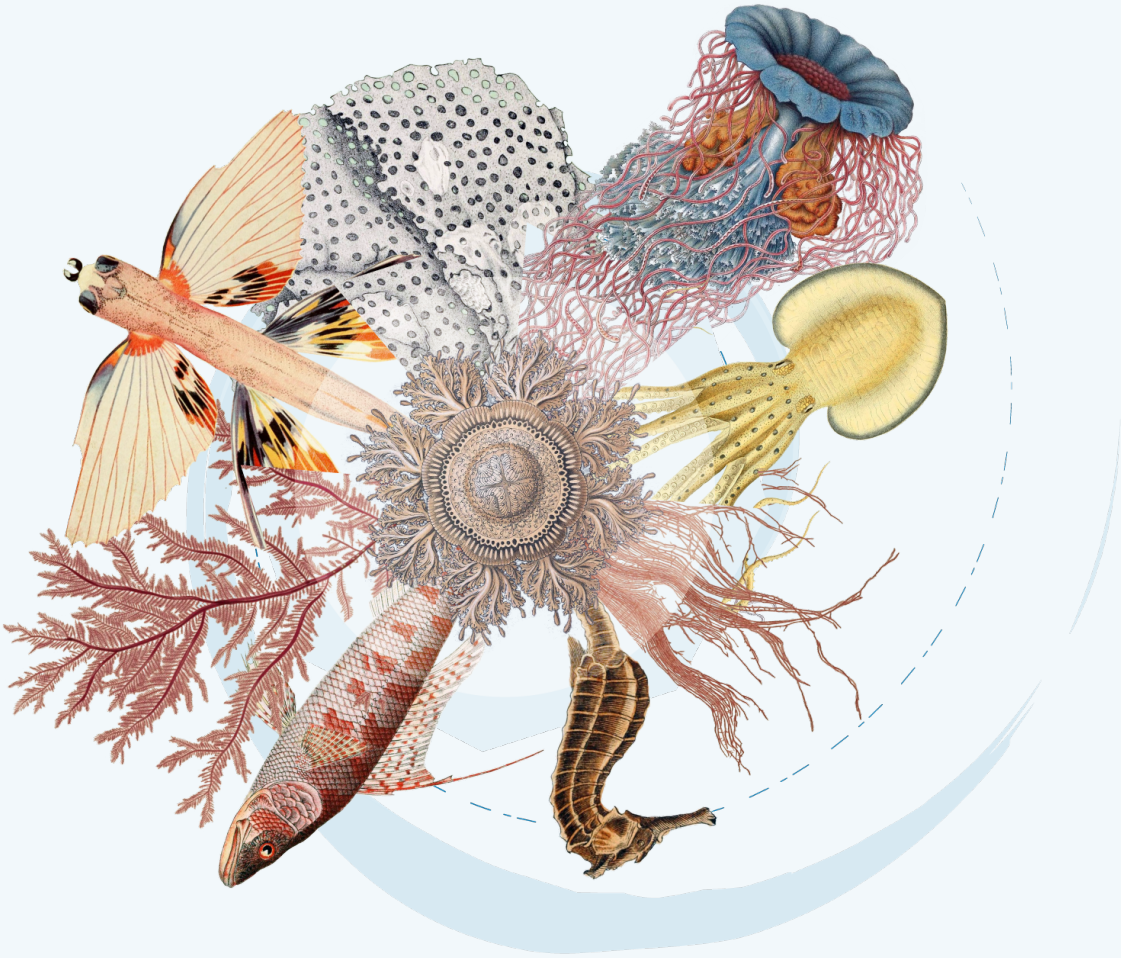


通過生物多樣性資訊平台發布DNA衍生資料

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

版本 1.3.0, 7 June 2023

GBIF Global Biodiversity Information Facility

OBIS OCEAN BIODIVERSITY INFORMATION SYSTEM

目錄

版權宣告

引用建議

Abarenkov K, Andersson AF, Bissett A, Finstad AG, Fossøy F, Grosjean M, Hope M, Jeppesen TS, Kõljalg U, Lundin D, Nilsson RN, Prager M, Provoost P, Schigel D, Suominen S, Svenningsen C & Frøslev TG (2023) Publishing DNA-derived data through biodiversity data platforms, v1.3. Copenhagen: GBIF Secretariat. https://doi.org/10.35035/doc-vf1a-nr22.

作者

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

貢獻者

此草案經與ELIXIR、iBOL、GGBN、GLOMICON和OBIS網絡成員進行了寶貴的討論後所整合。我們特別感謝 Andrew Bentley, Matt Blissett, Pier Luigi Buttigieg, Kyle Copas, Camila A. Plata Corredor, Gabriele Dröge, Torbjørn Ekrem, Birgit Gemeinholzer, Quentin Groom, Tim Hirsch, Donald Hobern, Hamish Holewa, Corinne Martin, Raissa Meyer, Chris Mungall, Daniel Noesgaard, Corinna Paeper, Tim Robertson, Maxime Sweetlove, Andrew Young, John Waller, Ramona Walls, John Wieczorek, Lucie Zinger在GBIF社群評論的過程中給予的貢獻和鼓勵。

授權

永久連結

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

文件版控

版本 1.3.0 發布於 7 June 2023。

此版本新增了有關海洋資料集和海洋生物多樣性信息系統 (OBIS) 的文字段落，以及小部分的文字編輯。

摘要

使用基因資訊來描述或鑑定物種分類群時，大多使用者會預設這些資料將會使用在分子生物或演化學的研究上。然而，同時擁有座標與時間資訊的基因序列是一個有價值的生物多樣性出現紀錄，相較其最初目的，可更廣泛的被使用。為了實現這一潛力，DNA衍生資料需要可以在生物多樣性資訊的平台上供搜尋。這份教學文件將帶領您了解公開 “帶有時間與座標資訊的序列”的生物多樣性資料的原理和方法。此文件涵蓋了特定的概要和術語、常見問題與合適的操作流程建議、且不討論各平台的細節。此文件將受益任何對使用各生物多樣性資訊平台 (包括國家級生物多樣性入口網站)有興趣的人更順利地公開DNA衍生資料。

前言

此文件的準備工作始於2019年在 biodiversity_next conference 研討會上各種面向的討論整合，例如：

- 未來科學平台環境經濟學項目 (Environomics Future Science Platform Project) 總結報告
- 部落格文章，內容關於ALA現在包含eDNA記錄
- ALA裡的環境DNA (eDNA)
- ALA eDNA資料模板
- 挪威寄存eDNA樣本與資料，包括憑證標本的標準
- 瑞典生物多樣性資訊建設機構 (SBDI) 的分子生物多樣性資料
- GBIF資源 (如何) 發布分子/序列/DNA資料到GBIF?
- GBIF裡的分子生物資料
- GBIF簡略和詳細發布資料指南
- 如何發布資料到GBIF，以及DwC/擴充表的欄位概述。

- 基因組生物多樣性工作小組 (Genomic Biodiversity Interest Group)

1. 介紹

1.1. 基本原理

在過去的20年裡，人們對分子生物方法在記錄地球上生命多樣性的能力有了更多的了解。儘管不是以一種觀察者會立即看到的方式–看似沒有生命的土壤和海水等平凡的基質卻充滿了生命。基於DNA的研究證明，真菌、昆蟲、卵菌、細菌和古細菌等生物群無處不在，儘管我們無法通過實體的方式觀察它們 (Debroas等人，2017)。分子生物方法的好處並不局限於微觀世界：有許多生物，例如某些魚類，即使可以實體觀察到與紀錄它們，但尋找這些生物需要高金錢與人力成本，而且對生物可能具有侵入性 (Boussarie等人，2018)。在這種情況下，DNA資料使我們能夠以最小的精力以非侵入性地的方式記錄這些生物的存在 (和過去存在)。 這些發展意味著我們並不總是需要在某個地點來記錄所有存在的生物體。所有生物，無論它們是否可以實體的被觀察到，在了解生物多樣性、生態學和生物保育方面都很重要。

DNA衍生資料使我們能夠記錄不顯眼或無法觀察到的物種，因為這些物種會在常用的田野調查、名錄、自然典藏收集等的方法被忽略。目前DNA方法的成熟度使我們已能更詳細的記錄這些生物的存在，並超過了對一般生物的觀察方式。然而，DNA方法仍有其自身的問題和偏差，重要的是利用現況來定義我們應該如何通過分子生物方法記錄和回報現行存在於某特定基質或位置的生物。由於各領域缺乏標準和方法，導致資料主體非常多樣化且基本上無法互用或做比較。標準化將有助於避免資料在其他領域被無效的呈現 (Berry等人，2021；Leebens-Mack等人，2006；Yilmaz等人，2011；Nilsson等人，2012；Shea等人，2023)。此外，清晰地記錄從原始序列reads到推斷物種觀察的計算處理方法，將在未來改進方法出現時能夠重新分析過去所使用的步驟。

物種的DNA衍生出現紀錄應盡可能標準化和可重複操作，無論檢測到的物種是否具有正式學名。在某些情況下，此類出現記錄將暗示所描述物種以前未知的地理和生態特性，從而豐富我們對這些物種的知識。在其他情況下，這些數據可能使我們能夠整合和視覺化有關當前未描述物種的資訊，從而加快它們的正式物種描述。這些未命名物種也能收集可用的相關資料，顯著地增加了GBIF和其他生物多樣性資訊平台對生物世界多種進行索引的方式，使所有人都可以使用這些知識並用於各種目的，包括生物多樣性保育。文獻估計指出所有現存物種中至少有85%未被描述 (Mora等人，2011；Tedesco等人，2014)。現有的資料標準是為已描述的少數類群所設計的。良好的處理DNA衍生資料將有助於描述所有生物的出現紀錄，無論該物種是否已被描述。

此文件描述DNA衍生資料的發布標準與方式，以便將此資料納入GBIF和其他生物多樣性平台。此文件不針對數據存取與序列資訊共享的議題發表任何意見，這是在生物多樣性公約 (CBD) 裡已廣泛討論的主題。然而值得注意的是，DNA分子條碼 (barcoding) 和DNA高通量份子條碼 (metabarcoding) 通常是基因或非編碼DNA片段，不適合用於商業開發。序列存檔於國際核苷酸序列數據庫協作 (INDSC) 是序列型研究中的普遍規範，但發布源自序列的出現紀錄不需要發布新序列。大多數的情況下，這些序列資料已經被放置在公開的基因資料庫中。因此，此文件描述從DNA衍生出現紀錄資料中的時空間與基於DNA的名稱資訊可能帶來的附加價值，而不是遺傳資訊本身的價值。除了處理序列衍生資料外，此文件還包括發布來自qPCR或ddPCR分析的物種出現資料的建議。

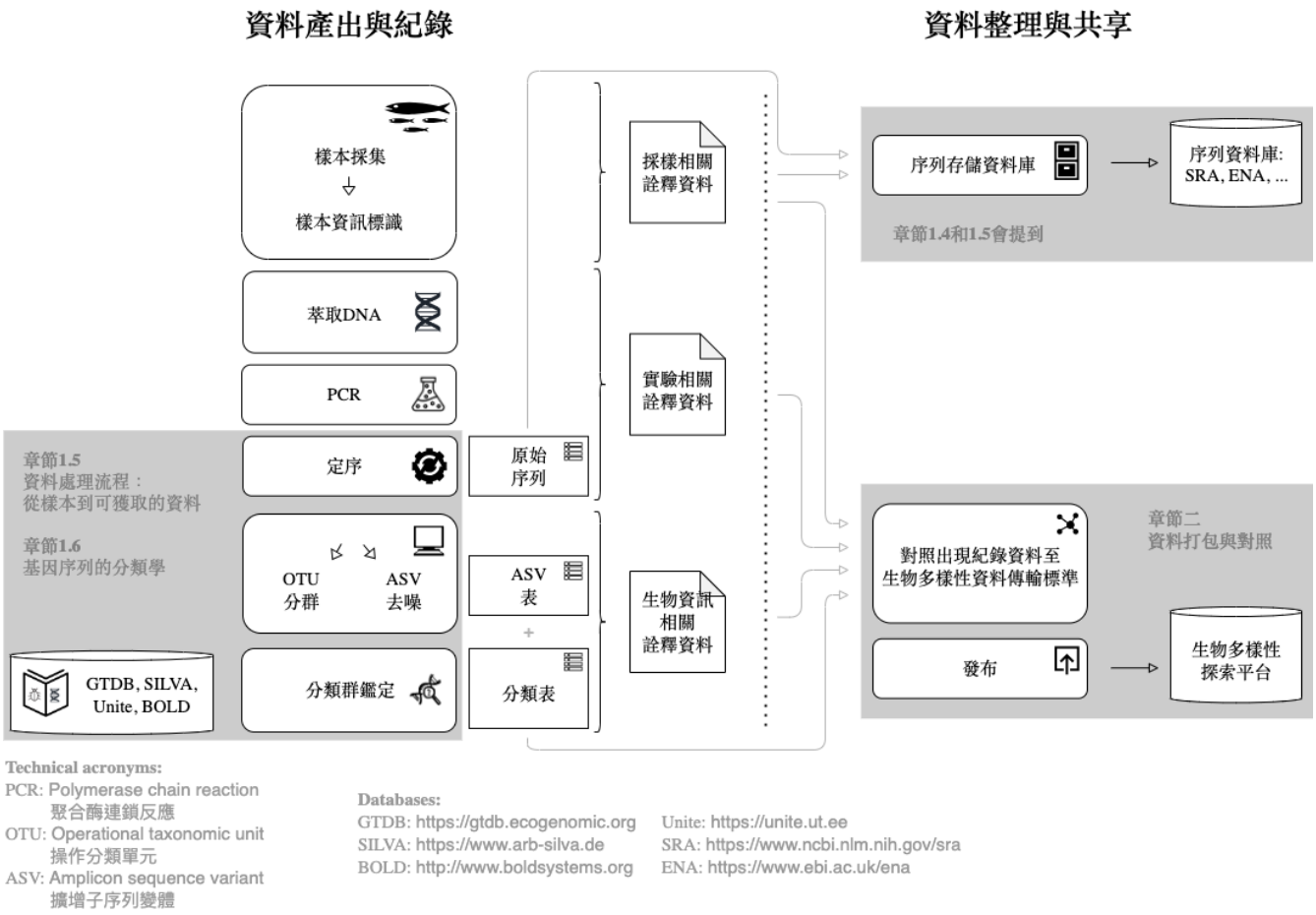
以開放和可重複性的方式回報DNA衍生的出現紀錄有很多好處。它增加了可引用性，凸顯了生物保育背景下相關的分類群，並有助於分類學和生態學知識。除此之外，它還使用了另一種方式來存儲未描述物種的出現記錄。當這個尚未被描述的物種分類最終與一個新的林奈名稱相聯時，所有相關的出現記錄將連結與可立即被使用。這些好處都為專業人士提供了強有力的理由來採用此文件中概述的做法，幫助他們凸顯出現存更廣的生物多樣性，加速其發現並將其納入生物保育和政策制定。

