A NEXT-GENERATION OF BIOMONITORING TO DETECT GLOBAL ECOSYSTEM CHANGE

EDITED BY: David Andrew Bohan, Dominique Gravel,
Alireza Tamaddoni-Nezhad, Corinne Vacher and Stéphane Robin

PUBLISHED IN: Frontiers in Ecology and Evolution and
Frontiers in Environmental Science

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-027-8
DOI 10.3389/978-2-88966-027-8

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

A NEXT-GENERATION OF BIOMONITORING TO DETECT GLOBAL ECOSYSTEM CHANGE

Topic Editors:
David Andrew Bohan, INRA Centre Dijon Bourgogne Franche-Comté, France
Dominique Gravel, Université de Sherbrooke, Canada
Alireza Tamaddoni-Nezhad, Imperial College London, United Kingdom
Corinne Vacher, Institut National de la Recherche Agronomique (INRA), France
Stéphane Robin, INRA Centre Versailles-Grignon, France

Citation: Bohan, D. A., Gravel, D., Tamaddoni-Nezhad, A., Vacher, C., Robin, S., eds. (2020). A Next-Generation of Biomonitoring to Detect Global Ecosystem Change. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-027-8

Table of Contents

frontiers
in Ecology and Evolution

ORIGINAL RESEARCH
published: 02 July 2019
doi: 10.3389/fevo.2019.00240

Field Evaluation of DNA Based Biodiversity Monitoring of Caribbean Mosquitoes

Sam P. Boerlijst[1,2]*, *Krijn B. Trimbos*[2], *Jordy G. Van der Beek*[1], *Klaas Douwe B. Dijkstra*[1], *Berry B. Van der Hoorn*[1†] *and Maarten Schrama*[1,2†]

[1] Naturalis Biodiversity Center, Leiden, Netherlands, [2] Department of Environmental Biology, Institute of Environmental Sciences, Leiden University, Leiden, Netherlands

Mosquito borne diseases pose a threat to human health worldwide. Disease risk is primarily determined by presence and abundance of vector species. A better understanding of mosquito diversity and abundance can direct improved vector control, but this requires a combination of monitoring techniques that yield both rapid and reliable information. Particularly improved larval detection is pivotal to move toward more targeted management with less environmental impact. Current detection methods rely strongly on manual labor and taxonomic expertise, which greatly limits the extent to which these methodologies can be employed. As such, insight in the efficiency of novel, high-throughput vs. traditional sampling techniques is required. We compared the effectiveness of a recently developed environmental DNA (eDNA) approach on water and sediment samples with other commonly used sampling techniques ("dipping" for larvae and adult trapping) in a field study on three Caribbean islands. All sampling methods were employed across a range of ecologically contrasting sites. Species identification was performed both morphologically and molecularly using an in-house developed reference database supplemented with sequences from BOLD and GenBank. Our analysis of water samples from 39 sites shows that eDNA sampling can be more reliable than dipping, yields a higher within-sample richness and produces a subset of the adult community in all sampled water types. Furthermore, for both adults and larvae, our identifications showed complete overlap between morphological and molecular approaches in 133 out of 134 samples. Overall, results from this study provide evidence that both our eDNA-based detection of larvae and our DNA-based identification of larvae and adults present methods that are, although more expensive, as reliable, and for some species even more reliable than the currently used methods. Additionally, our results highlight that a DNA approach can be used to identify larvae of early developmental stages, which generally lack important morphological characteristics. As such it allows for development of efficient disease control strategies, verification of management effectiveness and monitoring of population dynamics.

Keywords: ***Aedes aegypti***, **biomonitoring,** ***Culex quinquefasciatus***, **Dutch Leeward Isles, eDNA, mosquito communities, mosquitoes, vector-borne diseases**

OPEN ACCESS

Edited by:
David Andrew Bohan,
INRA Centre Dijon Bourgogne Franche-Comté, France

Reviewed by:
Konstantinos Demertzis,
Democritus University of Thrace, Greece
Cameron Ewart Webb,
University of Sydney, Australia

****Correspondence:***
Sam P. Boerlijst
samboerlijst@yahoo.com

[†] These authors share last authorship

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Ecology and Evolution

Received: *25 January 2019*
Accepted: *11 June 2019*
Published: *02 July 2019*

Citation:
Boerlijst SP, Trimbos KB, Van der Beek JG, Dijkstra KDB, Van der Hoorn BB and Schrama M (2019) Field Evaluation of DNA Based Biodiversity Monitoring of Caribbean Mosquitoes. Front. Ecol. Evol. 7:240. doi: 10.3389/fevo.2019.00240

INTRODUCTION

Mosquitoes (Culicidae) present a major risk for human health worldwide (Leslie et al., 2017). They cause hundreds of thousands of deaths annually due to their role as vector for vector-borne diseases (VBD) such as chikungunya, dengue, malaria, yellow fever, and zika (WHO, 2017b). Even though our understanding of these diseases is growing, case incidences are increasing (Risks et al., 2012; WHO, 2017a). Occurrence of VBD is limited by the presence, abundance, and dispersal of their respective vectors (Schaffner and Mathis, 2014). Due to land use change, climate change and increased global trade and travel (Lambin et al., 2001; Patz et al., 2004; Risks et al., 2012), distributions of vector species, especially those of exotic species such as *Aedes aegyti* and *Aedes albopictus* are shifting across all continents (Petric et al., 2014). This highlights the importance of monitoring tools that can provide reliable, high-throughput and up-to-date information on species distributions, both for larvae and adults, especially for surveillance near hotspots of travel and trade, such as harbors and airports.

Traditionally, methodologies used for studying the distribution of mosquito larvae and adults rely strongly on morphological identification. This, combined with methodological challenges, greatly limits the extent to which these techniques can be employed. For example, to identify localities where vector control is needed, presence and, ideally, larval habitats of (disease-vectoring) mosquitoes have to be confirmed. Larval identification methods play a crucial role in the detection of these larval habitats because mosquito populations are generally limited by the availability of suitable habitats (Frank et al., 1988; Rejmánková et al., 2013). The detection is usually performed using a "dipping" method (hereafter referred to as dipping), in which larvae and pupa are physically caught and identified (van der Berg and Schaffner, 2016). In doing so, larval habitats can be specifically targeted with mosquito control measures, thus minimizing the impact on the environment. However, dipping can be cumbersome, since the larvae and pupa dive upon visual and auditory disturbance (Becker et al., 2013). Depending on the species, the dive can last up to several minutes, increasing the required sampling effort and possibly decreasing detection probability. In addition to these methodological challenges, samples collected using dipping are typically identified morphologically, which is prone to unresolved or misidentification due to phenotypic plasticity and cryptic species (Jerde et al., 2011; Deiner et al., 2013; Fišer Pečnikar and Buzan, 2014; Mächler et al., 2014). Some characteristics, for example, are only apparent at a certain life stage or gender (Murugan et al., 2016; van der Berg and Schaffner, 2016). This is especially true for culicid larvae, since most characteristics are only visible on the fourth instar (Becker et al., 2013; ECDC, 2014). Likewise, adult sampling is widely used to detect the mosquito species community at a given locality, and is generally carried out using a variety of trapping methods (Becker et al., 2013). Adult sampling may yield higher diversity, since it is independent of larval habitat preference. After trapping, all individuals need to be collected at dawn, thus limiting the number of trapping sites. Afterwards, individuals are sorted to species. In general, identification is performed using morphological keys. This method is therefore almost entirely dependent on taxonomic expertise, which is becoming increasingly difficult to get by Mächler et al. (2014), particularly in the highly biodiverse ecosystems of the tropics. Furthermore (recent), identification keys based on morphology are not available for many regions, with a particular lack of keys in tropical areas where mosquito diversity and mosquito borne disease risk are highest (Rawlins et al., 2008; WHO, 2017b).

Molecular approaches based on environmental DNA (DNA that organisms shed into their environment, hereafter eDNA) or DNA could comprise valuable additions or even alternatives for both larval and adult morphological methods. For larval stages, an eDNA approach can have three main benefits. First, it can greatly reduce the time needed for species collection (Herder et al., 2014) and is non-invasive in the sense that it does not harm the species under investigation (Thomsen and Willerslev, 2015). Second, larvae can—in theory—be detected after adult emergence (Barnes et al., 2014; Schneider et al., 2016) thus allowing for repeated sampling with larger time intervals. Third, detection rates for eDNA may be higher than those of traditional techniques, which has already been observed for taxa such as fishes, amphibians and gastropods (Thomsen et al., 2012; Goldberg et al., 2013; Pilliod et al., 2013; Mächler et al., 2014). A growing number of eDNA studies target macro-invertebrates (Roussel et al., 2015), from which the necessity of eDNA collection based on target species ecology can be inferred (Deiner et al., 2015). Sediment dwelling organisms such as amphipods, for instance, tend to be hard to detect in aquatic samples (Mächler et al., 2014). Despite difficulties with other invertebrate taxa, we expect that eDNA of mosquito larvae can be reliably detected, since these larvae live and molt near the water surface (Rueda, 2008), which can be expected to result in a local accumulation of eDNA, thus allowing for successful eDNA collection (Schneider et al., 2016). For adult mosquitoes, molecular identification would also represent a valuable methodological addition, especially if employed in combination with high-throughput sequencing (also known as next-gen sequencing, hereafter NGS). All individuals within a sample (hereafter bulk sample) can then be analyzed simultaneously, allowing for rapid identification of large numbers of specimens, rendering it less labor intensive and time consuming (Batovska et al., 2016). Moreover, it holds the promise to be less prone to identification mistakes and might even result in a more resolved identification (Fišer Pečnikar and Buzan, 2014), which is especially important for vector control.

However, there are a number of caveats when it comes to using molecular identification. First of all, molecular methods currently remain more costly than traditional methods and may therefore prove to be less readily available for routine surveys. Also, environmental samples in particular are known to contain components that hinder DNA amplification, such as humic substances and polysaccharides (Herder et al., 2014). The amount and variety and therefore the influence of these PCR inhibitors varies across the different types of larval habitats (Wilson, 1997; Schrader et al., 2012). Also, the larval habitats vary in their abiotic properties (Becker et al., 2013), which influences DNA degradation (Strickler et al., 2015). Species in certain water body types may therefore be harder to detect. A better understanding of the factors that determine the reliability and usability of eDNA collection and molecular identification is therefore required to implement these techniques for adequate mosquito vector monitoring.

The aim of this research is threefold: (i) to explore whether eDNA-based assessments of larval communities in water bodies match with the morphological analysis, thus testing the reliability of eDNA-based assessments of larval mosquito communities; (ii) to explore whether DNA-based identification of adult and larvae matches the morphological identifications, thus determining the usability of molecular identification for high-throughput assessments; (iii) to determine whether DNA-based methods can be used across ecologically contrasting habitats by comparing relationships between larval mosquito community composition, and abiotic properties of the various habitats, using both traditional and DNA-based methods.

To this end, a comparative analysis was carried out to determine the effectiveness of eDNA sampling and molecular identification using a recently developed culicid-specific primer (Krol et al., 2019) vs. traditional sampling and morphological identification for detection of larval, pupal, and adult culicids. We used the mosquito communities of the Caribbean islands Saba, St. Eustatius and St. Maarten (Lesser Antilles) as a study system. These islands are ecologically diverse and have a relatively limited and relatively well-known species pool (Van der Kuyp, 1954), and therefore provide an ideal study system.

MATERIALS AND METHODS

All field work was conducted during April 2018 on the islands of St. Eustatius, St. Maarten and Saba. Samples of adults, larvae and eDNA were collected in the period of a single week for each of the three islands. For adults, a list of optimal trapping sites was gathered by consulting the local vector control units for knowledge on known larval habitats on each of the three islands. From this list 10 trapping sites per island were selected which cover all available habitats including the urban environment. Only for larvae, aquatic eDNA samples and sedimentary eDNA samples were collected; every water body that was encountered during intensive surveys on the island was included as a sample site. For each water body, we recorded the coordinates (**Figure 1**) and type (see below). Aquatic eDNA was collected at 36, 17, and 19 sites on St. Eustatius, Saba, and St. Maarten, respectively. Sedimentary eDNA was collected at 6 sites on St. Maarten only.

Traditional Sampling of Adults, Larvae, and Pupae

Adult Mosquitoes

Samples of adult mosquitoes, consisting of all individuals caught per trapping method, were collected using Mosquito magnets (Executive), BG-sentinels v2, resting traps, sticky traps, and human-landing catches at each of the sites (**Figure 1**). The Mosquito magnets were deployed at ground level because the high spatial coverage of the Mosquito magnet is designed to capture mosquitoes from the entire air column, thereby overcoming the stratifying effects of possible host preference (Andreadis and Armstrong, 2007; Harbach, 2007). Placement was ∼10 m leeward of larval habitats with a minimum distance of 100 meters between the trapping sites to allow for optimal spatial coverage (Harrington et al., 2005; Epopa et al., 2017; Medeiros et al., 2017). A similar approach was used for the BG-sentinel and resting trap (Burkett-Cadena, 2011). Sample collection encompassed 3 days to yield a representable composition of the mosquito community (Gorsich et al., pers. comm.). To minimize sampling bias which may arise from species-specific variation in lifestyle (as a result of e.g., varying flight times or feeding activity; Harbach, 2007; Panella et al., 2016), adult mosquitoes were collected during 24 h of continuous trapping. All traps were emptied between 5.30 and 7.30 a.m. to prevent the mosquitoes from drying out, which hinders morphological identification. BG-sentinels were baited with BG-lure and a sugar-yeast mixture of which the latter acted as CO_2 source as alternative for dry-ice which proved unobtainable on all of the islands. Mosquito magnets were baited with octenol and CO_2 using lure and propane combustion, respectively. The latter also served as an electricity source. The use of CO_2 and bait is expected to increase yield (Bhalala and Arias, 2009; Hoel et al., 2009; Kweka et al., 2013). All stationary traps were placed out of direct sunlight to prevent captured mosquitoes from drying out. Traps were shielded from rain and wind to prevent damage and optimize the efficiency of the octenol and CO_2. Human landing catches (ECDC, 2014) were performed each day at dusk.

Larvae and Pupae

Larvae and pupae were collected by dipping (Becker et al., 2013) in stagnant water bodies such as cisterns, ponds, rock pools, wells, tree holes, pots, and plant containers such as bromeliads. To test how well dipping and eDNA collection perform across a variety of conditions, whilst tackling habitat preference (Becker et al., 2013; Petric et al., 2014), we made an attempt to sample as many different types of water bodies as possible (Harbach, 2007; ECDC, 2014; Richardson and Richardson, 2014; Lebl et al., 2015). The risk of cross contamination was negligible because all sampling locations were spatially separated. Dipping was performed with either a 60 mm diameter sieve, 70 × 50 mm aquarium net or 25 mL pipet, depending on the accessibility of the water body. The larvae and pupae were stored in 96% ethanol. Sticky traps were used after dipping, just above surface level of water bodies that still carried water after sample collection.

eDNA-Based Sampling

Water Samples

Independent of whether larvae and/or pupae were found, aquatic eDNA was collected at all larval sites where >10 mL water could be sampled for a total of 32, 18, and 18 samples for St. Eustatius, Saba, and St. Maarten, respectively. Samples were taken by collecting surface water in steps of 25 mL with a PIPETBOY (Integra Bioscience) without agitating the water. Based on a pilot study, a maximum volume of 200 mL was used to prevent the filters from clogging. Because volume and subsample count varied between sites and are known

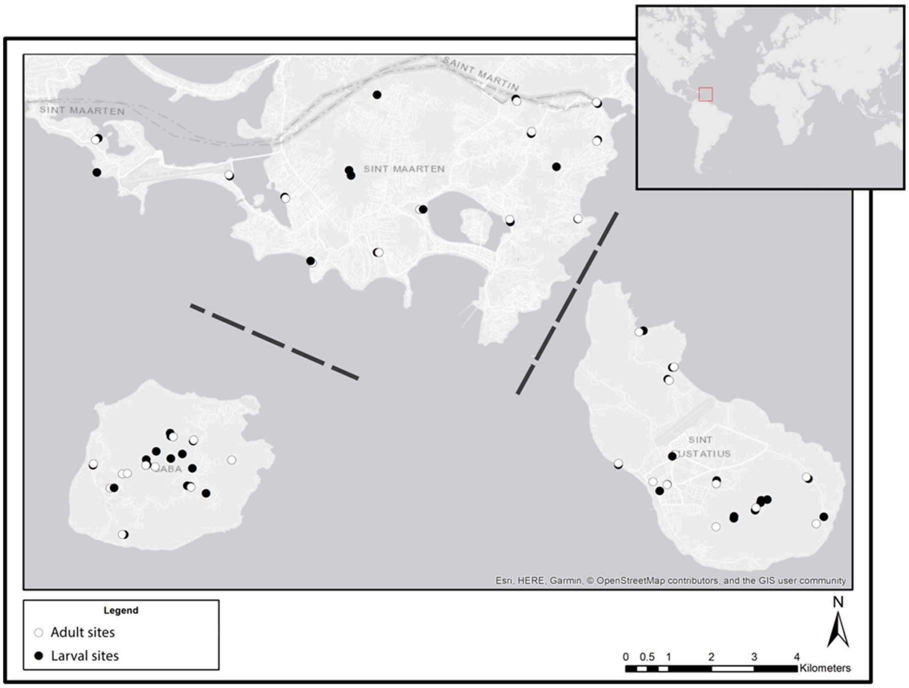


FIGURE 1 | Overview of all sampling sites at each of the three sampled Caribbean islands: St. Maarten (top), Saba (bottom left), and St. Eustatius (bottom right). White symbols indicate sites where adults were sampled, black symbols indicate sites where larval (dipping) and eDNA (water and sediment) samples were collected. Dashed lines indicate that the distance between islands is not to scale (Esri, 2011).

to influence detection probabilities (Turner et al., 2014), we recorded both parameters at each of the sites. The samples were stored at 4°C until further processing. Within 24 h after collection samples were filtered with a vacuum pump using a Sartorius polycarbonate filter holder and 47 mm 0.2 µm filter membrane (Sartorius-stedim). Filters were stored in 2 mL micro centrifuge tubes containing 900 µL Longmire solution (0.1 M TRIS, 0.1 M EDTA, 0.5% SDS, 10 mM NaCl; Williams et al., 2016) to prevent DNA degradation during storage and transport at ambient temperatures (Renshaw et al., 2015; Williams et al., 2016). After each sample, filter holders and pipets were cleaned by rinsing with bleach (2x) and water (3x) to prevent cross-contamination of samples by destroying the residual DNA. On every island, a negative control (tap water) was filtered and processed as if it was a sample to test for possible contamination between samples.

Sediment Samples

The water bodies were, whenever possible, sampled for sedimentary eDNA to allow for the collection of settled eDNA (Turner et al., 2014). Sediment collection consisted of filling a 15 mL falcon tube up to the 10 mL mark (equal to roughly 14 gr.) by scraping the entire depth profile perpendicular to the waterline in a transect ranging from 10 cm under water up to the waterline for 4–6 sub samples at 0.5 m distance from each other. Hereafter all subsamples per sampling location were merged. To each of the samples 5 mL of CTAB buffer was added and subsequently mixed by carefully inverting 2–3 times to prevent DNA degradation during storage and transport (Renshaw et al., 2015). The sediment samples were stored at 4°C until further processing.

Morphological Identification of Larvae and Adults

All larvae and adults from each traditional sample were morphologically identified. Identification was primarily conducted using keys and species descriptions by Belkin et al. (1970), Darsie et al. (2010), and Van der Kuyp (1954).

Construction of DNA Reference Database

An in-house developed reference database for the Cytochrome oxidase I gene (COI) of morphologically identified species was constructed (**Supplement 2** in Supplementary Material) to reduce the probability of misidentifications (Virgilio et al., 2010). This database contains species that were likely to occur on Saba, St. Eustatius and St. Maarten, based on the data from “mosquitocatalog.org,” but had insufficient public material on BOLD and GBIF. The dataset was constructed by Sanger sequencing with the primer set jgLCO1490 and jgHCO2198 (Geller et al., 2013). Sequences were obtained by barcoding specimens from the personal collection of Francis Schaffner. The sequences included the species *Aedes aegypti, Aedes busckii, Aedes serratus, Aedes taeniorhynchus, Aedes tentius, Aedes tortilis, Anopheles aquasalis, Culex atratus, Culex bisulcatus, Culex idottus, Culex nigripalpus, Culex quinquefasciatus, Deinocerites magnus, Haemagogus chrysochlorus, Haemagogus* *dyrisolchloratus, Isotomyia perturbans, Limatus durhami,* and *Psorophora ferox.*

DNA Extraction, Amplification, and Sequencing

For each traditional sample containing specimens, either adults or larvae, the right mid leg for adults or the proximal three segments of the larval abdomen were used for DNA extraction, so that the quality of the voucher specimen for morphological analysis was retained. The legs or abdominal segments of all specimens at a given location were merged per sample (hereafter called bulk sample), ground and DNA was extracted conform the Kingfisher's "Machery_Nagel_Tissue_96 KingFisher Flex" protocol for the NucleoMag 96 Tissue kit.

eDNA of water samples was extracted and purified using a PCI protocol developed by Renshaw et al. (2015) followed by a DNeasy blood and tissue kit extraction as clean-up (**Supplement 3** in Supplementary Material). This method combines the higher yield of the PCI (Deiner et al., 2015; Goldberg et al., 2016) and the blood and tissue kit inhibitor removal (Zhou et al., 1991) whilst being able to store it at room temperature (Renshaw et al., 2015). eDNA of sediment samples was extracted and purified using the FastDNA soil kit developed by MP Biomedicals. Verification of the quality and quantity of the DNA extracts was performed with a Trinean dropsense 96.

For both bulk samples and eDNA samples, DNA amplification of mini barcodes (154 bp) within the COI region was performed with IonCode labeled culicid primers developed in an earlier study: F: 5′-GGRKCHGGDACWGGDTGAAC-3′; R: 5′-RGATCAWACAAATAAAGGTAWTCGATC-3′ (Krol et al., 2019). Each PCR mixture (20 μL) contained 1.5 μL of DNA solution with 1 μL 10 pM forward and reverse primer in 10 μL 2x environmental master mix (Taqman environmental mix 2.0, Applied Biosystems, Foster City, CA, USA). PCR reactions were performed under thermocycler conditions of 10 min at 95°C, and 40 cycles of 15 s at 95°C, 30 s at 52°C, 40 s at 72°C and 5 m at 72°C in a Bio-rad C1000 touch^{TM} system.

After the PCR, product presence was visually confirmed by gelelectroforesis on E-gel (Invirtogen, Foster City, CA, USA). Samples with product were cleaned by mixing with 18 μL Nucleomag B-beads (Macherey-nagel GmbH & Co, Düren, Germany), incubated for 5 m and placing it on a magnetic rack. Supernatant was removed and the samples were washed two times with 100 μL 80% ethanol and left to air-dry. Thereafter, the samples were taken off the magnetic rack and resuspended in 25 μL Milli-Q. Subsequently DNA concentrations were quantified with the Qiagen© Qiaxcel and pooled equimolarly at 26 nM/L with the Qiagility. The pool was diluted to 30 pM/L and subsequent analysis was done conform the IonTorrent^{TM} IonPGM^{TM} Hi-Q^{TM} handbook using the BioAnalyzer, Ion OneTouch^{TM} 2 and Iontorrent on a 318^{TM} chip. All 97 bulk samples contained PCR product and were used for the Iontorrent run. Forty seven of 68 water samples and 6 of 10 sediment samples contained PCR product and were used for the Iontorrent run. Samples with undetectable amounts of DNA were not used to prevent dilution of the other samples.

Bioinformatics and Statistical Analysis

Initial assessment of the NGS data was performed with the software packages FastQC v0.11.7 (Andrews, 2010), cutadapt v1.16 (Martin, 2011), prinseq 0.20.4 (Magoč and Salzberg, 2011), FLASH v1.2.11 (Putra et al., 2015), Unoise v10.0.240 (Edgar, 2016), and Vsearch v2.4.3 (Rognes et al., 2016) integrated in the Naturalis Galaxy pipeline.

As in previous studies NGS data were filtered by clipping the primers and, in case of low data quality, by trimming the 3'side of the sequences based on a lower phred-score cut-off of 20 (Deiner et al., 2017). Operational taxonomic units (hereafter OTUs) were after removal of singletons, clustered with both Unoise (alpha 1.5) and Vsearch (threshold: 97% similarity). OTUs with without a read depth of 0.001% of the total amount of reads within at least one sample were discarded to remove artifactual sequences (Alberdi et al., 2018).

Both Unoise and Vsearch clustering algorithms were used since they were expected to yield dissimilar results due to the differences in clustering approach. Also, the authors of Unoise state that Unoise clustering with Iontorrent data may result in inflated abundance of incorrect reads due to sensitivity to barcoding errors. However, no difference in community per sample was detected between the two clustering methods for Bray-curtis similarity on presence-absence data using Past v3 (**Figure S1** in Supplementary Material; one-way ANOSIM, R 0.008531, $p > 0.1$). To reduce required computational power induced by the amount of OTUs, Unoise clustering, which resulted in far fewer OTUs, was used for further analysis.

The optimal Unoise alpha value was determined from the values 0.5, 1.0, 1.5, 2.0, and 3.0 by manually confirming the most parsimonious phylogeny whilst maintaining all Culicomorpha using the neighbor-joining phylogeny (**Datasheet S3** in Supplementary Material; ClustalW 2.1) and lowest common ancestor analysis (LCA) (**Supplement 5** in Supplementary Material; Megan 6.12.3). LCA was performed using the parameters: min score: 170; max expected: 0.01; min percent identity: 70; top percent: 5; min support percent: 0.0 (off); min support: 1.

Alignment was performed against the internally developed reference database, or if no reference ID > 95% is available against BOLD or thereafter Genbank. If no hit with ID > 95% was found, the optimal hit from the three databases was used. Genbank is used as last resort, since it is known to include misidentifications (Meier et al., 2006).

Hits >98% were accepted as species level identifications and hits >95% as genus level identifications since a culicid specific primer is used (Alberdi et al., 2018). Five Misidentifications were corrected for by manually comparing all species level identifications of species found outside their expected distribution.

The OTU table was transformed into a presence-absence matrix prior to the analysis, since the amount of reads cannot reliably be used as a proxy for biomass within species (Herder et al., 2014) and even more so across species (Goldberg et al., 2016).

