Plusieurs processus successifs sont réalisés permettant d'obtenir une comparaison des séquences génétiques présentes dans l'échantillon avec une ou plusieurs bibliothèques de séquences génétiques de référence disponibles.

Il faut signaler toutefois que des biais peuvent exister :

- Biais associés aux processus de récolte des échantillons, d'extraction et d'amplification de l'ADN entrainant des faux positifs (ex. contaminations dues aux prélèvements ou manipulations, séquences chimériques, etc.) ou des faux négatifs (groupes taxonomiques qui ne s'amplifient ou ne se séquencent pas et ne sont pas détectés) (Liu *et al.,* 2019)
- Biais associés aux erreurs d'affectation des espèces au niveau des bibliothèques de référence (Hofstetter *et al.,* 2019).

- **ADN environnemental**

L'ADN environnemental (ADNe) correspond aux fragments d'ADN relâchés par un organisme dans l'environnement comme l'eau, le sol, les sédiments ou l'air. Les organismes libèrent en effet dans les milieux où ils vivent de l'ADN (issu de sécrétions, excrétions, tissus perdus lors du renouvellement cellulaire *etc*.). Cet ADN peut être rapidement dégradé par des facteurs biotiques (bactéries, champignons, endonucléases, *etc*.) et abiotiques (radiations UV, acidité, température, *etc*.) ou persister dans le milieu en étant adsorbé sur des particules organiques ou inorganiques (souvent sous la forme de très courts fragments d'ADN). Après extraction de l'ADN contenu dans l'échantillon prélevé dans le milieu, les brins d'ADN sont amplifiés, et 2 approches peuvent être distinguées : l'approche spécifique et l'approche multispécifique (par metabarcoding de l'ADNe). Dans le cas du metabarcoding, les brins d'ADN amplifiés sont ensuite séquencés puis les

séquences obtenues sont comparées à une ou plusieurs bases de référence. Dans ces bases de référence figurent les séquences associées aux noms des espèces ou aux groupes taxonomiques. Ces différentes étapes permettent ainsi de mettre un nom sur les séquences obtenues à partir d'un échantillon prélevé sur le terrain.

Par abus de langage, le terme « ADNe » est aussi utilisé pour désigner la technique qui permet d'étudier la biodiversité grâce aux échantillons prélevés dans l'environnement dont sont extraits et analysés les fragments d'ADN. Cette technique moléculaire innovante est désormais opérationnelle pour certains écosystèmes d'eau douce, marins et pour un grand nombre de taxons. Depuis une dizaine d'année, les études utilisant l'ADNe pour étudier les espèces se développent fortement. Les applications pour l'acquisition de données de biodiversité sont en effet nombreuses : démarche d'inventaire et de suivi/surveillance, étude des communautés fonctionnelles des écosystèmes (ex. pollinisation), détection d'espèces rares (présentes en faible densité localement ou espèces menacées) ou cryptiques (ex. distinguables uniquement sur la base d'informations moléculaires), détection précoce de fragments d'ADN issus d'espèces invasives, réponse des écosystèmes face aux pressions anthropiques, suivis des espèces à enjeux … (Ruppert *et al.,* 2019). Le nombre de données collectées dans le cadre de ces études qui utilisent l'ADNe est très important et ne fera qu'augmenter à l'avenir.

B. Processus informatiques permettant une assignation taxonomique

Il existe de nombreux outils d'assignation taxonomique de séquences ADN. Ceux-ci peuvent être utilisés individuellement par un utilisateur disposant de sa base de référence propre, construite sur mesure pour ses besoins. Mais l'alternative la plus commune est leur intégration au sein de plateformes de stockage et de gestion de données moléculaires, comme BOLD ou GenBank. On peut ainsi citer le moteur d'identification (« ID engine ») de BOLD ou les outils BLAST de GenBank qui permettent de faire de l'assignation taxonomique avec comme avantage d'avoir accès à une très grande quantité de données, mais des risques bien plus élevés d'inclure des séquences affiliées à de mauvais taxons (dus à des erreurs d'identification, à l'évolution des connaissances taxonomiques, et à l'impossibilité de corriger les métadonnées associées aux séquences si on en n'est pas l'auteur).