1.2. 目標受眾

此文件針對多個目標受眾：與DNA有相關的研究計畫的學生、想要恢復或保存舊基因序列和豐富度資料的研究人員、不熟悉DNA衍生出現紀錄的生物多樣性資訊專家，以及不熟悉生物多樣性資訊平台但了解基因序列的生物資訊學家。此文件不直接針對在生物多樣性資訊平台中使用分子生物資料的人員，但此類使用者可能

對第1.7章節的資料輸出有興趣。作者們的目的是為針對生物多樣性資訊平台中發布基因序列衍生與相關屬性資料提供指引。

流程圖概述GBIF和國家級生物多樣性資訊平台 (包括建立在ALA上的平台) 等資料庫中發布擴增子 (amplicon) 衍生分子生物多樣性資料所涉及的處理步驟。此文件的重點主要放在定序後得到原始FASTQ序列的後續步驟。通過流程圖，並註意任何看起來熟悉或不清楚的步驟，讀者將能夠以此流程作參考並瀏覽此文件中所包含的內容。



圖表 *1. DNA*衍生生物多樣性資料在此文件中的流程圖總覽

我們盡力在此文件中提供對讀者有幫助的資訊，但延伸閱讀 (例如：GBIF資料發布快速指南) 在某些情況下可能會有幫助。

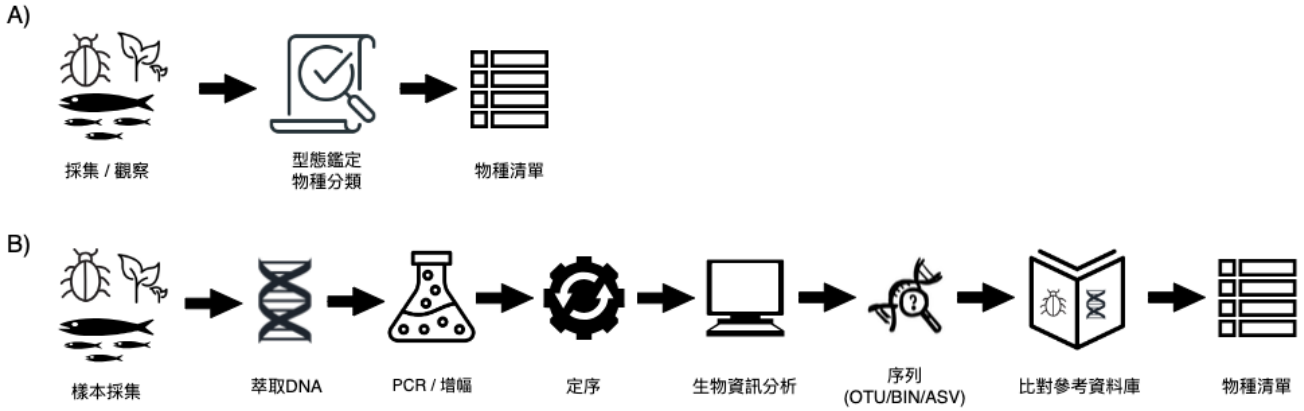
1.3. DNA衍生出現紀錄資料之簡介

DNA衍生的出現紀錄包括來自個體生物DNA的資訊、環境DNA的資訊 (eDNA，即從環境樣本中所萃取的DNA，Thomsen & Willerslev，2015) 和包含許多個體的大量樣本 (bulk samples，例如浮游生物樣本或來自Malaise的陷阱樣本，由許多個物種及多個個體所組成)。目前，最大量的DNA衍生出現紀錄來自eDNA。由於所有樣品來源的分析方法和最終成果在很大程度上相似，因此下面的討論將集中在eDNA (§ 2.1.1 和 § 2.1.2)，如適用於其他來源，也將特別提出。調查通常利用分類學和遺傳學上可參考的目標區間定序，但也可以使用在不直接產生DNA序列資料，如基於qPCR的方法 (§ 2.1.3 和 § 2.2.2)。此文件裡與DNA相關的所使用術語請查閱詞彙。

1.3.1. DNA衍生出現紀錄來源之一—環境DNA

環境DNA (eDNA) 這個詞自1987年來開始使用，當時首次用於描述沉澱物樣品中微生物的DNA (Ogram等人，1987)。現在eDNA更廣泛的被用於描述來自不同生物體DNA的混合物 (Taberlet等人，2018 和 2012)。因此，eDNA的定義為從特定環境樣本中萃取的所有DNA，不論其基質和包含的物種為何。它可以從多種來源中萃取，包括皮膚、毛細胞、唾液、土壤、糞便以及活的或最近死亡的生物體 (

Pietramellara等人，2009)。環境DNA通常足以代表樣本中的所有生物體。 然而實際上，環境樣本中DNA的存在取決於棲息地選擇、生物體體型、形態、活動量、用於採集DNA的採樣方法 (Taberlet等人，2018) 以及DNA的降解程度。



圖表 *2.* 這張示意圖呈現了採樣過程，比較了 *A)* 傳統生態/生物多樣性抽樣方法和 *B)* 以*eDNA*為基礎研究 *(*此圖以高通量分子條碼研究為例*)* 所收集的資料。對於*eDNA*研究來說，基因定序之前的大部分步驟將涉及使用技術性及生物性重複樣本，以便識別污染、假陽性以及假陰性的結果。這從而造就資料和詮釋資料的架構。研究會有可能包括這兩種類型的採樣模式。例如，如果 *B)* 中使用的「參考資料庫」不包含特定生物組相關的物種資訊，則有必要返回使用方法 *A)*。在有些情況下，也有可能出現在對照「參考資料庫」後產生了意想不到或不太可能的結果，則需要使用傳統方法進行進一步研究，以確定生物資訊分析鑑定的物種是否可以得到驗證。

因此，eDNA不是一種方法而是一種樣本類型，包括來自任何環境樣本的DNA，而不是來自目標個體的捕獲和定序。此類樣本類型包括水、土壤、沉澱物和空氣，還包括不以宿主DNA為目標的腸道內容物樣本和組織(植物/動物) (Taberlet等人，2018)。研究環境DNA有多種分析方法可用，大多分為兩大類：1) 目標為檢測特定生物體；2) 歸納與描述生物體的群落。不同的分析方法將會產生不同類型和數量的資料。在大多數DNA濃度較低的情況下，應使用技術性及生物性重複樣本來驗證所檢測到的物種。

多項研究顯示，eDNA用於水樣的分析可能比傳統方法更有可能發現稀有且難以調查的物種 (Thomsen等人，2012；Biggs，2015；Valentini等人，2016；Bessey等人，2020)，而在其他環境中也可能如此。儘管實際的生物體不再存在於那裡，但有時仍然可以檢測到DNA痕跡。因此，eDNA適合用來監測稀有的紅皮書名錄物種和外來物種，因這些物種通常密度較低，並且難以用常規方法檢測到。環境DNA方法也能夠檢測隱藏的生物體，尤其是那些肉眼無法目測到的微小生物體 (例如細菌和真菌)。此外，eDNA還可用於同時觀察許多物種，即用來描述整個生物群落或其主要組成 (Ekrem & Majaneva，2019)。

一些研究顯示，在環境樣本中，特定物種的DNA量與該物種在環境中的生物質之間存在某種關係。因此，環境DNA有潛力做為對生物質進行所謂的半定量估算 (semi-quantitative；間接目標)，無論是從環境樣本還是大量樣本中 (Takahara等人，2012；Thomsen等人，2012；Andersen等人，2012； Ovaskainen等人，2013；Lacoursière-Roussel等人，2016；Thomsen等人，2016；Valentini等人，2016；link: Fossøy等人，2019；Yates等人，2019；Doi等人，2017)。然而其他研究顯示，環境DNA的量與估算的族群密度之間幾乎沒有相關性 (Knudsen等人，2019)。PCR、定量、混合和其他偏差議題經常有爭議。例如脫皮、繁殖和大規模死亡可能會導致水中甲殼動物環境DNA水平增加，而濁度和水質惡劣則會減少可檢測到的環境DNA量 (Strand等人，2019)。因此，我們鼓勵資料發布者提供每個樣本每個OTU或ASV的read counts以及total read counts per sample，因為這可以讓資料使用者對於物種出現/無出現和 (相對) 豐度在使用資料時做參考並自行判斷是否適用。

1.3.2. DNA高通量分子條碼 (Metabarcoding)：序列衍生資料

由於DNA-metabarcoding的迅速發展，序列衍生資料目前正迅速的產生和增加中。此方法利用通用的引子，借助高通量定序 (High-throughtput sequencing, HTS，或稱次世代測序 (NGS))，為特定的生物群體產生數千至數百萬個短DNA序列。通過將每個DNA序列與GenBank (

Benson等人，2006)、BOLD (Ratnasingham等人，2007) 或UNITE (Nilsson等人，2019)等基因序列資料庫進行比對，可以將每個序列鑑定至物種或更高階層的分類單元。 DNA-metabarcoding可以用於源自陸地和水域環境的樣本，包括水、土壤、空氣、沉積物、生物膜、浮游生物、大體積樣本和糞便樣本，並可同時識別數百種物種 (Ruppert等人，2019)。

調查中使用序列或目標區間 (marker-based) 的定序並要成功鑑定物種，取決於那些已在形態上做了相關研究並已辨識的標本中所提取的基因序列。這些序列存在在基因序列資料庫中才能與調查所產出的新序列進行比對。分類的有效性取決於基因序列資料庫的完整性 (覆蓋率) 和可靠性，以及用於進行分類的分析工具。這些都是不斷在變化中的目標，因此在解釋結果時應謹慎且運用分類學專業知識 (§ 1.6)。所有經過驗證的 ASV (Amplicon Sequence Variants) (Callahan等人，2017)將有機會在未來對資料進行重分析讓其更精確、進行種內族群遺傳分析 (Sigsgaard等人，2019) 並且可能增加鑑定的準確性。基於這個原因，我們建議發布共享 (序列未分群的，unclustered) ASV資料。

1.3.3. 總體基因體學 (Metagenomic)：序列衍生資料

基因序列衍生的生物多樣性資料也可以從總體基因體 (Metagenomic) 的方法生成，即樣本中所有的DNA都進行定序 (Tyson & Hugenholtz，2005)且不使用特定的引子或分子條碼。從總體基因體序列中獲得的衍生序列生物多樣性資料可以與基因資料庫裡的已註解基因資料做比對並做註解 (如上所示)，也可以作為 (接近) 完整的總體組裝基因體 (Metagenome assembled genomes, MAGs)。雖然metabarcoding的方法在基因序列衍生生物多樣性資訊方面仍然佔主導地位，但總體基因體資料來源也變得越來越重要。這已證實在MAGs數量的迅速增加以及它們在系統分類學中的實用性 (Parks等人，2020)；但對於與MAGs分析相關迅速發展方法的討論超出了此文件的範疇。此文件主要圍繞metabarcoding作為發布序列衍生生物多樣性資料的概念和方法進行討論。雖然與MAGs在生物資訊分析上有差異，但最終結果 (一條序列，通常以contig/assembly的形式存在) 與metabarcoding資料概念一致 (即樣本收集，資料產生和工作處理流程的詮釋資料應該被記錄)。

1.3.4. qPCR/ddPCR：出現紀錄資料

eDNA樣本中特定物種檢測的大多數分析包括使用物種特定引子和qPCR (定量聚合酶鏈反應) 或ddPCR (微滴數位聚合酶鏈反應)。這些方法不生成DNA序列且出現紀錄完全取決於引子/assay的特異性。因此，對於這些來源的發布資料，有著如何驗證該嚴格性的建議與要求 (Bustin等人，2009；Huggett等人，2013)，以及對長期監測中可就緒應用的狀態要求 (Thalinger等人 2020)。使用qPCR分析eDNA樣本不需要大量的資源，並且可以在大多數DNA實驗室中進行。首次使用eDNA水樣本檢測入侵種的例子是利用qPCR來檢測美洲牛蛙 (*Rana catesbeiana*)(Ficetola等人 2008)。eDNA水樣本的qPCR分析可定期用於監測特定的魚類、兩棲動物、軟體動物、甲殼動物等物種，以及它們的寄生蟲 (Hernandez等人，2020，Wacker等人，2019，Fossøy等人，2019，Wittwer等人，2019)。因此，使用qPCR的eDNA檢測單一物種是重要的出現紀錄。