Because no mosquitoes were caught using the sticky traps and resting traps, these methods were excluded from further analysis. A total species list was composed using data from the Mosquito Magnet, BG-sentinel, human landing catches, dipping samples, eDNA water samples and eDNA sediment samples.

Differences in larval detections between dipping and eDNA samples and between different water body types were tested using a one-way ANOSIM: (9,999 permutations) and visualized with Non-metric multidimensional scaling (nMDS) plot using Bray-Curtis similarity in Past v3. The differences were further explored with binomial GLM with log-odds link function, to test for interaction effects between identification method and water body type and (interaction) effects of sample volume and subsample count (**Datasheet 2** in Supplementary Material).

Data from the Mosquito Magnet and BG-sentinel was highly unbalanced due to trap failure. Therefore, presence-absence counts of only the Mosquito Magnet, BG-sentinel samples, and dipping samples were used to calculate detection probabilities both overall and per species comparing morphological and molecular determinations via χ^2-test using R version 1.1.383 (**Datasheet 2** in Supplementary Material).

RESULTS

From all molecular data, 35 of the 255 OTUs were identified to species level. These mainly included species within the family Culicidae, but also other taxa in the Diptera order, and some taxa within the order Crustacea (**Supplement 5** in Supplementary Material). The OTUs identified as Culicidae clustered in accordance with the presumed phylogeny (**Datasheet 3** in Supplementary Material; Harbach, 2007), indicating that the species level identification at 98% identity was correct. The OTUs within the infraorder Culicomorpha were used for further analysis (**Datasheet 1** in Supplementary Material). None of the negative controls contained culicid DNA, thus indicating that no cross contamination occurred during sample processing.

Species Detected by Morphological Analysis

Morphological analysis of the larval samples yielded the following species: *Aedes aegypti, Ae. busckii, Culex bahamensis, Cx. bisulcatus, Cx. Quinquefasciatus,* and *Toxorhynchites guadeloupensis.* The adult samples also included the species *Aedes taeniorhynchus, Cx. nigripalpus, Deinocerites magnus,* and *Anopheles albimanus.* Molecular analysis of the eDNA and bulk samples resulted in a higher diversity (**Table 1**).

Species Detected With Molecular Analysis

Molecular analysis of the aquatic eDNA samples yielded the following species: *Aedes aegypti, Ae. busckii, Culex bahamensis, Cx. bisulcatus, Cx. pipiens molestus, Cx. quinquefasciatus, Cx. Renatoi,* and *Toxorhynchites* sp. The larval and adult samples also included the species *Anopheles albimanus, Aedes taeniorhynchus, Culex bidens, Cx. nigripalpus, Cx. pipiens pallens,* and *Deinocerites magnus.* The species *Cx. bidens* and *Cx. renatoi* are likely misidentifications: *Cx. bidens* = *Cx. nigripalpus, Cx. renatoi* = *Culex* sp., which we further elaborate on in the discussion. The corrected species have been used in further analysis.

TABLE 1 | Identified species using each of the different methods.

	Morphological		Molecular		
	Adult	Larvae	Adult	Larvae	eDNA
Aedes busckii		x		x	x
Ae. aegypti	x	x	x	x	x
Ae. taeniorhynchus	x		x		
Anopheles albimanus	x		x		
Culex bahamensis	x	x	x	x	x
Cx. bisulcatus	x	x	x	x	
Cx. nigripalpus	x		x		
Cx. quinquefasciatus	x	x	x	x	x
Deinocerites magnus	x		x		
Toxorhynchites guadeloupensis		x			
Tx. spp.				x	x
*Cx. pipiens molestus**					
*Cx. pipiens pallens**					
Cx. sp.					x

eDNA refers to the eDNA water samples. The species annotated with an "" fall within the Culex pipiens species complex and may actually be Culex quinquefasciatus as indicated in the discussion.*

Molecular analysis of the 6 samples of eDNA from sediment yielded no culicid DNA, which coincided with a lack of larvae and aquatic eDNA at the same locations. The samples did however contain DNA identified as chironomid *Chironomus calligraphus.*

eDNA Water vs. Dipping

In general, eDNA analysis of water samples resulted in a higher detection rate of larvae than dipping. Of the 68 aquatic samples, 39 contained eDNA of mosquitoes and of the latter 11 samples also contained larvae (**Figure 2**). Most of the species were detected equally well using both methods (χ^2-test: $p >$ 0.28). However, within these samples a significant difference in detection probability was detected for *Cx. bisulcatus* ($\chi^2 =$ 7.1842, $p < 0.05$) and *Cx. quinquefasciatus* ($\chi^2 = 20.651$, p < 0.001). *Cx. bisulcatus* was better detectable by morphology and *Cx. quinquefasciatus* by eDNA (**Table 1**). None of the sediment samples contained culicid eDNA. However, none of the larval samples and water samples taken at the same locations contained culicid eDNA, implying absence of mosquitoes in these water bodies.

The difference in detection chance for *Cx. quinquefasciatus* and (subsequently) Culex as a whole (NMDS 60% contribution) (**Figure S2** in Supplementary Material) resulted in a detected

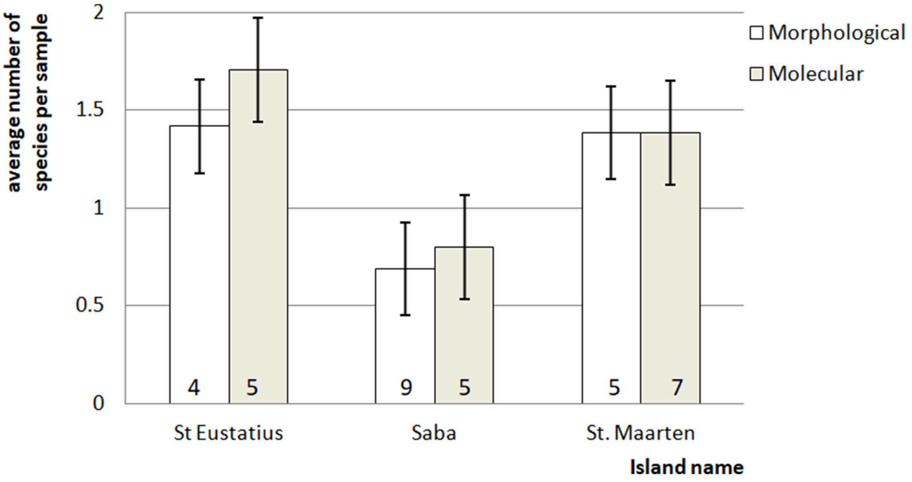


FIGURE 2 | Average number of mosquito species in the bulk samples ± standard error. The number at the bottom of each bar indicates for each island the total number of species detected for that particular method.

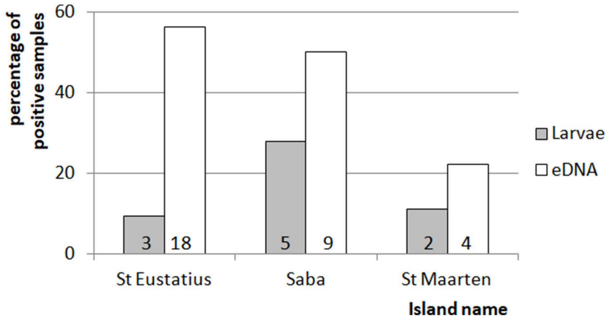


FIGURE 3 | Comparison of between the percentage of positive mosquito samples for the eDNA and traditional dipping method. Indicated is the percentage of times larvae were detected by dipping or eDNA water sampling for each of the islands. The numbers at the bottom of each of the bars indicate the number of detections for the corresponding method and island. The total amount of dipping and eDNA water samples was 32, 18, and 18 for St. Eustatius, Saba, and St. Maarten respectively.

difference between dipping and eDNA-based detections (one-way ANOSIM: $R = 0.4093$, $p < 0.001$). This is mainly caused by the fact that *Cx. quinquefasciatus* was detected over ten times more often in the eDNA samples than in the dipping samples.

Morphological Analysis vs. Molecular Analysis

The molecular and morphological analysis of larval and adult bulk samples performed very similar (**Figure 3**; **Table 1**). Differences between morphological and molecular analysis of the larval samples were found for *Cx. bahamensis* and *Cx. bisulcatus*. *Cx. bahamensis* was detected better with molecular analysis ($\chi^2 = 1.1781$, $p < 0.05$) whereas *Cx. bisulcatus* was detected better with morphological analysis ($\chi^2 = 7.1842$, $p < 0.001$). Differences between morphological and molecular analysis of the adult samples were found for *Cx. quinquefasciatus* and *Culex* overall. *Cx. quinquefasciatus* and *Culex* spp. were both detected better with molecular identification (χ^2-test $p < 0.05$) (χ^2-test $p < 0.05$) respectively.

Community Differences Per Water Body Type

Differences in species community between the different water body types were detected when comparing Bray-curtis similarity over the dipping and eDNA water samples (One-way ANOSIM: $R = 0.1991$, $p < 0.01$; **Table 2**), caused by the habitat types rock pool, plant container and artificial container. The separation is the largest between rock pool and artificial container ($R = 1$) and moderate between rock pool and plant container ($R = 0.2391$) and plant container and artificial container ($R = 0.2984$) (**Figure S2B**). These effects, when corrected for the influence of volume and subsample count, can be isolated as the result of the lower detection probability of *Cx. bisulcatus*, which is negatively correlated with the volume (logit ANOVA: Z-val $= -2.362$, $p < 0.05$) and *Cx.* sp. which shows a positive trend toward artificial containers (logit ANOVA: Z-val $= 1.799$, $p < 0.1$).

DISCUSSION

In our study, we used a range of sampling and processing methods for detection of Culicidae in a wide variety of habitats on three Caribbean islands: Saba, St. Maarten, and St. Eustatius. Our results suggest that our aquatic eDNA-based approach is as reliable and for certain species even more reliable than dipping. In contrast, eDNA originating from sediments did not result in detection of Culicidae. Although this suggests that this method may not be suitable, the lack of larval detection in the water and dipping samples taken at the same sites implies that no conclusions can be drawn about this method. Species identifications of larval and adult mosquitoes yielded very similar

TABLE 2 | Species composition per water body type for all three islands.

Water body type	*Ae. aegypti*	*Ae. busckii*	*Cx. bahamensis*	*Cx. bisulcatus*	*Cx. pipiens molestus*	*Cx. quinq.*	*Cx. sp.*	*Toxorhyn-chites spp.*
Artificial container	1 | 1 | 1	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	3 | 2 | 1	2 | 0 | 0	0 | 0 | 0
Cistern	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	1 | 1 | 0	0 | 0 | 0	0 | 0 | 0
Ditch	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0
Lake	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0
Plant container	0 | 0 | 0	0 | 1 | 0	0 | 0 | 0	1 | 3 | 0	0 | 0 | 0	2 | 0 | 1	0 | 0 | 0	0 | 2 | 0
Pond	0 | 1 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	1 | 1 | 1	0 | 0 | 0	0 | 0 | 0
Pool	0 | 0 | 0	0 | 0 | 0	0 | 0 | 1	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0
Rockpool	0 | 0 | 0	0 | 0 | 0	0 | 0 | 1	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0
Sink	1 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0	0 | 0 | 0
Treehole	0 | 0 | 0	0 | 1 | 0	0 | 0 | 0	0 | 1 | 0	0 | 0 | 0	4 | 0 | 0	0 | 0 | 0	0 | 0 | 0
Well	0 | 0 | 0	0 | 0 | 0	1 | 0 | 0	0 | 0 | 0	0 | 1 | 0	3 | 0 | 0	2 | 0 | 0	0 | 0 | 0

In each of the cells the order of the islands is: St. Eustatius | Saba | St. Maarten.

results when comparing morphological identification and DNA from bulk samples, but with some notable exceptions which are discussed below. Finally, we confirm the notion of Krol et al. (2019) that larval detection methods based on eDNA reveal a subset of the adult community, whilst confirming that this originates in part from inherent differences between larval and adult sampling. In our study species that were trapped as adult mosquitoes which were absent from the eDNA samples were also missing in the larval samples. In addition, some species detected in eDNA and dipping samples were absent in adult samples, suggesting that adult and larval sampling yield different yet complimentary parts of the mosquito puzzle.

Necessary Correction Steps in the DNA-Based Identification

Even though molecular identification provides promising results, the molecular analysis initially resulted in some misidentifications, for which manual corrections had to be made. These species included: (i) *Culex bidens*, (ii) *Culex pipiens* (var. *molestus* and *pallens*), and (iii) *Culex renatoi*. The former was identified using bulk sample DNA, the latter two were from aquatic eDNA samples. (i) *Culex bidens* is a species known only from South America, and is therefore an unexpected find. After re-evaluating the OTUs and corresponding BLAST results, it is likely that the OTUs identified as *Cx. bidens* are misidentified *Culex nigripalpus* DNA. This species has an identical identity and is congruent with the morphological identification from the sample that the OTUs originate from. (ii) *Cx. pipiens* var. *molestus* and *pallens* were found in samples where morphologically only *Cx. quinquefasciatus* was identified. However, *Cx. quinquefasciatus* is part of the *Culex pipiens* species complex (Harbach, 2012). Since these species are morphologically and molecularly almost identical (Laurito et al., 2013), it is possible that *Cx. pipiens* is actually present on the islands. (iii) The OTUs identified as *Cx. renatoi* were derived from aquatic eDNA samples from wells and artificial containers collected in St. Eustatius. They are highly dissimilar compared to the other known *Culex* species from the island (*Cx. bahamensis, Cx. bisulcatus* and *Cx. quinquefasciatus*; identity<93%). And, apart from the *Cx. renatoi* sequence, there are no sequences available that are similar enough for species identification. Since *Cx. renatoi* is a species that typically breeds in plant containers and is only known from South America, we consider this a misidentification. This might therefore be a new *Culex* species for the island, which has not yet been included in the BOLD and Genbank databases. Our results suggest that the reliability of molecular identification, and specifically that of aquatic eDNA sampling, is highly dependent on the quality of the reference library, thus re-emphasizing the previously identified need for more complete global databases (Deiner et al., 2017). All aforementioned misidentifications were corrected for prior to the analysis.

Unidentified OTUs

For all molecular data only 8% of the OTUs could be assigned with certainty to a mosquito species. Of the OTUs that could not be attributed to mosquito species, a large portion was found in environmental samples only (82.9%). The same is true for the OTUs that could not be identified to genus level (95.9%). Therefore, it is likely that these clusters were unidentifiable due to degradation of the DNA and due to the presence of DNA from organisms other than culicids, such as beetles, worms and amphibians. This is supported by the LCA analysis, which shows that a large portion of the unidentified OTUs could not be assigned at all or likely originate from crustaceans and other unknown taxa within the Diptera. Consequently, it is presumed that only a negligible amount of culicid DNA remained unidentified, which is also supported by the detection probabilities of the larval and aquatic samples. There were only two species that were difficult to identify using eDNA: *Culex bisulcatus* and *Toxorhynchites* spp. This is most likely caused by a lack of publicly available sequences. For *Cx. bisulcatus* only one sequence was available originating from our own reference library. DNA originating from bulk samples that, according to morphological identification, contained *Cx. bisulcatus*, clustered at 97% identity with this sequence. This indicates that genetic variability may indeed play a role in the inability to identify this species. Also for *Toxorhynchites* spp., a lack of reference sequences is the most likely explanation for the fact that none of the larvae of *Toxorhynchites* could be identified to species, given that BOLD and GenBank list only sequences for 18 of the 90 known species for this genus (mosquitocatalog.org). Overall, we are convinced that our molecular analysis yields an adequate representation of the observed species communities, both for bulk and aquatic eDNA samples, due to the high similarity when compared with the morphological identifications.

Detection of Larvae Using Dipping vs. eDNA

There were two major differences between sampling of mosquito larvae using dipping and our eDNA approach. First, the analysis of the aquatic samples resulted in a higher diversity (**Table S1** in Supplementary Material), which is congruent with previous research comparing traditional and eDNA-based detection (Deiner et al., 2017). Presumably this effect is caused by the inherent biases of the traditional methods (Deiner et al., 2017). Second, the eDNA sampling had a higher probability of detection, which is related to the first result. This difference was mainly caused by *Culex quinquefasciatus* DNA that was present in water bodies where no larvae were found. *Cx. quinquefasciatus* was 10 times more often present in aquatic eDNA samples than in the dipping samples. The water bodies where this species was detected included almost every water type, even plant containers. This is uncommon, but has been described before (Frank et al., 1988), thus confirming the generalist nature of the species. The reason for this discrepancy between dipping and eDNA-based detection may be 2-fold: (1) due to our inability to catch larvae using dipping: larvae can dive for several minutes (Becker et al., 2013) rendering them harder to catch, especially in water bodies with lower accessibility such as cisterns and wells, and (2) due to the persistence of eDNA in the aquatic environment. Larval development can be as short as 6–7 days (Becker et al., 2013), but eDNA can persist for weeks (Schneider et al., 2016) and is likely present at the water surface as the larvae spend most time there and also pupate at the water surface. In contrast to the water samples, none of the sediment samples contained culicid eDNA. Some of the samples did, however, contain DNA from chironomids, a closely related family within the infraorder Culicomorpha, indicating that detection was not hindered by PCR inhibition. Although this suggests that sediment samples can potentially be used for monitoring of Culicidae and (phylogenetically related) Diptera, our inability to collect samples at most sites illustrates that it may not be as straightforward as water samples. Overall, we conclude that reliability of aquatic eDNA sampling was higher than dipping, which is mainly due to the underestimation of presence of larvae of *Cx. quinquefasciatus*, one of the possible disease vectors on the islands.

Comparison Between Identification Using DNA Bulk Samples vs. Morphology

In general both identification methods performed comparably, but some differences between molecular and morphological identifications were found. Differences between morphological and molecular analysis of the larval samples were found for *Cx. bahamensis* and *Cx. bisulcatus*. *Cx. bahamensis* was detected more often with molecular analysis. One cause was that most captured larvae were early instars, which are unidentifiable morphologically. A portion of these was kept until they emerged to be identified as adult both morphologically and molecularly, which could be done successfully by both methods. *Cx. bisulcatus* was identified better by morphological analysis, which is likely due to a lack of reference sequences and will be elaborated on below. Differences between morphological and molecular analysis of the adult samples were found for *Cx. quinquefasciatus* and *Culex* overall. *Cx. quinquefasciatus* was detected more often by molecular analysis, which is reflected in the observed difference in detection probability for the *Culex* spp.

Differences in Subsample Count

There was no detected difference between the samples in relation to their subsample count.

This is counterintuitive, since eDNA is known to be heterogeneously distributed over the water bodies (Nathan et al., 2014). The cause for absence of this effect may be 2-fold. First, the result of the used subsample volume (25 mL), resulting in a bias toward samples with high subsample count, thereby countering the effects of eDNA heterogeneity. Secondly, we expect there is a correlated effect with habitat preference. The latter would also explain why this parameter still was included in the optimal GLMM model.

Differences Between Water Body Types

A difference in species communities was detected between the water body types originating from the differences between the types plant container and artificial container, rock pool and plant container and rock pool and artificial containers. This is in line with previous studies, showing the existence of species specific habitat preference (Andreadis and Armstrong, 2007; Abella-Medrano et al., 2015). When corrected for sample volume and subsample count, it becomes apparent that the detected difference is caused by *Cx. bisulcatus* which is negatively affected by higher volumes. This is congruent with the niche of the species, since it mainly inhabits tree holes and plant containers, which have small volumes. *Aedes aegypti*, shows a positive trend toward bigger water volumes and the *Culex* previously identified as *Cx. renatoi* shows a positive trend toward artificial containers. These effects are however not significant, which is likely the result of low sample numbers.

Further Research and Recommendations

Even though the highest mosquito diversity is thought to occur in the dry season (Abella-Medrano et al., 2015), sampling in the wet season may provide a more definitive answer due to the higher mosquito densities. This could also result in a more clear separation of the communities per water body type.

This study confirms the notion that eDNA collection should be tailored toward the ecology of the relevant species to account for the heterogeneous nature of the eDNA. Future research searching to extend the proposed methodology should therefore make a number of adaptations concerning the collection of aquatic eDNA, most notably sampling from specific parts of the watercolumn and collection of sufficient subsamples to cover the heterogeneity of eDNA over the larval habitats. Additionally there is a need for an adequate reference sequence library, as has been previously mentioned. We encourage sequencing of specimens from private/museum collections to supplement current references, which highlights the need for cooperation between institutions to locate and gain access to such material. Furthermore, to gain resolution within species complexes we recommend adding an additional locus to the analysis (e.g., CAD or 16S) (Schneider et al., 2016). Alternatively, other non-molecular approaches such as MALDI-TOF could be considered to augment the molecular analysis, as has been explored by Lawrence et al. (2019).

CONCLUSION

Results of our study provide evidence that the identification of mosquitoes based on aquatic eDNA using a novel culicid specific primer resulted in reliable detection of culicid larvae and overcomes some of the caveats surrounding dipping. Like dipping, aquatic eDNA collection result in the detection of a subset of the total community and should therefore be combined with adult trapping (e.g., human landing catches) for total culicid diversity assessments. Moreover, our results suggest that molecular identification could be a useful addition, particularly for rapid assessments of total diversity in a sample, overcoming some of the limitations in sample quality, developmental stage, and sampling effort of morphological identification in combination with dipping. In doing so, cryptic communities can be assessed without extensive prior taxonomic knowledge of the present species. However, molecular identification depends strongly on the quality of the reference databases. Therefore, a considerable amount of essential taxonomic work needs to be done before this method can become widely applicable in other regions. Completeness in respect to the expected species should therefore be assessed before implementing the method. Overall, results from this study provide evidence that both our eDNA-based detection of larvae and our DNA-based identification of larvae and adults present methods that are as reliable, and for some species even more reliable than the currently used methods. As such, it allows for development of efficient disease control strategies, verification of management effectiveness and monitoring of population dynamics.

AUTHOR CONTRIBUTIONS

SB, MS, KD, and BV conceived the idea for this study. SB carried out the measurements together with JV. SB carried out the molecular and bioinformatics work with the help of BV. SB wrote the first draft together with MS. KT and BV commented on the setup and assisted during writing of the paper and revisions.

ACKNOWLEDGMENTS

We are grateful to all people involved in the expedition. Special thanks go to Loes Busscher, Maud Kok, Delia-Maria Goilo, and Emily Khazan for their help during sampling and morphological identification. We would like to thank the laboratory technicians of Naturalis Biodiversity Center for their help during the lab work, especially Frank Stokvis and Kevin Beentjes. We are also grateful to the local governments and vector control units of the three islands for their advice on suitable sampling locations, especially Maria Henry, Gerald Davelaar, Teresa Leslie, Randall Johnson, and Koen Hulshof. We would like to thank Francis Schaffner for supplying mosquitoes for the reference database. Our thanks go to the ErasmusMC for enabling us to use the Mosquito Magnets and to the Gieskes Strijbis fund for funding this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2019.00240/full#supplementary-material

REFERENCES

Abella-Medrano, C. A., Ibáñez-Bernal, S., MacGregor-Fors, I., and Santiago-Alarcon, D. (2015). Spatiotemporal variation of mosquito diversity (Diptera: Culicidae) at places with different land-use types within a neotropical montane cloud forest matrix. *Parasites Vectors* 8:487. doi: 10.1186/s13071-015-1086-9

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., and Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods Ecol. Evol.* 9, 134–147. doi: 10.1111/2041-210X.12849

Andreadis, T. G., and Armstrong, P. M. (2007). A two-year evaluation of elevated canopy trapping for Culex mosquitoes and West Nile virus in an operational surveillance program in the northeastern United States. *J. Am. Mosq. Control Assoc.* 23, 137–148. doi: 10.2987/8756-971X(2007)23[137:ATEOEC]2.0.CO;2

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed September 24, 2018).

Barnes, M. A., Turner, C. R., Jerde, C. L., Renshaw, M. A., Chadderton, W. L., and Lodge, D. M. (2014). Environmental conditions influence eDNA persistence in aquatic systems. *Environ. Sci. Technol.* 48, 1819–1827. doi: 10.1021/es404734p

Batovska, J., Blacket, M. J., Brown, K., and Lynch, S. E. (2016). Molecular identification of mosquitoes (Diptera : Culicidae) in southeastern Australia. *Ecol. Evol.* 6, 3001–3011. doi: 10.1002/ece3.2095

Becker, N., Dahl, C., Bryant, B., Blair, C. D., Olson, K. E., Clem, R. J., et al. (2013). *Mosquitoes and Their Control. Insect Biochemistry and Molecular Biology*, Vol. 33. Berlin; Heidelberg: Springer Berlin Heidelberg.

Belkin, J. N., Heinemann, S. J., and Page, W. A. (1970). The *Culicidae* of Jamaica. *Contrib. Amer. Ent. Inst.* 6, 1–458. doi: 10.1007/s13398-014-0173-7.2

Bhalala, H., and Arias, J. R. (2009). The Zumba mosquito trap and BG-Sentinel trap: novel surveillance tools for host-seeking mosquitoes. *J. Am. Mosq. Control Assoc.* 25, 134–139. doi: 10.2987/08-5821.1

Burkett-Cadena, N. D. (2011). A wire-frame shelter for collecting resting mosquitoes. *J. Am. Mosq. Control Assoc.* 27, 153–155. doi: 10.2987/10-6076.1

Darsie, R. F., Taylor, D. S., Prusak, Z. A., and Verna, T. N. (2010). Checklist of the mosquitoes of the Bahamas with three additions to its fauna and keys to the adult females and fourth instars. *J. Am. Mosq. Control Assoc.* 26, 127–134. doi: 10.2987/09-5982.1

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., et al. (2017). Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol. Ecol.* 26, 5872–5895. doi: 10.1111/mec.14350

Deiner, K., Knapp, R. A., Boiano, D. M., and May, B. (2013). Increased accuracy of species lists developed for alpine lakes using morphology and cytochrome oxidase I for identification of specimens. *Mol. Ecol. Resour.* 13, 820–831. doi: 10.1111/1755-0998.12130

Deiner, K., Walser, J.-C., Mächler, E., and Altermatt, F. (2015). Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biol. Conserv.* 183, 53–63. doi: 10.1016/j.biocon.2014.11.018

ECDC (2014). *Guidelines for the Surveillance of Native Mosquitoes in Europe. Euro Surveillance : Bulletin Européen sur les Maladies Transmissibles = European Communicable Disease Bulletin, Vol. 17.*

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv.* 081257. doi: 10.1101/081257

Epopa, P. S., Millogo, A. A., Collins, C. M., North, A., Tripet, F., Benedict, M. Q., et al. (2017). The use of sequential mark-release-recapture experiments to estimate population size, survival and dispersal of male mosquitoes of the Anopheles gambiae complex in Bana, a west African humid savannah village. *Parasites Vectors* 10, 1–15. doi: 10.1186/s13071-017-2310-6

Esri. (2011). World Light Gray Base [basemap]. Available online at: https://www.arcgis.com/home/item.html?id=ed712cb1db3e4bae9e85329040fb9a49 (accessed October 10, 2018).