Dans le cas du metabarcoding, avec l'évolution de la technologie, des programmes informatiques ont été développés pour analyser les informations provenant des millions de données des séquenceurs de nouvelle génération (NGS) et les séquenceurs à haut débit. L'analyse de gros volumes de données de séquences doit se faire sur des serveurs de calculs dédiés, qui disposent à la fois les bases de séquences de référence et des outils bio-informatiques pour établir la meilleure correspondance entre les séquences obtenues et ces bases de référence. Ces pipelines, conçus par des spécialistes en bioinformatique (Coissac *et al.,* 2012) doivent être adaptés à chaque étude. Des pipelines sont disponibles sur Galaxy (ex. Dufresne *et al.,* 2019 ; Boyer *et al.,* 2016), R (ex. Keck & Altermatt, 2023) ou d'autres plateformes en ligne spécialisées (ex. Mitofish pour les poissons ; Sato *et al.,* 2018).

Remarque : La grande majorité des référentiels de séquences et entrepôts de données moléculaires externes utilisent leur propre standard de données. Pour l'acquisition de données non standardisées, il existe des outils comme ceux développés par **ISA-tools**. Ces outils reposent sur le format ISAtab. Ce format semble le meilleur moyen de standardiser des données génétiques car il sert de format "pivot" entre les formats existants de soumission de données dans des bases internationales telles que l'ENA. Cependant, les outils ISAtools ne sont plus développés depuis 2018. Désormais une API ISA nommée isaAPI est proposée et utilisable par un langage de programmation (pour plus d'information : isa-tools.org/software-suite.html).

Références :

Adamowicz, S., Boatwright, J., Chain, F., Fisher, B., Hogg, I., Leese, F., Lijtmaer, D., Mwale, M., Naaum, A., Pochon, X., Steinke, D., Wilson, J., Wood, S., Xu, J., Xu, S., Zhou, X. & Van der bank, M. 2019. Trends in DNA barcoding and metabarcoding. *Genome*, 62(3): v-viii.

Avise, J.C. 1994. *Molecular markers, natural history and evolution*. Chapman & Hall, New York. 511 pp.

Borland, E. & Kading, R. 2021. Modernizing the Toolkit for Arthropod Bloodmeal Identification. *Insects*, 12(1): 37.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. & Coissac E. 2016. Obitools: a unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources* 16(1): 176-82.

Coissac, E., Riaz, T. & Puillandre, N. 2012. Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, 21(8): 1834-1847.

Cristescu, M. 2014. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, 29(10): 566-571.

Dalmasso, A., Fontanella, E., Piatti, P., Civera, T., Rosati, S. & Bottero, M. 2004. A multiplex PCR assay for the identification of animal species in feedstuffs. *Molecular and Cellular Probes*, 18(2): 81-87.

Dettai, A., Gallut, C., Brouillet, S., Pothier, J., Lecointre, G. & Debruyne, R. 2012. Conveniently Pre-Tagged and Pre-Packaged: Extended Molecular Identification and Metagenomics Using Complete Metazoan Mitochondrial Genomes. *PLoS ONE*, 7(12): e51263.

Dufresne, Y., Lejzerowicz, F., Perret-gentil, L., Pawlowski, J. & Cordier, T. 2019. SLIM: a flexible web application for the reproducible processing of environmental DNA metabarcoding data. *BMC Bioinformatics*, 20(1): 88.

Gupta, N., Augustine, S., Narayan, T., O'Riordan, A., Das, A., Kumar, D., Luong, J. & Malhotra, B. 2021. Point-of-Care PCR Assays for COVID-19 Detection. *Biosensors*, 11(5): 141.

Hebert, P.D.N., Ratnasingham, S. & de Waard, J.R. 2003. Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proceeding of the Royal Society of London Part B.*, 270: S96–S99

Hofstetter, V., Buyck, B., Eyssartier, G., Schnee, S. & Gindro, K. 2019. The unbearable lightness of sequenced-based identification. *Fungal Diversity*, 96(1): 243-284.

Keck, F. & Altermatt, F. 2022. Management of DNA reference libraries for barcoding and metabarcoding studies with the R package refdb. *Molecular Ecology Resources*, 23(2): 511-518.

Liu, M., Clarke, L., Baker, S., Jordan, G. & Burridge, C. 2019. A practical guide to DNA metabarcoding for entomological ecologists. *Ecological Entomology*, 45(3): 373-385.

Ruppert, K., Kline, R. & Rahman, M. 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.

Sato, Y., Miya, M., Fukunaga, T., Sado, T. & Iwasaki, W. 2018. MitoFish and MiFish Pipeline: A Mitochondrial Genome Database of Fish with an Analysis Pipeline for Environmental DNA Metabarcoding. *Molecular Biology and Evolution*, 35(6): 1553-1555.

