MILJÖ | NATURVÅRDSVERKET FORSKNING

Reproducerbar analys av miljö-DNA i nationella övervakningsprogram

En kritisk granskning

Mats Töpel, Matthew Pinder

RAPPORT 7084 | DECEMBER 2022

Reproducerbar analys av miljö-DNA i nationella övervakningsprogram

Av Mats Töpel och Matthew Pinder

NATURVÅRDSVERKET

Beställningar
Ordertel: 08-505 933 40
E-post: natur@cm.se
Postadress: Arkitektkopia AB, Box 110 93, 161 11 Bromma
Internet: www.naturvardsverket.se/publikationer

Naturvårdsverket
Tel: 010-698 10 00
E-post: registrator@naturvardsverket.se
Postadress: Naturvårdsverket, SE-106 48 Stockholm
Internet: www.naturvardsverket.se

ISBN 978-91-620-7084-7
ISSN 0282-7298

Tryck: Arkitektkopia AB, Bromma 2022
Omslagsfoto: Jan Plue

Förord

I den här rapporten presenteras resultaten av syntesprojektet ”Reproducerbar analys av miljö-DNA i nationella övervakningsprogram”, ett av sex syntesprojekt som genomförts inom ramen för forskningssatsningen Digitalisering som stöd för en hållbar förvaltning.

Med satsningen ville Naturvårdsverkets och Havs- och vattenmyndigheten visa på digitaliseringens möjligheter i myndigheternas förvaltningsarbete.

Projektet har finansierats med medel från Naturvårdsverkets miljöforskningsanslag.
Författare är Mats Töpel vid IVL Svenska Miljöinstitutet och Matthew Pinder vid Institutionen för Marina Vetenskaper, Göteborgs Universitet. Författarna ansvarar för rapportens innehåll.

Stockholm i november 2022

Maria Ohlman,
Avdelningschef, Hållbarhetsavdelningen

Innehåll

Förord 3

Sammanfattning 5

Summary 7

1. Inledning 9
1.1 Miljö-DNA 9
1.1.1 Metabarkodning och metagenomik 10
1.1.2 Kvantitativ PCR-analys 11
1.1.3 Bioinformatisk analys 11
1.1.4 Sekvenseringsteknologier 11

2. Metod 14
2.1 Reproducerbarhet 14
2.2 Databaser 14
2.3 Söksträngar 15
2.4 Referenslitteratur 16
2.5 Urval av titlar 16
2.6 Analys av litteraturens innehåll 16

3. Resultat 17

4. Diskussion 19
4.1 Programversioner och analysparametrar 19
4.2 Referensdatabaser 20
4.3 Analyserad data 20
4.4 Enkla åtgärder för ökad reproducerbarhet 21
4.4.1 Kommandofil 21
4.4.2 Arbetsflödesmjukvara och versionshantering av mjukvara 22
4.4.3 Containertecknologi 22
4.4.4 Kryptografiska hashfunktioner 22
4.4.5 Versionshantering och onlinepublicering 23
4.5 Framtida forskningsbehov 24

5. Slutsatser och förslag 27
5.1 SBDI 27
5.2 NBIS 28
5.3 Förslag till framtida forskningsinriktningar 28

6. Tack 30

7. Källförteckning 31

8. Publikationer och data 39

Bilaga 1. Söksträngar 40

Bilaga 2. Jämförelselista 41

Sammanfattning

Syftet med detta projekt har varit att analysera problem och lösningar gällande reproducerbarhet rörande bioinformatisk analys av metabarkodningsdata från miljö-DNA (från engelskans Environmental DNA, eDNA). I detta syntesprojekt har det också sammanställts en lista med förslag på forskningsinriktningar som kan främja användningen av molekylära verktyg för övervakning, med särskild tonvikt på miljö-DNA och parallelliserad DNA-sekvensering (så kallad High-Throughput Sequencing, HTS) i nationella övervakningsprogram. Ett reproducerbart tillvägagångssätt för att analysera denna typ av data kommer att möjliggöra att datakällor av olika ursprung och kvalitet kan kombineras och analyseras tillsammans. Detta är särskilt relevant vid jämförelse av resultat från långa tidsserier, likt de som produceras inom nationella övervakningsprogram.

Miljöövervakning är en väsentlig del för att säkerställa ett hållbart utnyttjande av naturresurser. Nuvarande metoder för identifiering och övervakning av biologisk mångfald, särskilt analys av mikroorganismer som innebär tidskrävande och dyra mikroskopianalyser utförda av specialister, är ofta en flaskhals i analyskedjan från provtagning till dataanalys och utvärdering av miljöstatus. Dessutom kan de resultat som produceras av enskilda specialister vara svåra att återskapa, särskilt när resultatet bygger på observationer i fält, vilket gör analys av miljö-DNA till ett attraktivt komplement eller ersättning för dessa traditionella inventeringsmetoder.

Bioinformatisk analys av data från miljö-DNA är en digital process som i de flesta fall kan göras helt automatiserad. Förutsättningarna är därför goda för att skapa reproducerbara resultat, något som kommer att vara av största betydelse för storskalig användning av denna teknologi. För att undersöka i vilken utsträckning analyser av miljö-DNA är reproducerbara har vi därför genomfört en systematisk litteraturstudie av ett antal relevanta publikationer. Vi har även identifierat ett antal problem samt teknologiska lösningar som kan förbättra reproducerbarheten av bioinformatiska analyser.

Vi har undersökt reproducerbarheten av 67 undersökningar genom att definierat fyra kriterier som vi anser vara minimikrav och måste uppfyllas för att en bioinformatiska analys av data från miljö-DNA ska kunna reproduceras. Dessa är (1) programvarunamn och versioner, samt (2) analysparametrar har rapporterats, (3) referensdatabasen som används för taxonomisk klassificering är unikt definierad (t.ex. med namn och versionsnummer eller datum då den laddats ner), och slutligen (4) den data som analyserats har publicerats efter projektets slut. Vår studie visar att endast en tredjedel av de undersökta artiklarna uppfyller alla fyra kriterier, och därmed att de flesta av dessa analyser inte går att reproducera.

Många av de problem som forskare ställs inför när bioinformatik och storskalig DNA-sekvensering används för analys av biologisk mångfald, liknar de problem som finns inom andra områden, så som mjukvaruutveckling och molntjänster. Många fritt tillgängliga mjukvaruverktyg med öppen källkod har därför utvecklats för att lösa dessa problem. Ett antal av de undersökta miljö-DNA-projekten har använt flera av dessa teknologier, vilka inkluderar versionshantering av text (som underlättar distribution av reproducerbara databaser), containerteknologi (möjliggör reproducerbara arbetsflöden) och hashfunktioner (som kan säkra dataintegritet).

Utvecklingen inom analys av miljö-DNA har gjort stora framsteg under det senaste decenniet och är inom vissa områden (till exempel analys av biologisk mångfald i vatten)

redo att inkluderas i storskaliga övervakningsprogram. Men för att data och resultat ska fortsätta vara relevanta även när nya metoder för DNA-sekvensering och bioinformatisk analys utvecklas, måste reproducerbarhet vara en integrerad del av planering och genomförande av ett sådant program.

Övervakning av biologisk mångfald med hjälp av miljö-DNA kräver fortfarande utveckling och vår analys har identifierat fjorton kategorier av förslag på framtida forskningsinriktningar. Dessa inkluderar förbättring av tillgängliga referenssekvensdatabaser och utveckling av nya genetiska markörer för taxonomisk klassificering (inklusive hela mitokondrie-genom), kvantitativ analys av miljö-DNA, nya bioinformatiska verktyg och användning av nya sekvenseringsteknologier och längre sekvenser för bättre taxonomisk upplösning.

Summary

The aim of this synthesis project has been to analyse problems and solutions regarding reproducibility of bioinformatic analysis of environmental DNA (eDNA) data. During this project, we have also compiled a list of suggestions for future research that would advance the field of using molecular tools for monitoring, with special emphasis on implementing eDNA and High-Throughput Sequencing (HTS) in national monitoring programs. A reproducible approach to analysing this type of data will enable data sources of different origins and quality to be combined and analysed together, even as new technologies develop. This is particularly relevant when trying to produce comparable results from long time series, such as those obtained by national monitoring programs.

Environmental monitoring is an essential part of safeguarding the sustainable use of natural resources. Current methods for identification and monitoring of biodiversity, especially analysis of microorganisms that involve time-consuming and expensive microscopy analyses performed by specialists, are frequently a bottleneck in the analysis chain from sampling to data analysis and environmental status evaluation. In addition, the results produced by individual specialists may be difficult to reproduce, especially when field observations are required, making eDNA analysis an attractive complement or replacement.

The bioinformatic analysis of eDNA data is a digital process that in most cases can be almost totally automated, by using standard tools and methods that are openly available. Hence, bioinformatic analysis of eDNA has a great potential for generating reproducible results, something that will be of significance for any large-scale deployment of the technology. To investigate to what extent past and current eDNA analyses are actually generating reproducible results, we have conducted a systematic literature study of relevant publications.

As a way of determining the reproducibility of 67 bioinformatic analyses, we have defined four criteria that we consider the minimum requirements for being able to reproduce the bioinformatic analysis of eDNA data. These are (1) reporting of software names and versions, (2) reporting of analysis parameters, (3) unique identification of the reference database used for taxonomic classification (e.g. with name and version number or date of access), and (4) publication of the analysed data after the end of the project. Our study shows that only a third of the investigated publications meet all four criteria, meaning that most of these analyses cannot be reproduced.

Many of the problems faced by biodiversity researchers using computational methods and big data such as HTS are similar to problems found in other areas, such as software development and cloud computing. Hence, readily available open source software tools have been developed to address these issues. A number of the investigated eDNA projects have utilised several of these technologies, including version control of text (leading to reproducible databases, for example), containerisation of software (fully reproducible analysis pipelines) and hash functions (that can safeguard data integrity).

eDNA technology has seen great improvements over the last decade and is in many areas (for example biodiversity analysis in water) ready for inclusion in large scale monitoring campaigns. However, for the generated data and results to be relevant as new improvements are made available for steps such as DNA sequencing and analytical methods, reproducibility has to be an integral part of the planning stage when considering incorporation of eDNA methods into such campaigns.

Monitoring of biodiversity using eDNA still requires development in order to reach its full potential, and our analysis has identified fourteen categories of suggestions for future research. These include improvement of available reference sequence databases and development of new genetic markers for taxonomic classification (including sequencing of whole mitochondrial genomes), quantitative analysis of eDNA, new bioinformatic tools, and utilisation of long-read sequencing for better taxonomic resolution.

Based on our findings in this synthesis project, we have seen that the techniques required to facilitate reproducible eDNA studies are both available and to some extent in use in this field. With that in mind, we strongly believe that the widespread adoption of these techniques is a very feasible goal and would facilitate inclusion of eDNA metabarcoding into large-scale monitoring programmes.

1. Inledning

Det har under de senaste åren rapporterats om att det råder en reproducerbarhetskris inom flera vetenskapsområden (Baker, 2016; Samuel & König-Ries, 2021; Curty *et al.*, 2022). Speciellt i fall där resultat ska jämföras över tid, till exempel i ett nationellt övervakningsprogram, kan bristande dokumentation och transparens kring en undersökning vara särskilt problematiskt (Dully *et al.*, 2021). För att utröna om, och till vilken grad, denna reproducerbarhetskris även gäller för bioinformatisk analys av miljö-DNA data (från engelskans environmental DNA ofta förkortat eDNA), har det inom detta projekt genomförts en systematisk studie av relevant vetenskaplig litteratur samt myndighetsrapporter. Syftet med projekt har även varit att identifiera tekniska lösningar som kan underlätta reproducerbarhet av bioinformatiska analyser, och som lämpar sig att användas i övervakningsprogram för biologisk mångfald med hjälp av miljö-DNA och storskalig paralelliserad DNA sekvensering (High-Throughput Sequencing, HTS). Från det undersökta materialet har vi även gjort en sammanställning av förslag på framtida forskningsinriktningar som kan utveckla fältet och stärka framtida nationella övervakningsprogram.

Vi har valt att avgränsa undersökningen till reproducerbarhet av de bioinformatiska delarna av miljö-DNA analys. Anledningen till denna avgränsning är att en bioinformatisk analys är digital och därför kan göras helt eller till största delen reproducerbar. Teknologiska framsteg under de senaste årtiondena har dessutom gjort att kostnaden för DNA sekvensering sjunkit dramatiskt, vilket resulterat i att fler undersökningar nu inkluderar bioinformatisk sekvensanalys samt att större datamängder kan genereras per projekt. Tillsammans med utvecklingen av fält- och labb-protokoll har därför fördelarna med att inkludera miljö-DNA analys i nationella övervakningsprogram ökat, samtidigt som behovet av reproducerbarhet också ökar för att denna data ska komma bäst till nytta.