Fišer Pečnikar, Ž., and Buzan, E. V. (2014). 20 years since the introduction of DNA barcoding: from theory to application. *J. Appl. Genet.* 55, 43–52. doi: 10.1007/s13353-013-0180-y

Frank, J. H., Stewart, J. P., and Watson, D. A. (1988). Mosquito larvae in axils of the imported Bromeliad *Billbergia pyramidalis* in Southern Florida. *Fla. Entomol.* 71, 33–43. doi: 10.2307/3494890

Geller, J., Meyer, C., Parker, M., and Hawk, H. (2013). Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Mol. Ecol. Resour.* 13, 851–861. doi: 10.1111/1755-0998.12138

Goldberg, C. S., Sepulveda, A., Ray, A., Baumgardt, J., and Waits, L. P. (2013). Environmental DNA as a new method for early detection of New Zealand mudsnails (*Potamopyrgus antipodarum*). *Freshw. Sci.* 32, 792–800. doi: 10.1899/13-046.1

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., et al. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods Ecol. Evol.* 7, 1299–1307. doi: 10.1111/2041-210X.12595

Harbach, R. E. (2007). The *Culicidae* (Diptera): a review of taxonomy, classification and phylogeny. *Zootaxa* 1668, 591–638. doi: 10.1017/CBO9781107415324.004

Harbach, R. E. (2012). *Culex pipiens*: species versus species complex – taxonomic history and perspective. *J. Am. Mosq. Control Assoc.* 28, 10–23. doi: 10.2987/8756-971X-28.4.10

Harrington, L. C., Scott, T. W., Lerdthusnee, K., Coleman, R. C., Costero, A., Clark, G. G., et al. (2005). Dispersal of the dengue vector *Aedes aegypti* within and between rural communities. *Am. J. Trop. Med. Hyg.* 72, 209–220. doi: 10.4269/ajtmh.2005.72.209

Herder, J., Valentini, A., Bellemain, E., Dejean, T., van Delft, J., Thomsen, P. F., et al. (2014). *Environmental DNA - A Review of the Possible Applications for the Detection of (invasive) Species*. Stichting RAVON Report 2013 104, 21, 1789–1793. doi: 10.13140/RG.2.1.4002.1208

Hoel, D. F., Kline, D. L., and Allan, S. A. (2009). Evaluation of six mosquito traps for collection of *Aedes albopictus* and associated mosquito species in a suburban setting in North Central Florida[1]. *J. Am. Mosq. Control Assoc.* 25, 47–57. doi: 10.2987/08-5800.1

Jerde, C. L., Mahon, A. R., Chadderton, W. L., and Lodge, D. M. (2011). "Sight-unseen" detection of rare aquatic species using environmental DNA. *Conserv. Lett.* 4, 150–157. doi: 10.1111/j.1755-263X.2010.00158.x

Krol, L., Van der Hoorn, B., Gorsich, E. E., Trimbos, K., van Bodegom, P. M., and Schrama, M. (2019). How does eDNA compare to traditional trapping? Detecting mosquito communities in South-African freshwater ponds. *Front. Ecol. Evol.* doi: 10.3389/fevo.2019.00260

Kweka, E. J., Owino, E. A., Lee, M.-C., Dixit, A., Himeidan, Y. E., and Mahande, A. M. (2013). Efficacy of resting boxes baited with Carbon dioxide versus CDC light trap for sampling mosquito vectors: a comparative study. *Global Health Perspect.* 1, 11–18. doi: 10.5645/ghp2013.01.01.03

Lambin, E. F., Turner, B. L., Geist, H. J., Agbola, S. B., Angelsen, A., Folke, C., et al. (2001). The causes of land-use and land-cover change : moving beyond the myths. 11, 261–269. doi: 10.1016/S0959-3780(01)00007-3

Laurito, M., de Oliveira, T. M. P., Almirón, W. R., and Sallum, M. A. M. (2013). COI barcode versus morphological identification of Culex (Culex) (Diptera: Culicidae) species: a case study using samples from Argentina and Brazil. *Mem. Inst. Oswaldo Cruz* 108, 110–122. doi: 10.1590/0074-0276130457

Lawrence, A. L., Batovska, J., Webb, C. E., Lynch, S. E., Blacket, M. J., Jan, Š., et al. (2019). Accurate identification of Australian mosquitoes using protein profiling. *Parasitology* 146, 462–471. doi: 10.1017/S0031182018001658

Lebl, K., Zittra, C., Silbermayr, K., Obwaller, A., Berer, D., Brugger, K., et al. (2015). Mosquitoes (Diptera: Culicidae) and their relevance as disease vectors in the city of Vienna, Austria. *Parasitol. Res.* 114, 707–713. doi: 10.1007/s00436-014-4237-6

Leslie, T. E., Carson, M., Coeverden, E., van De Klein, K., Braks, M., and Krumeich, A. (2017). An analysis of community perceptions of mosquito-borne disease control and prevention in Sint Eustatius, Caribbean Netherlands. *Glob. Health Action* 10:1350394. doi: 10.1080/16549716.2017.1350394

Mächler, E., Deiner, K., Steinmann, P., and Altermatt, F. (2014). Utility of environmental DNA for monitoring rare and indicator macroinvertebrate species. *Freshw. Sci.* 33, 1174–1183. doi: 10.1086/678128

Magoč, T., and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957–2963. doi: 10.1093/bioinformatics/btr507

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17:10. doi: 10.14806/ej.17.1.200

Medeiros, M. C. I., Boothe, E. C., Roark, E. B., and Hamer, G. L. (2017). Dispersal of male and female *Culex quinquefasciatus* and *Aedes albopictus* mosquitoes using stable isotope enrichment. *PLoS Negl. Trop. Dis.* 11:e0005347. doi: 10.1371/journal.pntd.0005347

Meier, R., Shiyang, K., Vaidya, G., and Ng, P. K. L. (2006). DNA Barcoding and Taxonomy in Diptera : A Tale of High Intraspecific Variability and Low Identification Success. *Syst. Biol.* 55, 715–728. doi: 10.1080/10635150600969864

Murugan, K., Vadivalagan, C., Karthika, P., Panneerselvam, C., Paulpandi, M., Subramaniam, J., et al. (2016). DNA barcoding and molecular evolution of mosquito vectors of medical and veterinary importance. *Parasitol. Res.* 115, 107–121. doi: 10.1007/s00436-015-4726-2

Nathan, L. M., Simmons, M., Wegleitner, B. J., Jerde, C. L., and Mahon, A. R. (2014). Quantifying environmental DNA signals for aquatic invasive species across multiple detection platforms. *Environ. Sci. Technol.* 48, 12800–12806. doi: 10.1021/es5034052

Panella, N. A., Crockett, R. J. K., Biggerstaff, B. J., and Komar, N. (2016). The Centers for Disease Control and Prevention resting trap: a novel device for collecting resting mosquitoes. *J. Am. Mosq. Control Assoc.* 27, 323–325. doi: 10.2987/09-5900.1

Patz, J. A., Daszak, P., Tabor, G. M., Aguirre, A. A., Pearl, M., Epstein, J., et al. (2004). Unhealthy landscapes : policy recommendations on land use change and infectious disease emergence. *Environ. Health Perspect.* 1092, 1092–1098. doi: 10.1289/ehp.6877

Petric, D., Bellini, R., Scholte, E. J., Rakotoarivony, L. M., and Schaffner, F. (2014). Monitoring population and environmental parameters of invasive mosquito species in Europe. *Parasites Vectors* 7, 1–14. doi: 10.1186/1756-3305-7-187

Pilliod, D. S., Goldberg, C. S., Arkle, R. S., Waits, L. P., and Richardson, J. (2013). Estimating occupancy and abundance of stream amphibians using environmental DNA from filtered water samples. *Can. J. Fish. Aquat. Sci.* 70, 1123–1130. doi: 10.1139/cjfas-2013-0047

Putra, C. A., Hikmatullah, D., Prawiradilaga, D. M., and Harris, J. B. C. (2015). Surveys at Bagan Percut, Sumatra, reveal its international importance to migratory shorebirds and breeding herons. *Kukila* 18, 46–59. doi: 10.1093/bioinformatics/btr026

Rawlins, S. C., Hinds, A., and Rawlins, J. M. (2008). Malaria and its vectors in the Caribbean: the continuing challenge of the disease forty-five years after eradication from the islands. *West Indian Med. J.* 57, 462–469.

Rejmánková, E., Grieco, J., and Achee, N. (2013). *World' s Largest Science, Technology and Medicine Open Access Book Publisher*. INTECH.

Renshaw, M. A., Olds, B. P., Jerde, C. L., Mcveigh, M. M., and Lodge, D. M. (2015). The room temperature preservation of filtered environmental DNA samples and assimilation into a phenol-chloroform-isoamyl alcohol DNA extraction. *Mol. Ecol. Resour.* 15, 168–176. doi: 10.1111/1755-0998.12281

Richardson, B. A., and Richardson, M. J. (2014). Bromeliad invertebrate communities on Saba, Netherlands Antilles. *Caribbean Naturalist* 14, 1–12. Available online at: https://www.eaglehill.us/CANAonline/cana-n14-2014.shtml

Risks, P. H., Options, C., Medlock, J. M., Hansford, K. M., Schaffner, F., and Versteirt, V. (2012). A review of the invasive mosquitoes in europe, public health risks, and control options. 12, 435–447. doi: 10.1089/vbz.2011.0814

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *Peer J.* 4:e2584. doi: 10.7717/peerj.2584

Roussel, J.-M. M., Paillisson, J.-M. M., Tréguier, A., and Petit, E. (2015). The downside of eDNA as a survey tool in water bodies. *J. Appl. Ecol.* 52, 823–826. doi: 10.1111/1365-2664.12428

Rueda, L. M. (2008). Global diversity of mosquitoes (Insecta: Diptera: Culicidae) in freshwater. *Freshwater Animal Diversity Assessment*. eds E. V. Balian, C. Lévêque, H. Segers, and K. Martens (Dordrecht: Springer Netherlands), 595, 477–487.

Schaffner, F., and Mathis, A. (2014). Dengue and dengue vectors in the WHO European region: past, present, and scenarios for the future. *Lancet Infect. Dis.* 14, 1271–1280. doi: 10.1016/S1473-3099(14)70834-5

Schneider, J., Valentini, A., Dejean, T., Montarsi, F., Taberlet, P., Glaizot, O., et al. (2016). Detection of invasive mosquito vectors using environmental DNA (eDNA) from water samples. *PLoS ONE* 11:e0162493. doi: 10.1371/journal.pone.0162493

Schrader, C., Schielke, A., Ellerbroek, L., and Johne, R. (2012). PCR inhibitors - occurrence, properties and removal. *J. Appl. Microbiol.* 113, 1014–1026. doi: 10.1111/j.1365-2672.2012.05384.x

Strickler, K. M., Fremier, A. K., and Goldberg, C. S. (2015). Quantifying effects of UV-B, temperature, and pH on eDNA degradation in aquatic microcosms. *Biol. Conserv.* 183, 85–92. doi: 10.1016/j.biocon.2014.11.038

Thomsen, P. F., Kielgast, J., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., et al. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Mol. Ecol.* 21, 2565–2573. doi: 10.1111/j.1365-294X.2011.05418.x

Thomsen, P. F., and Willerslev, E. (2015). Environmental DNA - an emerging tool in conservation for monitoring past and present biodiversity. *Biol. Conserv.* 183, 4–18. doi: 10.1016/j.biocon.2014.11.019

Turner, C. R., Barnes, M. A., Xu, C. C. Y., Jones, S. E., Jerde, C. L., and Lodge, D. M. (2014). Particle size distribution and optimal capture of aqueous macrobial eDNA. *Methods Ecol. Evol.* 5, 676–684. doi: 10.1111/2041-210X.12206

van der Berg, H., and Schaffner, F. (2016). *Training Curriculum Invasive Mosquitoes and (re-)Emerging Vector-Borne Diseases in the WHO European Region*. WHO.

Van der Kuyp, E. (1954). Mosquitoes of the Netherlands Antilles and their hygienic importance. *Studies Fauna Curacao Other Caribbean Islands* 23, 36–114.

Virgilio, M., Backeljau, T., Nevado, B., and De Meyer, M. (2010). Comparative performances of DNA barcoding across insect orders. *BMC Bioinformatics* 11:206. doi: 10.1186/1471-2105-11-206

WHO (2017a). *WHO Malaria Report* 2017.

WHO (2017b). *Vector-Borne Diseases.* Geneva: World Health Organization (WHO*).* Available online at: https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases (accessed January 16, 2018).

Williams, K. E., Huyvaert, K. P., and Piaggio, A. J. (2016). No filters, no fridges : a method for preservation of water samples for eDNA analysis. *BMC Res. Notes* 9:298. doi: 10.1186/s13104-016-2104-5

Wilson, I. G. (1997). Inhibition and facilitation of nucleic acid amplification. *Appl. Environ. Microbiol.* 63, 3741–3751.

Zhou, Y. J., Estes, M. K., Jiang, X., and Metcalf, T. G. (1991). Concentration and detection of hepatitis A virus and rotavirus from shellfish by hybridization tests. *Appl. Environ. Microbiol.* 57, 2963–2968.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 10 July 2019
doi: 10.3389/fevo.2019.00260

How Does eDNA Compare to Traditional Trapping? Detecting Mosquito Communities in South-African Freshwater Ponds

Louie Krol[1,2]*, *Berry Van der Hoorn*[2], *Erin E. Gorsich*[3,4], *Krijn Trimbos*[1], *Peter M. van Bodegom*[1] *and Maarten Schrama*[1,2]

[1] Institute of Environmental Sciences, Leiden University, Leiden, Netherlands, [2] Naturalis Biodiversity Center, Leiden, Netherlands, [3] The Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research, University of Warwick, Coventry, United Kingdom, [4] School of Life Sciences, University of Warwick, Coventry, United Kingdom

Improved biomonitoring of mosquitoes requires an in-depth understanding on occurrences of both vector and non-vector species, in larval, and adult stages. Accurate descriptions of the ecological context in which mosquitoes thrive remain limited, particularly for larval stages. The aim of this study was to develop a mixed-amplicon eDNA approach to assess (i) whether mosquito larval communities of stagnant fresh-water bodies can be detected using a Culicidae-specific primer and (ii) how these results compare to traditional trapping of adult mosquitoes. Results from 32 ponds inside and outside Kruger National Park, South Africa show that our primer detected mosquito eDNA. However, it yielded only a subset of the species found using adult trapping methods. Particularly the less frequent and container-breeding species were not found. Our approach provides the first steps toward an eDNA-based method to assess the entire community of larval-stage mosquitoes. It may thereby overcome current taxonomic hurdles presented by morphological identification of larvae. As such, it holds great promise for biomonitoring and ecological studies of mosquitoes.

Keywords: eDNA, culicidae primer, mosquitoes, vector-borne diseases, biomonitoring, Kruger National Park, mosquito communities

OPEN ACCESS

Edited by:
David Andrew Bohan,
INRA Centre Dijon Bourgogne
Franche-Comté, France

Reviewed by:
Lingxin Chen,
Yantai Institute of Coastal Zone
Research (CAS), China
Gregor Devine,
QIMR Berghofer Medical Research
Institute, Australia

****Correspondence:***
Louie Krol
louie.krol@naturalis.nl

Specialty section:
This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Ecology and Evolution

Received: *09 February 2019*
Accepted: *20 June 2019*
Published: *10 July 2019*

Citation:
Krol L, Van der Hoorn B, Gorsich EE,
Trimbos K, Bodegom PM and
Schrama M (2019) How Does eDNA
Compare to Traditional Trapping?
Detecting Mosquito Communities in
South-African Freshwater Ponds.
Front. Ecol. Evol. 7:260.
doi: 10.3389/fevo.2019.00260

INTRODUCTION

Mosquitoes (order: Diptera, family: Culicidae) are known vectors for a wide variety of pathogens. The mosquito community composition is influenced by myriad of biotic and abiotic factors (e.g., resource availability, predation, temperature) that operate mostly at a local scale (Washburn, 1995; Reiter, 2001; Chase and Knight, 2003; Lafferty, 2009; Young et al., 2017; Schrama et al., 2018; Krol et al., 2019). A more comprehensive understanding of the drivers of mosquito community composition, may facilitate better management of mosquito communities (Beketov and Liess, 2007; Stresman, 2010). However, accurate descriptions of the impacts of these drivers on mosquito community composition remain limited, particularly those regarding larval habitats, largely due to logistic, and taxonomic challenges in identifying and quantifying mosquitoes (Cardoso et al., 2011; Ferraguti et al., 2016; Hunt et al., 2017).

A mixed amplicon metagenomics approach based on environmental DNA (eDNA) potentially allows for the simultaneous, DNA-based identification of an entire species community

(Taberlet et al., 2012), using species-specific traces of DNA (derived from feces, urine, hair, skin, or other cells). These traces can be extracted from various environments (e.g., water, air, soil) (Ficetola et al., 2008) and selected regions of these traces can be amplified using primers that bind to a well-known region on the genome. For animals, a standardized fragment of the mitochondrial cytochrome c oxidase 1 (CO1) is typically used for DNA-barcoding (Hebert et al., 2003). These, amplicons are then linked to known taxa, using a DNA barcoding reference database. A metagenomics approach to study entire mosquito communities would constitute a potentially powerful method to characterize mosquito communities and supplement traditional sampling approaches for determining larval communities.

Two aspects of the mosquito life history render them an ideal species group to develop an eDNA approach. First, mosquito larvae generally occur in high abundances (Hoekman et al., 2016). These high abundances likely increase the local amount of eDNA which increases the detection probability (Elbrecht et al., 2017). Second, the aquatic environment is required by mosquitoes to deposit eggs, for larvae to hatch and grow, and adults to emerge. For the vast majority of mosquito species, all pre-adult life stages are concentrated at the water surface, including adult emergence (Rejmánková et al., 2013). As a result, most eDNA is likely to be present in the upper part of the water column and the periphery of a given water body, an area that is generally most accessible for sampling. A previous study demonstrated that eDNA methods could detect invasive *Aedes* species in freshwater (Schneider et al., 2016). However, this study was limited to the detection of invasive *Aedes* species using a specific-primer targeting the 16S region, for which comprehensive species databases are currently non-existing. Moreover, in the same study, eDNA samples were isolated from small (and relatively clean) container habitats (Schneider et al., 2016), thus begging the question whether the method would also be applicable in larger, turbid environment of temporary ponds, or streams with a more complex community. Given that these challenges exist, it remains unknown if an eDNA approach can also be used to detect entire communities of mosquitoes and differences therein.

The aim of this study was therefore (i) to test if we were able to pick up the mosquito community composition of stagnant freshwater bodies using a customized local barcode database and a tailored family-specific mosquito eDNA primer on CO1, and (ii) to assess how our eDNA results relate to adult trapping methods. To this end, a field study was conducted, inside Kruger National Park (KNP), South-Africa, and in the fringing rural communities, by sampling both eDNA as well as adult mosquitoes in the same water body.

MATERIALS AND METHODS

Setup of the Field Study

A field study was conducted between 18 March and 10 May 2017 in and alongside Kruger National Park, South-Africa, at five locations (**Figure 1**). We sampled four paired locations (Punda Maria, Satara, Skukuza, and Malelane), each with one location inside the park (hereafter "inside") and one location outside the park (hereafter "outside"), and an additional unpaired location inside the park (Shingwedzi). Locations inside the park have far lower population densities of people and livestock and therefore differ in the degree of anthropogenic impact on the ecosystem, including the freshwater habitat (du Toit et al., 2003). At each location, we sampled three to four stagnant water bodies (depending on availability for adequate sampling; **Table S1**), which served as biological replicates. As a result, a total of 32 water bodies were sampled. Using a variety of trapping methods at 32 trapping sites across 4 regions, we trapped 3,918 adult female mosquitoes belonging to 43 species; **Table S1**). For more information about the adult trapping methodology at these water bodies, see **Electronic Appendix S1** and a detailed description in Gorsich et al. (2019). Our eDNA approach assessed the larval community in a single discrete water body at a single point in time whereas the adult trapping method assessed the adult community around a given water body during multiple trapping nights.

eDNA Field Sample Collection

A known challenge with eDNA sampling is that eDNA is not homogenously distributed (Turner et al., 2014). To improve the probability of detection, 30 subsamples per water body, each of 25 mL, were collected with a pipet controller (Integra Bioscience), and pooled into a 750 mL bottle. Each subsample was taken from the upper (0–5 cm) water layer along the shore line, approximately two meters apart. These were immediately stored in a cooling box, transferred to a fridge, and stored at 4°C until filtration within 24-h. The effects of using this method of initial eDNA preservation might have reduced the detection probability (Barnes et al., 2014). However, these effects have not been investigated in this study.

In the lab, eDNA was collected using a 250 mL Sartorius filtering tower (Sartorius-stedim), a mobile vacuum pomp (Datura Molecular Solutions), and 0.22-micron polyethersulfone (PES) filters (250 mL per filter, i.e., three filters per water body), with a diameter of 47 mm (Tisch Scientific) (Turner et al., 2014). To prevent cross-contamination, the Sartorius filtering tower was cleaned between water bodies with bottled water (to remove sand particles) and then soaked for 30 min in a 0.9% bleach solution to degrade remaining DNA. Prior to filtration of field samples, bottled water was used as a negative control to test for cross-contamination. After filtration, the filter was immediately placed in separate 2 mL centrifuge tubes and completely immersed in 900 µL Longmire buffer (100 mM Tris, 100 mM EDTA, 10 mM NaCl, 0.5% SDS, 0.2% sodium azide; Williams et al., 2016) and stored at 4°C in a fridge until DNA extraction. The advantage of Longmire buffer above CTAB buffer is that it preserves DNA at room temperature for at least 2 weeks (Renshaw et al., 2015).

eDNA Extraction

For the extraction and purification of DNA from the field samples, an established phenol-chloroform-isoamylalcohol (PCI) protocol for DNA extraction was used followed by a DNA purification step using the Qiagen DNeasy blood and tissue kit (Renshaw et al., 2015). Unfortunately, this method did not remove all PCR inhibitors, which may have negatively

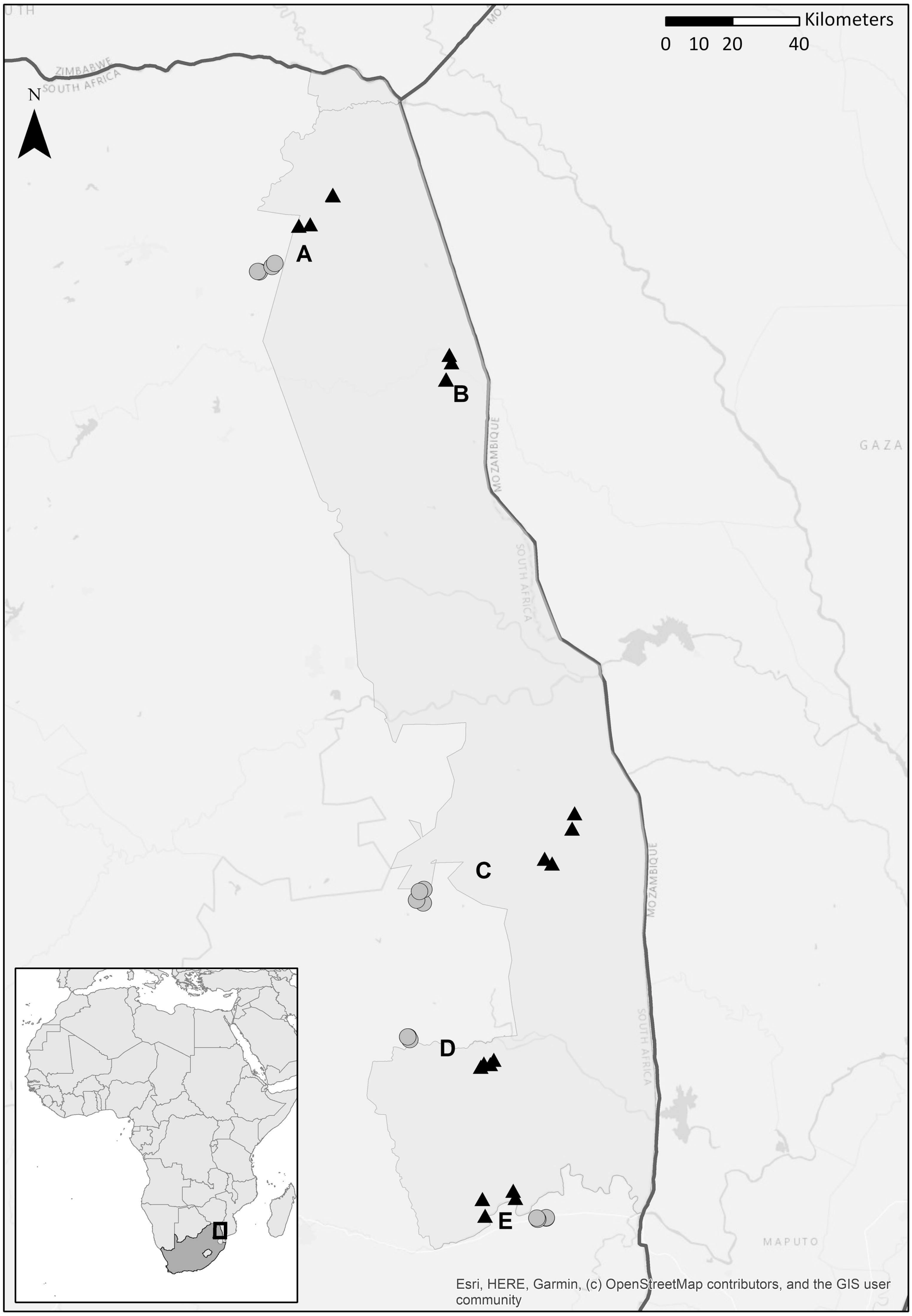

Kilometers
0 10 20 40
N
ZIMBABWE
SOUTH AFRICA
MOZAMBIQUE
GAZA
MAPUTO
A
B
C
D
E
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