Annexe 2 : MOLECULAIRE, la base de données moléculaires du MNHN

Cette base est composée de neuf tables. Le point d'entrée est la table PRELEVEMENT qui est liée aux vouchers informatisés dans les bases de données des collections du MNHN. Actuellement cette table est connectée aux collections des arthropodes terrestres, d'ichtyologie, des invertébrés non-arthropodes terrestres, des souches fongiques, de l'Herbier National et du réseau des herbiers associés, des mammifères et anatomie comparée, des vers parasites ainsi qu'à la base HORS_MNHN qui permet des gérer les données de spécimens conservés dans d'autres musées. Il est tout à fait envisageable de connecter cette table à d'autres bases de données si les besoins existent. La table PRELEVEMENT est également connectée à la table EXTRACTION. Cette table permet la gestion des données liées à l'extraction des ADN, notamment les méthodes employées, ainsi qu'à leur stockage. A la table EXTRACTION est connectée la table PCR, elle-même associée aux tables AMORCE_FOR (pour amorces forward) et AMORCE_REV (pour amorces reverse). Ces trois tables permettent de gérer les conditions et méthodes utilisées pour l'amplification des gènes ciblés. La table PCR est liée à la table CHROMATOGRAMME qui fait le lien vers la table SEQUENCE. Cette dernière assure la gestion des séquences consensus obtenues par interprétation des chromatogrammes via des URL pointant vers les fichiers fasta correspondants. Il est possible d'y qualifier les séquences (valides, contaminations, doublons, fragments, ...) et d'y stocker les identifiants GenBank (accession numbers) lorsque les séquences ont été publiées. La table SEQUENCES pointe également vers la table PRELEVEMENT permettant ainsi l'informatisation de séquences pour lesquelles les données d'extraction des ADN et de PCR ne sont pas disponibles. La table PRELEVEMENT est aussi rattachée à une dernière table, SPECIMENBOLD, pour permettre les exports nécessaires à la publication des données spécimens et des séquences associées vers la base de référence BOLD.

MOLECULAIRE est opérationnelle et simple de conceptualisation. Elle a également l'avantage d'être associée à plusieurs outils web qui permettent par exemple d'importer des séquences dans la base de données, de les extraire au format fasta (avec la possibilité de choisir les informations contenues dans le titre des séquences), de préparer le fichier d'exportation vers BOLD, *etc*... En revanche, elle a néanmoins deux défauts majeurs. Le premier est qu'elle n'est pas utilisée par tous les gestionnaires des collections, et que c'est un travail considérable d'incrémenter toutes les séquences disponibles et de les rattacher aux bons vouchers en collection. Car en plus de l'importance de la quantité de séquences publiées, les publications ne remontent pas forcément aux gestionnaires des collections. L'autre défaut, c'est qu'elle s'appuie sur des méthodes utilisées il y a une dizaine d'années (comme les séquences Sanger) et n'inclue pas les nouveaux outils et nouvelles technologies de séquençage (NGS). Des discussions sont en cours avec la DINSI pour une refonte du schéma conceptuel de la base ainsi que des outils web.