Bristande reproducerbarhet hos enskilda undersökningar resulterar onekligen i resultat med lägre tillförlitlighet. Dessa brister leder också till att jämförande studier och metaanalyser försvåras och att skillnader i resultat mellan undersökningar inte med säkerhet kan tillskrivas biologiska skillnader på den provtagna platsen, skillnader i analysmetod eller i de referenssekvenser som använts. Vi tror dock att de åtgärder vi rekommenderar i denna rapport kan avhjälpa dessa brister, vilket gör att miljö-DNA metoder kommer kunna bidra med tillförlitliga data till framtida nationella övervakningsprogram.

1.1 Miljö-DNA

Deoxyribonukleinsyra, DNA, är den molekyl som lagrar den genetiska kod som styr utveckling och funktion hos alla levande celler. Denna dubbelsträngade molekyl kan undkomma nedbrytning i miljön om rätt förutsättningar råder, vilket gör att genetiska spår sprids i miljön via bland annat döda hudceller, saliv och avföring från organismerna som lever där (Rees *et al.*, 2014). Dessa spår kan därmed tolkas och ge en bild av den biologiska mångfalden i den provtagna miljön, på artnivå (Deiner *et al.*, 2017), populationsnivå (Sigsgaard *et al.*, 2016; Weitemier *et al.*, 2021) eller individnivå (Farrell *et al.*, 2022).

Vid analys av dessa spår samlas ett prov från till exempel vatten, luft eller jord in och analyseras sedan genom sekvensering (se nedan) eller annan form av detektion som till exempel via polymeraskedjereaktion (PCR, se nedan). En miljö-DNA analys kan delas in i de tre distinkta men sammanlänkade delarna fältarbete, labbarbete och dataanalys, där den senare är en helt digital process, möjlig att helt eller till största delen automatisera och i förlängningen också reproducera.

Begreppet miljö-DNA (eDNA) användes första gången 1987 (Ogram *et al.*, 1987) men det var först när framsteg inom DNA-sekvensering lanserades tjugo år senare som teknologin blev tillgänglig för en bredare användargrupp. Ansträngningar för att sänka kostnader och öka sekvens-kvaliteten vid human-DNA sekvensering i medicinskt syfte har lett till ytterligare framsteg som även gynnat utvecklingen av analyser för miljö-DNA (Pompanon & Samadi, 2015).

Utveckling inom fältet är alltså till stor del driven av teknologisk utveckling inom angränsande områden. Förbättringar av provtagnings- och laboratorieteknik har gjort att ett bredare urval av provtyper nu kan analyseras, och förbättringar av sekvenseringsteknologier möjliggör användning av fler och längre markörgener (se nedan). Likaså möjliggör förbättringar av mjukvara för miljö-DNA analys, såväl som de databaser med referenssekvenser som används, att taxonomisk klassificering med högre upplösning nu är möjlig jämfört med tidigare (Yang *et al.*, 2017). För att miljö-DNA teknologi ska nå sin fulla potential och fullt ut användas inom miljöövervakning behöver dock frågor kring skalbarhet av analyser samt reproducerbarhet av resultat först lösas (Mousavi-Derazmahalleh *et al.*, 2021).

1.1.1 Metabarkodning och metagenomik

DNA-molekyler som samlats in från ett miljöprov kan antingen analyseras genom kvantifiering eller avläsas med så kallad DNA-sekvensering. Vid sekvensering avläses de delar som utgör DNA molekylen (kallade baser) och resulterar i en sekvens av adenin (A), guanin (G), cytosin (C) eller tymin (T). För att öka chansen att detektera en molekyl från en specifik art kan en metod kallad metabarkodning tillämpas. Med denna metod amplifieras (DNA-fragmenten mångfaldigas exponentiellt) och sekvenseras specifika DNA fragment från så kallade markörgener, gener som alla arter inom den undersökta organismgruppen har. Dessa sekvenser jämförs sedan med sekvenser av känt ursprung i en referensdatabas, och sekvensdata kan på så vis omvandlas till en lista med arter som detekterats i provet. Denna metod är alltså beroende av att (1) ett tillräckligt långt DNA-fragment sekvenseras så att artspecifika skillnader kan detekteras, (2) att sekvenserna har tillräckligt hög kvalitet så att sekvensfel inte resulterar i fel art-klassificering, och att (3) referensdatabasen innefattar arterna som förekommer i den provtagna miljön.

Ett annat tillvägagångssätt vid analys av miljöprov är att sekvensera den totala mängden DNA i provet, så kallad metagenomik. Detta möjliggör detektion av organismer vars markörgener är mindre kända, och därför svårare att amplifiera, samt att en större del av organismernas DNA kan analyseras.

1.1.2 Kvantitativ PCR-analys

En ofta använd metod för analys av miljö-DNA kallas kvantitativ PCR (qPCR). Med denna metod detekteras, och därefter kvantifieras, DNA molekylerna snarare än avläses på det sätt som görs vid DNA-sekvensering. Metoden lämpar sig för detektion av en eller ett fåtal arter eller sekvensvarianter i ett miljöprov, och genererar inte närmelsevis så stora datamängder som parallelliserad DNA-sekvensering gör. Reproduceringsproblematiken för denna typ av undersökningar skiljer sig därför från DNA-sekvenseringsprojekt varför de inte inkluderats i denna litteraturstudie.

1.1.3 Bioinformatisk analys

De sekvenseringsbaserade metoderna beskrivna ovan lämpar sig för undersökning av en eller flera specifika organismgrupper, och har gemensamt att de ofta genererar stora datamängder. För att denna data ska vara relevant och användbar i framtida jämförande analyser, till exempel i ett nationellt övervakningsprogram, krävs ofta specifik kompetens som ligger utanför generell bioinformatisk kompetens (se avsnittet *Enkla åtgärder för ökad reproducerbarhet*). Bioinformatik är ett tvärvetenskapligt forskningsområde som kombinerar biologi och datalogi och ofta appliceras på biologiska sekvensdata (https://www.ne.se/uppslagsverk/encyklopedi/lång/bioinformatik).

En bioinformatisk analys av ett metabarkodningsprov innebär ofta en kombination av följande analyssteg, (1) sekvensernas kvalitet undersöks, följt av (2) filtrering och trimning där hela eller delar av DNA-sekvenser av låg kvalitet exkluderas. Sekvenserna kan sedan analyseras ytterligare, för att (3) identifiera felaktiga sekvenser av hög kvalitet (så kallade falska positiva resultat), och kvarvarande data (4) jämförs sedan med sekvenser av känt ursprung i en referensdatabas (Taberlet *et al.*, 2018a). Efter denna analys har DNA-sekvenserna omvandlats till en taxonomisk lista av identifierade arter i provet. Samtliga dessa steg, inklusive val av sekvenseringsteknologi, analysprogram och utformning av referensdatabasen, har en större eller mindre påverkan på det slutliga resultatet, varför kännedom om vilka steg, och i vilken ordning som de utförts, är viktiga att känna till när resultatet från olika undersökningar jämförs (Santoferrara *et al.*, 2020). Detta kan till exempel gälla en jämförande studie över tid eller mellan geografiskt skilda platser. Skillnader som observeras mellan olika prover i en sådan jämförande studie kan alltså bero på biologiska olikheter eller skillnader i hur de olika proverna analyserats. En ofta ofrånkomlig skillnad mellan provresultat i en tidsserie-analys är att nya referenssekvenser hela tiden tillkommer i publika databaser (Figur 1), vilket möjliggör detektion av fler arter med hjälp av nyare versioner av en referensdatabas (Goodwin *et al.*, 2019).

1.1.4 Sekvenseringsteknologier

Sedan parallelliserad DNA sekvensering (dvs. metoder där miljontals molekyler sekvenseras samtidigt) introducerades under tidigt 2000-tal, så har möjligheterna att analysera miljö-DNA förändrats dramatiskt (Adamowicz *et al.*, 2017). De olika sekvenserings-teknologier som idag framför allt används inom miljö-DNA analys kan karaktäriseras baserat på tre egenskaper; mängden sekvenser som produceras, längden på sekvenserna, samt sekvensläsningarnas kvalitet. Framför allt längd (Adamowicz *et al.* 2017) samt kvalitet kan ha stor påverkan på analysresultatet i en miljö-DNA studie (Shokralla *et al.*, 2012). Forskare inom fältet har tidigare varit tvungna att kompromissa och antingen välja mellan längre sekvensläsningar eller

högre kvalitet, men nya framsteg inom sekvenseringsteknologi möjliggör nu att både ökad längd och högre kvalitet kan erhållas samtidigt. Användningen av dessa nya teknologier kan visa sig vara fördelaktig för metabarkodning, men kommer kräva ytterligare forskning och utveckling. Som ett exempel kan nämnas att längre DNA-sekvenser från miljö-DNA undersökningar kommer kräva att även längre referens-sekvenser tas fram för att teknologin ska kunna utnyttjas till fullo.

ILLUMINA

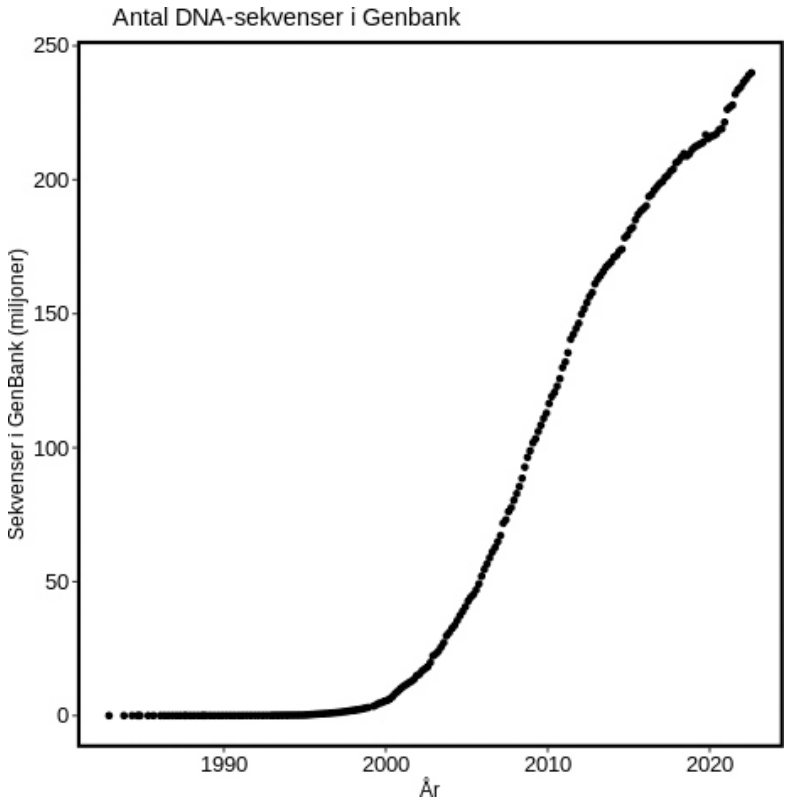
Illumina MiSeq lanserades 2011 (https://emea.illumina.com/company/news-center/press-releases/2011/1515239.html) och är fortfarande, trots att det är en relativt gammal plattform, den dominerande sekvenseringsteknologin för miljö-DNA analys. På senare år har dock Illumina NovaSeq 6000, som erbjuder liknande egenskaper som MiSeq, introducerats på marknaden. Fördelarna med dessa plattformar är att de kan producera många sekvenser av hög kvalitet. Nackdelen är att de genererar kortare sekvensläsningar (150-300 baser), vilket gör att endast en kortare del av en markörgen kan analyseras. Detta i sin tur kan leda till problem med taxonomisk upplösning i en undersökning, då längre DNA-molekyler kan behöva analyseras för att skilja närstående arter åt.

PACIFIC BIOSCIENCES (PACBIO)

I vår analys har vi identifierat få projekt som använt denna plattform för miljö-DNA analys. Denna teknologi erbjuder dock fördelen att kunna sekvensera betydligt längre DNA-molekyler än Illuminas plattformar, vilket möjliggör att en längre markörgen kan analyseras, och i sin tur att den taxonomiska upplösningen i en undersökning ökar.

OXFORD NANOPORE TECHNOLOGIES (ONT)

Ingen av de undersökta artiklarna inkluderar data från denna sekvenseringsplattform. Detta beror med största sannolikhet på att sekvenskvaliteten hos ONT tidigare varit för låg i jämförelse med sekvenser producerade av Illuminas och PacBios plattformar (Krehenwinkel *et al.*, 2019). Utveckling under senare år har dock gjort att kvaliteten har ökat avsevärt vilket gör att denna teknologi har potential att kunna användas för metabarkodning och metagenomik. En viktig fördelar med denna plattform är att den kan producera mycket långa sekvensläsningar vilket innebär att längre markörgener (till exempel hela mitokondrie- eller kloroplastgenom) kan analyseras, något som möjliggör högre taxonomisk upplösning i en del undersökningar. Ytterligare fördelar är att plattformen är både liten och billig vilket möjliggör att sekvensering kan utföras i fält, vilket ger mycket snabbare resultat än traditionella sekvenseringsplattformar (Tyler *et al.*, 2018; Chang *et al.*, 2020).