FIGURE 1 | Map depicting four paired locations [Punda Maria **(A)**, Malelane **(C)**, Skukuza **(D)**, and Satara **(E)**] and a single unpaired location [Shingwedzi **(B)**], each location has three or four stagnant water bodies. Five locations were situated inside Kruger National Park, South Africa (natural area, depicted with a black triangle); four were situated in the fringing rural communities (rural communities, depicted with a gray dot). See **Table S1** for the coordinates. Courtesy of Maarten van 't Zelfde.

impacted the following steps. We considered dilution of the sample undesirable (this would negatively impact the detection probability of less abundant species), and therefore used the OneStep-96^{TM} PCR inhibitor removal kit (Zymo Research) to remove remaining humic acids and other PCR-inhibitors. For the PCI protocol, the stored PES-filters containing the eDNA were incubated for 10 min at 65°C. After this, 900 μL phenol-chloroform-isoamylalcohol (PCI, 25:24:1) was added

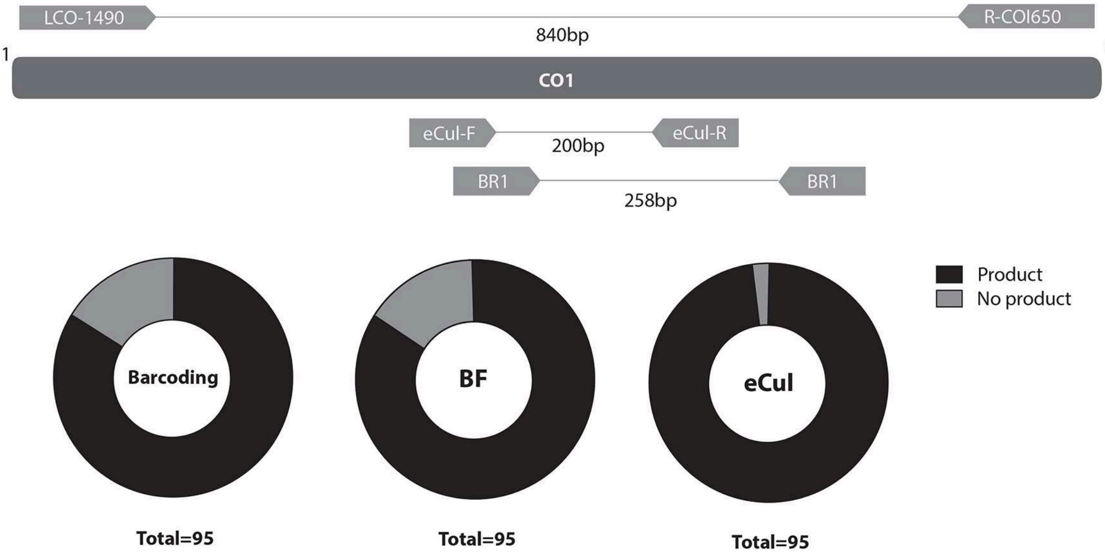


FIGURE 2 | Topography and PCR-efficacy of LCO-1490/R-COI650, BF1/BR1 (BF) and eCul-F/eCul-R (eCul). Primer pairs BF and eCul are shown with their respective position on the CO1-region, as amplified by the LCO-1490/R-COI650 primers. Note that the BF and eCul primers span an overlapping region, highlighting that this is a highly informative region. The black part of the pie chart shows the PCR-efficacy (e.g., the ability of the PCR protocol to generate a PCR fragment), indicating that the primer is able to pick-up the specimen and corresponding species or species complexes. Courtesy of Erik-Jan Bosch.

and vortexed until the PES-filter was completely disintegrated. The 2 mL tubes containing the disintegrated PES-filters were centrifuged at 15,000 × g for 5 min and 700 μL of the aqueous layer was transferred to a fresh 2 mL centrifuge tube. To this mixture, 700 μL chloroform-isoamylalcohol (CI, 24:1) was added, vortexed for 10 s, centrifuged at 15,000 × g for 5 min and 500 μL of the aqueous layer was transferred to a fresh 2 mL centrifuge tube. To this tube, 1.25 mL of ice-cold 96% ethanol and 20 μL 5M NaCl was added and precipitated for 20 min at −20°C, centrifuged at 15,000 × g for 10 min and liquid was decanted. Pellets were left to air-dry until no visible liquid remained (Laramie et al., 2015; Renshaw et al., 2015; Williams et al., 2016). This pellet was resuspended in 180 μL ATL-buffer and DNA extraction was continued with the DNeasy Blood & Tissue kit (Qiagen), for DNA purification, using the manufacture protocol. DNA was finally eluted in 200 μL AE-buffer. From this, 100 μL was transferred to OneStep-96^{TM} PCR inhibitor removal kit plates (Zymo Research). Inhibitors were removed following the protocol of the manufacturer. Samples belonging to the same water body were combined, which made a total of 32 eDNA-samples. The DNA quality and quantity of the eDNA mix-plates were measured with a DropSense96 (Trinean) spectrophotometer.

Construction of a Mosquito DNA Reference Database

For most mosquito species, reference sequences are not available. For example, of the 3,725 species of Culicidae known globally, barcodes are only available for 1,078 species (29%) in the barcoding of life database (BOLD) of which only 716 (19%) are in the public domain (database accessed 12-09-2017). For South Africa, a similar picture arises: 168 species are known within the subfamily Culicinae in South Africa for which only 45 species (26%) have publicly available sequences on BOLD (database accessed 31-08-2018, **Figure S1**). Therefore, during the field survey (**Electronic Appendix S1**), we collected 95 adult mosquito specimens from 38 taxa (**Table S2**) for DNA-barcoding. These morphologically identified specimens were used to construct a customized DNA reference database for the CO1 region, which reduces the probability of non-and misidentification, due to a lack of reference material (Virgilio et al., 2010). DNA extractions on all adult mosquito specimens were performed with the DNeasy Blood & Tissue kit (Qiagen), using the protocol provided by the manufacturer. An 840 bp fragment of the mitochondrial cytochrome c oxidase 1 (CO1) region was amplified using the primers LCO-1490 (forward) (Folmer et al., 1994) and R-COI650 (reverse) (Hemmerter et al., 2007). The reaction mix contained 3 μL 10x CoralLoad PCR-buffer (Qiagen), 0.5 μL 25 mM $MgCl_2$ (Qiagen), 1 μL 10 mg/mL BSA (Life), 0.5 μL 2.5 mM dNTP (Qiagen), 0.25 μL 5U *TaqPol* (Qiagen), 1 μL of 10 pMol/μL of each primer, 5 μL template DNA and 17.75 μL MQ (Ultrapure). The PCR was performed using a Bio-Rad C1000 thermocycler (Bio-Rad Laboratories) the amplification protocol was as follows: 94°C for 3 min, 45 cycles of 94°C for 30 s, 49°C for 45 s and 72°C for 45 s, then finally 72°C for 5 min (Batovska et al., 2016). After PCR, all reactions were visually assessed with an 2% electrophoresis agarose gel, stained with ethidium bromide. The amplicons were sequenced with Sanger sequencing at BaseClear (Leiden, the Netherlands), reads were assembled and annotated with Geneious, version R10 (Kearse et al., 2012).

eDNA Mosquito Specific Primer Design

A mosquito specific environmental DNA primer was designed for the CO1 region, based upon the sequences obtained during this study as well-upon all Culicidae species in BOLD and GenBank which were batch-downloaded (downloaded at 25-06-2017), and clustered into operational taxonomic units (OTUs) with PrimerMiner (Elbrecht and Leese, 2015, 2017a). Other genomic regions were also considered [e.g., CAD, ITS (Reidenbach et al., 2009; Batovska, 2016)]. However, we decided the use CO1 since most species barcodes are collected for the CO1 region. First, multiple degenerated primers were obtained and tested *in-silico* on the all compiled sequences using the Primer3 plug-in for Geneious, version R10 (Rozen and Skaletsky, 1999; Kearse et al., 2012). We selected the optimal primer pair based upon three criteria; primer efficacy, taxonomic resolution, and amplicon size (where smaller amplicons were preferred over larger amplicons because of their higher abundance). For the optimal pair: eCul-F, 5′GGRKCHGGDACWGGDTGAAC-3′ (forward) and eCul-R, 5′-GATCAWACAAATAAAGGTAWTCGATC-3′ (reverse), (hereafter "eCul primers"), 92% (1,050 of 1,135) of OTUs could be picked up, with a taxonomic resolution similar to the taxonomic resolution of the entire CO1 barcoding region, and an amplicon size of 200 bp. Upon primer sequence removal, a barcode of 154 bp remains. We did not further optimize the sequences of the primer.

In-situ and *in-vitro* Primer Evaluation and eDNA Sample PCR Processing

The BF1 and BR1 CO1 general freshwater metabarcoding primers (hereafter, BF primers; Elbrecht and Leese, 2017b) were included as a control for primer evaluation. All three primer pairs, i.e., the barcoding primer pair LCO-1490/R-COI650 and the metabarcoding primer pairs BF and eCul, were tested *in-situ* on the DNA of the 95 mosquito specimens representing 38 taxa (**Figure 2**) to assess amplification efficacy and efficiency.

The *in-vitro* primer evaluation on the 32 eDNA samples included only the eCul and the BF eDNA metabarcoding primers. The reaction mixes for both eDNA-primers contained 3μL 10x CoralLoad PCR-buffer (Qiagen), 0.5 μL 25 mM $MgCl_2$ (Qiagen), 1 μL 10 mg/mL BSA (Life), 0.5 μL 2.5 mM dNTP (Qiagen), 0.25 μL 5 U *TaqPol* (Qiagen), 1 μL of 10 pMol/μL of each primer, 3 μL template DNA and 17.75 μL MQ (Ultrapure). The PCR was performed using a Bio-Rad C1000 thermocycler (Bio-Rad Laboratories). The amplification protocol for the BF primers was as follows: 94°C for 3 min, 45 cycles of 94°C for 30 s, 41°C for 30 s, and 72°C for 20 s, then finally 72°C for 5 min. The amplification protocol for the eCul primers was as follows: 94°C for 3 min, 45 cycles of 94°C for 30 s, 60°C for 30 s, and 72°C for 20 s, then finally 72°C for 5 min. Per sample, a single replicate was used. After PCR, all reactions were visually assessed with an 2% electrophoresis agarose gel, stained with ethidium bromide. PCR-efficacy was assessed by presence-absence of a signal and PCR-efficiency was estimated based upon the relative signal intensity. All extraction and amplification negative controls were negative, indicating that there was no cross contamination.

Next-Generation Sequencing

Library preparation was performed with the NEBNext Fast DNA Library Prep Set for Ion Torrent (New England Biolabs) using only half of the described reaction volume. Amplicon concentration was assessed with capillary electrophoresis using the Qiaxcel (Qiagen) and concentration equalization was performed with the Qiagility pipetting robot (Qiagen). Subsequent analysis was done conform the IonPGM Hi-Q handbook with the Ion OneTouch2 (Life Technologies, Guilford, CT, USA) and BioAnalyzer (Agilent). The eDNA amplicons were sequenced on an Ion-Torrrent Personal Genome Machine (Life Technologies, Guilford, CT, USA) with an Ion 218C chip, at Naturalis Biodiversity Center (Leiden, the Netherlands). The output in FASTQ-format was processed using the Galaxy platform, on the Naturalis Galaxy instance (Blankenberg et al., 2010; Afgan et al., 2016). Initial assessment of the NGS data was performed with the PRINSEQ algorithm (Schmieder and Edwards, 2011). Sequences with a phred-score <20 on the 3′side of the sequence were removed. Only reads that contained both the forward and reverse primer, and those that had a minimal length of 200 bp for the eCuL-primers and 258 bp for the BF-primers, were used for further analysis. The primer sequences were not removed. Operational taxonomic units (OTUs) were generated using the VSEARCH algorithm (threshold: 97% similarity; minimal 2 reads) (Rognes et al., 2016). Only OTUs with >10 reads were used for further analysis. The sequences were queried with the BLAST-tool (Camacho et al., 2009) using the megablast algorithm, against the local copies of BOLD, NCBI/GenBank (downloaded at 14-02-2018) and our custom Culicidae KNP reference database, with a maximum e-value of 0.05, a minimum hit coverage of 80%, a minimum sequence identity of 80% and a maximum of 100 hits per sequences per database. We determined the lowest common ancestor from these BLAST-output files by clustering all hits with a bit-score differences lower than 8% from the best hit. All hits above a threshold for minimum hit coverage of 80% and a minimum sequence identity of 97% were described as a best hit. All LCA-output files were merged with OTU-tables and compared using MS Excel (version 16.14.1, for Macintosh). To test the identification accuracy, a phylogenetic analysis was performed on all OTUs that could be identified to family level (**Figure S2**), where accuracy implied that OTUs belonging to the same family, cluster together. Sequences were aligned by performing multiple sequence alignments, using the MAFFT v.7.222 plug-in for Geneious, version R10 (Katoh et al., 2002; Kearse et al., 2012) with a maximum of 1,000 iterations. The alignment was exported as a Nexus-file to Mesquite (Maddison and Maddison, 2018; Mesquite: a modular system for evolutionary analysis.V.3.31) and exported to the CIPRES science gateway v.3.3 (Miller et al., 2010) as a MrBayes Nexus-input file and run with MrBayes 3.2.2 on XSEDE (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) with the following parameters: (lset nst = 6 rates = invgamma; unlink, statefreq = (all), revmat = (all), shape = (all), pinvar = (all); prset, applyto = (all), ratepr = variable; mcmcp, ngen = 100000000, relburnin = yes, burninfrac = 0.25 printfreq = 1000, samplefreq = 1000, nchains = 4, savebrlens = yes). Posterior summarization and quality control was performed using Tracer V1.7.1 (Rambaut et al., 2018). Cladograms were visualized and annotated with FigTree v1.4.3 (Rambaut, 2012).

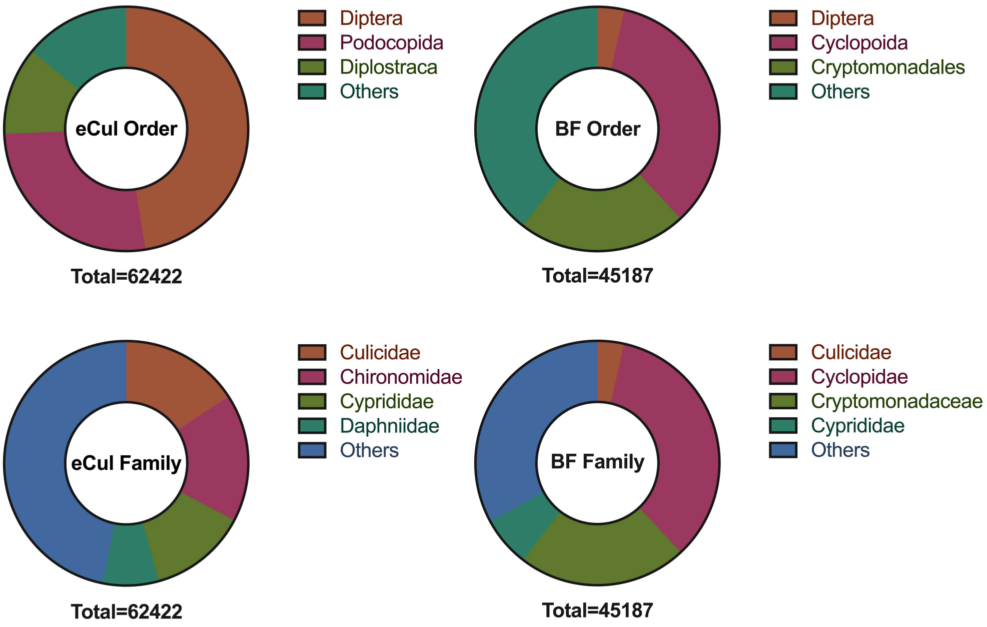


FIGURE 3 | Taxonomic range as picked-up by the BF1/BR1 (BF) eCul-F/eCul-R (eCul) primers, from the eDNA field samples, expressed in number of reads (total number of reads below each pie chart). Note that the number of reads represents the post-PCR distribution, and not the taxa abundance. The BF primers picked up a wide range of unrelated taxa, whereas the eCul primers picked up a narrow range of related taxa, most of which belonged to families within the infraorder Culicomorpha (order: Diptera).

Data Analysis

To compare the results of our eDNA approach with those from adult trapping, first we investigated the taxonomic resolution and accuracy of our morphologically and DNA-based species identification and how these relate. Second, we normalized the different measures of abundance (e.g., reads for eDNA vs. number of observations for adult trapping) for comparison of the different methods.

For the first step, we assessed the morphological identification accuracy and validated the reference database by querying all sequences with the BLAST-tool (Camacho et al., 2009). The megablast algorithm was used against the local copies of BOLD and NCBI/GenBank (downloaded at 14-02-2018) on the Galaxy platform (Blankenberg et al., 2010; Afgan et al., 2016), with a maximum e-value of 0.05, a minimum hit coverage of 70% and sequence identity of more than 97%. To assess if the eCul and BF primers were able to pick up all different taxa present in the reference database, an *in-silico* test was performed. The theoretical amplicons were extracted with Geneious, version R10 (Rozen and Skaletsky, 1999; Kearse et al., 2012) and an *in-silico* mock-community was constructed. This mock-community was analyzed in the same way as the real NGS data to assess the taxonomic resolution of our approach, which facilitated a comparison between both approaches by harmonizing the taxonomic resolution (**Table S3**).

For the second step, for both approaches and both primers (eCul and BF), presence-absence matrices were constructed, by transforming abundance data (number of reads or number of observations) to binary data (e.g., present 1 or absent 0). Both approaches and both primers were then compared with a Bray-Curtis similarity matrix. This also allowed comparing between the locations inside Kruger National Park and the locations in the fringing communities. For each location, the proportion of species found in the adult trapping that also was found with either primer was determined and plotted against the Bray-Curtis similarity results. The resulting graph (**Figure 4**) gives an indication of how the two approaches relate.

To visualize the detection probability (**Figure 5**), we calculated the average relative abundance (according to adult trapping) for each of the species and plotted this against the number of locations where we observed the corresponding larval stage. Using this method, we investigated the detection probability based on adult abundance and visualized potential insufficient sequencing depth. This graph mimics the theoretical abundance of template DNA during sequencing. To test for insufficient sequencing depth (e.g., the ability to detect the less abundant templates), a rarefaction analysis (Heck et al., 1975) of the sequencing data was performed, including only OTUs that were identified as Culicidae according to both the eCul and BF sequencing data (**Figure S3**).

Data analysis was conducted with RStudio (R version 3.2.1; R Core Team, 2016) using the Vegan package (Philip, 2009) and graphed using GraphPad Prism (version 7.00 for Macintosh), GraphPad Software, San Diego California USA.

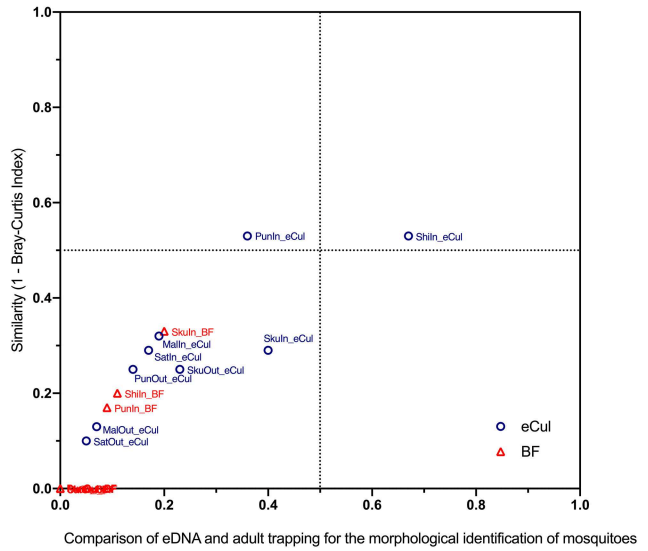


FIGURE 4 | Comparison of eDNA and adult trapping for the morphological identification of mosquitoes. Community similarity between adult trapping and eDNA (expressed as the distance in the Bray Curtis-index) was plotted against the proportion of species that were observed with the eCul or BF primers. In general, the eCul primers performed better than the BF primers, however overall similarity between the two approaches was low.

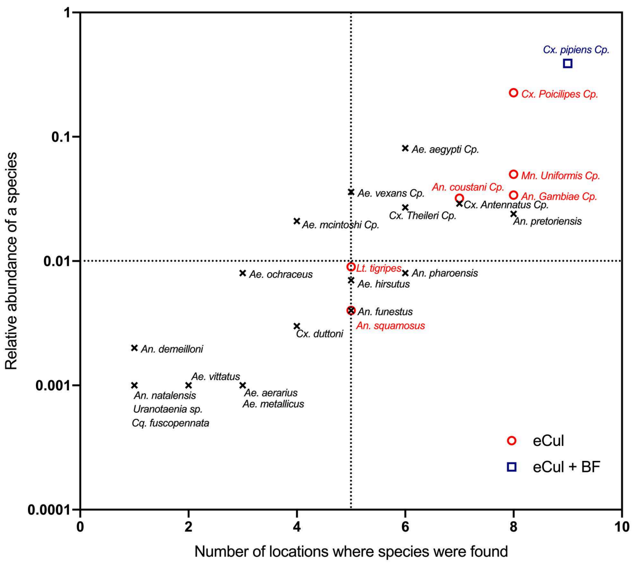


FIGURE 5 | Relationship between the relative species abundance (using eDNA) and the number of locations where these species were found (using adult trapping). Names in red indicate species that were found using both the eCul primer and adult trapping, blue species names indicate those that were detected by eCul and BF primers and adult trapping, black species names indicate species that were only detected using adult trapping. eDNA results always represent a subset of the adult populations. Specimens that could not be identified morphologically to species level were clustered into species complexes, indicated with "Cp".

RESULTS

Primer Evaluation and Comparison

In general, the eCul primer pair performed better than the BF and LCO primer pairs. The *in-silico* primer evaluation of the eCul and BF primers indicated that the taxonomic coverage (i.e., the proportion of species amplified of the target group) and taxonomic discrimination (i.e., the discrimination capacity at the species, genus or family level) were similar to the complete CO1 barcoding region. The adult mosquito specimens ($n = 87$) represented 38 taxa (**Table S2**). Our *in-silico* test based on the CO1 barcoding data also generated 38 OTUs. However, these OTUs do not always correspond to the level of species, either because of a lack of morphological differentiation between closely related species or because DNA was not informative enough to distinguish between species (**Table S3**).

The *in-situ* primer validation of the eCul primers showed the highest PCR-efficacy (97.9%) and efficiency in comparison with the LCO-1490/R-COI650 and BF primers (**Figure 2**). The LCO-1490/R-COI650 and BF primers generated a similar distribution in PCR-efficacy (84.2%) (**Table S2**). However, the signal intensity of the fragments on the gel was weak, indicating that the PCR-efficiency of the BF primers for mosquitoes was overall low, which is in line with previous results (Elbrecht and Leese, 2017b). The eCul and BF-primers are topographically overlapping (**Figure 2**), indicating that this region of CO1 is highly informative. The *in-vitro* primer evaluation showed that the eCul primers picked up more OTUs (expressed in number of reads) belonging to the order of Diptera (47.5%) when compared to the BF primers (3.6%) (**Figure 3**). Besides Diptera, the eCul primers also picked up a number of other taxa within the phylum Arthropoda, most notably Podocopida (26.9%) and Diplostraca (11.6%). Within the Diptera, the eCul primers picked up mainly OTUs belonging to families within the infraorder Culicomorpha of which Chironomidae (35.6%) was the most abundant family and Culicidae the second most abundant family (33.3%). In addition, a substantial proportion of the OTUs within the order Diptera (18.7%) could not be linked to any known family. The BF primers picked up taxa from a much wider taxonomic range, including the orders Cyclopoida (34.5%) and Cryptomonadales (22.4%) and taxa belonging to the kingdoms Fungi (0.7%) and Viridiplantae (1%). Within the Diptera, the BF primer pair picked up mainly OTUs belonging to the Culicidae family (98%), most of which were identified as belonging to the genera Culex (70%) and the remainder as Anopheles (30%).

Comparison Between Adult Trapping and eDNA Sampling

In general, the eCul primer pair performed better than the BF primers, although overall similarity between the two approaches was low. Using adult mosquito trapping, 38 mosquito taxa (species and species complexes) were identified in the field. This number of adult taxa was reduced to 25 with species complexes (**Table S3**). Using eDNA, 34 mosquito OTUs (representing six taxa) were picked-up with the eCul primers, and 10 mosquito OTUs (representing two taxa) were picked-up with the BF primers. A phylogenetic analysis of all sequences indicated that the assigned identities were correct for both the eCul as the

BF data (**Figure S2**). Our results show an OTU overlap of 28% between the eCul data and adult trapping and 4% with the BF data. The eCul primers were able to pick up more than 50% of the species detected with traps at two locations Satara (SatOut; 50%) and Shingwedzi (ShiIn; 67%), and at two locations the similarity was larger than 0.5 (**Figure 4**): Punda Maria (PunIn; 0.53) and Shingwedzi (ShiIn; 0.53). The BF primers were unable to recover more than 20% of the species detected with traps. The species that were detected with eDNA were generally also the most abundant species complexes in the traps (*Culex pipiens* species complex and *Culex poicilipes* species complex; **Table S3**), indicating that the eDNA method was more likely to pick up more abundant and common (found at more locations) species than rare species (**Figure 5**). Moreover, we observed a consistent dissimilarity between mosquito communities inside and outside the park for both sampling techniques, thus providing a first indication that an eDNA approach can be used to detect a shift in mosquito communities [average dissimilarity eDNA (eCul primer): 0.54 (± SE 0.04) vs. average dissimilarity adult trapping: 0.34 (± SE 0.11)]. The rarefaction analysis indicated a lack of sequencing depth for both the eCul and the BF sequencing data (**Figure S3**).

DISCUSSION

In this study we developed an eDNA approach based on a family specific primer and a local CO1 DNA reference database that was able to detect the most abundant species observed with traditional trapping methods. Even so, the eDNA method yielded a much smaller number of species than the adult trapping, which has implications for data interpretation and future work. We elaborate on these challenges in the following paragraphs.