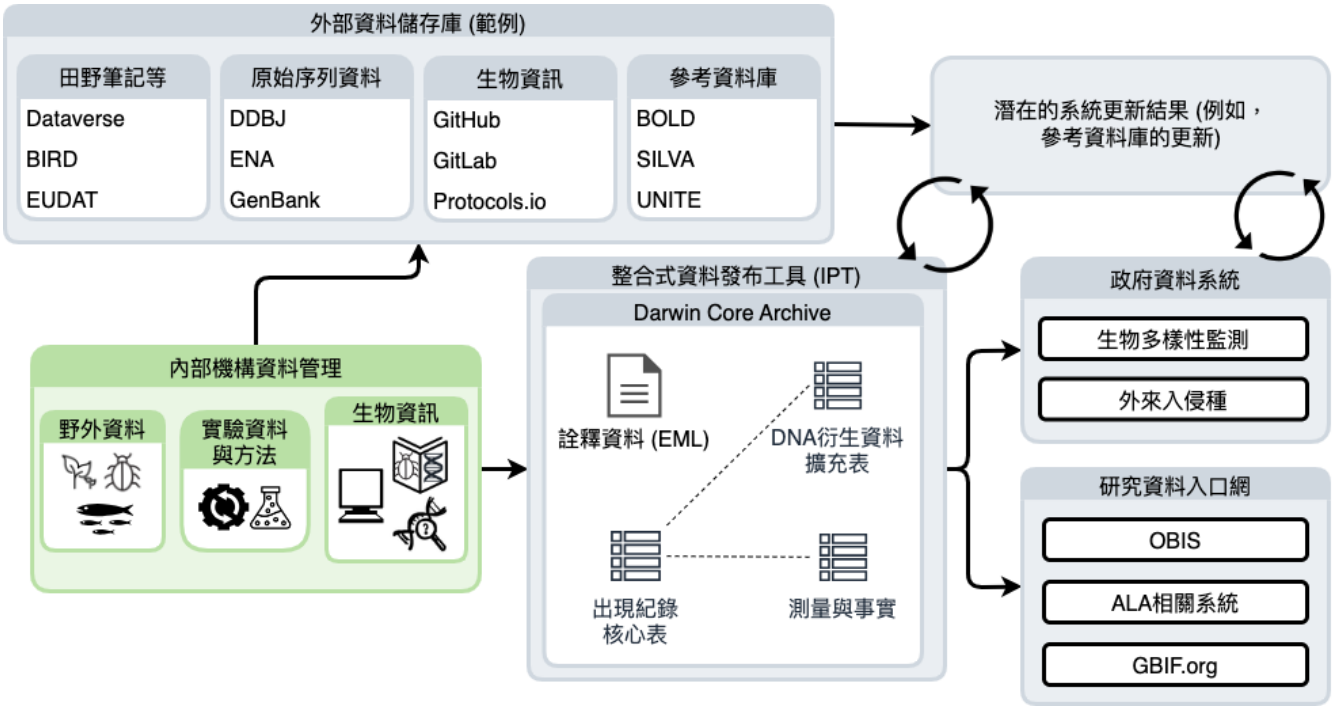
1.4. 生物多樣性發布介紹

發布生物多樣性資料的過程主要是根據FAIR原則 (Wilkinson等人，2016)，使物種出現資料可查找、可取用、可交換和可重複使用。生物多樣性資料平台有助於將基因序列資料被發現，並與其他類型的生物多樣性資料一起作為出現記錄進行共享，例如博物館收藏標本、公民科學觀測和傳統野外調查。每個原始資料的結構、管理和存儲將根據每個社群的需求而變化。生物多樣性資料平台通過使這些單獨資料集的相互兼容來支持資料的被發現、查詢和重複使用，以解決資料上分類、空間和其他不一致性。通過單一資料入口網做資料搜尋可以支持大規模的大數據研究、管理和政策。資料集之間的兼容性是通過標準化過程建立的。

通用的生物多樣性資料使用了一些資料標準，而基因序列資料則使用了另一套獨立的標準 (請參閱MIxS和GGBN)。此文件呈現了在為增加通用生物多樣性和基因資料標準兼容性上持續努力的成果。標準通常會突顯出最重要或最常用的字段 (fields) 子集。這些子集可以被稱為「核心」(cores)。目前，在GBIF和ALA網絡中發布資料的首選格式是使用 Darwin Core (DwC) 資料標準的達爾文核心集資料包 (DwC-A)。這是一個包含資料文件的壓縮文件

(zip)，以標準的逗號或制表符分隔的文本格式、一個描述資料資源的詮釋資料文件 (eml.xml)、以及一個說明資料包中包含的文件和字段結構的詮釋資料文件
(meta.xml)。標準化的打包確保了資料可以使用特定的資料交換方式在系統之間流通。此文件的第二章節提供了資料文件的對照（mapping）建議，而有關構建xml文件的方法和工具可以在此找到：TDWG、GBIF、和ALA。

標準化過程的主要部分是字段的對照 (mapping)，這是將原本資料的字段 (列)結構轉換為標準字段結構所必需的。標準化過程也可能影響每筆記錄中各個字段的內容，例如，座標重新轉換成共同常用的系統、重新排列日期或將字段內容對照為一組標準值，通常稱為控制詞彙
(vocabulary)。標準化的過程還提供了改善資料品質的機會，例如，填補遺漏、更正拼寫錯誤和多餘空格以及處理不一致使用的字段。這些修改提高了資料品質並增加了其可重用性。但同時，以任何狀態發布的資料都比未發布和無法查詢的資料更好。為了保留原始資料的原樣，標準化處理通常使用原始資料的副本或輸出的資料。



圖表 *3.* 呈現和發布*DNA*序列及相關詮釋資料的平台概要 *(*綠色框*)*，基於現有的系統和資料標準 *(*灰色框*)*。一個設想的系統 *(*基於機器對機器讀取資料*)* 定期更新結果 *(*白色框*)* 可以讀取並更新達爾文核心檔案或各種管理系統。各元素之間的資料傳輸 *(*黑色箭頭*)*
將需要各種程度的資料轉換和協調，可能包括機器或人工的品質評估。

一旦資料集通過標準化和資料品質處理流程，應該將其放置在可查詢的線上位置並與相關的詮釋資料相關聯。詮釋資料是關於資料集裡的資料資訊，包括描述資料集的關鍵參數讓其進一步提高其可發現性和重用性。詮釋資料應該包括其他重要資訊，如作者、DOI、組織單位以及其他資訊，以及有關資料集收集和管理的程序和方法資訊。我們鼓勵在EML文件的方法裡提供工作流程細節、版本描述和品質控管。

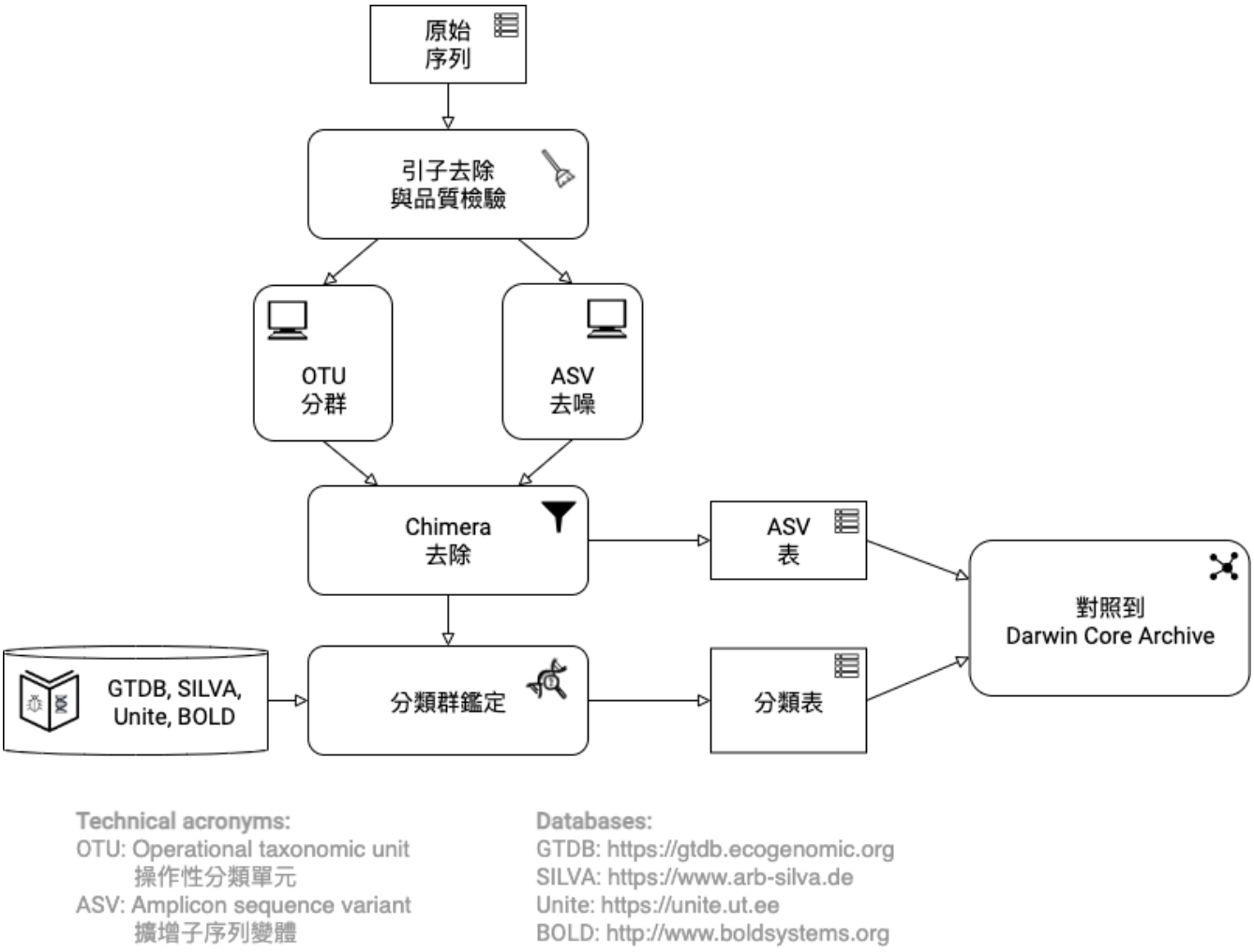
資料集及其相關的詮釋資料被每個資料入口網索引：使用戶可以在這過程通過API和線上查詢、篩選和處理資料。與期刊出版物不同的地方是，資料集可能是動態產品。即在相同標題與DOI的情況下，資料集可會不斷變化：更新成不同的版本、增加資料筆數和更改詮釋資料字段。

基因序列資料的持有者應該將基因序列上傳並存檔在原始序列資料存儲庫中，例如NCBI的SRA，EMBL的ENA，或DDBJ。這裡不涉及序列存檔的討論，不過 Penev等人 (2017) 提供了有關科學出版物中資料提交和指南的概述。ALA、GBIF和大多數國家生物多樣性入口網與平台不是原始序列和相關文件的存儲庫。我們在第二章節中強調了維持這些原始資料與衍生出現紀錄之間連結的重要性。

1.5. 資料處理流程：從樣本到可獲取的資料

不同定序技術 (Illumina、PacBio、Oxford Nanopore、Ion Torrent等) 使

metabarcoding資料生成方面的原則不同，令reads長度、錯誤配置文件以及序列是單端還是雙端等方面存在著差異。目前，　　Illumina短序列平台是最廣泛採用的，因此在這裡的描述以其為基準。然而，無論使用哪種定序技術，資料的生物資訊處理都遵循相同的一般原則（QC，去噪，分類）(Hugerth等人，2017，圖表2)



圖表 *4.* 生物資訊分析*metabarcoding*資料的流程。

DNA序列通常會先針對去除引子序列進行處理，然後根據所使用的定序方法，通常朝向序列的5'和3'端去除低品質碱基。不符合長度、整體質量、引子、標簽等要求的序列將被移除。

預處理過的序列可以通過與參考資料庫進行比對來鑑定分類單元。當參考資料庫不完整時，也可以在無進行分類識別的情況下對序列進行分群　　(clustering)，方法之一是根據它們的相似性將序列分群為操作性分類單元（OTUs；Blaxter等人，2005)。另一種方法是對序列進行去噪(denoise)，即明確檢測和排除PCR/序列的錯誤，以生成ASV；也稱為zero-radius　　OTU (zOTU)。去噪試圖糾正PCR和/或序列步驟中產生的錯誤，使得去噪後的序列是原始序列混合物中存在的真實生物序列。Paired-
end序列可以在合併之前對前向和反向序列個別進行去噪，或者在去噪前進行合併。結果集中的ASVs可能相差僅一個碱基，這表明了种内或种间的序列变异。在操作上，ASVs可以被認為是没有定義radius的OTUs。儘管去噪演算法通常非常出色，但它們並不能完全去除分裂或合併序列的問題。