Figur 1. Mängden DNA-sekvensdata som producerats och publicerats i publika databaser har ökat dramatiskt under de senaste årtiondena. Här visas antalet sekvenser i Genbank (varje punkt i grafen motsvarar en specifik version av databasen), en internationell databas för DNA- och protein-sekvensdata, som ofta används för taxonomisk klassifikation av miljö-DNA sekvenser. Den branta stigningen av kurvan i början av 2000-talet sammanfaller med att storskalig sekvensering av det mänskliga genomet inleddes, samt lansering av ny sekvenseringsteknologi som ledde till minskade kostnader för DNA-sekvensering.

(Källa: National Center for Biotechnology Information (NCBI); https://www.ncbi.nlm.nih.gov/genbank/statistics/; https://www.genome.gov/about-nhgri/Brief-History-Timeline)

2. Metod

2.1 Reproducerbarhet

Med reproducerbarhet menar vi i denna rapport möjligheten att återskapa ett bioinformatiskt resultat givet en beskrivning av programvara, data och referensmaterial som använts i en analys. Denna beskrivning kan utgöras av en metoddel i en vetenskaplig publikation, eller att den analyskod som använts gjorts tillgänglig. Då det inte varit möjligt att i denna undersökning försöka reproducera samtliga analyser i det undersökta materialet har vi i stället satt upp fyra kriterier som använts för att bedöma reproducerbarheten hos de inkluderade miljö-DNA analyserna:

- Namn och versionsnummer för programvaran som använts är dokumenterad.
- Analys-parametrar för samtliga analyssteg är angivna.
- Referensdatabasens namn och versionsnummer anges, alternativt en länk och/ eller unik identifierare är angiven.
- De analyserade DNA-sekvenserna är allmänt tillgänglig och angivna med en unik identifierare (så kallad FAIR data; Wilkinson *et al.*, 2016).

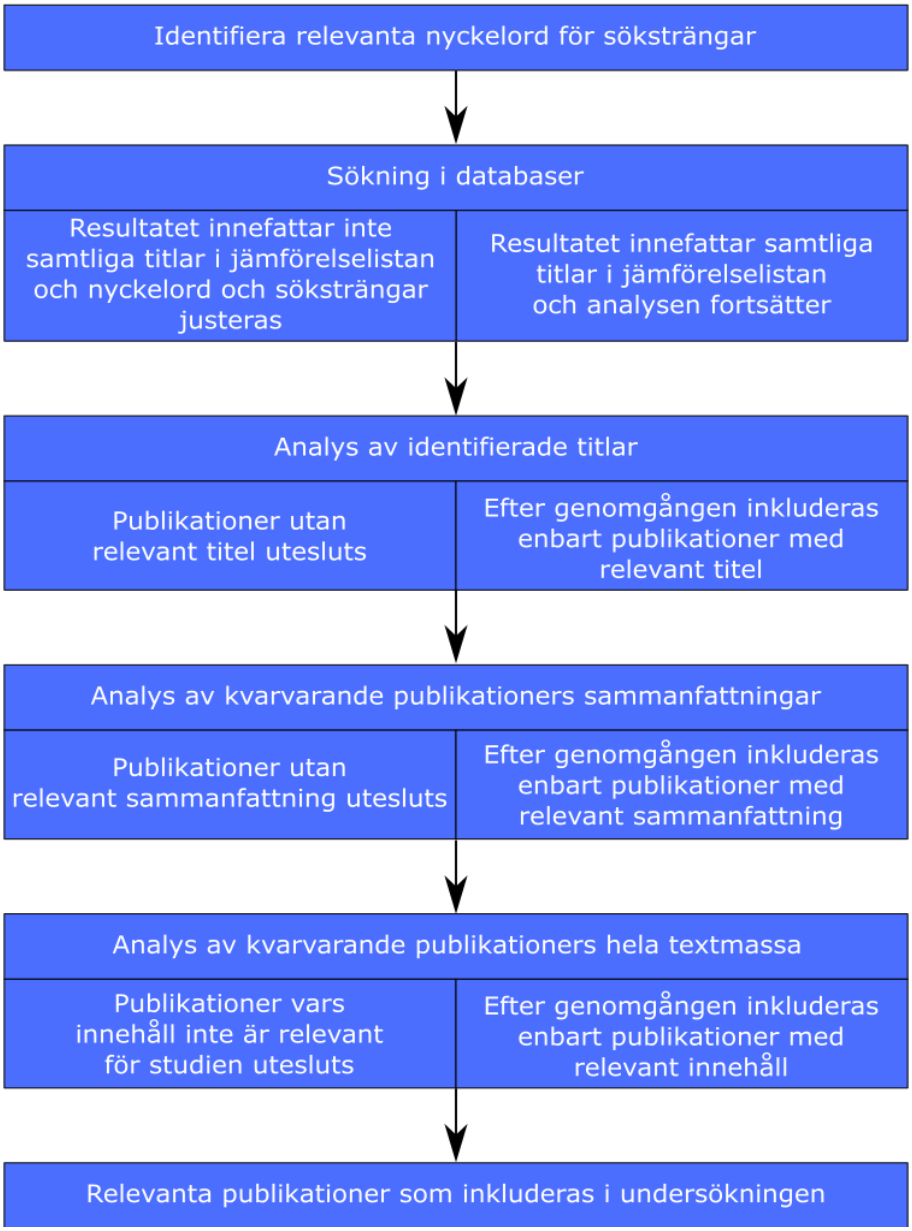
För att undersöka och belysa eventuella problem kring reproducerbarhet vid miljö-DNA analys har en systematisk litteraturstudie genomförts med en modifierad metod av Gough m.fl (2017; Figur 2).

2.2 Databaser

För att få en så heltäckande bild av ämnet gjordes litteratursökningar i följande databaser den 26:e april 2022:

- DIVA (https://www.diva-portal.org/)
- BASE (https://base-search.net/)
- Libris (https://libris.kb.se/)
- SwePub (https://swepub.kb.se/)
- PubMed (https://pubmed.ncbi.nlm.nih.gov/)
- Scopus (https://www.scopus.com/)

För att möjliggöra reproducerbarhet av denna analys, underlätta urvalet av titlar, samt att automatiskt exkludera duplicerade sökresultat, så gjordes alla sökningar via respektive databas Application Programming Interface (API). Datorkod för respektive databas-sökning skrevs i Python v3.8.3 och finns tillgänglig som Jupyter Notebooks via https://github.com/topel-research-group/Reproducible-analysis-of-eDNA-for-national-biodiversity-monitoring-programs.



Figur 2. Översikt av metoden som använts för urval av artiklar att inkludera i undersökningen. Baserad på metod av Gough et al. (2017).

2.3 Söksträngar

De respektive databaserna har olika inriktning på de dokument de lagrar, och den söksträng som användes för respektive databas har därför modifierats något, men utgår från följande grund, där de booleska operatorerna AND/OR använts för att begränsa eller vidga sökningen.

(“eDNA” OR “environmental DNA” OR “metabarcoding” OR “eRNA” OR “environmental RNA”) AND (“biodiversity” OR “species richness” OR “monitoring” OR “biomonitoring”) AND (“high throughput sequencing” OR “HTS” OR “throughput”)

I de flesta fall begränsades inte sökningen till vissa delar av dokumentet (till exempel titel eller nyckelord) utan gjordes på hela textmassan. Exakt söksträng som användes för respektive databas presenteras i Bilaga 1.

2.4 Referenslitteratur

För att begränsa sökresultatet till en överskådlig mängd litteratur och samtidigt säkerställa att relevanta titlar inkluderades, skapades innan databas-sökningen en referensliteraturlista med titlar som ansågs viktiga för denna undersökning (här kallad *jämförelselista*). Denna lista utformades så att den inkluderade titlar från vetenskaplig litteratur (Bista *et al.*, 2017; Deiner *et al.*, 2017; Tapolczai *et al.*, 2019) samt myndighetsrapporter (Hänfling *et al.*, 2016; Winding *et al.*, 2019; Meissner *et al.*, 2020; Norros *et al.*, 2022) som beskriver miljöövervakning med miljö-DNA och som på förhand identifierats som relevanta för denna undersökning (se Bilaga 2). Respektive söksträng justerades därefter så att samtliga titlar i jämförelselistan inkluderades i den slutgiltiga listan av analyserad litteratur.

2.5 Urval av titlar

För att reducera antalet undersökta publikationer samt för att säkerställa att enbart relevant material inkluderades i undersökningen gjordes en manuell bedömning av sökresultatet i två steg. I första steget kontrollerades samtliga titlar identifierade i den automatiska sökningen, och publikationer utan relevant titel exkluderades. I steg två lästes sammanfattningen av kvarvarande publikationer och ytterligare titlar exkluderades (till exempel undersökningar som enbart använt kvantitativ PCR-analys och därmed ställts inför en annan typ av reproduceringsproblematik jämfört med metabarkodningsstudier). I tillägg exkluderades även titlar när själva textmassan i dessa lästes och de då inte ansågs relevanta för denna undersökning. För att säkerställa en opartisk bedömning av respektive publikation i de två första urvalsstegen så inkluderades bara titel samt sammanfattning i de filer som undersöktes, så att till exempel namn på författare samt tidskriftens namn etc. inte skulle påverka urvalet.

2.6 Analys av litteraturens innehåll

Av de 134 titlar som identifierats som relevanta för denna undersökning bestod hälften av review-artiklar, rapporter och liknande och 67 innehöll egen analys av originaldata, dvs. data som producerats under respektive undersökning. Metodbeskrivningen i dessa 67 titlarna lästes sedan och de fyra kriterierna för reproducerbarhet undersöktes.

I en ytterligare analys sammanfattades även beskrivningar av framtida forskningsbehov som lyfts av författarna till de ursprungliga 134 undersökta artiklarna. Många forskningsförslag var identiska eller liknande mellan publikationer, och efter en sammanställning identifierades fjorton kategorier av förslag. Antalet publikationer med rekommendationer från vardera kategori summerades sedan.

Resultatet från dessa analyser samlades i ett kalkylblad som är tillgänglig via https://github.com/topel-research-group/Reproducible-analysis-of-eDNA-for-national-biodiversity-monitoring-programs.

3. Resultat

Efter att sökströngarna för respektive databas justerats så inkluderade sökresultatet samtliga titlar i jämförelselistan (se Bilaga 1 och 2). Antalet titlar som sökningarna resulterade i presenteras i Tabell 1. Då samma publikation kan registreras i flera databaser resulterade sökningen i ett antal duplikat och efter att dessa exkluderats så återstod 915 titlar. Relevant titel för denna undersökning identifierades hos 139 publikationer och av dessa uteslöts tre efter att sammanfattningen lästs. Ytterligare två titlar uteslöts efter att hela textmassan i dessa lästs. Av kvarvarande 134 publikationer innehöll 67 beskrivningar av miljö-DNA analys av originaldata, och resterande utgjordes av review-artiklar, myndighetsrapporter och liknande utan egen bioinformatisk analys (Tabell 2). De 67 original-artiklarna ingick i analysen av reproducerbarhet (Tabell 3), och samtliga 134 artiklar låg till grund för analysen av framtida forskningsbehov.

I ungefär hälften av publikationerna har den använda programvaran inte rapporterats på ett tillbörligt sätt och i ungefär 25% har inte analysparametrar angivits. Bland de undersökningar som inte tillräckligt utförligt presenterar använd referensdatabas ser vi framför allt tre typer av misstag. Antingen har enbart namnet på databasen rapporterats (t.ex. *Diat.barcode*) utan att versionsnummer eller datum då databasen laddats ner inkluderats, eller har enbart distributörens namn angivits (t.ex. *NCBI*) och inget specifikt databasnamn (t.ex. *NCBI nt* eller *NCBI nr*). En tredje kategori utgörs av individuellt utformade databaser, bestående av eget framtaget referensmaterial och/eller ett urval av allmänt tillgängliga sekvenser, utan någon referens till var denna specifika referensdatabas nu finns tillgängliga. Tio av de 67 studierna inkluderade heller inte information om var de analyserade DNA-sekvenserna finns tillgängliga.

Resultatet från sammanställningen av framtida forskningsbehov visas i Figur 3. I 109 av de 134 undersökta artiklarna presenterades förslag på framtida forskningsinriktningar. I tillägg till de fjorton kategorier vi identifierat, gavs även förslag på ett stort antal ytterligare inriktningar men då dessa enbart framfördes i ett fåtal publikationer presenteras här enbart de vanligast förekommande. Samtliga identifierade förslag finns presenterade på projektets Github-sida (se sektionen *Publikationer och Data* nedan).

Tabell 1. Antal publikationer som identifierades i respektive databas. Vissa publikationer identifierades i flera databaser.

Databas	Identifierade publikationer
DIVA	80
BASE	75
Libris	33
SwePub	33
PubMed	417
Scopus	653

Tabell 2. Antal undersökta publikationer

Totalt antal identifierade unika publikationer	Publikationer med relevant titel	Publikationer med relevant sammanfattning	Publikationer med relevant innehåll	Artiklar med orginaldata
915	139	136	134	67

Tabell 3. Resultat av reproducerbarheten hos 67 miljö-DNA studier. Tabellen visar antalet samt procent av alla publikationer som (1) specificerar namn och version av mjukvara som använts (2), analysparametrar som använts, (3) specificerar namn och versionsnummer för använd referensdatabas, samt (4) inkluderar unik identifierare eller länk till analyserad data. Sista kolumnen visar antal publikationer som uppfyller alla fyra kriterierna och i denna analys anses vara reproducerbara.