Primer Evaluation and Primer Comparison

In this study, presence of mosquito eDNA in South African ponds was detected using the general macroinvertebrate BF1 primers from Elbrecht and Leese (2017b) and a novel eCul mosquito primer. In contrast to other aquatic macroinvertebrate species, like Odonata (unpublished data), the eDNA assessment of mosquito species proved rather successful, which is likely a result of the particular lifestyle of mosquitoes. Mosquito eDNA concentration in the upper layer of the water may be higher than that of most other freshwater macroinvertebrates, because mosquitoes generally occur in high densities, spend a significant part of their life cycle close to the water surface (where eDNA samples are taken) and produce exuviae at the water surface before emerging as an adult mosquito. For example dragonflies and water beetles are generally less numerous and do not emerge at the water surface (Foster and Soluk, 2004; Jäch and Balke, 2008) and may therefore be more difficult to detect. The detection probability of our eCul primer was higher than when using the general BF primers for freshwater invertebrate taxa (Elbrecht and Leese, 2017b). This result highlights that the use of taxa-specific primers with a narrow taxonomic range greatly improves the probability of detection, and in general aligns with the idea that a primer needs to be suited to a question (Elbrecht and Leese, 2017b). However, not all OTUs could be identified to the species level, either because of shortcomings in the morphological or molecular identification process. The morphological issues were mostly restricted to a number of complexes within the genus *Anopheles* (e.g., *An. coustani s.l., An. gambiae s.l.*) which is a well-known problem (Gillies and Coetzee, 1987) and is normally resolved using PCR-based identification (Fanello et al., 2002). Challenges regarding the molecular identification were found for a number of species complexes within the genera *Culex* and *Aedes* (**Table S3**), which indicates that, by itself, the CO1 barcoding region might not be informative enough to differentiate between closely related species within these complexes. One way to overcome this issue in future studies is by targeting more than one regions on the mosquito mitochondrial genome [e.g., CAD, ITS, 16S (White et al., 1990; Reidenbach et al., 2009; Batovska, 2016; Schneider et al., 2016)]. Nevertheless, our results provide a proof-of-principle that eDNA-based methods hold great promise when it comes to using it for the detection of mosquito species communities across a range of freshwater habitats.

Comparison Between Adult Trapping and eDNA

Our eDNA approach provided only a subset of the species found when using traditional trapping methods and identification on morphology, with an overall moderate similarity between the two approaches. Particularly species with lower abundance were not readily retrieved using the eDNA method. Results of both adult trapping and eDNA methods indicate that the mosquito community composition differs between locations inside and outside Kruger National park—which may be related to the environmental differences resulting from the higher population densities outside compared to inside the park. In our study, the eDNA-based estimation of the community resulted in a greater dissimilarity between inside and outside locations compared to adult trapping. Possibly, this is due to missing the rarer species that occur both inside and outside Kruger National Park.

There are four main explanations why the community similarity between the two approaches differ. First, there is a possibility that the eDNA based approach did detect the majority of the mosquito species in the sampled water bodies, considering that our eDNA approach assesses the larval community *inside* a discrete water body whereas the adult trapping method assesses the adult community *around* a given water body. This suggests that the difference in community composition between eDNA and traps may partly be the result of adult mosquitoes being lured toward the traps from nearby breeding sites (e.g., nearby tires, buckets, and other artificial habitats). This view is strengthened by the absence of *Ae. aegypti* in the eDNA dataset, which was abundantly present in the adult traps (**Figure 5**). This species is known to breed almost exclusively in artificial habitats like plastic containers and car tires (Simard et al., 2005). Although this may partly explain the difference between the methods, there is no direct way to test this hypothesis, because mosquito larval communities were not sampled directly. Setting up controlled experiments with mixtures of species with varying abundance will likely resolve this issue.

Second, our sampling strategy might not be sufficient (30 spatially distributes subsamples of 25 mL per water body), decreasing the probability of detection, since it is known that eDNA is heterogeneously distributed (Nathan et al., 2014). Setting up experiments with more subsamples that are not mixed will likely resolve this issue.

Third, the effects of PCR-induced biases (e.g., sequencing depth and primer bias) can decrease the probability of detection (Elbrecht and Leese, 2015, 2017b), particularly for species with a relatively low abundance (Elbrecht et al., 2017). Indeed, our eDNA results suggest that our results may be suffering from a lack of sequencing depth (**Figure S3**). An inadequate sequencing depth can be one of the causes of missed species because highly abundant reads of non-target species might mask low-abundance sequences (Adams et al., 2013). Species belonging to more abundant genera, like *Culex* and *Mansonia* were readily detected whereas species belonging to less abundant genera like *Aedes* and *Anopheles* were less likely to be picked up (**Figure 5**). This picture was even more striking for the general freshwater macroinvertebrate BF primer, which picked up only *Culex* species belonging to the *Culex pipiens species complex* (**Figure 5**) with a very high adult abundance. These results therefore suggest that insufficient sequencing depth (**Figure S3**) might have reduced the probability of detection of less abundant species. Also, our results might suffer from the effects of overamplification and stochastic effects inherent to mixed amplicon PCR. The effects of overamplification can be mitigated by reducing the number of cycles and the stochastic effects by increasing the number of replicates from one to twelve. If the controlled experiments proposed above indeed show that we are missing the less abundant species, future work should address this gap by adopting methods that produce higher number of reads or by masking highly abundant species.

The effects of PCR induced biases (Elbrecht and Leese, 2015, 2017b) and PCR inhibitors [which environmental samples often contain (Jane et al., 2015)] were not assessed during this study. It is known that such biases and inhibitors may negatively affect the probability of detection, particularly for species with a relatively low abundance and/or biomass (Elbrecht and Leese, 2015; Elbrecht et al., 2017). This is further complicated by the unknown persistence of mosquito eDNA under a range of ecological conditions [e.g., biotic and abiotic degradation (Barnes et al., 2014; Strickler et al., 2015)] in a system where eDNA is spatially heterogeneously distributed (Nathan et al., 2014).

More work is therefore required to link the original amount of template DNA (pre-PCR) and the distribution of reads (post-PCR). The relative abundance of template DNA has to be included, for the use of more comprehensive indices of diversity. Current quantification methods (e.g., qPCR and ddPCR) are unsuitable for mixed amplicon metagenomic approaches (Doi et al., 2015). Fusion primers tagged with unique molecular identifiers (UMIs) (Kivioja et al., 2012) may provide the tools needed to address the effects of primer bias and PCR inhibition in eDNA and metabarcoding samples.

Fourth, the limited availability of a comprehensive and reliable reference database might reduce the detection probability. Accurate species-level identification of mosquitoes can be difficult, often leading to low taxonomic resolution, or misidentifications (Haase et al., 2010). This in turn decreases the accuracy of DNA-based approaches. During this study, not all adult mosquito specimens could be identified to species level (**Table S2**) and were clustered into species complexes (**Table S3**). The *in-silico* test to assess if all taxa present in the mock-community could be picked-up, yielded 38 OTUs, although not all 38 taxa could be identified. This might be due to insufficient taxonomic resolution of the CO1 target region or misidentification of the adult specimen (**Table S2**). There is likely room for improvement because it is unlikely that all cryptic species were included in our reference library. Furthermore, the underlying classification and phylogenic relationship of mosquitoes remains largely unresolved (Harbach, 2007; Reidenbach et al., 2009; Wilkerson et al., 2015). This highlights the need for taxonomic expertise to properly describe species based upon morphological and molecular evidence (Chan et al., 2014). More work is therefore needed to resolve the underlying classification and phylogeny of mosquitoes, in order to construct a comprehensive and reliable reference database.

Nevertheless, despite the unknowns listed above, the eDNA method detected the most abundant species, thus indicating its potential value in addition to the traditional sampling techniques, and, as such, provide a meaningful addition to the existing tool kit.

CONCLUDING REMARKS

This is the first study that applies an eDNA approach to determine the community composition of mosquitoes, based on water samples collected in the field. As such, it provides a proof-of-concept that eDNA-based methods can be used to better understand mosquito larval ecology and provides promising steps toward an eDNA-based biomonitoring of mosquito species communities. The comparison between adult and larval communities shows that less abundant adult species were not detected using our metabarcoding method. More research is needed to evaluate whether this mismatch is due to an overrepresentation of species from other nearby breeding sites or due to an incomplete eDNA-based survey. To improve differentiation between closely related species, eDNA-based surveys of the complete mosquito community require the identification of an additional informative region(s) on the mosquito genome [e.g., CAD, ITS or 16S (Reidenbach et al., 2009; Batovska, 2016; Schneider et al., 2016)]. Nevertheless, our results highlight that environmental DNA holds the potential to assess the larval community composition of mosquitoes quickly and reliably, provided that (i) samples are taken in accordance with the ecological context (i.e., life history traits), and (ii) a comprehensive and reliable local reference database and suitable primers are available. On the short term, given its ability to determine mosquito community composition based on larvae, eDNA is a promising complementary tool for monitoring species communities alongside existing adult and larval trapping methods.

AUTHOR CONTRIBUTIONS

LK, PB, and MS conceived the idea for this study. LK carried out the measurements together with MS and EG. LK carried out the molecular and bioinformatics work with the help of BV. LK wrote the first draft together with MS. PB, KT, EG, and BV commented on the setup and assisted during writing of the paper.

FUNDING

This study was supported by the Leiden University Fund (grant L17005) to LK, the Gratama Fund (grant 2016.08) which was awarded to MS, and a grant from the Gieskes Strijbis Foundation for the DNA Waterscan project of Naturalis Biodiversity Center.

ACKNOWLEDGMENTS

We want to thank Brianna Beechler for assisting with setting up the field study and Vicky Beckers, Nina Haver, Gijs van Nes, Karabo Moloi, Nondumiso Myataza, Skumbuzo Mdletshe, Herman Ntimane, and Thomas Mbokota for excellent help in the field. We thank Elza Duijm, Roland Butôt, Marcel Eurlings, and Ayesha Hassim for assistance in the laboratory. Scientific services and Veterinary wildlife services are gratefully acknowledged for accommodation and the permission to carry out the work at their laboratory facilities in Skukuza, KNP. Danny Govender was gratefully acknowledged for the permission to carry out this study in Kruger National Park.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2019.00260/full#supplementary-material

REFERENCES

Adams, R. I., Amend, A. S., Taylor, J. W., and Bruns, T. D. (2013). A unique signal distorts the perception of species richness and composition in high-throughput sequencing surveys of microbial communities: a case study of fungi in indoor dust. *Microb. Ecol.* 66, 735–741. doi: 10.1007/s00248-013-0266-4

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Cech, M., et al. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 44, W3–W10. doi: 10.1093/nar/gkw343

Barnes, M. A., Turner, C. R., Jerde, C. L., Renshaw, M. A., Chadderton, W. L., and Lodge, D. M. (2014). Environmental conditions influence eDNA persistence in aquatic systems (Supplemental Information). *Environ. Sci. Technol.* 48, 1819–1827. doi: 10.1021/es404734p

Batovska, J. (2016). Using next-generation sequencing for DNA barcoding: capturing allelic variation in ITS2. 7, 19–29. doi: 10.1534/g3.116.036145

Batovska, J., Blacket, M. J., Brown, K., and Lynch, S. E. (2016). Molecular identification of mosquitoes (Diptera: Culicidae) in southeastern Australia. *Ecol. Evol.* 6, 3001–3011. doi: 10.1002/ece3.2095

Beketov, M. A., and Liess, M. (2007). Predation risk perception and food scarcity induce alterations of life-cycle traits of the mosquito Culex pipiens. *Ecol. Entomol.* 32, 405–410. doi: 10.1111/j.1365-2311.2007.00889.x

Blankenberg, D., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Taylor, J., et al. (2010). Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Molecul. Biol.* Chapter 19:Unit 19.10.1–21. doi: 10.1002/0471142727.mb1910s89

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinform.* 10:421. doi: 10.1186/1471-2105-10-421

Cardoso, P., Erwin, T. L., Borges, P. A. V., and New, T. R. (2011). The seven impediments in invertebrate conservation and how to overcome them. *Biol. Conserv.* 144, 2647–2655. doi: 10.1016/j.biocon.2011.07.024

Chan, A., Chiang, L.-P., Hapuarachchi, H., Tan, C.-H., Pang, S.-C., Lee, R., et al. (2014). DNA barcoding: complementing morphological identification of mosquito species in Singapore. *Parasit. Vector.* 7:569. doi: 10.1186/s13071-014-0569-4

Chase, J. M., and Knight, T. M. (2003). Drought-induced mosquito outbreaks in wetlands. *Ecol. Lett.* 6, 1017–1024. doi: 10.1046/j.1461-0248.2003.00533.x

Doi, H., Uchii, K., Takahara, T., Matsuhashi, S., Yamanaka, H., and Minamoto, T. (2015). Use of droplet digital PCR for estimation of fish abundance and biomass in environmental DNA surveys. *PLoS ONE.* 10:e0122763. doi: 10.1371/journal.pone.0122763

du Toit, J., Rogers, K., and Biggs, C. H. (2003). *The Kruger Experience: Ecology and Management of Savanna Heterogeneity*. Washington, DC: Island Press.

Elbrecht, V., and 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* 10:324. doi: 10.1371/journal.pone.0130324

Elbrecht, V., and Leese, F. (2017a). PrimerMiner: an r package for development and in silico validation of DNA metabarcoding primers. *Methods Ecol. Evol.* 8, 622–626. doi: 10.1111/2041-210X.12687

Elbrecht, V., and Leese, F. (2017b). Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. *Front. Environ. Sci.* 5:11. doi: 10.3389/fenvs.2017.00011

Elbrecht, V., Peinert, B., and Leese, F. (2017). Sorting things out: assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecol. Evol.* 7, 6918–6926. doi: 10.1002/ece3.3192

Fanello, C., Santolamazza, F., and Della Torre, A. (2002). Simultaneous identification of species and molecular forms of the Anopheles gambiae complex by PCR-RFLP. *Med. Vet. Entomol.* 16, 461–464. doi: 10.1046/j.1365-2915.2002.00393.x

Ferraguti, M., Martínez-de la Puente, J., Roiz, D., Ruiz, S., Soriguer, R., and Figuerola, J. (2016). Effects of landscape anthropization on mosquito community composition and abundance. *Sci. Rep.* 6:29002. doi: 10.1038/srep29002

Ficetola, G. F., Miaud, C., Pompanon, F., and Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biol. Lett.* 4, 423–425. doi: 10.1098/rsbl.2008.0118

Folmer, O., Black, M., Hoeh, W., Lutz, R., and Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecul. Mar. Biol. Biotechnol.* 3, 294–299. doi: 10.1371/journal.pone.0013102

Foster, S. E., and Soluk, D. A. (2004). Evaluating exuvia collection as a management tool for the federally endangered Hine's emerald dragonfly, *Somatochlora hineana* Williamson (Odonata: Cordulidae). *Biol. Conserv.* 118, 15–20. doi: 10.1016/j.biocon.2003.06.002

Gillies, M. T., and Coetzee, M. (1987). A supplement to the Anophelinae of Africa south of the Sahara (Afrotropical Region). *Publicat. South African Inst. Med. Res.* 55, 1–143.

Gorsich, E. E., Beechler, B. R., Van Bodegom, P., and Govender, D. (2019). A comparative assessment of adult mosquito trapping methods to estimate spatial patterns of abundance and community composition in southern Africa. *bioRxiv [Preprint]*. doi: 10.1101/633552

Haase, P., Pauls, S. U., Schindehütte, K., and Sundermann, A. (2010). First audit of macroinvertebrate samples from an EU water framework directive monitoring program: human error greatly lowers precision of assessment results. *J. North Am. Benthol. Soc.* 29, 1279–1291. doi: 10.1899/09-183.1

Harbach, R. E. (2007). The culicidae (Diptera): a review of taxonomy, classification and phylogeny. *Zootaxa* 638, 591–638. doi: 10.1017/CBO9781107415324.004

Hebert, P. D., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proc. Biol. Sci.* 270, 313–321. doi: 10.1098/rspb.2002.2218

Heck, K. L., van Belle, G., and Simberloff, D. (1975). Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology* 56, 1459–1461. doi: 10.2307/1934716

Hemmerter, S., Šlapeta, J., van den Hurk, A. F., Cooper, R. D., Whelan, P. I., Russell, R. C., et al. (2007). A curious coincidence: mosquito biodiversity and the limits of the Japanese encephalitis virus in Australasia. *BMC Evolut. Biol.* 7:100. doi: 10.1186/1471-2148-7-100

Hoekman, D., Springer, Y. P., Barker, C. M., Barrera, R., Blackmore, M. S., Bradshaw, W. E., et al. (2016). Design for mosquito abundance, diversity, and phenology sampling within the National Ecological Observatory Network. *Ecosphere* 7:e01320. doi: 10.1002/ecs2.1320

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755. doi: 10.1093/bioinformatics/17.8.754

Hunt, S. K., Galatowitsch, M. L., and McIntosh, A. R. (2017). Interactive effects of land use, temperature, and predators determine native and invasive mosquito distributions. *Freshw. Biol.* 62, 1564–1577. doi: 10.1111/fwb.12967

Jäch, M. A., and Balke, M. (2008). Global diversity of water beetles (Coleoptera) in freshwater. *Hydrobiologia* 595, 419–442. doi: 10.1007/s10750-007-9117-y

Jane, S. F., Wilcox, T. M., Mckelvey, K. S., Young, M. K., Schwartz, M. K., Lowe, W. H., et al. (2015). Distance, flow and PCR inhibition: EDNA dynamics in two headwater streams. *Molecul. Ecol. Resour.* 15, 216–227. doi: 10.1111/1755-0998.12285

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., et al. (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods* 9, 72–74. doi: 10.1038/nmeth.1778

Krol, L., Gorsich, E. E., Hunting, E. R., Govender, D., van Bodegom, P. M., and Schrama, M. (2019). Eutrophication governs predator-prey interactions and temperature effects in Aedes aegypti populations. *Parasit. Vect.* 12:179. doi: 10.1186/s13071-019-3431-x

Lafferty, K. D. (2009). The ecology of climate change and infectious diseases. *Ecology* 90, 888–900. doi: 10.1890/09-1656.1

Laramie, M. B., Pilliod, D. S., Goldberg, C. S., and Strickler, K. M. (2015). "Environmental DNA sampling protocol - filtering water to capture DNA from aquatic organisms," in *U.S Geological Survey Techniques and Methods* (Reston, VA: U.S. Geological Survey), 15. doi: 10.3133/TM2A13

Maddison, W. P., and Maddison, D. R. (2018). *Mesquite: A Modular System for Evolutionary Analysis. Version 3.51*. Available online at: http://www.mesquiteproject.org

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). "Creating the CIPRES science gateway for inference of large phylogenetic trees," in *2010 Gateway Computing Environments Workshop (GCE)*, 1–8. doi: 10.1109/GCE.2010.5676129

Nathan, L. M., Simmons, M., Wegleitner, B. J., Jerde, C. L., and Mahon, A. R. (2014). Quantifying environmental DNA signals for aquatic invasive species across multiple detection platforms. *Environ. Sci. Technol.* 48, 12800–12806. doi: 10.1021/es5034052

Philip, D. (2009). VEGAN, a package of R functions for community ecology. *J. Veget. Sci.* 14, 927–930. doi: 10.1111/j.1654-1103.2003.tb02228.x

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Rambaut, A. and Drummond, A. J. (2012). *FigTree v1. 4. 3*.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in bayesian phylogenetics using tracer 1.7. *Systemat. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Reidenbach, K. R., Cook, S., Bertone, M. A., Harbach, R. E., Wiegmann, B. M., and Besansky, N. J. (2009). Phylogenetic analysis and temporal diversification of mosquitoes (Diptera: Culicidae) based on nuclear genes and morphology. *BMC Evolut. Biol.* 9:298. doi: 10.1186/1471-2148-9-298

Reiter, P. (2001). Climate change and mosquito-borne disease. *Environ. Health Perspect.* 109, 141–161. doi: 10.2307/3434853

Rejmánková, E., Grieco, J., Achee, N., and Roberts, D. R. (2013). *Ecology of Larval Habitats*. Anopheles Mosquitoes - New Insights into Malaria Vectors.

Renshaw, M. A., Olds, B. P., Jerde, C. L., Mcveigh, M. M., and Lodge, D. M. (2015). The room temperature preservation of filtered environmental DNA samples and assimilation into a phenol-chloroform-isoamyl alcohol DNA extraction. *Molecul. Ecol. Resour.* 15, 168–176. doi: 10.1111/1755-0998.12281

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *Peer J.* 4:e2584. doi: 10.7717/peerj.2584

Ronquist, F., and Huelsenbeck, J. P. (2003). MrBayes 3: bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574. doi: 10.1093/bioinformatics/btg180

Rozen, S., and Skaletsky, H. (1999). "Primer3 on the WWW for general users and for biologist programmers," in *Bioinformatics Methods and Protocols*, eds S. Misener, and S. A. Krawetz (Totowa, NJ: Humana Press), 365–386. doi: 10.1385/1-59259-192-2:365

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi: 10.1093/bioinformatics/btr026

Schneider, J., Valentini, A., Dejean, T., Montarsi, F., Taberlet, P., Glaizot, O., et al. (2016). Detection of invasive mosquito vectors using environmental DNA (eDNA) from water samples. *PLoS ONE* 11:e0162493. doi: 10.1371/journal.pone.0162493

Schrama, M., Gorsich, E. E., Hunting, E. R., Barmentlo, S. H., Beechler, B., and Bodegom, P. M. (2018). Eutrophication and predator presence overrule the effects of temperature on mosquito survival and development. *PLoS Neglected Trop. Dis.* 12, 1–13. doi: 10.1371/journal.pntd.0006354

Simard, F., Nchoutpouen, E., Toto, J. C., and Fontenille, D. (2005). Geographic distribution and breeding site preference of *Aedes albopictus* and *Aedes aegypti* (Diptera: culicidae) in Cameroon, Central Africa. *J. Med. Entomol.* 42, 726–731. doi: 10.1603/0022-2585(2005)042

Stresman, G. H. (2010). Beyond temperature and precipitation: ecological risk factors that modify malaria transmission. *Acta Trop*. 116, 162–72. doi: 10.1016/j.actatropica.2010.08.005

Strickler, K. M., Fremier, A. K., and Goldberg, C. S. (2015). Quantifying effects of UV-B, temperature, and pH on eDNA degradation in aquatic microcosms. *Biol. Conserv.* 183, 85–92. doi: 10.1016/j.biocon.2014.11.038

Taberlet, P., Coissac, E., Hajibabaei, M., and Rieseberg, L. H. (2012). Environmental DNA. *Molecul. Ecol.* 21, 1789–1793. doi: 10.1111/j.1365-294X.2012.05542.x

Turner, C. R., Barnes, M. A., Xu, C. C. Y., Jones, S. E., Jerde, C. L., and Lodge, D. M. (2014). Particle size distribution and optimal capture of aqueous macrobial eDNA. *Methods Ecol. Evol.* 5, 676–684. doi: 10.1111/2041-210X.12206

Virgilio, M., Backeljau, T., Nevado, B., and De Meyer, M. (2010). Comparative performances of DNA barcoding across insect orders. *BMC Bioinform.* 11:206. doi: 10.1186/1471-2105-11-206

Washburn, J. O. (1995). Regulatory factors affecting larval mosquito populations in container and pool habitats: implications for biological control. *J. Am. Mosquito Control Assoc.* 11(2 Pt 2), 279–283.

White, T., Bruns, T., Lee, S., and Taylor, J. (1990). "Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics," in *PCR Protocols A Guide to Methods and Applications*, eds M. A. Innis, D. H. Gelfand, J. J. Sninski, and T. J. White (San Diego, CA: Academic Press), 315–322.

Wilkerson, R. C., Linton, Y. M., Fonseca, D. M., Schultz, T. R., Price, D. C., and Strickman, D. A. (2015). Making mosquito taxonomy useful: a stable classification of tribe Aedini that balances utility with current knowledge of evolutionary relationships. *PLoS ONE* 10:e0133602. doi: 10.1371/journal.pone.0133602

Williams, K. E., Huyvaert, K. P., and Piaggio, A. J. (2016). No filters, no fridges: a method for preservation of water samples for eDNA analysis. *BMC Res. Notes* 9:298. doi: 10.1186/s13104-016-2104-5

Young, H. S., Wood, C. L., Kilpatrick, A. M., Lafferty, K. D., Nunn, C. L., and Vincent, J. R. (2017). Conservation, biodiversity and infectious disease: scientific evidence and policy implications. *Philosoph. Transac. R. Soc. Biol. Sci.* 372, 5–8. doi: 10.1098/rstb.2016.0124

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

PERSPECTIVE
published: 06 September 2019
doi: 10.3389/fevo.2019.00337

Gaps in DNA-Based Biomonitoring Across the Globe

*Katie M. McGee, Chloe V. Robinson and Mehrdad Hajibabaei**

Centre for Biodiversity Genomics, Department of Integrative Biology, University of Guelph, Guelph, ON, Canada

DNA-based methodology has proven to be a vital tool for ecosystem assessment and monitoring. Increasingly, high-throughput approaches such as DNA metabarcoding are being used to address more complex questions, including ecological network analyses through machine learning. Despite the technological advances which allow for such questions to be posed, there remains inherent limitations in studies utilizing DNA metabarcoding, referring to environmental sample type targeted, geographical coverage and lack of standardized field and laboratory procedures. Additionally, DNA reference databases are lacking information from taxa, resulting in unidentified sequences, and underrepresentation of some taxa. These issues need to be addressed to enable a more representative approach to ecosystem monitoring to allow for detection and monitoring of global ecosystem change.

Keywords: biomonitoring, DNA metabarcoding, next-generation sequencing, biodiversity, global, ecosystem, biomes

OPEN ACCESS

Edited by:
David Andrew Bohan,
INRA Centre Dijon Bourgogne Franche-Comté, France

Reviewed by:
Florian Leese,
University of Duisburg-Essen, Germany
Xavier Pochon,
Cawthron Institute, New Zealand
Tristan Cordier,
Université de Genève, Switzerland

****Correspondence:***
Mehrdad Hajibabaei
hajibabaei@gmail.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

Received: *04 June 2019*
Accepted: *23 August 2019*
Published: *06 September 2019*

Citation:
McGee KM, Robinson CV and Hajibabaei M (2019) Gaps in DNA-Based Biomonitoring Across the Globe. Front. Ecol. Evol. 7:337. doi: 10.3389/fevo.2019.00337

To better determine the global effects that the changing climate and anthropogenic damage have on the planets' ecosystems requires a more complete understanding of the global biodiversity than currently exists. However, this has been extremely difficult to ascertain and standardize due to the large number of taxa and the diversity of different geographic localities. More confounding is the reality that these natural and man-made changes are increasingly reshaping the global biodiversity and the associated ecosystem processes and services they provide (Díaz et al., 2015; Bohan et al., 2017). Unfortunately, to date, scientists studying the connections between biodiversity and ecosystem change in specific ecosystems have been poorly equipped to measure these relationships, and have tended to rely on the taxonomic identity and biomonitoring indicators collected from other, and perhaps distant areas, which may or may not be appropriate or accurate choices (Bohan et al., 2017).