用於生成定序library的PCR可能會產生chimera序列，即來自多個母序列的單個序列。這樣的序列可以通過生物資訊學方法檢測並移除，通常是在去噪之後所進行。

最後，將處理後的序列、OTUs或ASVs與已註解序列的資料庫　　(通常稱為參考庫，見§　　1.6)進行比對從而進行分類。與前幾個步驟一樣，這裡也有幾種替代方法可供選擇。其中大多數為將metabarcoding序列與參考序列進行比對或計算共享的k-mers (短精確序列) 數。

現存許多用於metabarcoding資料的開源生物資訊處理工具和演算法　　(QIIME2　　(Bolyen等人，2019)

，DADA2 (Callahan等人，2016)，SWARM (Mahé等人，2014)，USEARCH (Edgar，2010)，Mothur (Schloss等人，2009)，LULU (Frøslev等人，2017)，PROTAX (Somervuo等人，2016)，VSEARCH (Rognes等人，2016))。鑑於存在許多廣泛使用的流程方法，我們以下提出一些為提交到生物多樣性資料平台做資料分析的建議。這並不是表明這些是最佳或最適合所有目的的方法，而是試圖鼓勵提交相對標準化的資料讓平台之間進行比較。如果可以的話，應使用一個紀錄完整且維護良好的操作流程 (如nf-core/ampliseq pipeline)。詮釋資料應包括工作流程的詳細資訊和版本，納入在詮釋資料裡的方法步驟中或是在DNA衍生資料擴充表裡的SOP字段裡引用文獻 (見表格4中的mapping)。序列資料應存儲在適當的核苷酸檔案中 (NCBI的SRA: Leinonen等人，2011) 或 EMBL的ENA (Amid等人，2020))，並且提交到生物多樣性平台的資料應包括從 biosampleID (見§ 2.2)。使用這些sample ID將會降低資料重復的可能性，並確保序列資料在有重新分析的機會出現時能夠很容易地獲取，因為參考資料庫和生物資訊工具在不斷的改進中。這些操作的最終產品通常是每個樣本中個別OTU或ASV的計數文件以及對這些進行的分類學鑑定。這通常會以表格或BIOM (McDonald等人，2012) 的格式生成。OTU或ASV序列通常也以FASTA格式提供 (Pearson & Lipman，1988)。

1.6. 基因序列的分類學

序列的分類學鑑定是處理分子生物多樣性資料集的關鍵步驟，因為學名對於查詢和傳達有關觀察到的生物體的資訊至關重要。序列辨識的準確性和精確性將取決於可靠並具有生命樹所有分支的參考資料庫，這將需要來自分類學家和分子生態學家的共同努力。使用公開序列資料庫時應該知道它們存在各種缺陷，例如分類的可靠性和缺乏標準化的數據詞彙 (Hofstetter等人，2019；Durkin等人，2020)。

物種，如分類學家所描述，對於生物學至關重要。因此對生物多樣性進行歸納應該會利用分類研究的成果。然而，與DNA序列資料不同，分類學的產出並不總是容易拿來直接進行演算法或計算：傳統的分類學是一個人為驅動的過程，其中包括分類單元的手動確定、描述和命名的步驟，最終以符合國際命名法規的正式出版物告終。正如在前幾章中討論的那樣，基於DNA序列的調查非常有效地檢測難以觀察到的物種，並且通常會識別出目前超出傳統林奈分類知識範圍的生物體的存在。雖然此文件沒有涉及討論發布序列資料衍生的物種清單，但傳統分類學與eDNA工作之間的脫節是不可取的。因此我們向此文件的讀者提出以下建議。

由於分類學對於發現生物多樣性資料至關重要，因此強烈建議任何eDNA定序的工作都應該尋求在研究中包含相關的分類學專業知識。同樣，如果eDNA序列研究能夠將部分預算用於從以前未進行過序列化的模式標本或其他重要的參考材料生成和發布參考序列，則將非常有益。這些標本可以來自當地的標本館、博物館或生物收藏。分類學家們也可以通過始終在每個新物種描述中包含相關的DNA序列 (Miralles等人，2020) 以及對eDNA工作揭示的許多新生物體進行的貢獻 (如Tedersoo等人，2017)。

大多當前的生物多樣性資料平台都是以傳統的名稱列表和分類索引為基礎的。由於DNA序列衍生出現紀錄正迅速成為生物多樣性資料的重要來源，但此類資料的官方分類和命名的存在暫時還處在跟不上的現象。因此建議資料提供者和平台應繼續探索並將更彈性的分類形式納入其分類體系中。這包括分子參考資料庫 (例如GTDB、BOLD、UNITE)
承認並納入以前來自未成功分類生物的參考材料的序列資料。此外，我們建議其他常用的分子資料庫 (例如PR2、RDP、SILVA) 應該為分類單元開發穩定的識別碼並提供參考序列，以便將它們用作分類參考。

與傳統的分類學相比，將DNA序列分群到分類概念中依賴其相似性和其他資訊 (如親緣關係和概率) 的算法分析，以及一些人為編輯。OTUs在穩定性、參考序列和實物的存在、序列對齊和cut-off值、OTU識別碼 (如DOI) 等因素存在差異 (Nilsson等人，2019)。它們在規模上也存在差異：從局部研究或計畫特定的資料庫到能夠進行更廣泛的跨研究比較的全球資料庫。與正式描述在研究出版物中的林奈分類中心化和編碼化不同，OTUs分佈在多個不同的演變中的數位參考庫中，這些參考庫在分類焦點、條碼基因和其他因素上存在差異。通過將標準序列與已鑑定的參考標本關聯起來，BOLD和UNITE正在建立一個重要的對照層，將ASVs和OTUs與林奈分類相關聯。GBIF分類骨幹包括了對UNITE物種假設 (SHs) 和Barcode Index Numbers (BINs) 的識別碼，這使真菌和動物的OTU層級的物種出現紀錄可以進行分類鑑定並進行索引 (GBIF secretariat，2018，Grosjean，2019)。

eDNA分類鑑定的演算法通常會將每個獨特的序列根據某些分類群上最相關性和信心度的標準做鑑定。對於資訊較少的生物群體例如原核生物、昆蟲和真菌等，鑑定可能是一個非林奈式的暫時名稱 (cluster-based) 分類單元 (即相關SH或BIN的ID/編號)，而該分類單元可能代表一個物種或是在物種之上的階層。因地球上有許多未知、未鑒定和未描述的物種，

所以沒有任何參考資料庫包含所有物種。因此在過去30年中，對這一事實的普遍忽視已成為許多分類錯誤的源頭。

由於DNA衍生出現紀錄資料在與參考資料庫 (例如UNITE、BOLD) 進行比對的當時，參考資料庫沒有完整或未更新的分類索引，在將這些DNA衍生出現紀錄導入生物多樣性平台 (例如GBIF或OBIS)時，獲得的名稱/ID可能降低這些出現紀錄的分類的辨識率。然而，將這些OTU或ASV序列做發布將使未來使用者有機會將序列識別到更高分類階層，特別是參考資料庫會隨著時間的改善更新。在無法將序列做發布的情況下，我們主張將該物種的 (科學性或佔位符) 名稱 (例如BOLD BIN或UNITE SH) 加上序列的MD5校驗，作為唯一物種ID (§ 2.2, "資料對照 (mapping)")。MD5校驗是一種單向hash演算法，通常用於驗證文件完整性。它們將可以在無法得到序列本身的情況下，提供原始序列唯一且可重複的代表性。這在獲取敏感資料的情況下可能是必需的。MD5校驗可有效地查詢來確定其他eDNA工作中是否也得到了相同的序列，但它不能完全取代序列所以無法進行進一步的分析。類似BLAST的序列相似度搜索會無法在這裡派上用場，因為相差一個碱基的兩個序列將獲得完全不同的MD5校驗。

1.7. 資料輸出

將DNA衍生資料通過生物多樣性平台發布的目的是使這些資料能夠與其他生物多樣性資料類型結合並可重複使用。在準備資料發布時，保持這種可重複使用的目的非常重要。理想情況下，詮釋資料和資料應該以完整的方式講述一個完整的故事，以便新的、不知情的使用者可以在不需要任何額外諮詢或通信的情況下使用這些證據。生物多樣性資料平台提供搜索、篩選、瀏覽、視覺化、資料訪問和引用功能。對於metabarcoding資料，我們鼓勵使用者對資料進行適當的篩選，如配置最小絕對和相對read豐度的移除條件，這設定將會移除每個低於某個選定絕對值read數的OTU或ASV (使用organismQuantity字段)，也會移除掉singletons。除此之外，也可以移除相對read數低於選定閾值的出現紀錄。這是從檢測到的reads (organismQuantity) 和相應樣本中的總read數 (sampleSizeValue) 計算出來的 (§ 2.2.1)。使用者通常可以選擇資料輸出的格式 (例如DwC-A、CSV)，然後將資料加工、清理和轉換為所需的分析形式和格式。

在 GBIF.org 或通過 GBIF API，註冊用戶可以以下列三種方式搜尋、篩選和下載生物多樣性資料：

- 簡易格式 **(Simple)**：一種簡易的以制表符分隔的格式，僅包含由GBIF詮釋的資料版本，來自索引過程的結果。這適用於快速測試和直接導入試算表程式做開啟。
- 達爾文核心資料包 **(Darwin Core Archive)**：較豐富的資料格式，包括發布者提供的原始資料版本 (在GBIF進行索引和詮釋之前)。這種格式可以從下載的資料集中得到更多資訊，因為它也包含了詮釋資料和資料問題標記。
- 物種清單 **(Species list)**：一個簡單的表格格式，僅包括從資料集或查詢結果中已詮釋的物種名稱列表。

無論所選格式為何，每個GBIF用戶的下載動作都會收到一個可重複使用的查詢連結和一個包含DOI的資料引用。這種基於DOI的引用系統提供了一種辨識和表彰資料集及資料來源的方法，提高了資料的研究結果的可信度和透明度。遵循資料引用建議並使用DOI是至關重要的，因為良好的資料引用文化不僅是學術標準，而且還是一種為資料發布者帶來認可和激勵的強大機制。

2. 資料打包與對照 (mapping)

此章節主要介紹如何將您的資料輸出轉換為生物多樣性資料平台可索引的資料集。§ 2.1 將幫助您了解您手上資料的最佳對照架構。§ 2.2 詳細描述了這些對照。

本文件將生物多樣性資料發布的標準與基因DNA衍生的生物多樣性資料相結合 (圖表 5)。此流程呈現不同類型DNA衍生資料的對照建議。

資料打包和發布途徑會因平台而異，這會在文件中進行描述。目前，一種廣泛使用的資料文件打包方式是DwC-A，其中資料表是按照星芒架構排列，外部擴充表中的記錄 (行) 指向核心文件中的單個記錄 (圖表 5)。不同核心類型的文件 (例如出現紀錄和事件) 對應於不同類型資料集。儘管DNA衍生資料集通常是以事件為主，即數百甚至數千個已定量序列的出現紀錄可能源於單個採樣事件也共享

大多詮釋資料的屬性，但目前的建議是將資料發布為出現紀錄核心 (類別一或二)，並附上DNA衍生資料擴充表。這種方法彌補了DwC星芒架構的限制，既該架構不允許擴充文件中有任何出現紀錄資料(例如處理、分析後的條碼序列) 指向事件核心文件中的記錄。但是，我們建議為每個核心記錄加上一個eventID，以指示來自同一採樣事件的出現之間的關聯。