	Namn och program-version specificerad	Analys-parametrar specificerade	Referens-databas tillgänglig	Analyserad data tillgänglig	Reproducer-bara analyser
Antal	36	51	41	57	22
%	53.7	76.1	61.2	85.1	32.8

Figur 3. Fördelning av föreslagna framtida forskningsinriktningar som identifierats. Liknande förslag har här sammanställts i fjorton övergripande kategorier.

4. Diskussion

Resultatet av vår analys visar att problem med vetenskaplig reproducerbarhet som rapporterats från andra forskningsområden även förekommer inom detta fält (t.ex. Baker, 2016; Samuel & König-Ries, 2021; Curty *et al.*, 2022). I vårt undersökta material ingår även ett antal publikationer som beskriver liknande dåliga erfarenheter kring reproducerbarhet av bioinformatiska analyser (Dufresne *et al.*, 2019; Creedy *et al.*, 2022; Zaiko *et al.*, 2022). Detta gör att vi tror att vårt visserligen begränsade undersökningsmaterial är representativt för den problematik som finns kring reproducerbarhet av bioinformatik i miljö-DNA analyser. I tillägg till de fyra kriterier som undersökts skulle en djupare analys även kunna innefatta kvalitet och tillgänglighet till relevant metadata. Då en allmän standard för miljö-DNA sekvens-metadata ännu inte finns tillgänglig har vi inte kunnat definiera några sådana bedömningskriterier.

Flera författare påpekar att bioinformatisk analys kompliceras av att det finns ett närmast oändligt utbud av mjukvara och analysparametrar som kan kombineras på ett flertal olika sätt (Bailet *et al.*, 2020; Bowers *et al.*, 2021; Creedy *et al.*, 2022; Zaiko *et al.*, 2022). Vid denna typ av analys saknas ofta empiriska bevis för att ett specifikt resultat är det korrekta, vilket gör det omöjligt att avgöra vilken metod som kan anses vara bäst (Griffiths *et al.*, 2018; Li *et al.*, 2019). Valet av analysmetod kan därför anses vara av mindre betydelse (Deiner *et al.*, 2017), så länge som metodiken dokumenteras på ett tillräckligt utförligt sätt att den kan reproduceras vid ett senare tillfälle eller tillämpas i metaanalyser som inkluderar nya data (Tedersoo *et al.*, 2015).

4.1 Programversioner och analysparametrar

Misstagen att inte ange mjukvaruversionsnummer samt vilka analysparametrar som använts (även om de bara varit förinställda standardvärden) tolkar vi som en bristande förståelse för hur mycket dessa val kan påverka resultatet av en analys (Cashman *et al.*, 2018). På samma sätt som man ger en detaljerad beskrivning av ett laboratorieprotokoll i en material- och metodsektion när ett forskningsresultat publiceras, så är det också viktigt att ge detaljerade beskrivningar av de bioinformatiska verktyg som används i ett experiment. Till exempel, när man i ett laborativt experiment använder en viss kemikalie, är vanlig praxis att ange tillverkaren och mängden som används, eftersom dessa faktorer kan ha stor effekt på resultatet. På samma sätt är det lämpligt att ange version av programvaran samt de specifika inställningarna som använts.

Bioinformatisk programvara är inte statiska verktyg utan ofta under aktiv utveckling (Gardner *et al.*, 2022). Därför kan många versioner av ett givet verktyg finnas tillgängliga, vilka kan ge olika resultat beroende på de förbättringar som gjorts av programvaran (Garijo *et al.*, 2013; Piccolo & Frampton, 2016). Detta problem förvärras av antalet inställningar som denna typ av programvara kan ha. I många av artiklarna i vår studie presenterades ofta bara speciellt viktiga parameterinställningar. Bioinformatisk programvara kan dock ha dussintals inställningar, ofta satta till standardvärden, som kanske inte nämns av författarna. Dessa standardvärden kan ändras mellan versioner och kan därmed ha en påverkan på det slutliga resultatet (t.ex. https://www.ncbi.nlm.nih.gov/books/NBK131777/).

4.2 Referensdatabaser

En oroväckande trend i vårt resultat är att en stor del av publicerade miljö-DNA analyser inte tillräckligt utförligt dokumenterar det referensmaterial som använts för taxonomisk klassificering av sekvensdata. Medan de flesta av artiklarna presenterade vilka databaser som ingått i undersökningen, var det många som inte specificerade vilken version som använts, något som avsevärd kan påverka tolkningen av resultatet. Till exempel, om nya arter dyker upp i resultat som genereras vid en senare tidpunkt i en längre studie, beror detta på en verklig förändring i biologisk mångfald, eller att en sekvens för den nya arten lagts till i en senare version av referensdatabasen? Dessa frågor kan endast besvaras om databasversionen är känd. Medan många av artiklarna i vår studie använde väletablerade webbaserade databaser som BOLD, SILVA och NCBI nt/nr, som kan citeras med antingen versionsnummer eller datum för åtkomst, använde vissa också eget utformade databaser. Ett flertal olika digitala arkiv finns tillgängliga för lagring och publicering av denna typ av data, och i vårt undersökta material hittar vi referenser till webbplatser som GitHub (https://github.com/) och Zenodo (https://zenodo.org/). Det finns även bra lösningar för att säkerställa integriteten hos eget utformade databaser (se sektionen *Kryptografiska hashfunktioner*) och säkerställa att data inte förändras när filer delas via dessa plattformar.

4.3 Analyserad data

Av de fyra kriterier för reproducerbarhet som använts i denna undersökning uppfyllde flest författare kriteriet att analyserad data ska finnas tillgänglig efter publicering. Detta beror med största sannolikhet på att de flesta vetenskapliga tidskrifter numera ställer kvar på att data finns tillgänglig i någon form av publik databas vid publiceringstillfället. På samma sätt kräver även många forskningsfinansiärer att forskningsansökningar inkluderar en datahanteringsplan. Av de tio publikationerna som inte uppfyllde detta kriterium var tre äldre (från 2012, 2013 och 2015), och tre var inte sakkunniggranskade artiklar (två rapporter respektive en magisteravhandling). De återstående fyra publikationerna presenterade antingen en ny mjukvara eller ett arbetsflöde, och gav ofta referenser till åtminstone en del av den data som använts för publikationen.

Då den bioinformatiska analysen av sekvensdata är ett av de mest kritiska stegen i miljö-DNA analys (Deiner *et al.*, 2017), samt att variationsmöjligheterna av analysprogram, parametrar, referensmaterial och typ av sekvensdata är så stor, är hänsynen till reproducerbarhet extra viktig, men ofta förbisedd. Denna variation av analysmöjligheter spelar dock stor roll för utvecklingen av miljö-DNA fältet och möjliggör att nya insikter kan skapas från både nya och gamla data. Variationen utgör likväl en utmaning och en risk för att analysresultat och data kan bli inaktuella (Adamowicz et al. 2017; Zaiko *et al.*, 2022). Därför, för att den sekvensdata som kommer genereras i framtida nationella övervakningsprogram ska komma bäst till nytta bör extra vikt läggas vid att analys- och referensdata, analysmetoder och arbetsflöden som används görs tillgängliga på ett enkelt och reproducerbart sätt.

Tidsseriedata från till exempel ett övervakningsprogram ackumuleras till sin natur över tid, och om prover i en sådan tidsserie ska jämföras med varandra måste de kunna analyseras på samma sätt (Goodwin *et al.*, 2019). Detta gör det viktigt att korrekt dokumentera detaljerna i analysen enligt ovan (program och deras versioner,

de parametrar som används, den underliggande databasen). Att ha dessa analysdetaljer tillgängliga i en form som automatiskt kan tillämpas på nya data (som till exempel en kommandofil eller en mjukvarucontainer som beskrivs nedan) har inte bara fördelen att möjliggöra konsekvent analys över hela tidsserien, utan minskar även risken för mänskliga fel i varje analyssteg, samt leder till en automatisk dokumentation av hur analysen utförts.

4.4 Enkla åtgärder för ökad reproducerbarhet

Traditionellt utgör metoddelen i en vetenskaplig publikation den enda källan som beskriver hur en bioinformatisk analys utförts, men vår undersökning visar är att detta ofta inte är tillräckligt för att reproducera ett resultat (se även Goldberg *et al.*, 2016; Nicholson *et al.*, 2020). I den undersökta litteraturen presenteras dock ett antal teknologier och metoder som om rätt tillämpade avsevärt kan öka reproducerbarheten av bioinformatiska analyser. Vi har identifierat ett antal olika teknologier som ännu inte fullt utnyttjas inom detta fält, men som med lätthet skulle kunna förbättra reproducerbarheten av bioinformatiska analyser avsevärt, och som lämpar sig väl att implementera i framtida nationella övervakningsprogram.

4.4.1 Kommandofil

Av de steg som ingår i ett miljö-DNA projekt är den bioinformatiska analysen den som enklast kan göras reproducerbar, något som tyvärr inte återspeglas i vårt analysresultat. Den absoluta merparten av de analysprogram som används för DNA-sekvensanalys är kommandoradsprogram, dvs. program som startas i en virtuell datorterminal (hädanefter kallad terminal). Denna terminal kan vara till exempel MobaXterm (https://mobaxterm.mobatek.net/), iTerm (https://iterm2.com/) eller GNOME terminal (https://help.gnome.org/users/gnome-terminal) och finns tillgängliga för samtliga operativsystem. I och med att interaktionen med terminalen är helt textbaserad kan samtliga kommandon, inkluderande namn på analysprogram samt samtliga använda programparametrar, enkelt sparas i en textfil som möjliggör enkel reproduktion av analysresultaten, samt samtidigt automatisk dokumentation av de steg som ingått i analysen. Att frångå manuell inmatning av kommandon och i stället använda sig av kommandofiler underlättar inte bara själva arbetet utan även kvaliteten och reproducerbarheten hos en analys (Deiner *et al.*, 2017). Flera artiklar som ingår i vårt analyserade material länkar till kommandofiler som använts i de respektive analyserna, i programmeringsspråk som bash (t.ex. Atherton & Jondelius, 2020), R (t.ex. Ritter *et al.*, 2020) och Python (t.ex. Ríos-Castro *et al.*, 2021). En annan liknande teknologi som användes i de undersökta artiklarna är Jupyter Notebook (Harper *et al.*, 2018). Detta är ett typ av interaktivt dokument som kan användas i en webbläsare för att utföra analyser (flera olika programmeringsspråk stöds), samt kan inkludera dokumenterande text, analysresultat och figurer.

4.4.2 Arbetsflödesmjukvara och versionshantering av mjukvara

Ett något mer avancerat sätt att dokumentera, distribuera och utföra sekvensanalyser på är att använda speciell mjukvara utvecklad för att bygga arbetsflöden. Exempel på sådan mjukvara är Snakemake (Mölder *et al.*, 2021) och Nextflow (Di Tommaso *et al.*, 2017). Vår litteraturstudie har även identifierat ett mjukvaruprojekt specifikt inriktad på arbetsflöde vid analys av metabarkodnings data (Mousavi-Derazmahalleh *et al.*, 2021). Welzel *et al.* (2020) har i sitt arbetsflöde kallat Natrix även integrerat Conda (https://conda.io/), ett system för versionshantering av mjukvara. Dessa två system tillsammans möjliggör att arbetsflöden inkluderande både ordningsföljden på analyssteg, samt specifik version av mjukvara som används, på ett lätt sätt kan dokumenteras och att analysen dessutom enkelt kan reproduceras vid ett senare tillfälle.

4.4.3 Containertecknologi

Tekniska lösningar som kommandofiler, Snakemake eller Conda är utmärkta för att säkerställa reproducerbara bioinformatiska analyser, men har gemensamt att de kräver att användaren själv installerar de programvaror som krävs för själva analysen av data. Detta kan vara en mer eller mindre komplicerad process då programvaror kan behöva kompileras från källkod, eller kan vara beroende av ytterligare mjukvara som kan vara komplicerad att installera. Dessa problem kan kringgås genom att kommandofiler, arbetsflöden, analysprogram och även data görs tillgänglig i en så kallad mjukvarucontainer, en form av virtualisering på operativsystemnivå. Exempel på denna typ av teknologi är Docker (Merkel, 2014) som ofta används i molntjänster och mjukvaruutveckling, samt Singularity (Kurtzer *et al.*, 2017) som fokuserar på reproducerbarhet och IT-säkerhet. Denna typ av lösning erbjuder ypperliga möjligheter att reproducera analyser och resultat men är enligt vår undersökning fortfarande relativt ovanligt förekommande i miljö-DNA analys. Inom det bioinformatiska fältet är dock denna teknologi inte ovanlig, och fyra av publikationerna i vår analys presenterar ny mjukvara tillgänglig i containrar, antingen Docker (SLIM [Dufresne *et al.*, 2019]; Natrix [Welzel *et al.*, 2020]) eller Singularity (Anacapa [Curd *et al.*, 2019]; eDNAFlow [Mousavi-Derazmahalleh *et al.*, 2021]).