DNA metabarcoding utilizes bulk samples such as soil, water, and benthos to extract DNA (termed environmental DNA, eDNA) and generate sequence data for standard taxonomic marker genes (e.g., DNA barcodes) via high-throughput sequencing (Porter and Hajibabaei, 2018b). By streamlining and scaling-up biodiversity data generated, DNA metabarcoding provides the ability to increase the amounts of assessment of the status of biodiversity associated with ecosystem change that can occur across a wide range of global ecosystems (Ruppert et al., 2019). The approach is cost-effective, easy to implement, and provides a robust and comprehensive dataset of taxa from environmental samples, making DNA metabarcoding an important tool of choice for future fundamental research and large-scale biodiversity monitoring programs (Zinger et al., 2019). Moreover, DNA metabarcoding provides an important component to be used with the ecological network analyses and machine learning algorithms that are rapidly advancing to enhance the capacity to detect global ecosystem change through biodiversity assessment (Bohan et al., 2017; Cordier et al., 2019). The complex relationships between changes in nodes and links, and their impact on ecosystem functions should be understood at the network level if we are to develop more robust biomonitoring (Bohan et al., 2017). That said, there are still various barriers that need to be overcome in order to accurately and effectively detect such global ecosystem change, regardless of how quickly these technologies and analyses advance.

DNA metabarcoding has been used to assess eukaryotic and prokaryotic communities, to answer ecological questions such as identifying soil microbiome communities associated with nitrogen-fixing tree species in secondary tropical forests (McGee et al., 2019), assessing bioindicators of river health through macroinvertebrate biomonitoring (Hajibabaei et al., 2011; Dowle et al., 2016) and investigating the effects of oil spills on coastal biodiversity (Xie et al., 2018). Robust experimental design is vital to ensure reproducibility and the ability to draw sound ecological conclusions from the data (Fahner et al., 2018; Zinger et al., 2019). Type I and Type II errors are common with DNA-based biomonitoring, and to overcome this, firstly the sampling design needs to be effective at capturing the full taxonomic diversity or the ecological processes being investigated (Zinger et al., 2019). Secondly, the laboratory and bioinformatic workflow should be optimized to reduce sampling, extraction, amplification, or sequencing bias (Fahner et al., 2018; Ruppert et al., 2019; Zinger et al., 2019). For detecting biodiversity changes, both the taxonomic reference database (for taxonomic annotation of sequences), and environmental sample type (as a proxy for biodiversity) need to be efficient and suitable for detection of target taxa (Ruppert et al., 2019). Geographic variability of environmental sample types also needs to be taken into consideration, to provide the most inclusive representation of taxa, which is vital for detecting biodiversity change within different ecosystems.

Ecological network analyses are becoming an increasingly popular approach to study how ecosystems respond to change and the functional implications of these responses. Typically, network analyses are able to link together species indicators, gathered via DNA metabarcoding methods and others, and functions/interactions to represent a totality of nodes as an ecosystem model (Bohan et al., 2017; Laroche et al., 2018). Network structures can elucidate environmental shifts from stable ecosystem states (Beisner et al., 2003; Bohan et al., 2017; Derocles et al., 2018) through changes that occur in species composition and manifest in an ecological network. These ecological network analyses can potentially explain and possibly predict why stable states in ecology can persist over a period of time (Carpenter et al., 2001; Scheffer et al., 2001; Beisner et al., 2003; Bohan et al., 2017), in order to aid advancements in global biomonitoring. Network analyses, combined with machine learning algorithms, provide a standardized and sensitive method at a high resolution to foster a general understanding of the current state of ecosystem function across the globe (Vacher et al., 2016; Bohan et al., 2017; Derocles et al., 2018).

However, even if we advance the technologies behind these network and machine learning methods, the reference databases for taxonomic identification, sample type, and geographical location remain as the most influential limitations to advancing an understanding of detecting global ecosystem change. Next-generation biomonitoring involves the isolation of DNA from samples including freshwater (Valentini et al., 2016; Muha et al., 2017; Harper et al., 2019), salt/brackish water (Lobo et al., 2017; Aylagas et al., 2018; Hansen et al., 2018), benthos (Hajibabaei et al., 2011; Turner et al., 2015; Aylagas et al., 2016; Robinson et al., 2019; Salonen et al., 2019), soil (Andersen et al., 2012; Yoccoz et al., 2012; Fahner et al., 2016; McGee et al., 2019), permafrost (Bellemain et al., 2013; Zielińska et al., 2017; Zimmermann et al., 2017), passive biomass collection efforts such as malaise traps (Morinière et al., 2016; Adamowicz et al., 2019), and more recently air (Kraaijeveld et al., 2015; Ferguson et al., 2019). Within these different types of environmental samples, there are taxa which are either unique to a particular sample type or can be detected across a breadth of environments, which ultimately influences the ecological questions that can be addressed with each type of environmental sample (Ruppert et al., 2019). In a brief, robust Web of Knowledge hit search from the last 5 years (2015–2019), using various search terms to show where various sample types are popularly collected, or sample type (i.e., water, soil, benthos), suggested that samples may be substantially lacking in various geographical regions. Overall, tropic* returned the greatest number of searches for environmental DNA/eDNA/metabarcoding studies ($n = 319$), followed by Arctic/Antarctic/polar ($n = 262$), and then temperate ($n = 188$; **Table 1**). What this brief hit search does not highlight is the lack of geographic coverage within some geographic regions. For example, despite temperate returning the fewest searches for environmental DNA/eDNA/metabarcoding studies, the range of sample localities is vaster than for both the tropics and arctic/Antarctic/polar regions. Studies returned for the temperate region include localities such as Asia, United Kingdom, Canada and France, whereas for Antarctic for example, the studies are concentrated around remote field stations on the Antarctic peninsula. In terms of sample type, soil environmental DNA/eDNA studies return more searches in temperate locations, whereas permafrost and benthos/sediment return a greater percentage of searches from arctic/Antarctic/polar regions (**Table 2**; **Figure 1**). Water, river/stream/pond/lake and seawater/marine return relatively even percentage of searches across the three geographic regions (**Table 2**; **Figure 1**).

Often, one type of environmental sample is collected in an attempt to answer broad ecological questions regarding an ecosystem, such as a watershed (Dickie et al., 2018). However, this is problematic and can lead to bias in terms of taxa recovered (Baird and Hajibabaei, 2012; Taberlet et al., 2018). In addition to the geographic location of sample collection, sample type is a large bottleneck in terms of taxa recovered (**Figure 2**). For example, recent studies have found that eDNA samples from freshwater are a poor substitute for bulk-benthos samples for assessing macroinvertebrate community assemblages (Macher et al., 2018; Hajibabaei et al., 2019). Furthermore, the terminology surrounding the types of environmental sample is inconsistent across the literature, with variations of "eDNA" and "bulk-tissue DNA" used interchangeably (Dickie et al., 2018). Often aquatic-based DNA monitoring samples are referred to as "eDNA" (e.g., Valentini et al., 2016; Deiner et al., 2017),

TABLE 1 | Results of Web of Knowledge searches returned for different environmental DNA/eDNA/metabarcoding, search terms (found anywhere in the article), associated with different sample types (water, river/stream/lake/pond, benthos/sediment, soil, seawater/marine, and permafrost), or geographic region (tropic*, temperate, Arctic/Antarctic/polar).

First term	Operator	Second term	Search results
Environmental DNA	OR	eDNA	13,873
eDNA	OR	Metabarcoding	1,931
eDNA	SAME	Metabarcoding	184
DNA	SAME	Metabarcoding	796
Tropic*	AND	Environmental DNA OR eDNA OR metabarcoding	319
Temperate	AND	Environmental DNA OR eDNA OR metabarcoding	188
Arctic ORAntarctic ORpolar	AND	Environmental DNA OR eDNA OR metabarcoding	262
River OR stream OR lake OR pond	AND	Environmental DNA OR eDNA OR metabarcoding	1,113
Benthos OR sediment	AND	Environmental DNA OR eDNA OR metabarcoding	551
Soil	AND	Environmental DNA OR eDNA OR metabarcoding	1,132
Seawater OR marine	AND	Environmental DNA OR eDNA OR metabarcoding	1,025
Permafrost	AND	Environmental DNA OR eDNA OR metabarcoding	24

Searches were conducted using Boolean Operators "OR" (find records containing any of the terms), "AND" (find records containing all terms) and "SAME" (terms that must occur within the same sentence), restricted to the last 5 years (2015–2019).

TABLE 2 | Results of Web of Knowledge searches returned for different environmental DNA/eDNA/metabarcoding, search terms (found anywhere in the article), associated with different sample types (water, river/stream/lake/pond, benthos/sediment, soil, seawater/marine, and permafrost) for each geographic region (tropic*, temperate, Arctic/Antarctic/polar).

First term	Operator	Second term	Operator	Third term	Search results
Tropic*	AND	Environmental DNA OR eDNA OR metabarcoding	AND	Water	78
	AND	Environmental DNA OR eDNA OR metabarcoding	AND	river OR stream OR lake OR pond	51
	AND	Environmental DNA OR eDNA OR metabarcoding	AND	Benthos OR sediment	21
	AND	Environmental DNA OR eDNA OR metabarcoding	AND	Soil	42
	AND	Environmental DNA OR eDNA OR metabarcoding	AND	Seawater OR marine	50
	AND	Environmental DNA OR eDNA OR metabarcoding	AND	Permafrost	0
Temperate	AND	Environmental DNA OR eDNA OR metabarcoding	AND	Water	49
	AND	Environmental DNA OR eDNA OR metabarcoding	AND	River OR stream OR lake OR pond	25
	AND	Environmental DNA OR eDNA OR metabarcoding	AND	Benthos OR sediment	10
	AND	Environmental DNA OR eDNA OR metabarcoding	AND	Soil	38
	AND	Environmental DNA OR eDNA OR metabarcoding	AND	Seawater OR marine	34
	AND	Environmental DNA OR eDNA OR metabarcoding	AND	Permafrost	1
Arctic OR Antarctic OR polar	AND	Environmental DNA OR eDNA OR metabarcoding	AND	Water	73
	AND	Environmental DNA OR eDNA OR metabarcoding	AND	River OR stream OR lake OR pond	51
	AND	Environmental DNA OR eDNA OR metabarcoding	AND	benthos OR sediment	45
	AND	environmental DNA OR eDNA OR metabarcoding	AND	Soil	49
	AND	Environmental DNA OR eDNA OR metabarcoding	AND	Seawater OR marine	66
	AND	Environmental DNA OR eDNA OR metabarcoding	AND	Permafrost	8

Searches were conducted using Boolean Operator "AND" (find records containing all terms) and "OR" (find records containing any of the terms), restricted to the last 5 years (2015–2019).

whereas sediment/benthos or soil samples are termed "bulk-tissue DNA" (Hatzenbuhler et al., 2017; Hajibabaei et al., 2019; Harper et al., 2019), despite these types of sample all referring to DNA which is isolated from an environmental sample (Dickie et al., 2018). This lack of consistency is particularly challenging when attempting to amalgamate literature and compare studies from different research groups and for effectively communicating results of DNA-based studies to non-specialists. Going forward, it would be greatly beneficial to have a consistent and shared ontology across the environmental DNA and metabarcoding community in terms of environmental sample type. Although eDNA could provide an all-encompassing term for analysis of DNA from environmental samples, it is important to provide complementary information about sample type (e.g., soil, water, and benthos) and technology used for detection in all scientific/technical communication. To fully investigate the current uses of DNA-based terminology, an in-depth review would be necessary, which is beyond the scope of this paper. Ultimately, different types of environmental samples, with their varying associated terminologies, are likely to reflect specific

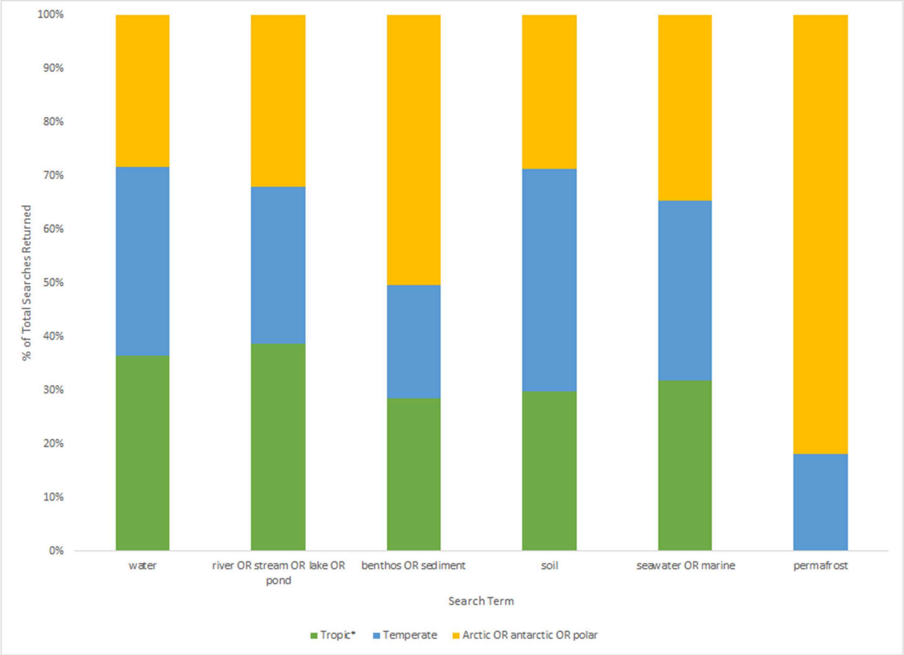


FIGURE 1 | Histogram displaying the percentage of total results returned (tropic*: $n = 242$, temperate: $n = 157$, arctic/Antarctic/polar: $n = 292$; based on **Table 2**) for each sample type search term (water, river/stream/lake/pond, benthos/sediment, soil, seawater/marine, and permafrost) for each geographic region.

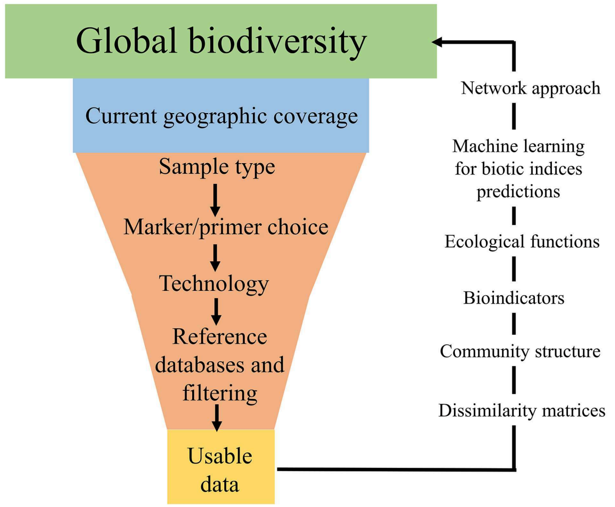


FIGURE 2 | Infographic displaying the "bottlenecks" associated with global DNA metabarcoding data generation.

communities of taxa based on factors such as life histories, season and geographic location (Thomsen and Willerslev, 2015; Dickie et al., 2018), and if global ecological questions are to be addressed using next-generation biomonitoring, sample design will need to incorporate the processing of multiple sample types for accurate assessments of biodiversity.

In addition, there is a substantial degree of variation within metabarcoding as to the sequencing technology implemented for data generation (Bleidorn, 2016; Evans et al., 2016; Elbrecht and Steinke, 2019; Singer et al., 2019; Zinger et al., 2019). As of 2015, there were 13 different PCR-based NGS technologies (Pavan-Kumar et al., 2015), with Illumina® MiSeq currently the prominent NGS platform for processing biomonitoring data (Bleidorn, 2017). In terms of sequencing, different environmental sample types require varying degrees of sequencing breadth and depth (Porter and Hajibabaei, 2018b; Singer et al., 2019). Tropical forest soils are considered to be one of the most diverse ecosystems on the planet, in comparison to alpine mountain lakes, which have vastly different biological richness (Schluter and Pennell, 2017; Dumbrell, 2019). For example, two separate studies looking at microbial community structure in tropical soils and alpine lakes, produced a large difference in sequence reads for the two environments (tropical soil 16s: 1.3 million; alpine lake: 184,273; Filker et al., 2016; Dopheide et al., 2019). In addition, detection of whole communities as opposed to fewer taxa will require a greater sequencing depth (Porter and Hajibabaei, 2018b). Similar to environmental sample type, the sequencing process of DNA-based biomonitoring is often referred to as "NGS," "High-throughput sequencing (HTS)," and "Second-generation sequencing (2GS)" (Dickie et al., 2018; Divoll et al., 2018; Zinger et al., 2019); this varying use of terminology again adds another level of inconsistency to DNA-based biomonitoring. Referring to a consistent term for this sequencing technique, similar to the ontology discussed for sample terminology, would be beneficial. As many companies, such as illumina®, which produce sequencing equipment, often refer to this sequencing technology as "next-generation sequencing," therefore it would be logical to maintain consistency with this term (von Bubnoff, 2008; Quail et al., 2012). As with sample terminology, it is necessary to provide complementary information regarding the technological processes (i.e., high-throughput targeted sequencing). Since January 2016, there have been a few publications referring to the use of Illumina®'s newest high-capacity platform, NovaSeq (Singer et al., 2019) in metabarcoding studies, which have highlighted the higher performance of this new technology in comparison to both the HiSeq and MiSeq, with NovaSeq detecting 40% more metazoan families in metabarcoded sea water samples in comparison to the MiSeq (Singer et al., 2019). The implementation of new technology brings to light the need for evaluating available technologies to address biomonitoring needs for a given system with the main limitation being the taxonomic coverage achieved per sample (Divoll et al., 2018). For example, MiSeq may provide optimal solution to tackle biodiversity in freshwater systems or specific taxonomic assemblages whereas NovaSeq would be a better platform for more complex situations such as oceanic samples. Suboptimal use of data generation platforms could lead to misrepresentation of taxonomic information and can be problematic when considering the implications of this on the ecological conclusions already having been drawn from metabarcoding-based biomonitoring data (Zinger et al., 2019).

Environmental sample choice and implementation of different sequencing platforms are not the only sources of taxa detection bias (**Figure 2**). There are numerous bioinformatic pipelines for processing samples, which vary greatly across studies (Alberdi et al., 2018) and appropriate clustering/filtering thresholds can lead to mis-classification and thus bias in the taxa detected (Hajibabaei et al., 2016; Alberdi et al., 2018; Zinger et al., 2019). In addition, the most prominent bottleneck in terms of recovering present taxa in an environmental sample is incomplete DNA reference databases (**Figure 2**; Zaiko et al., 2015; Elbrecht et al., 2017; Stat et al., 2017). Commonly used, both the BOLD (Barcode of Life Datasystem) and GenBank databases regularly lack reference sequences and/or have conflicting taxonomic assignments for the species (Ammon et al., 2018). Reference database incompletion causes inability to identify all DNA sequences in a sample and means some taxonomic groups are underrepresented (Creer et al., 2010; Ratnasingham and Hebert, 2013; Porter and Hajibabaei, 2018a), which highlights the current substantial gap in global biodiversity knowledge (Zaiko et al., 2015). If DNA-based biomonitoring is to be an effective, reliable tool for assessing biodiversity on a global scale, efforts need to be primarily concentrated toward better curation and updating of DNA reference records, as well as continued barcoding of taxonomically identified specimens to improve the quality and quantity of information in DNA databases (Hajibabaei et al., 2016; Elbrecht et al., 2017; Stat et al., 2017; Zinger et al., 2019).

In essence, what will dominate the database in terms of sequence data for various biota will be based on what has been collected from the temperate areas more so than the tropics and polar regions. Thus, for example, how will soil scientists (and others) be able to effectively identify organisms in their soil samples based on databases from other regions? More importantly, some of these areas that need to be further sampled are those that are experiencing drastic intensities of climate pattern changes. This also describes the need for more seasonal studies over periods of time to assess the variability in climate patterns across the globe. If we are to detect ecosystem change globally, more comprehensive work involving biomonitoring and DNA metabarcoding/eDNA will be needed to generate consensus data, generate the metadata, and start analyzing trends across the globe.

With advancing technologies and methodologies such as implementing machine learning and neural networks pertaining to ecological status and modeling, as has been described elsewhere (Díaz et al., 2015; Bohan et al., 2017; Derocles et al., 2018), we still need to increase the information in a database to identify particular organisms of interest and from more geographical locations across the globe, for biomonitoring, and more robust experimental designs rather than straight survey-based approaches to draw sound ecological conclusions (Zinger et al., 2019). Yet, if sample types are inherently variable due to geographical location and/or sample type across the globe, how can we ever expect these taxonomic databases to accurately reflect a global perspective of ecosystem, in order to effectively and accurately detect global ecosystem change. By collecting samples from more geographical locations where the representation is lacking, collecting a wider array of sample types, and constructing the replicated ecological networks of ecological interactions, together, will provide useful standards of global ecosystem information, dramatically enhancing the ability to assess the taxa within global ecosystems, and understanding how these respond to climate change and other forms of ecosystem damage. We propose that combining the use of these technologies would greatly enhance the capacity to better predict how various ecosystems respond to environmental change at local, regional and global levels.

AUTHOR CONTRIBUTIONS

MH, KM, and CR conceived the idea and co-wrote the manuscript. CR conducted the literature search.

FUNDING

Funding was provided from Government of Canada through Canada First Excellence Research project, Food from Thought and Genome Canada & Ontario Genomics Institute project STREAM.

REFERENCES

Adamowicz, S. J., Boatwright, J. S., Chain, F., Fisher, B. L., Hogg, I. D., Leese, F., et al. (2019). Trends in DNA barcoding and metabarcoding. *Genome* 62, v–viii. doi: 10.1139/gen-2019-0054

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., and Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods Ecol. Evol.* 9, 134–147. doi: 10.1111/2041-210X.12849

Ammon, U., von, Wood, S. A., Laroche, O., Zaiko, A., Tait, L., Lavery, S., et al. (2018). Combining morpho-taxonomy and metabarcoding enhances the detection of non-indigenous marine pests in biofouling communities. *Sci. Rep.* 8:16290. doi: 10.1038/s41598-018-34541-1

Andersen, K., Bird, K. L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kjær, K. H., et al. (2012). Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Mol. Ecol.* 21, 1966–1979. doi: 10.1111/j.1365-294X.2011.05261.x

Aylagas, E., Borja, Á., Muxika, I., and Rodríguez-Ezpeleta, N. (2018). Adapting metabarcoding-based benthic biomonitoring into routine marine ecological status assessment networks. *Ecol. Indic.* 95, 194–202. doi: 10.1016/j.ecolind.2018.07.044

Aylagas, E., Mendibil, I., Borja, Á., and Rodríguez-Ezpeleta, N. (2016). Marine sediment sample pre-processing for macroinvertebrates metabarcoding: mechanical enrichment and homogenization. *Front. Mar. Sci.* 3, 1–12. doi: 10.3389/fmars.2016.00203

Baird, D. J., and Hajibabaei, M. (2012). Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Mol. Ecol.* 21, 2039–2044. doi: 10.1111/j.1365-294X.2012.05519.x

Beisner, B. E., Haydon, D. T., and Cuddington, K. (2003). Alternative stable states in ecology. *Front. Ecol. Environ.* 1, 376–382. doi: 10.1890/1540-9295(2003)001[0376:ASIE]2.0.CO;2

Bellemain, E., Davey, M. L., Kauserud, H., Epp, L. S., Boessenkool, S., Coissac, E., et al. (2013). Fungal palaeodiversity revealed using high-throughput metabarcoding of ancient DNA from arctic permafrost. *Environ. Microbiol.* 15, 1176–1189. doi: 10.1111/1462-2920.12020

Bleidorn, C. (2016). Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *System. Biodivers.* 14, 1–8. doi: 10.1080/14772000.2015.1099575

Bleidorn, C. (2017). "Sequencing Techniques," in *Phylogenomics: An Introduction*, ed C. Bleidorn (Cham: Springer International Publishing), 43–60. doi: 10.1007/978-3-319-54064-1_3

Bohan, D. A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A. J., and Woodward, G. (2017). Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. *Trends Ecol. Evol.* 32, 477–487. doi: 10.1016/j.tree.2017.03.001

Carpenter, S. R., Cole, J. J., Hodgson, J. R., Kitchell, J. F., Pace, M. L., Bade, D., et al. (2001). Trophic cascades, nutrients, and lake productivity: whole-lake experiments. *Ecol. Monogr.* 71, 163–186. doi: 10.1890/0012-9615(2001)071[0163:TCNALP]2.0.CO;2

Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., and Pawlowski, J. (2019). Embracing environmental genomics and machine learning for routine biomonitoring. *Trends Microbiol.* 27, 387–397. doi: 10.1016/j.tim.2018.10.012

Creer, S., Fonseca, V. G., Porazinska, D. L., Giblin-Davis, R. M., Sung, W., Power, D. M., et al. (2010). Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Mol. Ecol.* 19, 4–20. doi: 10.1111/j.1365-294X.2009.04473.x

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., et al. (2017). Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol. Ecol.* 26, 5872–5895. doi: 10.1111/mec.14350

Derocles, S. A. P., Bohan, D. A., Dumbrell, A. J., Kitson, J. J. N., Massol, F., Pauvert, C., et al. (2018). "Chapter one - biomonitoring for the 21st century: integrating next-generation sequencing into ecological network analysis," in *Advances in Ecological Research Next Generation Biomonitoring: Part 1*., eds D. A. Bohan, A. J. Dumbrell, G. Woodward, and M. Jackson (Cambridge, MA: Academic Press), 1–62. doi: 10.1016/bs.aecr.2017.12.001

Díaz, S., Demissew, S., Carabias, J., Joly, C., Lonsdale, M., Ash, N., et al. (2015). The IPBES Conceptual Framework — connecting nature and people. *Curr. Opin. Environ. Sust.* 14, 1–16. doi: 10.1016/j.cosust.2014.11.002