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

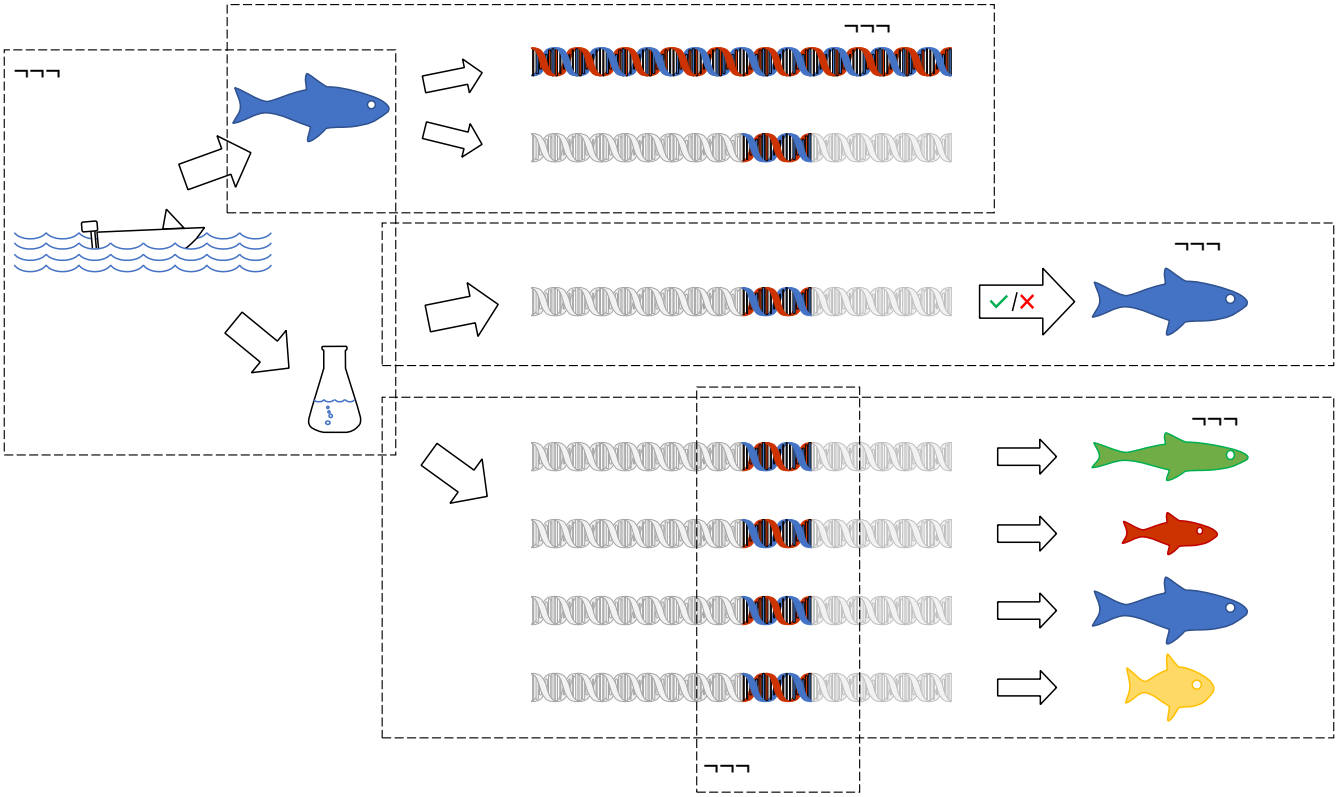
圖表 *5.* 第*1.4*章節中圖表*3*的*DwC-A / IPT*的放大圖。核心的選擇主要是將資料匹配到生物多樣性資料平台的資料導入機制 *(ingestion)* 中。大多數資料可以制定為出現紀錄、事件或分類單位的核心，但由於只有核心可以擴充，這將會影響其選擇。例如，如果使用事件核心包裝資料，則無法將*DNA*序列擴充表作為出現紀錄。

2.1. 資料類別

以這份文件的目的，我們將資料分為五個類別，通過一個關鍵的ID字段 (*eventID*) 進行關聯，該字段相當於一般生物多樣性資料的標準，並包括與DNA相關的字段 (見 § 2.2, "資料對照(mapping)")。這五個類別旨在反映研究生物多樣性裡最常用的分子方法，分別是 (一) DNA衍生出現紀錄、(二) 多重引證出現紀錄、(三) 目標物種檢測、(四) 物種基因代號、和 (五) 僅有詮釋資料的資料集。使用下列的決策樹來引導至該類別。

表格 *1. DNA*衍生資料類別的決策樹。

<table>
<tr><td colspan="4">? 您的資料來源是 (meta)barcoding 或是 qPCR?</td></tr>
<tr><td colspan="3">(Meta)barcoding
↓
? 資料包含數位化的基因物質或序列，並與地點和時間相關聯嗎？</td><td>qPCR
↓
類別三
目標物種檢測</td></tr>
<tr><td colspan="2">是
↓
?
遺傳物質是特定生物或群體的唯一證據嗎？</td><td>否
↓
? 資料集是基於DNA的名錄清單嗎？</td><td></td></tr>
<tr><td>是
↓
類別一
DNA衍生出現紀錄</td><td>否
↓
類別二
多重引證出現紀錄</td><td>是
↓
類別四
物種基因代號</td><td>否
↓
類別五
僅有詮釋資料</td></tr>
</table>



圖表 *6.* 類別一至五的視覺化呈現。

2.1.1. 類別一：DNA衍生出現紀錄

此類別涉及到的資料是指DNA序列或PCR檢測為唯一用來證明特定生物體或群落存在的證據。換句話說，這些資料無法追溯到可觀察的標本。這是許多總體基因體學、metabarcoding和eDNA研究的情況。

DNA衍生出現紀錄資料集範例

- MGnify (2019) Impact of rainforest transformation on phylogenetic and functional diversity of soil prokaryotic communities in Sumatra (Indonesia). Sampling event dataset https://doi.org/10.15468/osp7hi accessed via GBIF.org on 2020-04-16.
- MGnify (2020) Marine metagenomes from the bioGEOTRACES project. Sampling event dataset https://doi.org/10.15468/oifcho accessed via GBIF.org on 2020-04-16.
- Bessey C, Jarman SN, Berry O et al. (2020) Maximizing fish detection with eDNA metabarcoding. Environmental DNA: 1–12. https://doi.org/10.1002/edn3.74 (Atlas of Living Australia website at https://collections.ala.org.au/public/show/dr14581. Accessed 24 June 2020)

請參考 § 2.2.1 以了解如何格式化和分享這些資料集。有關達爾文核心出現紀錄資料集可通過達爾文核心出現紀錄資料集的DwC-A模板和出現紀錄資料的品質要求獲得。

2.1.2. 類別二：多重引證出現紀錄

如果某些遺傳物質可以與之觀察或標本相關聯，我們將此類資料歸類為「多重引證出現紀錄」。這情形下，出現紀錄的該序列並不是唯一的證據來源。我們總是可以將資訊追溯到有標本證明或觀察到的生物體，例如該類別包括分子條碼資料集和一些具有參考材料的DNA metabarcoding資料集。有關分子條碼的更多資訊，請參閱Centre for Biodiversity Genomics, University of Guelph (2021年)。

多重引證出現紀錄資料集範例

- The International Barcode of Life Consortium (2016) International Barcode of Life project (iBOL). Occurrence dataset https://doi.org/10.15468/inygc6 accessed via GBIF.org on 2020-04-16.

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

請參考 § 2.2.1 以了解如何格式化和分享這些資料集。有關達爾文核心出現紀錄資料集可通過達爾文核心出現紀錄資料集的DwC-A模板和出現紀錄資料的品質要求獲得。

2.1.3. 類別三：目標物種檢測 (qPCR/ddPCR)

此類別涉及使用特定 (qPCR/ddPCR) assay檢測環境樣本中目標生物的DNA序列存在 (或不存在)的資料。在這種情況下，出現記錄甚至可能不包含序列資料，因為該檢測方法確定了該物種的「存在」。通過qPCR/ddPCR的分析來進行目標物種的檢測，許多研究也報告了特定樣本中該特定物種的「不存在」。不存在資料高度依賴於特定assay的檢測限制，以及野外和實驗室的操作方法。DNA-metabarcoding資料存在假陰性和假陽性結果的問題，所以重要的是發布足夠的資訊以評估該記錄。

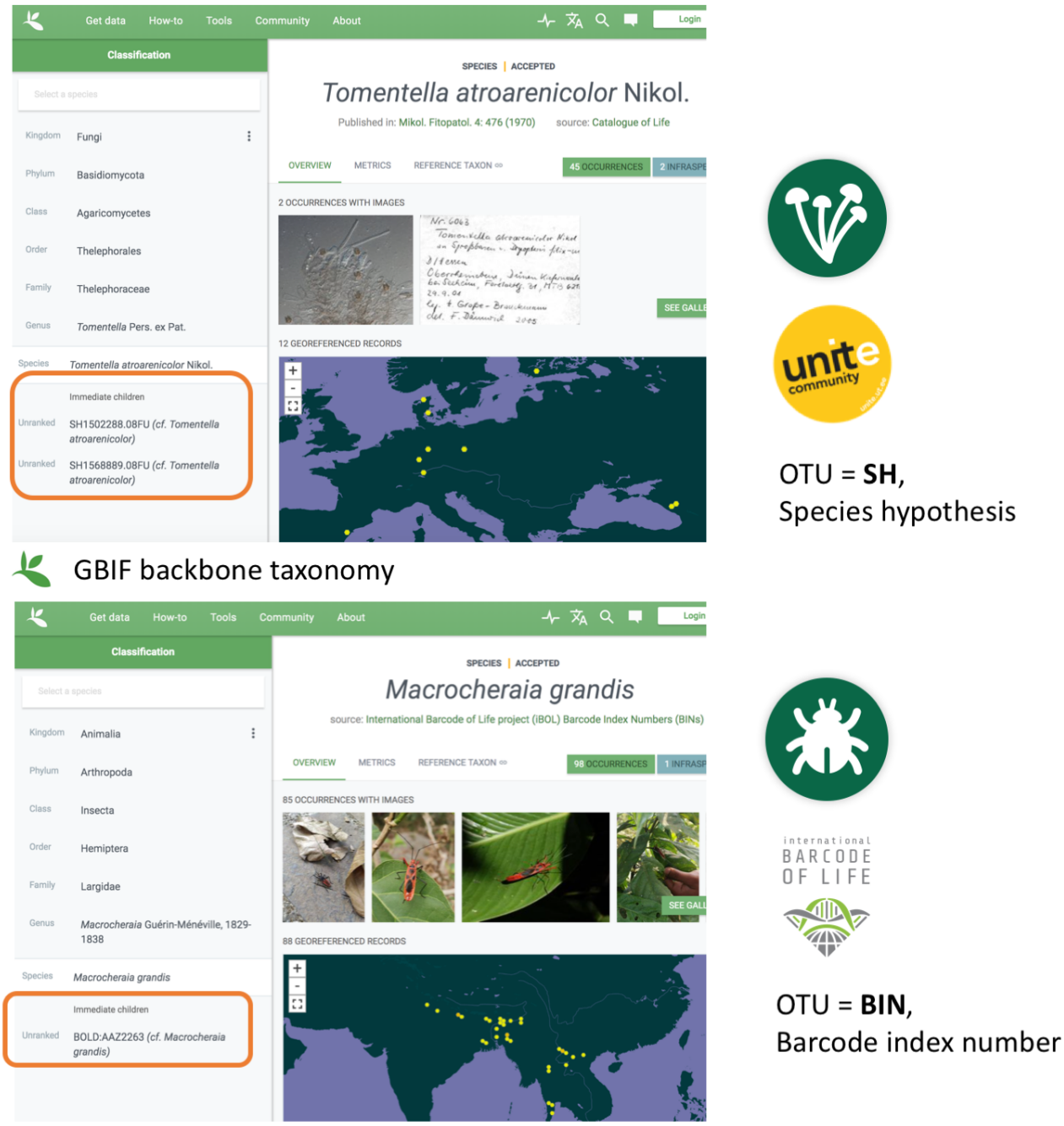
目標物種出現紀**錄**資料集範例

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

請參考 § 2.2.1 以了解如何格式化和分享這些資料集。有關達爾文核心出現紀錄資料集可通過達爾文核心出現紀錄資料集的DwC-A模板和出現紀錄資料的品質要求獲得。

2.1.4. 類別四：物種基因代號

這個類別對應於DNA衍生的代號名稱，是從序列分群 (clustering) 或去噪 (denoising，基於錯誤校正的模型)中衍生出來的，例如穩定的非林奈操作分類單元 (OTU)、擴增子序列變體 (ASV) 和Barcode Index Numbers (BINs) — 換句話說，任何與林奈分類系外定義的分類群或臨時名稱相關的參考。許多計畫會產生特定研究的OTU資料庫，儘管從技術上來說將其發布為物種清單雖是可行的，但它們對於資料連結或詮釋幾乎沒有價值；因此，我們不鼓勵這類型的資料通過生物多樣性資料平台來做發布。然而，被廣泛採用、穩定、全球化、數位化及可參考的OTU包含到林奈分類的骨幹中對於索引未命名的「黑暗」生物多樣性至關重要。GBIF在將這些大型全球OTU資料庫整合到GBIF分類骨幹中積累了豐富的經驗，這使得可以次資料在具有學名最接近的分類下顯示該OTU (圖表 7)。



圖表 *7.* 網頁呈現來自*UNITE (*主要是真菌，上方*)* 和來自*BOLD (BINs)(*主要是節肢動物，下方*)* 的*OTUs (SHs)* 在*GBIF*分類骨幹中以具有學名的對應分類。通過單一訪問點，多個個別觀察到的隱藏生物多樣性會隨著非基因來源的證據一起變得可發現。

物種基因代號名錄範例

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

此文件未提供全球OTU名錄清單/資料庫 (類別四) 的對照建議，並且不建議將 (計畫或研究特定的) OTU資料庫發布為物種清單。有關如何格式化和共享OTU物種清單的指南，請參見以下DwC-A 物種清單模板和物種清單的資料品質要求中的達爾文核心指南。MIxS物種清單的一般指南。如需有關如何將全球OTU資料庫對照到GBIF分類骨幹中的建議，請聯繫 GBIF help desk。

2.1.5. 類別五：僅有詮釋資料的資料集

詮釋資料 (metadata) 是關於資料的資料，是對資料集的廣泛描述，內容包括作者、作者所屬機構、資料集的原始研究目的、DOI (或多個DOI)、物種分類範圍、時間範圍和地理範圍等。也包括有關實驗和定序方法的資訊。詮釋資料也可包括目前未數位化的資料集或典藏的相關資訊。

僅有詮釋資料的資料集範例

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

對僅有詮釋資料的DNA衍生資料集 (類別五)，對照建議與其他僅有詮釋資料的資料集相同，此文件不提供僅有詮釋資料資料集的對照建議。請參考生物多樣性資料入口網的建議，並特別注意所需和建議的詮釋資料。對田野、實驗室和生物資訊學步驟的描述應盡可能詳細。GBIF的資料集頁面會呈現在EML檔案中您描述的方法和步驟（Frøslev T, Ejrnæs R, 2018。BIOWIDE eDNA Fungi dataset. Danish Biodiversity Information Facility. Occurrence dataset https://doi.org/10.15468/nesbvx accessed via GBIF.org on 2021-07-06)。但如果某平台或發表已經描述了結構化且可能更詳細的方法（例如 protocols.io 或 NEON protocols collection)，則可以直接通過MIxS SOP字段提供連結 (見§ 2.2.1)。