4.4.4 Kryptografiska hashfunktioner

Hashkodning (https://www.ne.se/uppslagsverk/encyklopedi/lång/hashkodning) är ett sätt att enkelt säkerställa integriteten hos data, och verifiera att innehållet i en datafil inte har ändrats. Metoden används därför med fördel för att försäkra att till exempel kommandofiler eller filer innehållande sekvensdata efter nedladdning från en server fortfarande är en exakt kopia av originalfilen (se t.ex. Elbrecht *et al.*, 2017; Harper *et al.*, 2018; Curd *et al.*, 2019). Metoden kan även appliceras på egenhändigt sammansatt referenssekvensdata som används för taxonomisk klassificering av DNA-sekvenser. Filnamn och versionsnummer för dessa referensdatabaser kan enkelt ändras av misstag, men tillhörande hashkod kommer endast förändras om själva innehållet i filen ändras. Metoden kan tillämpas på alla typer av filer, och har använts för att verifiera data som skapats i denna undersökning, där varje delresultat presenteras i text-filformat med tillhörande hashkod. Detta för att öka transparensen i metoden vi använt samt för att säkerställa reproducerbarhet av resultatet (https://github.com/topel-research-group/Reproducible-analysis-of-eDNA-for-national-biodiversity-monitoring-programs).

4.4.5 Versionshantering och onlinepublicering

Under tidigt 2000-tal introducerade Jean-Claude Bradley ett nytt sätt att distribuera vetenskapliga resultat kallat "Open Notebook Science" (Bradley, 2007). Filosofin bakom detta initiativ bygger på tanken att vetenskapliga framsteg snabbare kan göras om resultat, utfallet av experiment eller kännedom om vem som arbetar med vilka frågor, kan distribueras mer eller mindre i realtid. Delvis samma tanke förekommer även inom mjukvaruutveckling där mjukvara licensierad med öppen källkod ofta görs tillgänglig via webben. För att möjliggöra detta har en hel infrastruktur kring versionshantering av filer och enkel distribuering av dessa vuxit fram inom IT-sektorn. Exempel på denna typ av infrastrukturer är versionshanterings-programmen Git (https://git-scm.com/) och SVN (https://subversion.apache.org/) samt distributionskanalerna GitHub (https://github.com/), GitLab (https://gitlab.com/) och BitBucket (https://bitbucket.org/). Dessa infrastrukturer lämpar sig även väldigt väl för data och arbetsflöden, samt andra typer av filer som genereras i miljö-DNA analyser och i vår undersökning har vi identifierat 17 av 67 undersökningar (25%) som använt Git och GitHub för detta. Att arkivera kommandofiler och resultat i realtid, eller efter att en undersökning slutförts, möjliggör därför större transparens och reproducerbarhet i ett projekt (Deiner *et al.*, 2017). Ytterligare en fördel med dessa distributionskanaler är att de tillhandahåller unika identifierare för de uppladdade filerna, vilket underlättar rapportering (och framtida nedladdning) av korrekt version av programvara och data. På liknande sätt kan unika, versionsspecifika identifierare erhållas i form av DOI:er (Digital Object Identifier; t.ex. när du använder Zenodo), eller versionshanterade accessionsnummer för referenssekvenser i databaser såsom de som NCBI tillhandahåller.

En förutsättning för att de förslag som presenteras ovan ska fungera är att öppna filformat används för lagring, samt att fri- eller öppen mjukvara (Free/Libre/Open source software, FLOSS) används för analyser, arbetsflöden, versionshantering etc. Detta gör att data och resultat inte blir låsta till ett specifikt mjukvaruprojekt som kan försvinna över tid, och möjliggör att infrastruktur för distribution av filer kan drivas vidare i egen regi om så skulle ske.

Ett reproducerbart tillvägagångssätt ökar inte bara tillförlitligheten hos en enskild analys utan underlättar även avsevärt möjligheten att utföra jämförande studier som inkluderar data från flera olika källor. En analysmetod som dokumenterats i en kommandofil, där mjukvaruversioner hanteras av Conda eller distribueras som en mjukvarucontainer kan ofta enkelt appliceras på nya data eller återanvändas när en nyare och mer komplett referensdatabas blir tillgänglig. Detta tillvägagångssätt ger flera fördelar genom att (1) fungera som automatisk dokumentation av alla mjukvara och inställningar som används; (2) enkelt möjliggöra analys av ny data, och 3) underlätta delning av metoder och resultat.

Focus för denna undersökning har varit reproducerbarhet och standardisering av bioinformatisk sekvensanalys då denna process är digital och möjlig att på ett enkelt sätt göras reproducerbar. Miljö-DNA projekt innehåller även fält- och labb-moment som även dessa behöver kvalitetssäkras och göras reproducerbara, något som är viktigt att ta med i planeringen för framtida forskningsprogram.

4.5 Framtida forskningsbehov

I den undersökta litteraturen presenteras ett antal förslag på forskningsinriktningar och teknologiska förbättringar som förväntas avancera tillämpningen av miljö-DNA analyser. Bland dessa finns även förslag på nya områden där metodiken visat stor potential men där mer forskning krävs. Då det undersökta textmaterialet lämpar sig väl för en sammanfattning av framtida forskningsbehov inkluderar vi här en sammanställning.

Ett ofta förekommande påpekande i många artiklar är bristen på *referenssekvenser för taxonomisk klassificering* av DNA-sekvenser identifierade i ett miljö-DNA prov (se t.ex. Taberlet *et al.*, 2012; Deagle *et al.*, 2018; Porter & Hajibabaei, 2018; Beng & Corlett, 2020). Tillgång till tillförlitligt referensmaterial kan vara avgörande för resultatet av en analys varför framtida forskningsutlysningar bör inkludera finansiering för DNA-sekvensering av referensmaterial från muséer, biobanker och fältinsamlingar. Som exempel på omfattningen av problemet kan nämnas att för en så väl studerad organismgrupp som kärlväxter (där https://artportalen.se/ rapporterar förekomst av 3783 arter i Sverige) finns idag en referenssekvens av den ofta använda markörgener matK endast för 606 arter (NCBI nt, https://www.ncbi.nlm.nih.gov/nucleotide/, 2022-03-09). Inte bara sekvenser från fler arter efterfrågas, utan även för *nya genetiska markörer*. Detta då nya markörer kan ha bättre taxonomisk upplösning för vissa organismgrupper, till exempel genom att vara längre och då dessutom bättre utnyttja teknologiska framsteg inom DNA-sekvensering (se nedan). Ett konkret förslag gällande genetisk markör är *sekvensering av hela mitokondriegenom* hos arter som ska undersökas med metabarkodning.

Utvecklingen av protokoll för provtagning och analys av miljö-DNA har kommit längst för akvatiska miljöer (Clare *et al.*, 2021). Potentialen har även visat sig vara stor för *andra miljötyper, såsom luft* (Karlsson *et al.*, 2020; Clare *et al.*, 2021; Lynggaard *et al.*, 2022) *och jord* (Carrasco-Puga *et al.*, 2021; Fernandez Nuñez *et al.*, 2021). Utveckling av metoder för dessa miljötyper kommer utöka antalet arter som kan övervakas med miljö-DNA och metabarkodning, varför ytterligare forskning i dessa miljöer behövs. Flera studier efterfrågar även *utveckling och förbättring av provtagningsstrategier*, dvs. ett behov av att identifiera hur många prover som behövs för ett representativt resultat, hur stort område ett miljö-DNA prov representerar, årstidsskillnader i DNA-förekomst etc. (Keck *et al.*, 2017; McGee *et al.*, 2019). Man bör även titta på hur miljö-DNA kan kombineras med traditionella metoder för biologisk övervakning (Beng & Corlett, 2020).

I nuläget kan miljö-DNA analys anses till största delen vara en kvalitativ analysmetod där förekomst av artspecifikt DNA i ett prov kan konfirmeras eller avfärdas (Elbrecht & Leese, 2015; Liu *et al.*, 2020). Möjligheten att även kunna omvandla DNA-signalen till ett *kvantitativt mått* av till exempel antal individer av en viss art på en viss plats har visat sig vara svårt (Taberlet *et al.*, 2018b), då olika stora individer eller arter släpper ifrån sig olika mycket DNA, avstånd mellan provpunkt och detekterad individ varierar etc. Vidare forskning är därför nödvändig för att även kunna kvantifiera antalet individer i en population med hjälp av miljö-DNA. En angränsande frågeställning är hur metoden kan användas för att skilja på molekyler från döda respektive levande individer. Detta gäller bland annat vid analys av sedimentprover, där både levande organismer och DNA från döda organismer ackumuleras (Pawlowski *et al.*, 2014). För denna typ av undersökning kan i stället RNA-molekyler användas, så kallad *eRNA*. RNA-molekyler skapas kontinuerligt i levande celler och är mindre

stabila i miljön än DNA vilket gör dem lämpliga för att påvisa biologisk aktivitet och förekomsten av levande celler (Barnes & Turner, 2016).

Ett steg som ofta inkluderas i ett metabarkodningsprojet är amplifiering av DNA-molekyler med PCR. Detta kan göra att mindre vanliga DNA-molekyler enklare kan detekteras under sekvenseringen men innebär även vissa risker, då amplifiering med PCR kan vara mer effektiv för vissa typer av molekyler och därmed ge en skev bild av variationen i ett prov. Metoden ökar även risken för sekvenseringsfel vilket kan resultera i artificiella sekvensvarianter, och en överskattning av den biologiska mångfalden i ett prov (Coissac *et al.*, 2012; Kennedy *et al.*, 2020). Flera författare förespråkar därför användningen av *PCR-fria metoder för metabarkodning* och att mer forskning och utveckling bör inrikta sig på detta (t.ex. Taberlet *et al.*, 2012; Kennedy *et al.*, 2020; Liu *et al.*, 2020).

Relaterat till detta är önskemål om fler metoder för att bedriva *populationsgenetiska studier med hjälp av miljö-DNA* (Tsuji *et al.*, 2020). Detta kan även kopplas till förslaget att utöka referensbibliotek med mitokondrie-genom sekvenser, en typ av data som lämpar sig för populationsgenetiska studier, och som nyligen tillämpats på undersökningar av fiskpopulationer (Weitemier *et al.*, 2021). Här finns det alltså möjlighet att adressera fler forskningsbehov genom att i framtida utlysningar finansiera sekvensering av fler mitokondriella genom.

Förslag presenteras även om att skapa *standardiserade jämförelse-material* av DNA-prover (så kallade mock communities) samt sekvensdata som kan användas för att verifiera labbmetoder samt bioinformatiska algoritmer som utvecklas eller används i olika labb (Elbrecht *et al.*, 2017; Martoni *et al.*, 2022). Önskemål har även uttryckts för införande av *standarder för laborativa- samt bioinformatiska moment* (Goodwin *et al.*, 2017; Rivera *et al.*, 2020). Dock uttrycker flera författare att försiktighet bör iakttas gällande krav på standardisering av protokoll, då detta kan ha en hämmande effekt på utveckling och innovation inom fältet (Piper *et al.*, 2019). Enkla förslag i linje med att standardisera miljö-DNA analys, utan att hämma utvecklingen inom fältet, är till exempel *standardiserad dokumentation samt metadata* som görs tillgänglig med sekvensdata (Tedersoo *et al.*, 2015; Goodwin *et al.*, 2017).

Den tekniska utvecklingen inom DNA-sekvensering har skapat ett behov av *nya bioinformatiska analysmetoder*. Detta inkluderar utveckling av maskininlärningsmetoder som visat sig framgångsrika vid ekologisk statusklassificering av miljöprover (Cordier *et al.*, 2017), och där bristande tillgång till referenssekvensdata kan hanteras med taxonomifria analysmetoder (Apothéloz-Perret-Gentil *et al.*, 2017). Metoderna har även föreslagits för undersökning av *ekologiska nätverk där biotiska interaktioner modelleras* med DNA-data (Vacher *et al.*, 2016). Exempel på interaktioner som kan undersökas med miljö-DNA är växter och deras pollinatörer (Zinger *et al.*, 2020), organismers mag-tarmflora samt näringsvävar (Kennedy *et al.*, 2020).

Teknisk utveckling har även gjort att laboratorieutrustning minskat i storlek, samt automatiserats till den grad att det nu är möjligt att utföra *DNA-sekvensering direkt på platsen för provtagningen*. Det finns dock behov för ytterligare forskning och utveckling inom detta fält (Piper *et al.*, 2019), där utvecklingen av automatiska provtagare, sensorer och mobila sekvenseringsplattformar har möjlighet att dramatiskt förändra förutsättningarna för biologisk övervakning (Krehenwinkel *et al.*, 2019). I ett pågående forskningsprogram från Norska Miljödirektoratet (avslutas i februari 2023) undersöks möjligheter med automatisk miljöövervakning och miljö-DNA, och en eventuell svensk utlysning på samma tema bör först utvärdera utfall och erfarenheter från denna.