Dickie, I. A., Boyer, S., Buckley, H. L., Duncan, R. P., Gardner, P. P., Hogg, I. D., et al. (2018). Towards robust and repeatable sampling methods in eDNA-based studies. *Mol. Ecol. Resour.* 18, 940–952. doi: 10.1111/1755-0998.12907

Divoll, T. J., Brown, V. A., Kinne, J., McCracken, G. F., and O'Keefe, J. M. (2018). Disparities in second-generation DNA metabarcoding results exposed with accessible and repeatable workflows. *Mol. Ecol. Resour.* 18, 590–601. doi: 10.1111/1755-0998.12770

Dopheide, A., Xie, D., Buckley, T. R., Drummond, A. J., and Newcomb, R. D. (2019). Impacts of DNA extraction and PCR on DNA metabarcoding estimates of soil biodiversity. *Methods Ecol. Evol.* 10, 120–133. doi: 10.1111/2041-210X.13086

Dowle, E. J., Pochon, X., Banks, J. C., Shearer, K., and Wood, S. A. (2016). Targeted gene enrichment and high-throughput sequencing for environmental biomonitoring: a case study using freshwater macroinvertebrates. *Mol. Ecol. Resour.* 16, 1240–1254. doi: 10.1111/1755-0998.12488

Dumbrell, A. J. (2019). Size matters in regulating the biodiversity of tropical forest soils. *Mol. Ecol.* 28, 525–527. doi: 10.1111/mec.14996

Elbrecht, V., and Steinke, D. (2019). Scaling up DNA metabarcoding for freshwater macrozoobenthos monitoring. *Freshw. Biol.* 64, 380–387. doi: 10.7287/peerj.preprints.3456v4

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

Evans, D. M., Kitson, J. J. N., Lunt, D. H., Straw, N. A., and Pocock, M. J. O. (2016). Merging DNA metabarcoding and ecological network analysis to understand and build resilient terrestrial ecosystems. *Funct. Ecol.* 30, 1904–1916. doi: 10.1111/1365-2435.12659

Fahner, N. A., McCarthy, A., Barnes, J. G., Singer, G., and Hajibabaei, M. (2018). Experimental design considerations for assessing marine biodiversity using environmental *DNA* 6:e26814v1. doi: 10.7287/peerj.preprints.26814v1

Fahner, N. A., Shokralla, S., Baird, D. J., and Hajibabaei, M. (2016). Large-scale monitoring of plants through environmental DNA metabarcoding of soil: recovery, resolution, and annotation of four DNA markers. *PLoS ONE* 11:e0157505. doi: 10.1371/journal.pone.0157505

Ferguson, R. M. W., Garcia-Alcega, S., Coulon, F., Dumbrell, A. J., Whitby, C., and Colbeck, I. (2019). Bioaerosol biomonitoring: sampling optimization for molecular microbial ecology. *Mol. Ecol. Resour.* 19, 672–690. doi: 10.1111/1755-0998.13002

Filker, S., Sommaruga, R., Vila, I., and Stoeck, T. (2016). Microbial eukaryote plankton communities of high-mountain lakes from three continents exhibit strong biogeographic patterns. *Mol. Ecol.* 25, 2286–2301. doi: 10.1111/mec.13633

Hajibabaei, M., Baird Donald, J., Fahner Nicole, A., Beiko, R., and Brain Golding, G. (2016). A new way to contemplate Darwin's tangled bank: how DNA barcodes are reconnecting biodiversity science and biomonitoring. *Philos. Trans. R. Soc. B Biol. Sci.* 371:20150330. doi: 10.1098/rstb.2015.0330

Hajibabaei, M., Porter, T. M., Robinson, C. V., Baird, D. J., Shokralla, S., and Wright, M. (2019). Watered-down biodiversity? A comparison of metabarcoding results from DNA extracted from matched water and bulk tissue biomonitoring samples. *Sci. Rep. bioRxiv [preprint]*. doi: 10.1101/575928

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A. C., and Baird, D. J. (2011). Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE* 6:e17497. doi: 10.1371/journal.pone.0017497

Hansen, B. K., Bekkevold, D., Clausen, L. W., and Nielsen, E. E. (2018). The sceptical optimist: challenges and perspectives for the application of environmental DNA in marine fisheries. *Fish Fisheries* 19, 751–768. doi: 10.1111/faf.12286

Harper, L. R., Buxton, A. S., Rees, H. C., Bruce, K., Brys, R., Halfmaerten, D., et al. (2019). Prospects and challenges of environmental DNA (eDNA) monitoring in freshwater ponds. *Hydrobiologia* 826, 25–41. doi: 10.1007/s10750-018-3750-5

Hatzenbuhler, C., Kelly, J. R., Martinson, J., Okum, S., and Pilgrim, E. (2017). Sensitivity and accuracy of high-throughput metabarcoding methods for early detection of invasive fish species. *Sci. Rep.* 7:46393. doi: 10.1038/srep46393

Kraaijeveld, K., Weger, L. A., de, García, M. V., Buermans, H., Frank, J., Hiemstra, P. S., et al. (2015). Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. *Mol. Ecol. Resour.* 15, 8–16. doi: 10.1111/1755-0998.12288

Laroche, O., Pochon, X., Tremblay, L. A., Ellis, J. I., Lear, G., and Wood, S. A. (2018). Incorporating molecular-based functional and co-occurrence network properties into benthic marine impact assessments. *FEMS Microbiol. Ecol.* 94:fiy167. doi: 10.1093/femsec/fiy167

Lobo, J., Shokralla, S., Costa, M. H., Hajibabaei, M., and Costa, F. O. (2017). DNA metabarcoding for high-throughput monitoring of estuarine macrobenthic communities. *Sci. Rep*. 7:15618. doi: 10.1038/s41598-017-15823-6

Macher, J.-N., Vivancos, A., Piggott, J. J., Centeno, F. C., Matthaei, C. D., and Leese, F. (2018). Comparison of environmental DNA and bulk-sample metabarcoding using highly degenerate cytochrome c oxidase I primers. *Mol. Ecol. Resour.* 18, 1456–1468. doi: 10.1111/1755-0998.12940

McGee, K. M., Eaton, W. D., Shokralla, S., and Hajibabaei, M. (2019). Determinants of soil bacterial and fungal community composition toward carbon-use efficiency across primary and secondary forests in a costa rican conservation area. *Microb. Ecol.* 77, 148–167. doi: 10.1007/s00248-018-1206-0

Morinière, J., Araujo, B. C., de, Lam, A. W., Hausmann, A., Balke, M., Schmidt, S., et al. (2016). Species identification in malaise trap samples by DNA barcoding based on NGS technologies and a scoring matrix. *PLoS ONE* 11:e0155497. doi: 10.1371/journal.pone.0155497

Muha, T. P., Rodríguez-Rey, M., Rolla, M., and Tricarico, E. (2017). Using environmental DNA to improve species distribution models for freshwater invaders. *Front. Ecol. Evol.* 5, 1–7. doi: 10.3389/fevo.2017.00158

Pavan-Kumar, P., Gireesh-Babu, A., and Lakra, W. (2015). DNA Metabarcoding: a new approach forrapid biodiversity assessment. *J. Cell Sci. Mol. Biol.* 2:111.

Porter, T. M., and Hajibabaei, M. (2018a). Over 2.5 million COI sequences in GenBank and growing. *PLoS ONE* 13:e0200177. doi: 10.1371/journal.pone.0200177

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

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., et al. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13:341. doi: 10.1186/1471-2164-13-341

Ratnasingham, S., and Hebert, P. D. N. (2013). A DNA-based registry for all animal species: the barcode index number (BIN) system. *PLoS ONE* 8:e66213. doi: 10.1371/journal.pone.0066213

Robinson, C. V., Leaniz, C. G., and de Consuegra, S. (2019). Effect of artificial barriers on the distribution of the invasive signal crayfish and Chinese mitten crab. *Sci. Rep.* 9:7230. doi: 10.1038/s41598-019-43570-3

Ruppert, K. M., Kline, R. J., and Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: a systematic review in methods, monitoring, and applications of global eDNA. *Global Ecol. Conser.* 17:e00547. doi: 10.1016/j.gecco.2019.e00547

Salonen, I. S., Chronopoulou, P.-M., Leskinen, E., and Koho, K. A. (2019). Metabarcoding successfully tracks temporal changes in eukaryotic communities in coastal sediments. *FEMS Microbiol. Ecol.* 95:fiy226. doi: 10.1093/femsec/fiy226

Scheffer, M., Carpenter, S., Foley, J. A., Folke, C., and Walker, B. (2001). Catastrophic shifts in ecosystems. *Nature* 413, 591–596. doi: 10.1038/35098000

Schluter, D., and Pennell, M. W. (2017). Speciation gradients and the distribution of biodiversity. *Nature* 546, 48–55. doi: 10.1038/nature22897

Singer, G. A. C., Fahner, N., Barnes, J., McCarthy, A., and Hajibabaei, M. (2019). Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. 9:5991. doi: 10.1038/s41598-019-42455-9

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., et al. (2017). Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Sci. Rep.* 7:12240. doi: 10.1038/s41598-017-12501-5

Taberlet, P., Bonin, A., Zinger, L., and Coissac, E. (2018). *Environmental DNA: For Biodiversity Research and Monitoring.* New York, NY: Oxford University Press. doi: 10.1093/oso/9780198767220.001.0001

Thomsen, P. F., and Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biol. Conserv.* 183, 4–18. doi: 10.1016/j.biocon.2014.11.019

Turner, C. R., Uy, K. L., and Everhart, R. C. (2015). Fish environmental DNA is more concentrated in aquatic sediments than surface water. *Biol. Conserv.* 183, 93–102. doi: 10.1016/j.biocon.2014.11.017

Vacher, C., Tamaddoni-Nezhad, A., Kamenova, S., Peyrard, N., Moalic, Y., Sabbadin, R., et al. (2016). "Chapter 1 - learning ecological networks from next-generation sequencing data," in *Advances in Ecological Research Ecosystem Services: From Biodiversity to Society, Part 2.,* eds G. Woodward and D. A. Bohan (Cambridge, MA: Academic Press), 1–39. doi: 10.1016/bs.aecr.2015.10.004

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., et al. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Mol. Ecol.* 25, 929–942. doi: 10.1111/mec.13428

von Bubnoff, A. (2008). Next-generation sequencing: the race is on. *Cell* 132, 721–723. doi: 10.1016/j.cell.2008.02.028

Xie, Y., Zhang, X., Yang, J., Kim, S., Hong, S., Giesy, J. P., et al. (2018). eDNA-based bioassessment of coastal sediments impacted by an oil spill. *Environ. Pollut.* 238, 739–748. doi: 10.1016/j.envpol.2018.02.081

Yoccoz, N. G., Bråthen, K. A., Gielly, L., Haile, J., Edwards, M. E., Goslar, T., et al. (2012). DNA from soil mirrors plant taxonomic and growth form diversity. *Mol. Ecol.* 21, 3647–3655. doi: 10.1111/j.1365-294X.2012.05545.x

Zaiko, A., Samuiloviene, A., Ardura, A., and Garcia-Vazquez, E. (2015). Metabarcoding approach for non-indigenous species surveillance in marine coastal waters. *Mar. Pollut. Bull.* 100, 53–59. doi: 10.1016/j.marpolbul.2015.09.030

Zielińska, S., Kidawa, D., Stempniewicz, L., Łoś, M., and Łoś, J. M. (2017). Environmental DNA as a valuable and unique source of information about ecological networks in Arctic terrestrial ecosystems. *Environ. Rev.* 25, 282–291. doi: 10.1139/er-2016-0060

Zimmermann, H. H., Raschke, E., Epp, L. S., Stoof-Leichsenring, K. R., Schwamborn, G., Schirrmeister, L., et al. (2017). Sedimentary ancient DNA and pollen reveal the composition of plant organic matter in Late Quaternary permafrost sediments of the Buor Khaya Peninsula (North-Eastern Siberia). *Biogeosciences* 14, 575–596. doi: 10.5194/bg-14-575-2017

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., et al. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Mol. Ecol.* 28, 1857–1862. doi: 10.1111/mec.15060

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

ORIGINAL RESEARCH
published: 29 October 2019
doi: 10.3389/fevo.2019.00409

Diatom DNA Metabarcoding for Biomonitoring: Strategies to Avoid Major Taxonomical and Bioinformatical Biases Limiting Molecular Indices Capacities

Kálmán Tapolczai[1,2*†], *François Keck*[3,4†], *Agnès Bouchez*[4,5], *Frédéric Rimet*[4,5], *Maria Kahlert*[3] *and Valentin Vasselon*[6†]

[1] Premium Postdoctoral Research Program, Hungarian Academy of Sciences, Budapest, Hungary, [2] Department of Limnology, University of Pannonia, Veszprém, Hungary, [3] Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden, [4] INRA, UMR CARRTEL, Thonon-les-Bains, France, [5] University of Savoie Mont-Blanc, UMR CARRTEL, Le Bourget du Lac, France, [6] AFB, Pôle R&D "ECLA", INRA, UMR CARRTEL, Thonon-les-Bains, France

OPEN ACCESS

Edited by:
Dominique Gravel,
Université de Sherbrooke, Canada

Reviewed by:
Anastasija Zaiko,
Cawthron Institute, New Zealand
Tristan Cordier,
Université de Genève, Switzerland

***Correspondence:**
Kálmán Tapolczai
tapolczai.kalman@gmail.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Ecology and Evolution

Received: *14 June 2019*
Accepted: *10 October 2019*
Published: *29 October 2019*

Citation:
Tapolczai K, Keck F, Bouchez A, Rimet F, Kahlert M and 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. doi: 10.3389/fevo.2019.00409

Recent years provided intense progression in the implementation of molecular techniques in a wide variety of research fields in ecology. Biomonitoring and bioassessment can greatly benefit from DNA metabarcoding and High-Throughput Sequencing (HTS) methods that potentially provide reliable, high quantity and quality standardized data in a cost- and time-efficient way. However, DNA metabarcoding has its drawbacks, introducing biases at all the steps of the process, particularly during bioinformatics treatments used to prepare HTS data for ecological analyses. The high diversity of bioinformatics methods (e.g., OTU clustering, chimera detection, taxonomic assignment) and parameters (e.g., percentage similarity threshold used to define OTUs) make inter-studies comparison difficult, limiting the development of standardized and easy-accessible bioassessment procedures for routine freshwater monitoring. In order to study and overcome these drawbacks, we constructed four *de novo* indices to assess river ecological status based on the same biological samples of diatoms analyzed with morphological and molecular methods. The biological inventories produced are (i) morphospecies identified by microscopy, (ii) OTUs provided via metabarcoding and hierarchical clustering of sequences using a 95% similarity threshold, (iii) individual sequence units (ISUs) via metabarcoding and only minimal bioinformatical quality filtering, and (iv) exact sequence variants (ESVs) using DADA2 denoising algorithm. The indices based on molecular data operated directly with ecological values estimated for OTUs/ ISUs/ ESVs. Our study used an approach of bypassing taxonomic assignment, so bias related to unclassified sequences missing from reference libraries could be handled and no information on ecology of sequences is lost. Additionally, we showed that the indices based on ISUs and ESVs were equivalent, outperforming the OTU-based one in terms of predictive power and accuracy by revealing the hidden ecological information of sequences that are otherwise clustered in the same OTU (intra-species/intra-population variability). Furthermore, ISUs, ESVs, and morphospecies indices provided similar estimation of site ecological status, validating that ISUs with limited bioinformatics treatments may be used for DNA freshwater monitoring. Our study is a proof of concept where taxonomy- and clustering-free approach is presented, that we believe is a step forward a standardized and comparable DNA bioassessment, complementary to morphological methods.

Keywords: bioassessment, biomonitoring, diatoms, high-throughput sequencing, metabarcoding

INTRODUCTION

In the past several decades, use of bioindicator organisms has been widely adopted in order to monitor, maintain or develop the quality of water bodies worldwide. The permanently evolving field of freshwater ecology made freshwater biomonitoring an efficient tool, incorporated in national and international water management frameworks like the Clean Water Act (CWA; U.S. Congress, 1972) or the Water Framework Directive (WFD; European Commission, 2000).

The EU WFD uses four groups of organisms (i.e., Biological Quality Elements—BQEs) to assess ecological quality by comparing the community structure of the impacted sites to the community structure of a reference site considered not to be—or slightly—affected by anthropogenic alterations (Pardo et al., 2012). Experts faces several challenges depending on the BQE they are working on but they are all limited by the following factors: time consuming and costly sampling and preparation procedures, differences in expert knowledge and equipment for taxa identification, thus disharmony in taxa inventories among laboratories (Kahlert et al., 2012).

The recent intense development of DNA metabarcoding and High-Throughput Sequencing (HTS) techniques has set a new milestone in biomonitoring (Baird and Hajibabaei, 2012; Leese et al., 2016; Keck et al., 2017). Instead of the identification based on morphological features, this technique employs standard gene markers to identify taxa-specific sequences in the organism's DNA, serving as a barcode (Hebert et al., 2003). This method allows the simultaneous identification of multiple taxa from multiple environmental samples (Taberlet et al., 2012), being more time- and cost-effective than the classical methods, providing a fine-scale taxonomic characterization of communities, often revealing hidden diversity (Lindeque et al., 2013).

However, in order to use metabarcoding techniques as a routine biomonitoring tool, the standardization of the method is required. Extensive studies have been produced analyzing and suggesting solutions for the biases at each step of the metabarcoding process including the sampling, DNA extraction (Vasselon et al., 2017a), choice of the marker gene (Kermarrec et al., 2013) or the choice of the applied HTS technology (Loman et al., 2012; Shokralla et al., 2012). A further bias that can hamper the comparability among different analyses is the large variety of bioinformatic pipelines used to process HTS data. One step particularly critical is the clustering of the raw sequence data into operational taxonomic units (OTUs).

Molecular OTUs are widely used to describe microbial communities using HTS amplicon sequencing as proxies for species, generally using the 97% similarity threshold proposed initially for the 16S rRNA gene by Stackebrandt and Goebel (1994). Sequence clustering aims to reduce the noise in the data and provide a granularity close to that of the species delimitation. Clustering with a high sequence similarity threshold to create OTUs increases the risk of giving ecological sense to sequence errors and artifacts (Chen et al., 2013). However, Edgar (2018) proposed a recent update of this threshold to ∼ 99–100% and several studies advocate the use of denoised DNA reads to avoid the biases linked to the choice of the clustering algorithm and the similarity threshold (Tikhonov et al., 2015; Edgar, 2016; Callahan et al., 2017). Exact Sequence Variant (ESV) are considered as unique DNA reads with biological meaning and they offer several advantages for community analyses compared to OTUs, including computational tractability, reproducibility of analysis and the possibility to perform meta-analyses from different studies (Callahan et al., 2017). Although the use of ESVs is an important step toward a better characterization of intraspecific genetic diversity (Elbrecht et al., 2018; Forster et al., 2019), they still rely on denoising algorithm that may introduce biases and lead to non-reproducible results (Nearing et al., 2018). In order to have the most reproducible bioinformatic treatments, one can also work with the so-called Individual Sequence Units (ISUs), composed by ESV and erroneous sequences that correspond to PCR and sequencing errors, thus applying strictly limited bioinformatic filters.

Biological indices based on the microalgae group, diatoms (Bacillariophyta phylum), are frequently used by scientists and environmental managers to assess the ecological status of ecosystems and their response to local pressures and global change. Numerous diatom indices are based on a simple equation that weights the ecological optimum of each detected species by its abundance and ecological tolerance (Zelinka and Marvan, 1961). In traditional indices, species ecological profiles (optimum and tolerance) are often directly obtained from a large set of data by using simple statistical descriptors of location (e.g., weighted mean) and dispersion (e.g., weighted standard deviation) along a pollution gradient. Interestingly, the morphospecies as a taxonomic unit of the traditional indices can be replaced by molecular taxonomic units, allowing to train a new type of biological indices, the so-called taxonomy-free indices (Apothéloz-Perret-Gentil et al., 2017; Tapolczai et al., 2019). So far, taxonomy-free indices have proved to be an efficient strategy to keep biological information carried by OTUs whose taxonomic assignment is otherwise not possible because of the incomplete reference library. Here, we propose to apply a similar strategy on ISU and ESV data, in order to limit the bias associated to OTU clustering and to improve the reproducibility and standardization of biomonitoring indices.

In our study, we compare different strategies to use biological data and provide quality assessment indices. Four diatom indices were developed, based on the ecological optimum and tolerance of morphospecies/OTUs/ISUs/ESVs calculated along an integrated environmental gradient, using the classical Zelinka-Marvan equation (1961). The first index is based on taxa inventory obtained via the microscopic identification of morphospecies and their relative abundance. The second index is based on OTU inventory without taxonomic assignment, obtained via DNA metabarcoding and hierarchical *de novo* clustering with 95% sequence similarity. The third strategy tested was the development of an index using ISUs. In this case no taxonomic assignment, nor clustering into OTUs were used. The last strategy used the DADA2 denoising algorithm to select ESVs from ISUs.

We hypothesize that by avoiding taxonomic assignment, clustering and denoising of ISUs, relevant sources of biases in the metabarcoding process are removed. This strategy is a step toward a standardized metabarcoding-based bioassessment without losing the indicator efficiency of the quality index and enabled to propose an easy transferable bioinformatics tool for stakeholders in charge of freshwater management.

MATERIALS AND METHODS

Study Site and Sampling Network

The routine survey of the WFD monitoring network is carried out by French offices responsible for the monitoring and water quality assessment of rivers, including national agencies, and private consultancies. They are in charge of the chemical and biological surveys following the WFD recommendations, meaning that they realize the acquisition of physico-chemical parameters and the description of benthic diatom assemblages for each site.

Based on those available information, among the hundreds of French river sites annually surveyed for water quality assessment in the context of the WFD, 76 sites were selected from the 2016 national monitoring campaign following several criteria : (i) sites showed a gradient of pressure (e.g., organic pollution, turbidity, nutrients, etc.) allowing to obtain a water quality gradient from pristine to polluted conditions, (ii) sampling sites are distributed within the country (different river types), far enough to limit potential effects of non-independence among sampling sites during quality index construction, (iii) benthic diatom samples are available to perform morphological and molecular approaches and (iv) information on physico-chemical parameters is available for all the samples.

Physical and Chemical Parameters

For each site, environmental data were extracted from the French national database “Naïades” (http://www.naiades.eaufrance.fr/) over a period of 70 days (60 days preceding and 10 days following the biological sampling). For each environmental variable, we computed the mean value of all the records available during this time window. It resulted in a table of 76 rows (samples) and 15 columns (variables) without missing values. The environmental parameters kept are dissolved oxygen (O_2), oxygen saturation (O_2 sat), pH, Conductivity, nitrate (NO_3^-), nitrite (NO_2^-), ammonium (NH_4^+), total Kjeldahl nitrogen (TKN), total phosphorus (TP), phosphate (PO_4^{3-}), temperature (T), total suspended solids (TSS), total organic carbon (TOC), biological oxygen demand (BOD_5), and turbidity (Turb) (**Table S1**).

Diatom Sampling, Biofilm Sample Preparation, and Morphological Analysis

For each site, the biofilm containing benthic diatom communities was sampled from at least five submerged stones collected from the lotic parts of the rivers following European standards (European Committee for Standardization, 2016). The upper surface of the stones were scrubbed using a clean toothbrush at each sampling site and mixed into a tray. The samples were homogenized by manual shaking and divided into two subsamples, one for microscopic identification performed by national offices and one sent to our lab for molecular analysis. The subsamples were transferred into 50 mL Falcon tubes and preserved using 96% ethanol for a final ethanol concentration of at least 70% and stored at room temperature under dark conditions until preparation for morphological analysis and DNA extraction (performed within 6 months).

For the microscopic analysis, diatom samples were treated using 40% H_2O_2 and HCl according to the European standard (European Committee for Standardization, 2014). Permanent slides were prepared by mounting the cleaned diatom samples. Morphological analysis was carried out using microscope with 1,000x magnification objective. A minimum of 400 diatoms valves were determined using up to date identification literature.

Diatom DNA Metabarcoding

The preserved biofilm samples were homogenized by manual shaking and a volume of 2 mL of each sample was used as starter for DNA extraction. The samples were first centrifuged at 17,000 g during 30 min in order to remove the supernatant containing ethanol. Total genomic DNA was extracted from the remaining pellet using the Sigma-Aldrich GenEluteTM-LPA DNA precipitation protocol as described previously (e.g., Vasselon et al., 2017a) in a final elution volume of 30 μL.

PCR amplification of diatom communities was performed by targeting a short fragment (312 bp) of the Ribulose Bisphosphate Carboxylase Large subunit (*rbc*L) plastid gene, a DNA marker commonly used for diatom metabarcoding on lake and river samples (Rivera et al., 2018; Bailet et al., 2019; Chonova et al., 2019; Mortágua et al., 2019). The primer pair used to amplify the 312 bp *rbc*L region corresponds to the equimolar mix of 3 forward primer (Diat_rbcL_708F_1, Diat_rbcL_708F_2, Diat_rbcL_708F_3) and 2 reverse primers (R3_1, R3_2) as described in Vasselon et al. (2017b). Forward and reverse primers carry the 5′-CTTTCCCTACACGACGCTCTTCCGATCT-3′ and 5′-GGAGTTCAGACGTGTGCTCTTCCGATCT-3′ tails used to prepare Illumina libraries with a dual-step PCR approach (PCR1 and PCR2). For the PCR1, each DNA sample was amplified in triplicate in a final volume of 25 μL using the tailed *rbc*L primers and the Takara LA Taq® polymerase with PCR1 reaction mix and conditions detailed in the **Table S2**.

The 3 PCR1 replicates prepared for each DNA sample were pooled together and sent to the "GenoToul Genomics and Transcriptomics" facility (GeT-PlaGe, Auzeville, France) which performed: (i) the purification of PCR1 amplicons; (ii) the PCR2 amplification using PCR1 purified amplicon as template and Illumina-tailed primers allowing to add dual-index specific to each sample; (iii) the preparation of the final pool corresponding to an equimolar mix of the 76 PCR2 dual-indexed amplicons; (iv) the sequencing of the final pool on an Illumina MiSeq platform using the V3 paired-end sequencing kit (250 bp × 2).