2.2. 資料對照 (mapping)

核心 (core) 文件紀錄有關「什麼、在哪裡和何時」的資料，而擴充 (extension) 文件用於描述特定類型觀察的細節。我們建議使用DNA衍生資料擴充文件來補充來自barcoding、metabarcoding (eDNA)、或qPCR/ddPCR的出現紀錄相關資料。DNA衍生資料擴充文件由Genomic Standards Consortium (GSC)制定的Minimum information standards，並且由ENA用於提交eDNA樣本詮釋資料。我們正在遵循並參與由TDWG的Sustainable DwC-MIxS interoperability task group提出的指南。為了提高索引和搜索效果，我們選擇將一些MIxS術語拆分，例如區分正向和反向引子序列和名稱。此外，一些來自GGBN標準的字段以及來自MIQE (minimum information for the publication of quantitative real-time PCR) 指南的字段已被包括在內，使其適用於廣泛範圍的DNA衍生資料。

初步在準備資料進行發布時，您應確保您的字段名稱/列標題符合達爾文核心資料標準。在許多情況下這很簡易，如將您的“lat”或“latitude”字段重命名為“decimalLatitude”。然而，達爾文核心標準非常彈性，一些術語的使用方式會因資料類型而異。例如，字段`organismQuantity`和`organismQuantityType`可用於描述個體數量、百分比生物量或Braun-Blanquet量表上的分數、以及樣本中ASV的read數量。因此，我們在這裡提供了必填字段和建議字段的表，並提供了描述和範例（表格1、表格 2、表格3和表格 4)。將DNA衍生資料發布在出現紀錄核心集的建議源於希望這些發布的序列可以作為出現紀錄的證據。另外，也可使用出現紀錄核心集和擴充文件的其他字段 (例如擴展測量或事實擴充 (eMoF))。當序列來自一個生物體 (例如寄生蟲、腸道內容物、附生生物等) 時，此出現紀錄與宿主的相關聯可以通過達爾文核心的資源關係擴充來實現（例如 https://www.gbif.org/zh-tw/species/143610775/verbatim)。 最重要的建議是盡可能使用全球獨一 (如果可用) 和其他永久識別碼來標識所有資料字段和參數(在下表的所有ID字段中)。

2.2.1. Mapping metabarcoding (eDNA) 和barcoding資料

此部分呈現類別一與二在mapping上的建議。

表格 *2.* 給出現紀錄核心集的字段建議，用於*Metabarcoding*類型的資料

字段名稱	範例	描述	填寫建議
basisOfRecord	MaterialSample	資料記錄的具體性質 - 是dcterms:type的子類型。對於DNA衍生的出現紀錄資料，(見類別一和類別三)，使用MaterialSample。對於多重引證出現紀錄，視情況使用PreservedSpecimen或LivingSpecimen。	必填
occurrenceID	urn:catalog:UWBM:Bird:89776	出現紀錄的唯一識別碼，允許不同版本在的資料集中以及通過資料下載和應用中可被識別。可以是全球唯一辨識碼或特定於資料集的辨識碼。	必填
eventID	urn:uuid:a964765b-22c4-439a-jkgt-2	與事件 (在特定地點和時間發生的事情) 相關的資訊的識別碼。可以是全球唯一辨識碼或特定於資料集的辨識碼。	強烈建議
eventDate	2020-01-05	記錄事件的日期。推薦的最佳做法是使用符合ISO 8601-1:2019的日期格式。有關更多資訊，請參考 https://dwc.tdwg.org/terms/#dwc:eventDate	必填
recordedBy	"Oliver P. Pearson | Anita K. Pearson"	記錄原始出現紀錄的人、組織或團體的名稱 (連接和分隔)。推薦的最佳做法是用垂直線 (|) 分隔每個值。包含有關觀察者的資訊有助於提高科學可重複性(Groom等人，2020)。	強烈建議
organismQuantity	33	樣本裡OTU或ASV的read數量。	強烈建議
organismQuantityType	DNA sequence reads	應填寫「DNA sequence reads」	強烈建議
sampleSizeValue	1233890	樣本中的總read數。這紀錄很重要因它允許計算樣本中每個OTU或ASV的相對豐度。此數字最好在通用處理 (品質控制、ASV去噪、chimera去除等) 之後，但在從資料集中手動/選擇性刪除非目標OTU或ASV之前做計算。稀釋化抽樣 (rarefraction，在樣本間均勻地重新抽樣到相同的定序深度) 並不必要或建議。	強烈建議
sampleSizeUnit	DNA sequence reads	應填寫「DNA sequence reads」	強烈建議

字段名稱	範例	描述	填寫建議
materialSampleID	https://www.ncbi.nlm.nih.gov/biosample/15224856 https://www.ebi.ac.uk/ena/browser/view/SAMEA3724543 urn:uuid:a964805b-33c2-439a-beaa-6379ebbfcd03	MaterialSample的ID (不是特定於材料樣本的數位化記錄)。如果從核苷酸檔案中獲取了生物樣本ID (biosample ID)，則使用生物樣本ID。如果沒有唯一識別碼，則可以從記錄中ID的組合構造出一個最可能全球唯一的識別碼。	強烈建議
samplingProtocol	UV light trap	在採樣事件期間使用的方法名稱、參考或描述。https://dwc.tdwg.org/terms/#dwc:samplingProtocol	建議
associatedSequences	https://www.ncbi.nlm.nih.gov/nuccore/MK405371	與出現紀錄相關的遺傳序列資訊的ID (出版物、全球唯一的識別碼、URI) 列表 (連接並分隔)。可用於連接到原始條形碼和/或相關基因組序列，例如在公開資料庫中。	建議
identificationRemarks	RDP annotation confidence (at lowest specified taxon): 0.96, against reference database: GTDB	種的鑑定過程的具體規范，理想情況下應包括應用算法和參考資料庫的資訊，以及對鑑定結果的信心評分。	建議
identificationReferences	https://www.ebi.ac.uk/metagenomics/pipelines/4.1 https://github.com/terrimporter/CO1Classifier	用於鑑定的參考 (出版物、全球唯一的識別碼、URI) 列表 (連接並分隔)。推薦的做法是用空格、垂直線 (|) 和空格分隔列表中的值。	建議

字段名稱	範例	描述	填寫建議
decimalLatitude	60.545207	位置的地理緯度 (以十進制度為單位，使用大地基準面的空間參考系統)。正值在赤道以北，負值在赤道以南。值介於-90和90之間，包括-90和90。	強烈建議
decimalLongitude	24.174556	位置的地理經度 (以十進制度為單位，使用大地基準面的空間參考系統)。正值在格林威治子午線以東，負值在格林威治子午線以西。值介於-180和180之間，包括-180和180。	強烈建議
taxonID	ASV:7bdb57487bee022ba30c03c3e7ca50e1	eDNA資料建議使用序列的MD5 hash，開頭再加上“ASV:”。參考 § 1.6。	如沒提供DNA_sequence，強烈建議填寫
scientificName	*Gadus morhua* L. 1758, BOLD:ACF1143	最接近已知的分類單位 (物種或更高階層) 的學名，或來自BOLD或UNITE的OTU ID。	必填
kingdom	Animalia	更高階層分類	強烈建議
phylum	Chordata	更高階層分類	建議
class	Actinopterygii	更高階層分類	建議
order	Gadiformes	更高階層分類	建議
family	Gadidae	更高階層分類	建議
genus	*Gadus*	更高階層分類	建議

表格 *3. DNA*衍生資料擴充中 *(*部分*)* 建議用於*metabarcoding*資料的字段

字段名稱	範例	描述	填寫建議
DNA_sequence	TCTATCCTCAATTATAGGTCATAATTCACCATCAGTAGATTTAGGAATTTTCTCTATTCATATTGCAGGTGTATCATCAATTATAGGATCAATTAATTTTATTGTAACAATTTTAAATATACATACAAAAACTCATTCATTAAACTTTTTACCATTATTTTCATGATCAGTTCTAGTTACAGCAATTCTCCTTTTATTATCATTA	DNA序列 (ASV)。序列的詮釋取決於發布時當下使用的技術和資料庫。因此，最客觀的處理方式是該序列可以在將來重新詮釋。	強烈建議
sop	https://www.protocols.io/view/emp-its-illumina-amplicon-protocol-pa7dihn	用於組裝和/或詮釋基因組、總體基因體或環境序列的標準操作流程。或使用一個有完整記錄方法的引用，例如protocols.io	建議
target_gene	16S rRNA, 18S rRNA, ITS	marker-based研究的目標基因或marker名稱	強烈建議
target_subfragment	V6, V9, ITS2	基因或marker子片段的名稱。主要是要識別，例如16S rRNA基因的高變異性V6區域等marker的特殊區域	強烈建議
pcr_primer_forward	GGACTACHVGGGTWTCTAAT	用於擴增目標基因或子片段序列的前向PCR引子序列。	強烈建議
pcr_primer_reverse	GGACTACHVGGGTWTCTAAT	用於擴增目標基因或子片段序列的反向PCR引子序列。	強烈建議
pcr_primer_name_forward	jgLCO1490	前向PCR引子的名稱	強烈建議

字段名稱	範例	描述	填寫建議
pcr_primer_name_reverse	jgHCO2198	反向PCR引子的名稱	強烈建議
pcr_primer_reference	https://doi.org/10.1186/1742-9994-10-34	引子的參考文獻	強烈建議
env_broad_scale	forest biome [ENVO:01000174]	等同於**MIxS v4**中的**env_biome** 在此字段中，列出樣品或標本來源的主要環境系統。識別的系統應具有較粗略的空間廣度，以提供採樣地點的一般環境背景描述 (例如，您是否在沙漠還是雨林中) 。我們建議使用ENVO的生物群系類的子類： http://purl.obolibrary.org/obo/ENVO_00000428	建議
env_local_scale	litter layer [ENVO:01000338]	等同於**MIxS v4**中的**env_feature** 在此字段中，列出樣品或標本來源的當地環境中的實體或實體，且您認為會對您的樣品或標本具有重要因果影響。請使用ENVO中存在的並且比您env_broad_scale輸入的空間廣度更小的術語。	建議
env_medium	soil[ENVO:00001998]	等同於**MIxS v4**中的**env_material** 在此字段中，列出在採樣之前環繞樣品或標本的環境或基質 (使用 (|) 做分隔)。請使用ENV的環境材料類的一個或多個子類： http://purl.obolibrary.org/obo/ENVO_00010483	建議
lib_layout	Paired	等同於**MIxS v4**中的**lib_const_meth** 指定是單端、成對或其他reads配置	建議
seq_meth	Illumina HiSeq 1500	使用的定序方法/平台	強烈建議
otu_class_appr	"dada2; 1.14.0; ASV"	定義OTUs或ASVs時使用的方法/演算法和分群等級 (如適用)	強烈建議
otu_seq_comp_appr	"blastn;2.6.0+;e-value cutoff: 0.001"	用於將OTUs或ASVs分配為“物種”名稱的工具和閾值	強烈建議
otu_db	"Genbank nr;221", "UNITE;8.2"	用於將OTUs或ASVs做分類辨識的參考資料庫 (即未作為當前研究的一部分生成的序列)	強烈建議

2.2.2. Mapping ddPCR / qPCR 資料

此部分呈現類別三在mapping上的建議。

表格 *4.* 給*ddPCR/qPCR*資料類型的出現紀錄核心集的字段建議

字段名稱	範例	描述	填寫建議
basisOfRecord	MaterialSample	資料所記錄的具體性質 - dcterms:type 的子類。對於DNA衍生的出現紀錄 (參見類別一和類別三)，請使用MaterialSample。	必填
occurrenceStatus	Present, Absent	有關生物在特定位置的存在或不存在的說明。	必填
eventID	urn:uuid:a964765b-22c4-439a-jkgt-2	與事件 (在特定地點和時間發生的事情) 相關資訊的ID。可以是全GUI或特定於資料集的ID。	強烈建議
eventDate	2020-01-05	記錄事件的日期。建議寫法是使用符合ISO 8601-1:2019的日期。有關更多詳情請查看 https://dwc.tdwg.org/terms/#dwc:eventDate	必填
recordedBy	"Oliver P. Pearson | Anita K. Pearson"	記錄原始出現紀錄的人、組織或機構的姓名列表 (串聯並分隔)。建議的寫法是使用垂直條 (|) 分隔值。包含觀察者的名字可提高科學的可重複性 (Groom等人，2020)。	強烈建議
organismQuantity	50	樣本中陽性droplets/chambers的數量	ddPCR, dPCR 來源資料強烈建議使用
organismQuantityType	ddPCR droplets dPCR chambers	分割類型 partition type	ddPCR, dPCR 來源資料強烈建議使用
sampleSizeValue	20000	Accepted partitions 的數量 (n), 例如 accepted ddPCR droplets 或 dPCR chambers	ddPCR, dPCR 來源資料強烈建議使用
sampleSizeUnit	ddPCR droplets dPCR chambers	分割類型 partition type, 應與organismQuantityType中的值相等	ddPCR, dPCR 來源資料強烈建議使用