Den sekvenseringsteknologi som kommit längst inom utvecklingen av mobila plattformar, och som redan tillämpas i fältundersökningar (Krehenwinkel *et al.*, 2019) är MinION från Oxford Nanopore. Denna plattform har möjlighet att sekvensera mycket långa DNA-fragment (så kallad *long-read sequencing*), vilket i sin tur kräver anpassat referensmaterial för att utnyttja teknologin till fullo. Denna typ av sekvenseringsteknologi förmodas få stor påverkan på miljö-DNA området (Goodwin *et al.*, 2019), och återkopplar tydligt till behovet av utveckling av nya och längre referenssekvenser för att uppnå den taxonomiska upplösning som krävs för effektiv övervakning av biologisk mångfald i ett nationellt övervakningsprogram.

5. Slutsatser och förslag

Vår tolkning av vad som framkommit i denna undersökning är att krav på att publicera data samt att inkludera datahanteringsplaner i forskningsansökningar har haft en positiv effekt, och varit gynnsam för reproducerbarheten av miljö-DNA analyser. Vi tror därför att liknande krav på att en "reproducerbarhetsplan" inkluderas i ansökningar till framtida forskningsprogram bör ha en lika positiv effekt, och att denna plan beskriver hur bioinformatiska analyser kommer dokumenteras, distribueras och göras reproducerbara. I tillägg, som komplement till de metodbeskrivningar som inkluderas i publikationer bör krav ställas på att analyskod och arbetsflöden, referensmaterial med unika identifierare eller verifierat med hashfunktioner görs tillgängliga tillsammans med analysresultatet.

Vi noterar även att det just nu pågår ett antal initiativ i andra nordiska länder som syftar till att undersöka, utveckla och implementera miljö-DNA i nationella övervakningsprogram. Som exempel kan nämnas Miljöministeriet i Finlands arbete för att rutinmässigt tillämpa molekylära metoder i övervakningsprogram senast 2025 (Norros *et al.*, 2022), samt Miljödirektoratet i Norge som undersöker potentialen för automatisk provtagning och analys av miljöprover. Även Danmark har varit tidiga med att utvärdera tillämpningar av miljö-DNA i nationella övervakningsprogram (Winding et al. 2019), och i utveckling och standardisering av protokoll för övervakning av specifika arter (Andersen et al. 2018). Lärdomar från dessa program, samt pågående svenska initiativ för att implementera DNA-metoder inom miljöövervakning (https://www.naturvardsverket.se/om-miljoarbetet/forskning/miljoforskning/forskningssatsningar-natur/dna-metoder-inom-miljoovervakning/) kommer kunna hjälpa utformningen av nya forskningsprogram. Vi vill även lyfta fram den svenska infrastruktur (idag fokuserad på forskning snarare än miljöövervakning) som redan finns i drift och som potentiellt kommer kunna vara viktig för utformningen av ett nationellt övervakningsprogram som inkluderar miljö-DNA analys.

5.1 SBDI

Swedish Biodiversity Data Infrastructure (SBDI, https://biodiversitydata.se/) arbetar bland annat med att göra biodiversitetsdata, inklusive miljö-DNA data, tillgänglig för forskare, myndigheter och allmänheten. I detta arbete ingår även standardisering av format för metadata för metabarkodnings-sekvenser, en ofta förekommande datatyp för miljö-DNA, och vars kvalitet kommer vara av största vikt i framtida övervakningsprogram. Organisationen utvecklar även analysverktyg med stark inriktning mot reproducerbarhet. Dessa verktyg använder bland annat arbetsflödesmjukvara, containerteknologi och versionshantering som identifierats som några av de tekniska lösningar som förenklar bioinformatisk reproducerbarhet (https://nf-co.re/ampliseq). SBDI erbjuder också möjligheten att dela biologisk mångfalds-data då de utgör den svenska noden i *the Global Biodiversity Information Facility* (GBIF). Förutom de data- och metadatastandarder som GBIF tillämpar ger organisationen även data en unik Digital Object Identifier (DOI, se rekommendationer under *Versionshantering och onlinepublicering*). Molekylära data kan både delas och sökas via den svenska ASV-portalen som SBDI driver (https://asv-portal.biodiversitydata.se/). Portalen har potentiellt många fördelar för ett nationellt övervakningsprogram, bland annat genom att

sekvenser som laddas upp klassificeras på nytt när referensdatabaser uppdateras, vilket säkerställer kvalitet och jämförbarhet mellan olika projekt som samlar in miljö-DNA-data. Projektet lagrar dock inte ursprungssekvenserna, utan de lagras i internationella sekvensdatabaser (till exempel The European Nucleotide Archive, https://www.ebi.ac.uk/ena/browser/home).

5.2 NBIS

National Bioinformatics Infrastructure Sweden (NBIS, https://nbis.se/) är en nationell forskningsinfrastruktur under Science for Life Laboratory (SciLifeLab, https://www.scilifelab.se/). Organisationen erbjuder bioinformatiskt stöd till svenska forskare, både i form av engagemang i enskilda projekt, och med utveckling av bioinformatiska verktyg samt genom utbildning. Av speciellt intresse för denna rapport är de kurser som specifikt handlar om datahantering och reproducerbarhet (https://nbis.se/training/events.html). NBIS kompetens kring datahantering bör även inkluderas i framtida övervakningsprogram för att underlätta och möjliggöra tillgänglighet, kvalitet och relevans hos den data som produceras.

5.3 Förslag till framtida forskningsinriktningar

De fjorton kategorier av efterfrågade forskningsinriktningar som identifierats i denna undersökning är alla mer eller mindre sammanlänkade, och i vissa fall till och med helt beroende av varandra för att vara framgångsrika. Som exempel kan nämnas att för att kunna tillgodogöra sig fördelarna med längre DNA-sekvensläsningar behövs även längre referenssekvenser tas fram. Vidare ser vi att de föreslagna forskningsinriktningarna faller inom olika övergripande kategorier:

Biologiska tillämpningar

- Utveckling av metoder för kvantitativ miljö-DNA-analys (efterfrågas i 22 publikationer)
- Undersökning av biologiska interaktioner och ekologiska nätverk (8)
- Analys av miljö-DNA från olika miljötyper (7)
- Populationsgenetiska studier med miljö-DNA (6)
- Analys av miljö-RNA, eRNA (6)

Teknisk utveckling

- Utveckling av nya bioinformatiska verktyg, metoder och modeller (17)
- Long-read sekvensering (8)
- Utveckling av metoder och protokoll för miljö-DNA analys i fält (8)
- Utveckling av PCR-fria metoder för DNA-sekvensering (7)

Förbättring av nuvarande metoder

- Utveckling och förbättring av provtagningsstrategier (21)
- Standardisering av protokoll och referensmaterial (21)
- Utveckling av dokumentation, metadata och reproducerbarhet (10)

Förbättrade referenssekvenser

- Utökade referensdatabaser för taxonomisk klassificering av metabarkodnings-sekvenser (37)
- Sekvensering av hela mitokondriegenom (5)

Ett framtida forskningsprogram kan adressera flera av dess förslag genom att fokusera på någon av de föreslagna biologiska tillämpningarna (till exempel utveckling av kvantitativa metoder) eller teknisk utveckling inom miljö-DNA området (till exempel analys av miljö-DNA med artificiell intelligens) och samtidigt kräva att finansierade projekt bidrar till förbättring av nuvarande metoder och referensmaterial.

6. Tack

Vi vill tacka Niclas Engene och Erland Lettevall från Havs- och vattenmyndigheten, samt Johan Wulff och Ola Inghe från Naturvårdsverket som i diskussion med oss hjälpt till att utforma inriktningen för detta projekt. Vi vill även rikta ett tack till Maria Prager från Swedish Biodiversity Infrastructure (SBDI) och SciLifeLab, samt Niklas Jareborg från National Bioinformatic Infrastructure Sweden (NBIS) som diskuterat bioinformatiska lösningar och infrastruktur tillgängliga för det svenska forskarsamhället. Tack även till Kersti Karltorp för hjälp med projektets utformning och metod, samt Tomas Larsson (NBIS) och Mikael Dahl (IVL) som kritiskt granskat texten med fokus på analysens metod, identifierade bioinformatiska lösningar samt slutsatser.

7. Källförteckning

Adamowicz SJ, Hollingsworth PM, Ratnasingham S, Van Der Bank M, Cristescu ME (2017) *International Barcode of Life: Focus on big biodiversity in South Africa*. Genome 60:875–879. https://doi.org/10.1139/gen-2017-0210

Andersen JH, Kallenbach E, Thaulow J, Hesselsøe M, Bekkevold D, Hansen BK, Jacobsen LMW, Olesen CA, Møller PR, Knudsen SW (2018) *Development of species-specific eDNA-based test systems for monitoring of non-indigenous species in Danish marine waters*. NIVA Denmark Report. 77 pp. http://hdl.handle.net/11250/2573117

Apothéloz-Perret-Gentil L, Cordonier A, Straub F, Iseli J, Esling P, Pawlowski J (2017) *Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring*. Mol Ecol Resour 17:1231–1242. https://doi.org/10.1111/1755-0998.12668

Atherton S, Jondelius U (2020) *Biodiversity between sand grains: Meiofauna composition across southern and western Sweden assessed by metabarcoding*. Biodivers Data J 8:1–40. https://doi.org/10.3897/BDJ.8.E51813

Bailet B, Apothéloz-Perret-Gentil L, Baričević A, Chonova T, Franc A, Frigerio J-M, Kelly M, Mora D, Pfannkuchen M, Proft S, Ramon M, Vasselon V, Zimmermann J, Kahlert M (2020) *Diatom DNA metabarcoding for ecological assessment: Comparison among bioinformatics pipelines used in six European countries reveals the need for standardization*. Sci Total Environ 745:140948. https://doi.org/10.1016/j.scitotenv.2020.140948

Baker M (2016) *1,500 scientists lift the lid on reproducibility*. Nature 533:452–454. https://doi.org/10.1038/533452a

Barnes MA, Turner CR (2016) *The ecology of environmental DNA and implications for conservation genetics*. Conserv Genet 17:1–17. https://doi.org/10.1007/s10592-015-0775-4

Beng KC, Corlett RT (2020) *Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges and prospects*. Biodivers Conserv 29:2089–2121. https://doi.org/10.1007/s10531-020-01980-0

Bista I, Carvalho GR, Walsh K, Seymour M, Hajibabaei M, Lallias D, Christmas M, Creer S (2017) *Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity*. Nat Commun 8:14087. https://doi.org/10.1038/ncomms14087

Bowers HA, Pochon X, von Ammon U, Gemmell N, Stanton JAL, Jeunen GJ, Sherman CDH, Zaiko A (2021) *Towards the optimization of eDNA/eRNA sampling technologies for marine biosecurity surveillance*. Water 13:1113. https://doi.org/10.3390/w13081113

Bradley JC (2007) *Open notebook science using blogs and wikis*. Nat Prec. https://doi.org/10.1038/npre.2007.39.1

Carrasco-Puga G, Díaz FP, Soto DC, Hernández-Castro C, Contreras-López O, Maldonado A, Latorre C, Gutiérrez RA (2021) *Revealing hidden plant diversity in arid environments*. Ecography 44:98–111. https://doi.org/10.1111/ecog.05100

Cashman M, Cohen MB, Ranjan P, Cottingham RW (2018) *Navigating the Maze: The Impact of Configurability in Bioinformatics Software*. In: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering. Association for Computing Machinery, New York, NY, USA, pp 757–767. https://doi.org/10.1145/3238147.3240466

Chang JJM, Ip YCA, Bauman AG, Huang D (2020) *MinION-in-ARMS: Nanopore Sequencing to Expedite Barcoding of Specimen-Rich Macrofaunal Samples From Autonomous Reef Monitoring Structures*. Front Mar Sci 7:448. https://doi.org/10.3389/fmars.2020.00448

Clare EL, Economou CK, Faulkes CG, Gilbert JD, Bennett F, Drinkwater R, Littlefair JE (2021) *eDNAir: Proof of concept that animal DNA can be collected from air sampling*. PeerJ 9:e11030. https://doi.org/10.7717/peerj.11030

Coissac E, Riaz T, Puillandre N (2012) *Bioinformatic challenges for DNA metabarcoding of plants and animals*. Mol Ecol 21:1834–1847. https://doi.org/10.1111/j.1365-294X.2012.05550.x

Cordier T, Esling P, Lejzerowicz F, Visco J, Ouadahi A, Martins C, Cedhagen T, Pawlowski J (2017) *Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning*. Environ Sci Technol 51:9118–9126 https://doi.org/10.1021/acs.est.7b01518

Creedy TJ, Andújar C, Meramveliotakis E, Noguerales V, Overcast I, Papadopoulou A, Morlon H, Vogler AP, Emerson BC, Arribas P (2022) *Coming of age for COI metabarcoding of whole organism community DNA: Towards bioinformatic harmonisation*. Mol Ecol Resour 22:847–861. https://doi.org/10.1111/1755-0998.13502