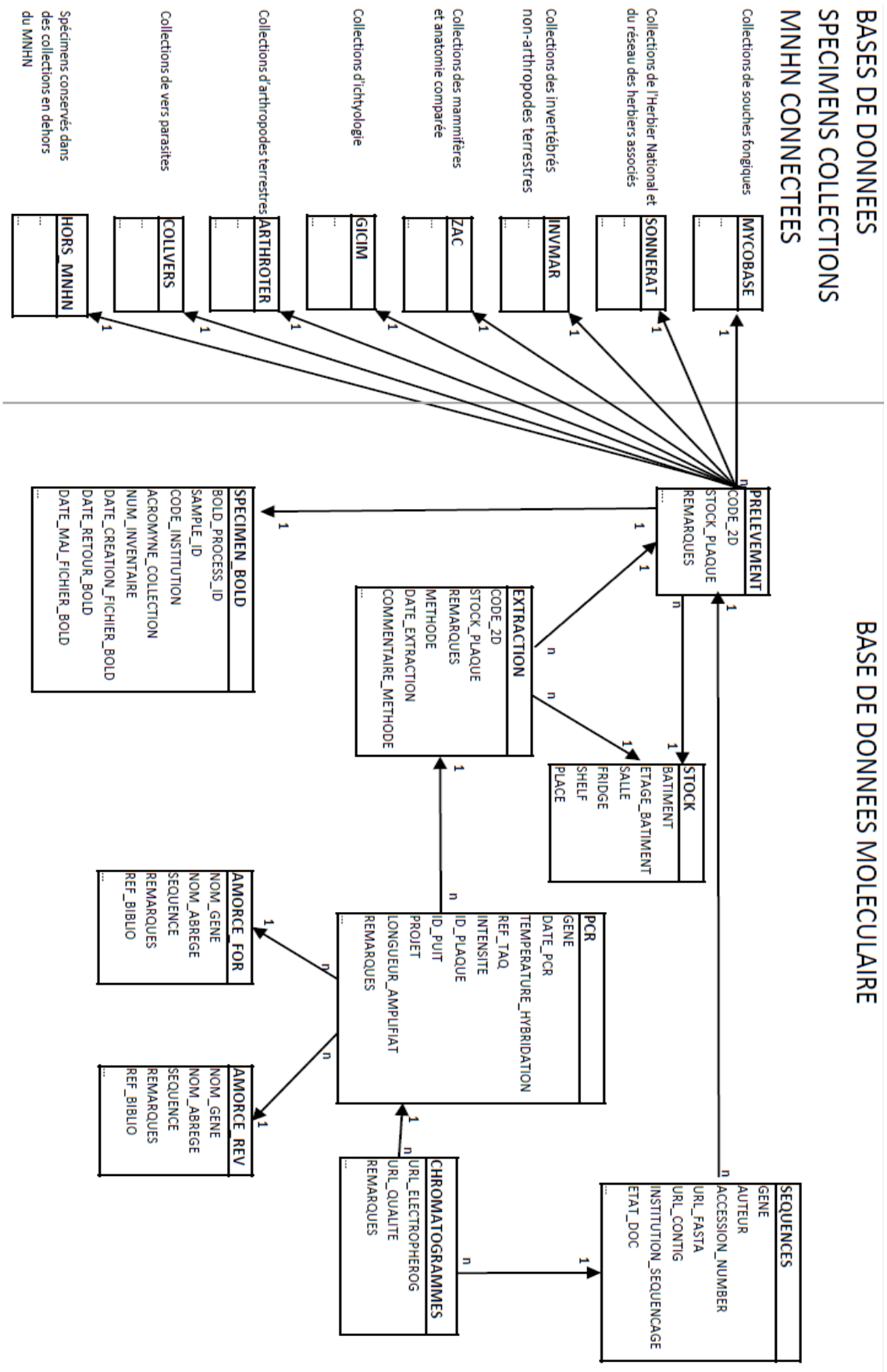

Schéma conceptuel de la base MOLECULAIRE du MNHN montrant les tables utilisées ainsi que les liens entre elles

RÉSUMÉ

Le référentiel, dont il est question dans ce document, est un nouvel outil national mettant à disposition des séquences génétiques pour :

- Améliorer et objectiver la **fiabilité des assignations taxonomiques** dérivées des analyses d'échantillons environnementaux ou de spécimens par des techniques moléculaires ;
- Améliorer la **qualité des données d'occurrence provenant d'analyses moléculaires** portant sur les espèces présentes dans les territoires français, en **faciliter l'accès, l'interopérabilité et leur réutilisation** ;
- **Mutualiser les efforts de construction, de maintenance et d'alimentation des bases de séquences de référence**, en proposant un référentiel public et partagé.

Coordonné par PatriNat en collaboration avec un consortium scientifique qui a contribué à sa construction et sa validation, ce document a pour objectifs principaux de :

1. Identifier et caractériser les besoins et les enjeux relatifs à la création d'un référentiel de séquences génétiques relevant du système d'information sur la biodiversité (SIB), en associant les parties prenantes, dont le Muséum national d'Histoire naturelle,

2. Faire un état des lieux des dispositifs existants pouvant contribuer à la constitution de ce référentiel,

3. Proposer une préfiguration du référentiel de séquences génétiques à travers :
   - Une liste de critères de qualité à prendre en compte,
   - Un périmètre taxonomique et un périmètre génétique et des standards sur lesquels s'appuyer,
   - Une préfiguration du schéma de l'architecture du référentiel,
   - Une gouvernance du référentiel.

Si les groupes prioritaires sont ceux ciblés par des politiques publiques, le référentiel concerne les espèces Eucaryotes (Hexagone, Corse et Outre-mer) présentes ou ayant été présentes sur le territoire national (natives et introduites) et les espèces non présentes en France mais pour lesquelles la France met en place un dispositif spécifique de surveillance. La création de ce référentiel national représente ainsi un enjeu pour l'inventaire, le suivi et la surveillance de la biodiversité et *in fine* pour la conservation de la biodiversité.

PatriNat (OFB-MNHN-CNRS-IRD)
Centre d'expertise et de données sur le patrimoine naturel
Jardin des Plantes
CP41 – 36 rue Geoffroy Saint-Hilaire
75005 Paris
www.patrinat.fr