字段名稱	範例	描述	填寫建議
materialSampleID	https://www.ncbi.nlm.nih.gov/biosample/15224856 urn:uuid:a964805b-33c2-439a-beaa-6379ebbfcd03	MaterialSample的ID (不是特定於材料樣本的數位化記錄)。如果從核苷酸檔案中獲取了生物樣本ID (biosample ID)，則使用生物樣本ID。如果沒有識別碼，則可以從記錄中ID的組合構造出一個。	強烈建議
samplingProtocol	UV light trap	在採樣事件期間使用的方法名稱、參考或描述。https://dwc.tdwg.org/terms/#dwc:samplingProtocol	建議
decimalLatitude	60.545207	位置的地理緯度 (以十進制度為單位，使用大地基準面的空間參考系統)。正值在赤道以北，負值在赤道以南。值介於-90和90之間，包括-90和90。	強烈建議
decimalLongitude	24.174556	位置的地理經度 (以十進制度為單位，使用大地基準面的空間參考系統)。正值在格林威治子午線以東，負值在格林威治子午線以西。值介於-180和180之間，包括-180和180。	強烈建議
scientificName	*Gadus morhua* L. 1758, BOLD:ACF1143	最接近已知的分類單位 (物種或更高階層) 的學名，或來自BOLD或UNITE的OTU ID。	必填
kingdom	Animalia	更高階層分類	強烈建議
phylum	Chordata	更高階層分類	建議
class	Actinopterygii	更高階層分類	建議
order	Gadiformes	更高階層分類	建議
family	Gadidae	更高階層分類	建議
genus	*Gadus*	更高階層分類	建議

表格 *5.* DNA衍生資料擴充 *(*部分*)* 建議用於*ddPCR/qPCR*資料的字段

字段名稱	範例	描述	填寫建議
sop	https://www.protocols.io/view/protocol-for-dna-extraction-and-quantitative-pcr-d-vwie7ce https://doi.org/10.17504/protocols.io.vwie7ce	用於組裝和/或詮釋基因組、總體基因體或環境序列的標準操作流程。可使用有完整記錄方法的引用，例如 protocols.io	強烈建議
annealingTemp	60	PCR annealing 階段的反應溫度。	如果annealingTemp有填寫則必填
annealingTempUnit	攝氏度		強烈建議
pcr_cond	initial denaturation:94_3; annealing:50_1;elongation:72_1.5;final elongation:72_10;35	以 "initial denaturation:94degC_1.5min; annealing=..." 的形式描述 PCR的反應條件和階段。	強烈建議
probeReporter	FAM	使用的fluorophore (reporter)。探針在擴增目標DNA中合并。聚合 酶活性會降解與模板結合的探針，探針會從中釋放螢光物並與quencher斷開接近，從而產出螢光。	強烈建議
probeQuencher	NFQ-MGB	使用的quencher類型。當被螢光團的光源激發時，quencher分子會消除螢光物發出的螢光，只要螢光物和quencher在接近狀態下，會抑制任何螢光信號。	強烈建議
ampliconSize	83	擴增子 (amplicon) 的长度，以鹼基對 (base pair) 為單位。	強烈建議
thresholdQuantificationCycle	0.3	螢光信號在cycle之間的閾值。	qPCR: 強烈建議
baselineValue	15	目標擴增的螢光信號低於非來自真實目標扩增的背景螢光信號的循環數。	qPCR: 強烈建議

字段名稱	範例	描述	填寫建議
quantificationCycle	37.9450950622558	達到閾值以上的螢光信號所需的循环数。Quantification cycle (Cq), threshold cycle (Ct), crossing point (Cp), 和 take-off point (TOP) 指的是即時儀器的相同值。使用 quantification cycle (Cq), 建議根據 RDML (Real-Time PCR Data Markup Language) 資料標準	
automaticThresholdQuantificationCycle	no	儀器的閾值是否為儀器默認或手動輸入	
automaticBaselineValue	no	儀器的基準值是否為儀器默認或手動輸入	
contaminationAssessment	no	是否進行了DNA或RNA的污染評估。	
estimatedNumberOfCopies	10300	每μl的目標分子數。每個分區的平均copy數可以 ?=m/n 來計算，期中n為分區數，m為所有分區的總體積的估計copy數。	
amplificationReactionVolume	22	PCR反應體積	
amplificationReactionVolumeUnit	μl	PCR反應體積的單位。許多儀器需要準備比實際分析體積更大的初始樣品體積。	
pcr_analysis_software	BIO-RAD QuantaSoft	用於分析d(d)PCR運行的軟體。	
experimentalVariance		鼓勵進行多重生物重復以評估總實驗差異。當進行單一dPCR實驗時，必須從二項 (或適當等效) 分佈中計算出僅由計數誤差引起的差異的最小估計。	
target_gene	16S rRNA, 18S rRNA, nif, amoA, rpo	基於marker-based研究的目標基因或marker名稱	強烈建議
target_subfragment	V6, V9, ITS	基因或marker子片段的名稱。主要是要識別，例如16S rRNA基因的高變異性V6區域等marker的特殊區域	強烈建議
pcr_primer_forward	GGACTACHVGGGTWTCTAAT	用於擴增目標基因或子片段序列的前向PCR引子序列。	強烈建議
pcr_primer_reverse	GGACTACHVGGGTWTCTAAT	用於擴增目標基因或子片段序列的反向PCR引子序列。	強烈建議

字段名稱	範例	描述	填寫建議
pcr_primer_name_forward	jgLCO1490	前向PCR引子的名稱	強烈建議
pcr_primer_name_reverse	jgHCO2198	反向PCR引子的名稱	強烈建議
pcr_primer_reference	https://doi.org/10.1186/1742-9994-10-34	引子的參考文獻	強烈建議
env_broad_scale	forest biome [ENVO:01000174]	等同於**MIxS v4**中的**env_biome** 在此字段中，列出樣品或標本來源的主要環境系統。識別的系統應具有較粗略的空間廣度，以提供採樣地點的一般環境背景描述 (例如，您是否在沙漠還是雨林中) 。我們建議使用ENVO的生物群系類的子類： http://purl.obolibrary.org/obo/ENVO_00000428	建議
env_local_scale	litter layer [ENVO:01000338]	等同於**MIxS v4**中的**env_feature** 在此字段中，列出樣品或標本來源的當地環境中的實體或實體，且您認為會對您的樣品或標本具有重要因果影響。請使用ENVO中存在的並且比您env_broad_scale輸入的空間廣度更小的術語。	建議
env_medium	soil [ENVO:00001998]	等同於**MIxS v4**中的**env_material** 在此字段中，列出在採樣之前環繞樣品或標本的環境或基質 (使用 (|) 做分隔)。請使用ENV的環境材料類的一個或多個子類： http://purl.obolibrary.org/obo/ENVO_00010483	建議
concentration	67.5	DNA濃度 (重量 ng/體積 µl)，參考 http://terms.tdwg.org/wiki/ggbn:concentration	建議
concentrationUnit	ng/µl	濃度量化的單位，參考 http://terms.tdwg.org/wiki/ggbn:concentrationUnit	建議
methodDeterminationConcentrationAndRatios	Nanodrop, Qubit	濃度量化的方法，參考 http://terms.tdwg.org/wiki/ggbn:methodDeterminationConcentrationAndRatios	建議
ratioOfAbsorbance260_230	1.89	在260nm和230nm波長的吸收比例，評估DNA純度 (通常為次要指標測量EDTA、碳水化合物、苯酚)，(DNA樣本)。 參考http://terms.tdwg.org/wiki/ggbn:ratioOfAbsorbance260_230	建議
ratioOfAbsorbance260_280	1.91	在280nm和230nm波長的吸收比例，評估DNA純度 (通常為次要指標測量EDTA、碳水化合物、苯酚)，(DNA樣本)。 參考	建議

字段名稱	範例	描述	填寫建議
samp_collect_device	biopsy, niskin bottle, push core	用於收集樣本的方法或設備	建議
samp_mat_process	filtering of seawater, storing samples in ethanol	從環境中採樣中和後，對樣本進行的任何處理。該字段接受OBI，有關OBI (v 2018-02-12) 術語的瀏覽器，請參考 http://purl.bioontology.org/ontology/OBI	建議
samp_size	5 litre	收集的樣本的量或大小 (體積、質量或面積)	建議
size_frac	0-0.22 micrometer	樣本處理中使用的過濾孔徑大小	建議
pcr_primer_lod	51	Assay最低可檢測目標的水平	強烈建議
pcr_primer_loq	184	Assay最低可檢測的copy數量	強烈建議

2.3. 海洋資料集和海洋生物多樣性信息系統 (OBIS)

在處理來自海洋環境的資料集時，除了GBIF之外，也建議將資料發布在海洋生物多樣性信息系統(OBIS)。OBIS是一個全球性的生物多樣性資料庫，是IOC-UNESCO的一部分，專門提供與海洋生物相關的可靠並可查找的資料。與GBIF和ALA一樣，OBIS使用DwC-A格式進行資料索引和發布。同時將海洋資料集通過其他生物多樣性資料庫和OBIS發布，資料可以觸及更廣泛的受眾以及在海洋生物多樣性領域工作的多個社群，因為OBIS中的資料常應用在聯合國的相關工作。海洋資料發布到OBIS時所需的資訊與GBIF有些不同的小差異，這些差異帶出了資料嚴格品質管控，提高了海洋資料集資料的可信度。

為了確保一致的分類學命名方式，OBIS使用世界海洋物種名錄 (WoRMS)作為唯一的分類學骨幹。這也適用於從基因資料衍生的出現記錄；從WoRMS資料庫中，連結出現紀錄學名到WoRMS資料庫的scientific name ID是高度推薦的做法。如果未提供scientific name ID，OBIS將在資料詮釋過程中嘗試將學名與WoRMS比對，但應避免這樣做。未列在WoRMS中的學名是可以被接受的，WoRMS將會在後續資料詮釋過程中進行審查並可能納入學名到資料庫裡。未分類的序列建議分類為“incertae sedis”，並將WoRMS的 `scientificNameID` 輸入為urn:lsid:marinespecies.org:taxname:12。這將確保GBIF和OBIS可正確的詮釋資料。此外，建議將所使用的序列資料庫 (例如BOLD中的Barcode index numbers：BINs) 的序列識別碼添加到出現紀錄核心表的 `taxonConceptID` 字段中。通過這種方式，OBIS將保留基於WoRMS的分類學骨架，同時實現與不同的參考序列資料庫的連結。如果來自參考資料庫的名稱不是嚴格的學名，可以添加為 `verbatimIdentification`。物種學名的自動分類通常可以通過WoRMS分類比對工具和R套件(如worrms和taxize) 來完成。將來OBIS計劃定期搜索和更新已發布序列的分類，因此也強烈建議記錄序列與每個出現相關聯的資訊。

OBIS資料提交中的另一個必填字段是地理坐標。OBIS進行了與海洋數據相關的額外品質檢查；例如：海洋物種的坐標不應在陸地上、深度值應在合理的範圍內。最後值得一提的是，OBIS還支持使用擴充測量或事實(eMoF)。該擴充允許以彈性和標準化的方式將環境數據、採樣資訊與事件或出現紀錄相關聯，以及將生物測量與出現紀錄相關聯。OBIS具有eDNA metabarcoding資料集的資料模板當作範例，可參考 https://github.com/iobis/dataset-edna 。

表格 *6. OBIS*對*DNA-based*出現紀錄的紀錄要求和建議。此表突出了與發布到*GBIF*時字段值和要求的重要差異。這裡以藍鯨 *(Balaenoptera musculus)* 的*DNA*檢測為例進行說明。

字段名稱	範例 **(OBIS)**	描述	填寫建議
scientificName	Balaenoptera musculus	學名，最好按照WoRMS資料庫裡的名稱。這與GBIF不同。GBIF的建議為使用所用分類方法對應的物種名稱。	必填
scientificNameID	urn:lsid:marinespecies.org:taxname:137090	“Balaenoptera musculus” 的WoRMS資料庫中的scientific name ID。	強烈建議
taxonConceptID	NCBI:txid9771	與NCBI分類資料庫中的 Balaenoptera musculus 相關聯的NCBI ID。如果使用BOLD進行辨識，也可以是 BIN-ID，或來自其他資料庫的其他ID。	建議
verbatimIdentification	Balaenoptera musculus	對應於NCBI ID (Balaenoptera musculus) (或其他 ID) 的名稱。這不一定與學名中的值對應。	建議

表格 *7. OBIS*對無法在任何分類階層上歸類學名的序列的紀錄要求和建議。

字段名稱	範例 **(OBIS)**	描述	填寫建議
scientificName	incertae sedis	OBIS建議用於未知序列的學名。當序列/分類未知時，請使用此名稱。這與GBIF不同，後者建議即使序列資料庫所使用的科學名稱不嚴謹，也應使用該名稱。	必填
scientificNameID	urn:lsid:marinespecies.org:taxname:12	OBIS建議使用根據WoRMS資料庫裡 “incertae sedis”的 scientific name ID於未知序列。當序列/分類未知時，請使用此ID。	強烈建議
taxonConceptID	NCBI:txid1899546	其他分類學資料庫中的ID，例如序列參考資料庫。	強烈建議
verbatimIdentification	Phototrophic eukaryote	對應於分類概念ID的外部資料庫中的生物名稱。	建議