Curd EE, Gold Z, Kandlikar GS, Gomer J, Ogden M, O'Connell T, Pipes L, Schweizer TM, Rabichow L, Lin M, Shi B, Barber PH, Kraft N, Wayne R, Meyer RS (2019) *Anacapa Toolkit: An environmental DNA toolkit for processing multilocus metabarcode datasets*. Methods Ecol Evol 10:1469–1475. https://doi.org/10.1111/2041-210X.13214

Curty RG, Lee JS, Chang W, Kao TH, Jeng W (2022) *Practicing What is Preached: Exploring Reproducibility Compliance of Papers on Reproducible Research*. In: Smits M (ed) Information for a Better World: Shaping the Global Future. iConference 2022. Lecture Notes in Computer Science, vol. 13192. Springer International Publishing, Cham, pp 255–264. https://doi.org/10.1007/978-3-030-96957-8_23

Deagle BE, Clarke LJ, Kitchener JA, Polanowski AM, Davidson AT (2018) *Genetic monitoring of open ocean biodiversity: An evaluation of DNA metabarcoding for processing continuous plankton recorder samples*. Mol Ecol Resour 18:391–406. https://doi.org/10.1111/1755-0998.12740

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L (2017) *Environmental DNA metabarcoding: Transforming how we survey animal and plant communities*. Mol Ecol 26:5872–5895. https://doi.org/10.1111/mec.14350

Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C (2017) *Nextflow enables reproducible computational workflows*. Nat Biotechnol 35:316-319. https://doi.org/10.1038/nbt.3820

Dufresne Y, Lejzerowicz F, Perret-Gentil LA, Pawlowski J, Cordier T (2019) *SLIM: A flexible web application for the reproducible processing of environmental DNA metabarcoding data*. BMC Bioinformatics 20:1–6. https://doi.org/10.1186/s12859-019-2663-2

Dully V, Balliet H, Frühe L, Däumer M, Thielen A, Gallie S, Berrill I, Stoeck T (2021) *Robustness, sensitivity and reproducibility of eDNA metabarcoding as an environmental biomonitoring tool in coastal salmon aquaculture – An inter-laboratory study*. Ecol Indic 121:107049. https://doi.org/10.1016/j.ecolind.2020.107049

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:e0130324. https://doi.org/10.1371/journal.pone.0130324

Elbrecht V, Vamos EE, Meissner K, Aroviita J, Leese F (2017) *Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring*. Methods Ecol Evol 8:1265–1275. https://doi.org/10.1111/2041-210X.12789

Farrell JA, Whitmore L, Mashkour N, Rollinson Ramia DR, Thomas RS, Eastman CB, Burkhalter B, Yetsko K, Mott C, Wood L, Zirkelbach B, Meers L, Kleinsasser P, Stock S, Libert E, Herren R, Eastman S, Crowder W, Bovery C, Anderson D, Godfrey D, Condron N, Duffy DJ (2022) *Detection and population genomics of sea turtle species via noninvasive environmental DNA analysis of nesting beach sand tracks and oceanic water*. Mol Ecol Resour 22:2471–2493. https://doi.org/10.1111/1755-0998.13617

Fernandez Nuñez N, Maggia L, Stenger PL, Lelievre M, Letellier K, Gigante S, Manez A, Mournet P, Ripoll J, Carriconde F (2021) *Potential of high-throughput eDNA sequencing of soil fungi and bacteria for monitoring ecological restoration in ultramafic substrates: The case study of the New Caledonian biodiversity hotspot*. Ecol Eng 173:106416. https://doi.org/10.1016/j.ecoleng.2021.106416

Gardner PP, Paterson JM, McGimpsey S, Ashari-Ghomi F, Umu SU, Pawlik A, Gavryushkin A, Black MA (2022) *Sustained software development, not number of citations or journal choice, is indicative of accurate bioinformatic software*. Genome Biol 23:56. https://doi.org/10.1186/s13059-022-02625-x

Garijo D, Kinnings S, Xie Li, Xie Lei, Zhang Y, Bourne PE, Gil Y (2013) *Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome*. PLoS One 8:e80278. https://doi.org/10.1371/journal.pone.0080278

Goldberg CS, Turner CR, Deiner K, Klymus KE, Thomsen PF, Murphy MA, Spear SF, McKee A, Oyler-McCance SJ, Cornman RS, Laramie MB, Mahon AR, Lance RF, Pilliod DS, Strickler KM, Waits LP, Fremier AK, Takahara T, Herder JE, Taberlet P (2016) *Critical considerations for the application of environmental DNA methods to detect aquatic species*. Methods Ecol Evol 7:1299–1307. https://doi.org/10.1111/2041-210X.12595

Goodwin KD, Thompson LR, Duarte B, Kahlke T, Thompson AR, Marques JC, Caçador I (2017) *DNA sequencing as a tool to monitor marine ecological status*. Front Mar Sci 4:1–14. https://doi.org/10.3389/fmars.2017.00107

Goodwin KD, Muller-Karger FE, Djurhuus A, Allen LZ, Allen AE, McCrow JP, Canonico Hyde G (2019) *Molecular approaches for an operational marine biodiversity observation network*. In: Sheppard C, editor. World Seas: An Environmental Evaluation Volume III: Ecological Issues and Environmental Impacts. Second edition. Elsevier Ltd. p. 613–631. https://doi.org/10.1016/B978-0-12-805052-1.00032-2

Gough D, Oliver S, Thomas J (editors) (2017) *An introduction to systematic reviews (second edition)*. SAGE Publications, Los Angeles.

Griffiths BS, de Groot GA, Laros I, Stone D, Geisen S (2018) *The need for standardisation: Exemplified by a description of the diversity, community structure and ecological indices of soil nematodes*. Ecol Indic 87:43–46. https://doi.org/10.1016/j.ecolind.2017.12.002

Harper LR, Lawson Handley L, Hahn C, Boonham N, Rees HC, Gough KC, Lewis E, Adams IP, Brotherton P, Phillips S, Hänfling B (2018) *Needle in a haystack? A comparison of eDNA metabarcoding and targeted qPCR for detection of the great crested newt (Triturus cristatus)*. Ecol Evol 8:6330–6341. https://doi.org/10.1002/ece3.4013

Hänfling B, Handley LL, Read D, Winfield I (2016) *eDNA-based metabarcoding as a monitoring tool for fish in large lakes.* Report - SC140018/R. Environment Agency, Horizon House, Deanery Road, Bristol, BS1 5AH. ISBN: 978-1-84911-386-1. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/575833/A_DNA_based_monitoring_method_for_fish_in_lakes_-_report.pdf

Karlsson E, Johansson AM, Ahlinder J, Lundkvist MJ, Singh NJ, Brodin T, Forsman M, Stenberg P (2020) *Airborne microbial biodiversity and seasonality in Northern and Southern Sweden*. PeerJ 8:e8424. https://doi.org/10.7717/peerj.8424

Keck F, Vasselon V, Tapolczai K, Rimet F, Bouchez A (2017) *Freshwater biomonitoring in the Information Age*. Front Ecol Environ 15:266–274. https://doi.org/10.1002/fee.1490

Kennedy SR, Prost S, Overcast I, Rominger AJ, Gillespie RG, Krehenwinkel H (2020) *High-throughput sequencing for community analysis: the promise of DNA barcoding to uncover diversity, relatedness, abundances and interactions in spider communities*. Dev Genes Evol 230:185–201. https://doi.org/10.1007/s00427-020-00652-x

Krehenwinkel H, Pomerantz A, Prost S (2019) *Genetic biomonitoring and biodiversity assessment using portable sequencing technologies: Current uses and future directions*. Genes 10:858. https://doi.org/10.3390/genes10110858

Kurtzer GM, Sochat V, Bauer MW (2017) *Singularity: Scientific containers for mobility of compute*. PLoS One 12:e0177459. https://doi.org/10.1371/journal.pone.0177459

Li J, Hatton-Ellis TW, Lawson Handley L-J, Kimbell HS, Benucci M, Peirson G, Hänfling B (2019) *Ground-truthing of a fish-based environmental DNA metabarcoding method for assessing the quality of lakes*. J Appl Ecol 56:1232–1244. https://doi.org/10.1111/1365-2664.13352

Liu M, Clarke LJ, Baker SC, Jordan GJ, Burridge CP (2020) *A practical guide to DNA metabarcoding for entomological ecologists*. Ecol Entomol 45:373–385. https://doi.org/10.1111/een.12831

Lynggaard C, Bertelsen MF, Jensen C V., Johnson MS, Frøslev TG, Olsen MT, Bohmann K (2022) *Airborne environmental DNA for terrestrial vertebrate community monitoring*. Curr Biol 32:701-707. https://doi.org/10.1016/j.cub.2021.12.014

Martoni F, Piper AM, Rodoni BC, Blacket MJ (2022) *Disentangling bias for non-destructive insect metabarcoding*. PeerJ 10:e12981. https://doi.org/10.7717/peerj.12981

McGee KM, Robinson C V., Hajibabaei M (2019) *Gaps in DNA-Based Biomonitoring Across the Globe*. Front Ecol Evol 7:1–7. https://doi.org/10.3389/fevo.2019.00337

Meissner K, Aroviita J, Baattrup-Pedersen A, Buchner D, Ekrem T, Friberg N, Johnson R, Leese FK, Majaneva M, Ólafsson JS, Schartau AK, Elbrecht V (2020) *Metabarcoding for use in Nordic routine aquatic biomonitoring: a validation study* Nordic Council of Ministers, Nordic Council of Ministers Secretariat, Nordisk Arbejdsgruppe for Biologisk Mangfoldighed (NBM). Copenhagen: Nordisk Ministerråd, 2020, 538. , p. 101. ISBN: 978-92-893-6806-3, 978-92-893-6807-0. https://www.norden.org/en/publication/metabarcoding-use-nordic-routine-aquatic-biomonitoring

Merkel D (2014) *Docker: lightweight Linux containers for consistent development and deployment*. Linux J 239:2. https://dl.acm.org/doi/10.5555/2600239.2600241

Mousavi-Derazmahalleh M, Stott A, Lines R, Peverley G, Nester G, Simpson T, Zawierta M, De La Pierre M, Bunce M, Christophersen CT (2021) *eDNAFlow, an automated, reproducible and scalable workflow for analysis of environmental DNA sequences exploiting Nextflow and Singularity*. Mol Ecol Resour 21:1697–1704. https://doi.org/10.1111/1755-0998.13356

Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J (2021) *Sustainable data analysis with Snakemake [version 2; peer review: 2 approved]*. F1000Research 10:33. https://doi.org/10.12688/f1000research.29032.2

Nicholson A, McIsaac D, MacDonald C, Gec P, Mason BE, Rein W, Wrobel J, de Boer M, Milián-García Y, Hanner RH (2020) *An analysis of metadata reporting in freshwater environmental DNA research calls for the development of best practice guidelines*. Environ DNA 2:343–349. https://doi.org/10.1002/edn3.81

Norros V, Laamanen T, Meissner K, Iso-Touru T, Kahilainen A, Lehtinen S, Lohtander-Buckbee K, Nygård H, Pennanen T, Ruohonen-Lehto M, Sirkiä P, Suikkanen S, Tolkkinen M, Vainio E, Velmala S, Vuorio K, Vihervaara P (2022) *Roadmap for implementing environmental DNA (eDNA) and other molecular monitoring methods in Finland - Vision and action plan for 2022–2025*. Reports of the Finnish Environment Institute, Biodiversity centre. ISBN 978-952-11-5482-9. https://helda.helsinki.fi/handle/10138/342992

Ogram A, Sayler GS, Barkay T (1987) *The extraction and purification of microbial DNA from sediments*. J Microbiol Methods 7:57–66. https://doi.org/10.1016/0167-7012(87)90025-X

Pawlowski J, Lejzerowicz F, Esling P (2014) *Next-generation environmental diversity surveys of foraminifera: Preparing the future*. Biol Bull 227:93–106. https://doi.org/10.1086/BBLv227n2p93

Piccolo SR, Frampton MB (2016) *Tools and techniques for computational reproducibility*. Gigascience 5:s13742-016-0135–4. https://doi.org/10.1186/s13742-016-0135-4

Piper AM, Batovska J, Cogan NOI, Weiss J, Cunningham JP, Rodoni BC, Blacket MJ (2019) *Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance*. Gigascience 8:1–22. https://doi.org/10.1093/gigascience/giz092

Pompanon F, Samadi S (2015) *Next generation sequencing for characterizing biodiversity: promises and challenges*. Genetica 143:133–138. https://doi.org/10.1007/s10709-015-9816-7

Porter TM, Hajibabaei M (2018) *Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis*. Mol Ecol 27:313–338. https://doi.org/10.1111/mec.14478

Rees HC, Maddison BC, Middleditch DJ, Patmore JRM, Gough KC (2014) *The detection of aquatic animal species using environmental DNA – a review of eDNA as a survey tool in ecology*. J Appl Ecol 51:1450–1459. https://doi.org/10.1111/1365-2664.12306

Ríos-Castro R, Romero A, Aranguren R, Pallavicini A, Banchi E, Novoa B, Figueras A (2021) *High-Throughput Sequencing of Environmental DNA as a Tool for Monitoring Eukaryotic Communities and Potential Pathogens in a Coastal Upwelling Ecosystem*. Front Vet Sci 8:1–18. https://doi.org/10.3389/fvets.2021.765606