3. 前景

目前對通過生物多樣性資料平台公開DNA衍生資料的關注非常高，而且很可能需求量會增加。我們的目標是在這裡提供對照建議並保持有效並且緩慢的更改，即使生物多樣性資料平台的打包和索引可能發展得更快。作者有意識到但尚未參閱 BOLD手冊、BIOM格式和 http://edamontology.org/page 。

我們建議ALA和GBIF類似的資料平台開始採用支持更複雜的資料關聯和層次的格式。例如，無摩擦資料格式和更具領域特定性的生物觀察矩陣 (BIOM) 格式。後者被幾個生物資訊工具使用 (QIIME2、Mothur 、USEARCH 等)，可能有助於發布者跳過將資料轉換為DwC-A格式的步驟。比當前的DwC星芒架構更彈性的資料格式會對於允許不同階層的採樣事件和材料樣本，以及將序列資料附加到採樣事件中的個別出現紀錄中至關重要。

生物多樣性資料平台還需要能夠讓研究人員輕鬆地將DNA衍生的出現紀錄資料包含或排除在其查詢結果中。上述建議的資料格式可以為特定出現記錄所基於的證據類型提供更豐富的分類機會。然而，目前在 “BasisOfRecord” 的詞彙表中缺乏適當的值來表示這些資料類型。我們建議，作為一個務實的即時解決方案，“BasisOfRecord” 應該增加 “DNA” 、“DNA-derived” 或類似的值。正如上文所述，DNA衍生的資料可能來自有充分文檔記錄的採樣或個體生物、可能由保存的實體材料支持或不支持、可能是由基因序列分析或其他DNA檢測方法(例如qPCR) 產生的。生物多樣性資料平台和TDWG應提供區分這些資料類型以及其來源的方法。

我們還建議資料平台將實際序列或至少這些序列的MD5開發索引功能，以便在資料集之間進行ASV的搜尋。如果提供了ASV，則MD5應由生物多樣性資料平台生成；如果沒有提供ASV，則MD5應是強制性提供的。

正如在 § 1.6 和 § 2.1.4 中提到的那樣，我們鼓勵生物多樣性資料平台繼續努力將相關的分子分類參考資料庫納入其分類骨幹中。

其他方法和技術的廣泛應用，例如Oxford Nanopore、PacBio和shotgun定序，很可能會使此文件進行調整更新，以適應特定的新資料和詮釋資料相關字段的需求。

詞彙

Atlas of Living Australia (ALA)

ALA是一個從多個來源聚合澳大利亞的生物多樣性資料網絡的平台，使其對任何人都可查詢和重複使用(請見 https://www.ala.org.au/about-ala/) 。ALA開發的開放式基礎架構平台也被其他國家用於建立自己的國家生物多樣性資料平台 (https://living-atlases.gbif.org/)。

擴增子序列變體 (Amplicon Sequence Variant, ASV)

高通量定序和去噪处理生成的唯一DNA序列，被預定为代表生物上真實的序列變體 (sequence variant)。參考 Operational Taxonomic Unit (OTU) 和 (Callahan等人 2017)。

應用程式開發介面 **(API)**

一組用于不同主機應用之間互通和資料傳輸的方法和工具。

Barcode Index Numbers (BINs)

動物中的cytochrome c oxidase I (COI) 基因序列分群產生的物種層級操作分類單元 Operational Taxonomic Units (OTUs) 。每個BIN都被分配了一個全球唯一的識別碼，並在 Barcode of Life Data System (BOLD) 資料庫內可被搜尋。

Barcode of Life Data System (BOLD)

BOLD 是由Guelph的Centre for Biodiversity Genomics代表 (IBOL) 所維護的參考資料庫。它收集了關於真核生物物種標本的條碼參考和序列資料，尤其是動物的COI序列，并維護了(BIN; Ratnasingham & Hebert, 2013) 系統。接近物種層級的OTUs的辨識碼來自序列分群後高度相似的序列。

生物多樣性資料平台

從各種來源 (如自然史收藏、公民科學、生態監測以及基因序列) 衍生的生物多樣性資料的線上資源，可以是全球性 (GBIF) 或國家級 (ALA)。

序列分群 **(Clustering)**

在分類學中，根據某種相似性準則將生物分群在一起的過程。見 Operational Taxonomic Unit。

群落 **(bulk) DNA**

來自大量樣本 (bulk) 的DNA (例如浮游生物樣本或由許多物種的幾個個體組成的Malaise陷阱樣本)。此文件將bulk樣本DNA納入了eDNA概念中。

達爾文核心集檔案格式 **(DwC-A)**

達爾文核心標準 Darwin Core (DwC) standard 編制的生物多樣性資料打包的壓縮 (ZIP) 文件格式。基本上是一組相互關聯的CSV文件和一個XML檔，描述了包含的文件和資料列以及它們之間的相互關係。

達爾文核心 **(DwC)** 標準

由生物多樣性資訊標準 (TDWG) 社群所發源的用於共享和發布生物多樣性資料的標準。原則上，這是一組用於描述生物多樣性觀察的術語，例如採樣事件、出現紀錄和分類單元。目前的達爾文核心術語描述在快速參考指南中。

資料詞彙

一組首選術語或概念，具有特定的、明確定義的含義和相互關係，有助於資料交換和重複使用。

ddPCR (droplet digital 聚合酶連鎖反應**)**

Droplet digital PCR。測量樣本中標記DNA的絕對量 (copy數) 的方法。也參考 qPCR。

去噪 **(Denoising)**

在metabarcoding中用於將真實生物序列 (見 ASVs) 與由PCR增幅和序列錯誤引起的偽序列變異分離的方法。

數位物件識別碼 **(DOI)**

用於唯一識別 (和定位) 數位化資訊物件的永久性參考，例如生物多樣性資料集或科學出版物。

DNA 條形碼與元條形碼 **(**擴增子定序**)**

使用短、標準化的DNA片段通過序列化來識別個體生物的方法。Metabarcoding將條形碼技術與高通量DNA序列化結合，使用通用引子來放大和序列化eDNA樣本中的大量生物群。

DNA分子標記 **(marker)**

DNA片段作為某種特性 (例如分類歸屬) 的標記。可能是基因或基因的一部分，但不一定是。

DNA高通量分子條碼資料庫

包含先前研究過的生物的DNA序列 (DNA條碼) 的資料庫。理想情況下的參考序列是從已描述和研究充分的物種的個體中生成的，以模式標本最為理想，或者高一級的分類階層 (例如屬、科)，但也可能來自eDNA定序。盲目的信任 “參考序列” 是不建議的。

DNA探針 **(probe)**

一段短與帶有螢光標記的合成單股DNA片段，在PCR過程中結合到目標DNA的特定區域 (標記)。增加了特異性使其可以與引子一起在qPCR和ddPCR中使用以檢測和定量一個基因標記。

歐洲生物資訊研究機構 **(EMBL-EBI)**

生物資訊學研究和服務的政府間組織，隸屬於歐洲分子生物學實驗室 (EMBL)，通過 European Nucleotide Archive (ENA) 提供包括 (原始) 序列reads和assembly資料。

環境**DNA (eDNA)**

環境樣本中的DNA，例如土壤、水、空氣或宿主生物體。常用的定義為環境DNA是從環境樣本中獲得的基因物質 (DNA)，沒有明顯的生物來源材料的證據 (Thomsen and Willerslev，2015)。

European Nucleotide Archive (ENA)

歐洲核苷酸序列資料庫，涵蓋原始序列、序列組裝資訊和功能註解。包括序列檔案 Sequence Read Archive (SRA)，由歐洲生物資訊學研究所 (EMBL-EBI) 管理，作為國際核苷酸序列資料庫合作項目 the International Nucleotide Sequence Database Collaboration (INSDC) 的一部分。

FASTQ

一种以文字的格式用於存儲从高通量定序 High-throughput sequencing (HTS) 中得到的分子序列及相關質量度。單個ASCII字符分別表示每個序列位置base call (已識別的核苷酸) 和質量分數。

全球生物多樣性資訊機構 **(GBIF)**

國際網絡和研究基礎建設，主要致力於提供全球生物多樣性資料的開放和流通。

Global Genome Biodiversity Network (GGBN)

致力於有效分享和使用基因體生物多樣性樣本及相關詮釋資料的國際機構網絡，推廣符合達爾文核心相容之GGBN資料標準。

全球定位系統 **(Global Positioning System, GPS)**

由美国空軍太空司令部營運的衛星導航系統。

高通量定序 **(High-throughput sequencing, HTS)**

不同技術且大規模並行的進行定序，從基因材料的製備中產生數百萬個DNA序列reads，也被稱為次世代定序 (NGS)。相較傳統的Sanger定序為針對特定DNA片段進行定序。

資料整合

從不同來源 (例如本地資料庫、文字檔文件或試算表) 將資料導入到共同的目標系統，例如線上生物多樣性資料平台 biodiversity data platform 以進行存儲和進一步分析的過程。通常包括擷取 (extraction)、轉換 (transform) 和載入 (load) 等步驟。

索引

根據特定的架構或結構對資料進行組織，使資料更容易的被查詢和呈現。

International Nucleotide Sequence Database Collaboration (INSDC)

日本DNA資料庫 (DDBJ)、EMBL 和 NCBI 的聯合合作，旨在提供全球公眾查詢核苷酸序列資料及相關資訊。

總體基因體學 **(Metagenomics)**

無經過PCR的混合樣品中隨機基因組片段的定序。

Minimum Information about any (x) Sequence (MIxS) 標準

由基因體標準協會 (GSC) 制定的用於序列詮釋資料的一系列標準 (清單)。

分子操作分類單元 **(molecular Operational Taxonomic Unit, mOTU)**

參考 Operational Taxonomic Unit (OTU)。

National Center for Biotechnology Information (NCBI)

美國國家醫學圖書館 (NLM) 的一個部門，擁有重要的生物資訊資源，如DNA序列的GenBank資料庫，以及高通量序列資料的 Sequence Read Archive (SRA)。

次世代定序 **(Next Generation Sequencing, NGS)**

參考 High-throughput sequencing (HTS)。

出現紀**錄**

在特定時間及特定地點之生物 (即 http://rs.tdwg.org/dwc/terms/Organism 的概念) 的存在記錄。

操作分類單元 **(Operational Taxonomic Unit, OTU)**

基於特定DNA標記序列相似性的生物分群，用於分類。例如，在UNITE中有 Species Hypothesis 和在Barcode of Life Data System (BOLD) 中有 Barcode Index Numbers。Amplicon Sequence Variants (ASVs) 可被視為 zero radius OTUs (zOTUs) 的類比。

聚合**酶**連鎖反應 **(Polymerase Chain Reaction, PCR)**

特定DNA (或RNA) 序列片段的快速增幅和檢測技術。增幅區域由反應中使用的引子 PCR primers 決定。

執行流程

在生物資訊學裡一組演算法或工具被應用於預定義的工作流程中，用於處理例如高通量定序 High-throughput sequencing (HTS) 資料。

引子 **(PCR primers)**

簡短的合成單股DNA片段，它們與目標DNA (標記) 的選定區域結合，以在 <pcr,PCR> 中啟動複製過程。需一對引物才能讓聚合酶酶能夠擴增所選標記。

定量聚合**酶**連鎖反應 **(quantitative Polymerase Chain Reaction, qPCR)**

定量型 PCR。測量樣本中特定標記的相對DNA量的方法。參考 ddPCR。

樣本

材料 (水、土壤、腸道內容物等) 用於分析。

序列比對

使用生物資訊方法來比較和排列兩個或以上分子 (DNA、RNA或protein) 序列來測因演化等過程照成的相似性。

物種假**說 (Species Hypothesis, SH)**

UNITE資料庫與序列管理的環境中，真菌的物種分類階級 Operational Taxonomic Unit (OTU)。

標本

單個動物、植物、真菌等生物，用作其物種或模式標本在科學研究或展示的參考。

次世代基因定序資料**檔**案 **(Sequence Read Archive, SRA)**

次世代定序 (NGS) 基因資料倉儲庫，維運 the National Center for Biotechnology Information (NCBI)、the European Bioinformatics Institute (EMBL-EBI) 與 DNA Data Bank of Japan (DDBJ)，包含原始 (未denoise) 的定序結果和 sequence alignments 結果。the European Nucleotide Archive (ENA) 裡三大組成的其中一項，過去的名稱為 Short Read Archive。

特定範圍定序

使用核酸探針所萃取的DNA片段做定序。

UNITE

UNITE是一個基於真核核糖體ITS區域的序列管理系統。所有公開的序列都被分群為物種假設 (Species hypothesis, SHs)，並給予DOI。SH比對服務會輸出各種資訊，包括eDNA樣本中存在的物種、這些物種是否可能是未描述的新物種、它們在其他研究是否也被發現、這些物種是否對一個地區來說是外來種，以及它們是否受到威脅。DOI與 PlutoF平台 和GBIF的分類骨架互相連結，因此會附帶有一個分類名稱 (taxon name)。UNITE中使用的資料由

PlutoF管理。資料通過一系列標準，主要包括 Darwin Core, MIxS, 和 DMP Common Standard；部分則是 EML, MCL，和 GGBN。PlutoF主要通過CSV和FASTA格式輸出資料。PlutoF也可以用於在GBIF中發布資料 (使用DwC格式) 並準備GenBank提交文件。此外，您還可以從您的資料中下載物種清單，以及以具有階層結構的JSON檔案格式下載您的計畫。

Zero radius otu (zOTU)

參考 ASV。

參考文獻

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

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

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

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

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