Ritter CD, Dunthorn M, Anslan S, de Lima VX, Tedersoo L, Nilsson RH, Antonelli A (2020) *Advancing biodiversity assessments with environmental DNA: Long-read technologies help reveal the drivers of Amazonian fungal diversity*. Ecol Evol 10:7509–7524. https://doi.org/10.1002/ece3.6477

Rivera SF, Vasselon V, Bouchez A, Rimet F (2020) *Diatom metabarcoding applied to large scale monitoring networks: Optimization of bioinformatics strategies using Mothur software*. Ecol Indic 109:105775. https://doi.org/10.1016/j.ecolind.2019.105775

Samuel S, König-Ries B (2021) *Understanding experiments and research practices for reproducibility: an exploratory study*. PeerJ 9:e11140. https://doi.org/10.7717/peerj.11140

Santoferrara L, Burki F, Filker S, Logares R, Dunthorn M, McManus GB (2020) *Perspectives from Ten Years of Protist Studies by High-Throughput Metabarcoding*. J Eukaryot Microbiol 67:612–622. https://doi.org/10.1111/jeu.12813

Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) *Next-generation sequencing technologies for environmental DNA research*. Mol Ecol 21:1794–1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x

Sigsgaard EE, Nielsen IB, Bach SS, Lorenzen ED, Robinson DP, Knudsen SW, Pedersen MW, Jaidah M Al, Orlando L, Willerslev E, Møller PR, Thomsen PF (2016) *Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA*. Nat Ecol Evol 1:4. https://doi.org/10.1038/s41559-016-0004

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) *Towards next-generation biodiversity assessment using DNA metabarcoding*. Mol Ecol 21:2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Taberlet P, Bonin A, Zinger L, Coissac E (2018a) *DNA metabarcoding data analysis*. In: Taberlet P, Bonin A, Zinger L, Coissac E, editors. Environmental DNA: For Biodiversity Research and Monitoring. Oxford University Press. p. 65–84. https://doi.org/10.1093/oso/9780198767220.003.0008

Taberlet P, Bonin A, Zinger L, Coissac E (2018b) *The future of eDNA metabarcoding*. In: Taberlet P, Bonin A, Zinger L, Coissac E, editors. Environmental DNA: For Biodiversity Research and Monitoring. Oxford University Press. p. 144–150. https://doi.org/10.1093/oso/9780198767220.003.0019

Tapolczai K, Keck F, Bouchez A, Rimet F, Kahlert M, Vasselon V (2019) *Diatom DNA Metabarcoding for Biomonitoring: Strategies to Avoid Major Taxonomical and Bioinformatical Biases Limiting Molecular Indices Capacities.* Front Ecol Evol 7:409. https://doi.org/10.3389/fevo.2019.00409

Tedersoo L, Ramirez KS, Nilsson RH, Kaljuvee A, Kõljalg U, Abarenkov K (2015) *Standardizing metadata and taxonomic identification in metabarcoding studies*. Gigascience 4:1–4. https://doi.org/10.1186/s13742-015-0074-5

Tsuji S, Maruyama A, Miya M, Ushio M, Sato H, Minamoto T, Yamanaka H (2020) *Environmental DNA analysis shows high potential as a tool for estimating intraspecific genetic diversity in a wild fish population*. Mol Ecol Resour 20:1248–1258. https://doi.org/10.1111/1755-0998.13165

Tyler AD, Mataseje L, Urfano CJ, Schmidt L, Antonation KS, Mulvey MR, Corbett CR (2018) *Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications*. Sci Rep 8:10931. https://doi.org/10.1038/s41598-018-29334-5

Vacher C, Tamaddoni-Nezhad A, Kamenova S, Peyrard N, Moalic Y, Sabbadin R, Schwaller L, Chiquet J, Smith MA, Vallance J, Fievet V, Jakuschkin B, Bohan DA (2016) *Learning Ecological Networks from Next-Generation Sequencing Data*. In: Woodward G, Bohan DAB, editors. Ecosystem Services: From Biodiversity to Society, Part 2. Vol. 54. Academic Press. p. 1–39. https://doi.org/10.1016/bs.aecr.2015.10.004

Weitemier K, Penaluna BE, Hauck LL, Longway LJ, Garcia T, Cronn R (2021) *Estimating the genetic diversity of Pacific salmon and trout using multigene eDNA metabarcoding*. Mol Ecol 30:4970–4990. https://doi.org/10.1111/mec.15811

Welzel M, Lange A, Heider D, Schwarz M, Freisleben B, Jensen M, Boenigk J, Beisser D (2020) *Natrix: a Snakemake-based workflow for processing, clustering, and taxonomically assigning amplicon sequencing reads*. BMC Bioinformatics 21:1–14. https://doi.org/10.1186/s12859-020-03852-4

Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, 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*. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18

Winding A, Bang-Andreasen T, Hansen LH, Panitz F, Krogh PH, Krause-Jensen D, Stæhr P, Nicolaisen M, Hendriksen NB, Sapkota R, Santos S, Andersen LW (2019) *eDNA in environmental monitoring.* Aarhus University, DCE – Danish Centre for

Environment and Energy, 40 pp. Technical Report No. 133. http://dce2.au.dk/pub/TR133.pdf

Yang J, Zhang X, Zhang W, Sun J, Xie Y, Zhang Y, Burton Jr. GA, Yu H (2017) *Indigenous species barcode database improves the identification of zooplankton*. PLoS One 12:e0185697. https://doi.org/10.1371/journal.pone.0185697

Zaiko A, Greenfield P, Abbott C, von Ammon U, Bilewitch J, Bunce M, Cristescu ME, Chariton A, Dowle E, Geller J, Ardura Gutierrez A, Hajibabaei M, Haggard E, Inglis GJ, Lavery SD, Samuiloviene A, Simpson T, Stat M, Stephenson S, Sutherland J, Thakur V, Westfall K, Wood SA, Wright M, Zhang G, Pochon X (2022) *Towards reproducible metabarcoding data: Lessons from an international cross-laboratory experiment*. Mol Ecol Resour 22:519–538. https://doi.org/10.1111/1755-0998.13485

Zinger L, Donald J, Brosse S, Gonzalez MA, Iribar A, Leroy C, Murienne J, Orivel J, Schimann H, Taberlet P, Lopes CM (2020) *Advances and prospects of environmental DNA in neotropical rainforests*. Adv Ecol Res 62:331–373. https://doi.org/10.1016/bs.aecr.2020.01.001

8. Publikationer och data

Data, analyskod och resultat samt källkod för figurer inkluderade i denna rapport finns tillgänglig via https://github.com/topel-research-group/Reproducible-analysis-of-eDNA-for-national-biodiversity-monitoring-programs.

Resultatet från denna undersökning har även presenterats under Gothenburg Global Biodiversity Center (GGBC) föreläsningsserie, samt vid konferensen *Incorporating an evolutionary approach in conservation management* arrangerad av Centrum för Marin Evolutionärsbiologi (CeMEB) i oktober 2022.

Bilaga 1. Söksträngar

Respektive söksträng som använts för de olika databaserna.

DIVA

("eDNA" OR "environmental DNA" OR "metabarcoding" OR "eRNA" OR "environmental RNA") AND ("biodiversity" OR "species richness" OR "monitoring" OR "biomonitoring")

BASE

dctitle:eDNA+monitoring+dctype:+1

Libris, SwePub och PubMed

("eDNA" OR "environmental DNA" OR "metabarcoding" OR "eRNA" OR "environmental RNA") AND ("biodiversity" OR "species richness" OR "monitoring" OR "biomonitoring") AND ("high throughput sequencing" OR "HTS" OR "throughput")

Scopus

TITLE-ABS-KEY (("eDNA" OR "environmental DNA" OR "metabarcoding" OR "eRNA" OR "environmental RNA") AND ("biodiversity" OR "species richness" OR "monitoring" OR "biomonitoring") AND ("high throughput sequencing" OR "HTS" OR "throughput"))

Bilaga 2. Jämförelselista

Publikationer som innan litteratursökningen identifierats som relevanta för denna analys. Sökströngarna i Bilaga 1 justerades så att samtliga titlar i denna lista inkluderades i det slutliga sökresultatet.

Rapporter

Hänfling B, Handley LL, Read D, Winfield I (2016) *eDNA-based metabarcoding as a monitoring tool for fish in large lakes.* Report - SC140018/R. Environment Agency, Horizon House, Deanery Road, Bristol, BS1 5AH. ISBN: 978-1-84911-386-1. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/575833/A_DNA_based_monitoring_method_for_fish_in_lakes_-_report.pdf
(Rapport från Brittiska Miljömyndigheten)

Meissner K, Aroviita J, Baattrup-Pedersen A, Buchner D, Ekrem T, Friberg N, Johnson R, Leese FK, Majaneva M, Ólafsson JS, Schartau AK, Elbrecht V (2020) *Metabarcoding for use in Nordic routine aquatic biomonitoring: a validation study* . Nordic Council of Ministers, Nordic Council of Ministers Secretariat, Nordisk Arbejdsgruppe for Biologisk Mangfoldighed (NBM). Copenhagen: Nordisk Ministerråd, 2020, 538, p. 101. ISBN: 978-92-893-6806-3, 978-92-893-6807-0. https://www.norden.org/en/publication/metabarcoding-use-nordic-routine-aquatic-biomonitoring

(Rapport från Nordiska ministerrådet)

Norros V, Laamanen T, Meissner K, Iso-Touru T, Kahilainen A, Lehtinen S, Lohtander-Buckbee K, Nygård H, Pennanen T, Ruohonen-Lehto M, Sirkiä P, Suikkanen S, Tolkkinen M, Vainio E, Velmala S, Vuorio K, Vihervaara P (2022) *Roadmap for implementing environmental DNA (eDNA) and other molecular monitoring methods in Finland - Vision and action plan for 2022–2025*. Reports of the Finnish Environment Institute, Biodiversity centre. ISBN 978-952-11-5482-9. https://helda.helsinki.fi/handle/10138/342992

(Rapport från Finska Miljöinstitutet)

Winding A, Bang-Andreasen T, Hansen LH, Panitz F, Krogh PH, Krause-Jensen D, Stæhr P, Nicolaisen M, Hendriksen NB, Sapkota R, Santos S, Andersen LW (2019) *eDNA in environmental monitoring.* Aarhus University, DCE – Danish Centre for Environment and Energy, 40 pp. Technical Report No. 133. http://dce2.au.dk/pub/TR133.pdf

(Rapport från Danska Nationalt Center för Miljö och Energi)

Vetenskaplig litteratur

Bista I, Carvalho GR, Walsh K, Seymour M, Hajibabaei M, Lallias D, Christmas M, Creer S (2017) *Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity*. Nat Commun 8:14087. https://doi.org/10.1038/ncomms14087

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L (2017) *Environmental DNA metabarcoding: Transforming how we survey animal and plant communities*. Mol Ecol 26:5872–5895. https://doi.org/10.1111/mec.14350

Tapolczai K, Keck F, Bouchez A, Rimet F, Kahlert M, Vasselon V (2019) *Diatom DNA Metabarcoding for Biomonitoring: Strategies to Avoid Major Taxonomical and Bioinformatical Biases Limiting Molecular Indices Capacities*. Front Ecol Evol 7:409. https://doi.org/10.3389/fevo.2019.00409

RAPPORT 7084 | DECEMBER 2022 ISBN 978-91-620-7084-7 ISSN 0282-7298

Reproducerbar analys av miljö-DNA i nationella övervakningsprogram

En kritisk granskning

Rapporten analyserar problem och lösningar kring reproducerbarhet vid analys av miljö-DNA. Reproducerbarhet gör det möjligt att analysera data från olika källor, ursprung och kvalitet vilket är särskilt viktigt vid jämförelse av resultat från långa tidsserier, likt de som produceras inom nationella övervakningsprogram.

Nuvarande metoder för identifiering och övervakning av biologisk mångfald inom miljöövervakning är tidskrävande och dyra. Resultat som produceras av enskilda specialister kan vara svåra att återskapa, särskilt när resultatet bygger på observationer i fält. Analys av miljö-DNA skulle dock kunna vara ett bra komplement eller ersättning för traditionella inventeringsmetoder.

Övervakning av biologisk mångfald med hjälp av miljö-DNA kräver fortfarande utveckling, och i rapporten presenteras därför förslag på framtida forskning som kan främja användningen av miljö-DNA i nationella miljöövervakningsprogram.

Projektet har finansierats med medel från Naturvårdsverkets miljöforskningsanslag som finansierar forskning till stöd för Naturvårdsverkets och Havs- och vattenmyndighetens kunskapsbehov.

NATUR VÅRDS VERKET

Naturvårdsverket, 106 48 Stockholm. E-post: registrator@naturvardsverket.se, www.naturvardsverket.se
Besöksadress Stockholm: Virkesvägen 2. Besöksadress Östersund: Forskarens väg 5, hus Ub.
Beställningar tel: +46 8-505 933 40, e-post: natur@cm.se, www.naturvardsverket.se/publikationer