Bioinformatics

Initial Bioinformatic Steps for ISUs and OTUs

The GeT-PlaGe sequencing platform assembled the MiSeq paired-end reads into full-length DNA sequences (paired sequences overlap >140 bp and mismatches <0.1%) and performed the demultiplexing of the 76 samples, providing 1 fastq file per sample. All the bioinformatics treatments were performed using Mothur software v1.39.5 (Schloss et al., 2009). Initial bioinformatic steps were applied to keep good quality DNA reads using the *trim.seqs()* command and the following parameters: a sequence length of 263 ± 10 bp (*rbc*L barcode length without primers), a Phred quality score ≥23 over a moving window of 25 bp, 0 ambiguities ("N"), a maximum homopolymer length of 8 bp, a maximum of 1 mismatch in the primer sequence. Remaining DNA reads were dereplicated into ISUs with the *unique.seqs()* command and the resulting files processed with two distinct bioinformatic strategies in order to prepare the final ISU and OTU tables used for the construction of water quality indices, as shown in **Figure 1**.

Preparation of ISU Table

Even if the *rbc*L primers used for metabarcoding were designed to be diatom specific, the presence of degenerated bases in the primer sequence may introduced non-target organism amplification (Linhart and Shamir, 2002) In order to perform the most objective comparison between diatom morphospecies and ISUs water quality indices developed in this study, "non-diatom" ISUs must be removed as they can interfere, positively or negatively, on the predictive power of the ISU index. Thus, we used the *classify.seqs()* command (default parameters, cutoff = 75%) with the "diat.barcode" reference database (version v7: 23-02-2018, https://doi.org/10.15454/HYRVUH) to provide a taxonomy to each ISU and we applied the *remove.lineage()* command to remove the non-Bacillariophyta (phylum) ISUs ("Bacillaryophyta_unclassified" ISUs were also discarded).

The ISU abundance distributions along the environmental gradient were used to develop the ISU index (Idx_ISU), meaning that ISUs with low abundance and rare ISUs were automatically removed during the index development [see section Calculation of diatoms indices (Idx_morph, Idx_OTU, Idx_ISU, Idx_ESV)]. We decided to use the *split.abund()* command in order to keep only ISUs represented by at least 50 reads among the 79 samples. By this way, spurious ISUs were removed and the computing power required to create Idx_ISU was reduced, without affecting its efficiency.

Preparation of OTU Table

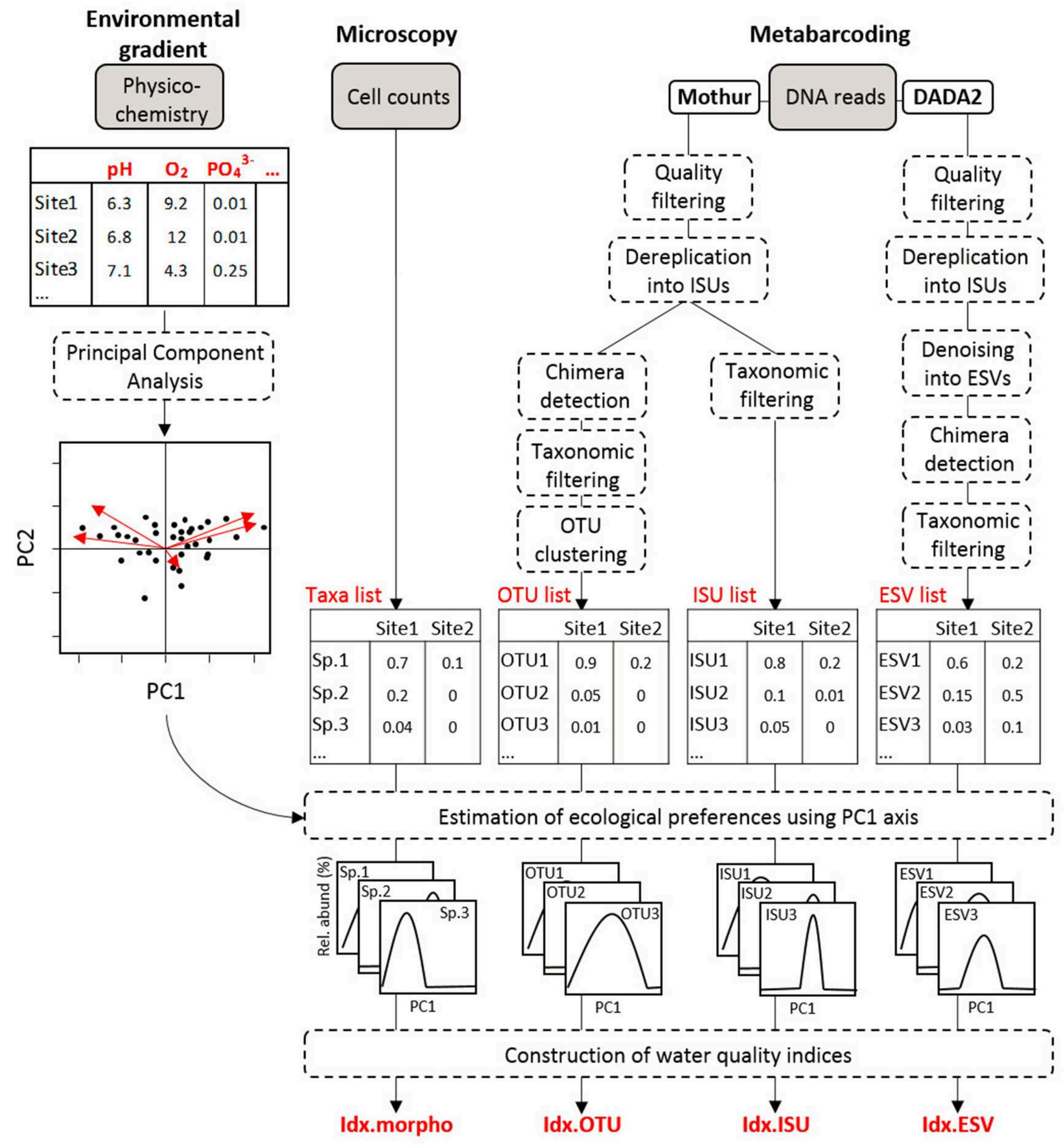
Using the files produced after the *unique.seq()* command (see section Initial Bioinformatic Steps for ISUs and OTUs), OTU table was created following the bioinformatic workflow detailed by Vasselon et al. (2017a) with some adjustments: (i) ISUs were aligned using the *align.seqs()* command and poorly aligned reads were removed using the command *screen.seqs(start=28, optimize=end, criteria=90)*; (ii) we used the *pre.cluster()* command to denoise sequencing errors by preclustering rare ISUs with related more abundant ones (1 bp threshold); (iii) detection of chimeras was performed using the *chimera.vsearch()* command; (iv) removal of "non-diatom" ISUs was performed as presented above (section Preparation of ISU Table) using the *classify.seqs()* and the *remove.lineage()* commands; (v) ISUs represented by <3 reads were removed with the *split.abund()* command; (vi) a similarity distance matrix of ISUs was created with the command *dist.seqs()*; (vii) OTU clustering was performed using the *cluster.split()* command applying the furthest neighbor method with a 95% similarity threshold.

Preparation of ESV Table

The software package DADA2 was used to infer ESVs from demultiplexed MiSeq reads (one R1 and one R2 fastq file per sample) following the methods described by Callahan et al. (2016). The DADA2 pipeline adapted to diatom metabarcoding data and applied in this study is available on Github (https://github.com/fkeck/DADA2_diatoms_pipeline) and includes : (i) for each sample, primers sequences are removed from R1 and R2 reads using cutadapt (Martin, 2011); (ii) the R1 and R2 reads are truncated to 200 and 170 nucleotides, respectively in order to remove last poor quality nucleotides; (iii) R1 and R2 reads with 0 ambiguities ("N") and a maximum of expected errors (maxEE) of 2 are conserved; (iv) after dereplication of R1 and R2 reads into ISUs, ESVs are selected based on the error rates model determined by the DADA2 denoising algorithm and paired reads merged into one sequence; (v) chimeric ESVs are removed; (vi) ESVs are taxonomically assigned using the DADA2 default parameters with an adapted version of the "diat.barcode" reference database (available on https://www6.inra.fr/carrtel-collection/Barcoding-database/Database-download); (vii) finally, a taxonomic filtering is applied in order to remove the non-Bacillariophyta (phylum) ESVs ("Bacillaryophyta_unclassified" ESVs were also discarded).

Correlation Between Community Data Tables

Prior to indices development, the correlation between morphospecies, OTU, ISU and ESV tables was assessed using the Procrustes superimposition method (Peres-Neto and Jackson, 2001). Non-metric multidimensional scaling (NMDS) on Bray-Curtis distances was used to derive a three-dimensional configuration of each table. The pairwise matching between NMDS ordinations was then measured using Procrustes correlation and tested by permutations (999 repetitions). Analyses were conducted with the *metaMDS* and *protest*



	pH	O_2	PO_4^{3-}	...
Site1	6.3	9.2	0.01	
Site2	6.8	12	0.01	
Site3	7.1	4.3	0.25	
...				

	Site1	Site2
Sp.1	0.7	0.1
Sp.2	0.2	0
Sp.3	0.04	0
...		

	Site1	Site2
OTU1	0.9	0.2
OTU2	0.05	0
OTU3	0.01	0
...		

	Site1	Site2
ISU1	0.8	0.2
ISU2	0.1	0.01
ISU3	0.05	0
...		

	Site1	Site2
ESV1	0.6	0.2
ESV2	0.15	0.5
ESV3	0.03	0.1
...		

FIGURE 1 | Overview of the analyses. The diagram indicates the steps to compute the four indices based on microscopic data (Idx_morph) and on metabarcoding data (Idx_OTU, Idx_ESV, and Idx_ISU).

functions of the R package "vegan" (R Development Core Team, 2008; Oksanen et al., 2016).

Index Development

Definition of the Reference Pressure Gradient

Principal component analysis (PCA) was executed using the *prcomp* function in R (Venables and Ripley, 2002) to study the structure of the 76 samples and their relationship to the environmental variables (**Figure 2**). Logarithmic transformation was applied on the environmental variables to ensure the normal distribution of data required for the PCA. The first principal component (PC1) represents the reference pressure gradient, i.e., the position of the samples along this gradient represent their reference quality. These values were then multiplied by −1 and then calibrated on a scale from 0 to 20, so that higher values representing better reference quality. Multiplication by −1 was necessary because higher values on the original PC1 were associated with high concentration of the variables, referring to "poor" quality.

Calculation of Diatoms Indices (Idx_morph, Idx_OTU, Idx_ISU, Idx_ESV)

The development of the four diatom indices followed the methodology described in Tapolczai et al. (2019). Both

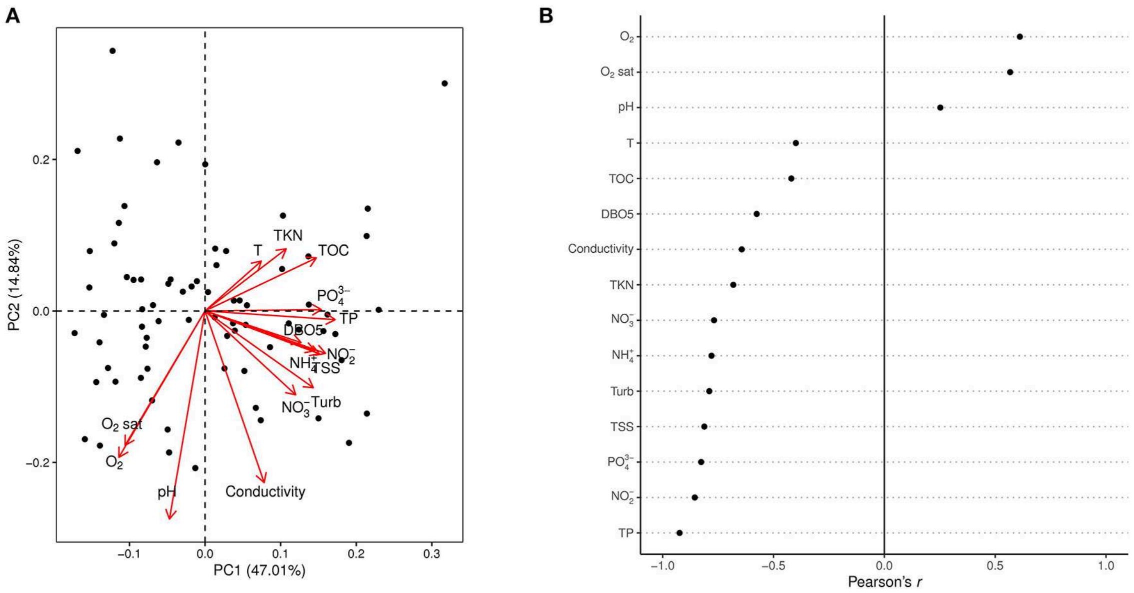


FIGURE 2 | Principal component analysis of the environmental variables. The PCA biplot **(A)** shows the projection of the sites (black dots) and the variable loadings on the first two principal components (PC1 and PC2). The dotplot **(B)** indicates the correlation (Pearson's r) of each environmental variable with the first principal component (PC1) that was used as the reference gradient for the indices. High and low values indicate strong positive and negative correlation respectively, while values close to zero indicate weak correlations between PC1 and the given parameter.

morphospecies inventory obtained via microscopic identification and sequence reads inventory obtained via HTS were transformed into relative abundances in order to ensure a comparable quantification among samples.

The four datasets according to the biological inventories (morphospecies, OTU, ISU and ESV lists) were randomly divided into: (i) a training datasets containing the randomly selected 75% of the samples, including their position along PC1 and their associated morphospecies (and OTUs, ISUs, ESVs) relative abundances; (ii) a test dataset containing the remaining 25% of the samples. Therefore, the indices could be tested on an independent dataset that was not included in the index development. This cross validation approach to randomly select training and test datasets was executed 100 times to measure the average and standard deviation of the values of the four indices at each sample instead of a single measure that could bias the results. This resulted in 100 indices tested for each of the four index types (Idx_morph, Idx_OTU, Idx_ISU, Idx_ESV) (400 indices in total).

Ecological profiles of the morphospecies, OTUs, ISUs, and ESVs in the training datasets were defined by modeling their relative abundances in the samples along PC1. Rare morphospecies, OTUs, ISUs, and ESVs were removed from the data tables and only those present in more than 5% of the samples in the training dataset were kept. This arbitrary limit, well-established in previous studies (Stenger-Kovács et al., 2007; Bere et al., 2014; Tapolczai et al., 2019), was necessary to keep a minimum number of samples based on which robust ecological profiles are ensured.

Weighted averages and standard deviations of the profiles were calculated to estimate the ecological optimum (s) and the tolerance (v) values. The Zelinka-Marvan equation (Zelinka and Marvan, 1961) was adapted to our data to define the four indices:

$$Idx\_morph/OTU/ISU/ESV = \frac{\sum_{j=1}^{n} a_j s_j v_j}{\sum_{j=1}^{n} a_j s_j}$$

where Idx_morph/OTU/ISU/ESV are the indices based on morphospecies, OTUs, ISUs, and ESVs, respectively; a_j is the relative abundance of morphospecies/OTU/ISU/ESV j; s_j is the sensitivity or optimum of morphospecies/OTU/ISU/ESV j; and v_j is the indicator value or tolerance of morphospecies/OTU/ISU/ESV j in the sample. Sensitivity and indicator values for each morphospecies, OTUs and ISUs were calculated using their abundance values plotted as functions of the samples' PC1 values. The two ecological values (sensitivity and indicator) comprised a database that was used together with the relative abundance values of the morphospecies, OTUs, ISUs, and ESVs in the samples for which the indices were calculated. Data of the training dataset was used to define these profiles. Idx_morph, Idx_OTU, Idx_ISU, and Idx_ESV for each sample in the test dataset were calculated and correlated with their corresponding PC1 values.

The R script used for the index development and data analysis is uploaded and freely available on Zenodo repository (https://doi.org/10.5281/zenodo.3463043).

Index Comparison

In order to assess and compare the performance of the four quality indices developed, several metrics were used. Correlation coefficients of the linear models fitted on mean data of quality values per sites were compared and significance tests were performed using the "cocor" package in R (R Development Core Team, 2008; Diedenhofen and Musch, 2015). The residuals of the four regression models were compared with Wilcoxon signed rank test and Bonferroni correction (Hollander and Wolfe, 1973) in order to measure the prediction performance of the models. Stability of the indices were estimated by comparing the standard deviation of index values per sites originated from the 100 iterations to select training and test datasets. It was tested with Wilcoxon signed rank test with Bonferroni correction.

RESULTS

Reference Gradient

The first principal component of the PCA (**Figure 2A**), explaining the 47.01% of the total variation in the dataset was used as the reference gradient for the indices. Poor quality is associated with those parameters indicating higher nutrient concentration, organic matter concentration, turbidity. Good quality is represented by well-oxygenated waters. pH, oxygen concentration, oxygen saturation and conductivity are the main factors responsible for the distribution of sites on the second principal component, explaining the 14.84% of the total variation. All environmental variables correlated significantly with PC1 ($p < 0.05$) with Pearson's correlation coefficients (r) presented on **Figure 2B**.

Morphological Identification

A total of 355 diatom taxa were identified via microscopic analysis from which 321 at species level. The average number of taxa identified per sample was 28 ($SD = 10$) with a minimum of 4 and with a maximum of 56. Based on our criteria to remove rare taxa, the number of taxa kept in the training datasets and used for index development varied between 110 and 141 depending on the random selection of training and test datasets, with a mean taxa number of 122 (**Table S3**).

HTS Results

The 76 samples selected for this study were part of a MiSeq (2 × 250 bp) sequencing run composed of 284 *rbc*L diatoms libraries from freshwater biofilm samples and were analyzed in a joined analysis of 464 samples. In order to allow the bioinformatic reproducibility of our study, the global dataset corresponding to the 464 samples fastq files used for the Mothur and DADA2 bioinformatics analysis are available on the Zenodo repository system (https://doi.org/10.5281/zenodo.3244156). We will only present the results obtained for 76 samples studied here.

The sequencing platform performed the demultiplexing and the contig steps, providing one fastq file per sample which generated a total of 3,071,693 DNA reads for the 76 samples with an average of 40,417 reads per sample (min = 23,140; max = 67,292). After the application of the bioinformatic procedure to generate the OTU table, 1,426,272 DNA reads remained and were clustered into 856 OTUs (95% similarity threshold) with an average of 122 OTUs per sample (min = 49; max = 236) (**Table S4**). For the generation of the ISU table, bioinformatic procedure conserved 2,008,452 DNA reads corresponding to a total of 21,241 ISUs with an average of 2,214 ISUs per sample (min = 344; max = 4,244) (**Table S5**). Regarding the ESV table, DADA2 bioinformatic procedure conserved 2,852,542 DNA reads corresponding to a total of 1,266 ESVs with an average of 96 ESVs per sample (min = 31; max = 186) (**Table S6**). Detailed information regarding the effect of bioinformatic procedures on DNA reads are summarized in **Table S7**.

Morphospecies/OTU/ISUs/ESVs Community Structure Comparison

Basic summary data of the four biological inventories are presented in **Table 1**. Logically, both the total and mean richness per sample was much higher using molecular data; the number of OTUs detected (856) was almost 2.5 times higher than the total morphospecies richness (355). ESV richness was 1,266 in total and ISU richness was several fold higher with a total and mean richness per sample of 21,241 and 2,214 ISUs, respectively. Here we note that rarefaction was not used to set all the samples to the same read number as it was not mandatory for indices development. Values of morphospecies, OTUs, ESVs, and ISUs were converted into relative proportions in the different biological tables for the different indices development. Richness values are provided just as descriptive information and not for comparison.

The four tables (morphospecies, OTU, ISU, and ESV) were all found to be correlated with each other (all $p < 0.001$). The strongest correlation was measured between the ISU and ESV tables (Procrustes correlation = 0.99). Both ISU and ESV tables were strongly correlated with the OTU table (0.87 and 0.86, respectively) and with the morphospecies table (0.78 and 0.77, respectively). Finally, the lowest correlation was found between the OTU and the morphospecies tables (Procrustes correlation = 0.67).

Distribution of Ecological Values

Ecological values (sensitivity and indicator) derived from the abundance distribution of the four kinds of biological units were defined (**Tables S8–S11**) and their distribution is presented in **Figure 3**. The general pattern for the four data types is similar to each other with a quasi-normal distribution of sensitivity values and a right skewed distribution pattern of the indicator values. Morphospecies inventory consists of the fewest data points while ISU database contains the most. Consequently, morphospecies inventory involves higher relative abundances of taxa than the abundances of unique sequences. Both the relative abundance per OTUs and the number of OTUs are between the morphospecies and ISU inventories.

ISU composition and abundance within OTUs were further analyzed in order to reveal hidden ecological information and the results are presented on **Figure 4** in the case of the ten most abundant OTUs. Within some OTUs (e.g., OTU00001, OTU00002, OTU0003, OTU00007, OTU00008, OTU00009, and OTU00010) the frequency distribution of ISU sensitivity and indicator values follow a unimodal pattern in which ecological values of the most abundant ISUs are very close to the ones of the OTU it belongs to. However, in other cases (OTU00004, OTU00005, and OTU00006), OTUs contain more abundant ISUs

TABLE 1 | Summary table indicating the number of distinct morphospecies, OTUs, ESVs, and ISUs in the entire dataset and the training datasets.

		Morphospecies	OTUs	ESVs	ISUs
Entire dataset	Total richness	355	856	1,266	21,241
	Mean richness per sample	28 (*SD* = 10)	122 (*SD* = 32)	96 (*SD* = 32)	2,214 (*SD* = 725)
	Minimum richness	4	49	31	344
	Maximum richness	56	236	186	4,244
Without rare species/OTUs/ISUs/ESVs	Richness	110–133	442–498	432–491	14,641–15,756

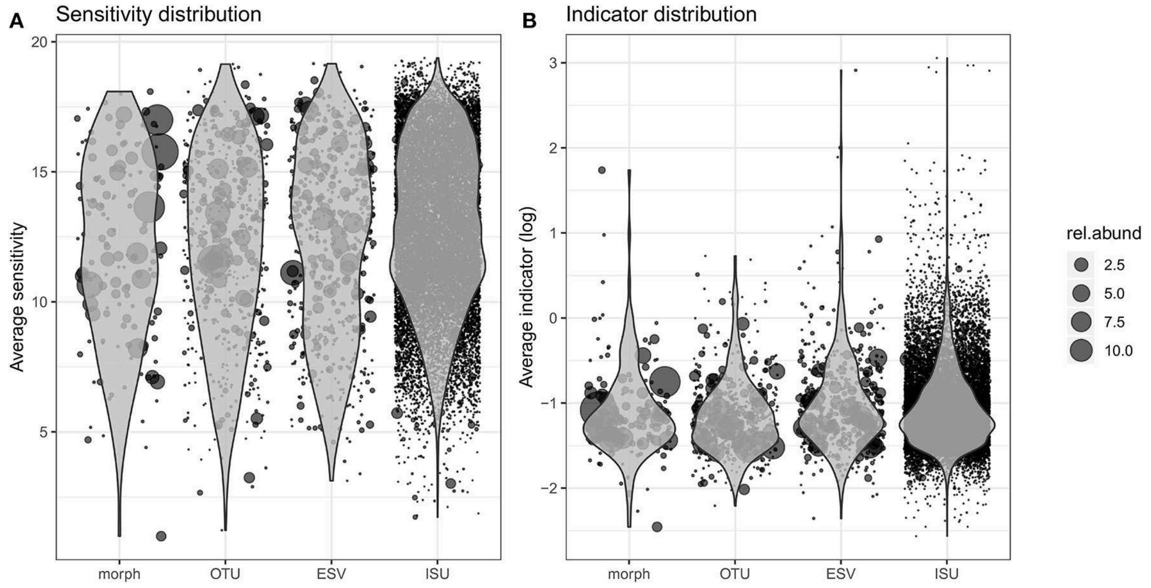


FIGURE 3 | Distribution of the sensitivity **(A)** and indicator **(B)** values estimated for each morphospecies, OTU, ESV, and ISU during the training procedure. The position of each dot corresponds to its average sensitivity/indicator values (over 100 estimates) and the size indicates its relative abundance.

whose ecological values differ from the one of the OTU they belong to.

Comparison of Indices' Values

The performance of the four indices (Idx_morph, Idx_OTU, Idx_ISU, and Idx_ESV) was assessed by fitting a linear model using the "lm" function in R (Chambers, 1992; R Development Core Team, 2008) on the relationship between the calculated index values and their corresponding reference pressure gradient values (PC1) (**Figures 5A–D**). The relationship was significant for each index ($p < 0.01$) with regression coefficient values of 0.84, 0.76, 0.84, and 0.84 for Idx_morph, Idx_OTU, Idx_ISU, and Idx_ESV, respectively. R^2 values for the correlation between Idx_ISU and PC1, Idx_ESV and PC1 and finally, Idx_morph and PC1 were significantly higher than the R^2 values of the correlation between Idx_OTU and PC1 ($p < 0.05$). The slope of the linear model however differed from the $m = 1$ value at each case, with slope values of $m = 0.49$, 0.45, 0.49, and 0.53 for Idx_morph, Idx_OTU, Idx_ESV, and Idx_ISU, respectively (**Table 2**). The Wilcoxon-test to compare prediction performance showed significantly higher MSE values (i.e., weaker prediction) for Idx_OTU (MSE = 8.73) than for all the other indices and both Idx_morph (MSE = 6.85) and Idx_ISU (MSE = 6.75) performed better in this aspect than Idx_ESV (MES = 6.98). Wilcoxon-test for the prediction instability assessed by the mean standard deviation due to the cross validation step showed that Idx_morph is more stable (mean SD from CV = 6.85) than Idx_OTU, Idx_ESV, and Idx_ISU (mean SD from CV= 0.58, 0.51, and 0.58, respectively).

DISCUSSION

De novo Construction of Morphological and Molecular Diatom Indices

In this study we developed, tested, and compared diatom indices based on morphospecies identified with microscopy and molecular taxonomic units based on metabarcoding. Similar studies aiming to develop quality indices using such approaches have been already conducted but their number is quite few (Apothéloz-Perret-Gentil et al., 2017; Cordier et al., 2017, 2018). Significant correlations in our study between the reference gradient and the predicted quality notes proved the validity of our approach the model developed on the training dataset using cross validation method could successfully be used on the test dataset. From a further aspect, we carried out a comparison of index performances based on molecular and microscopical

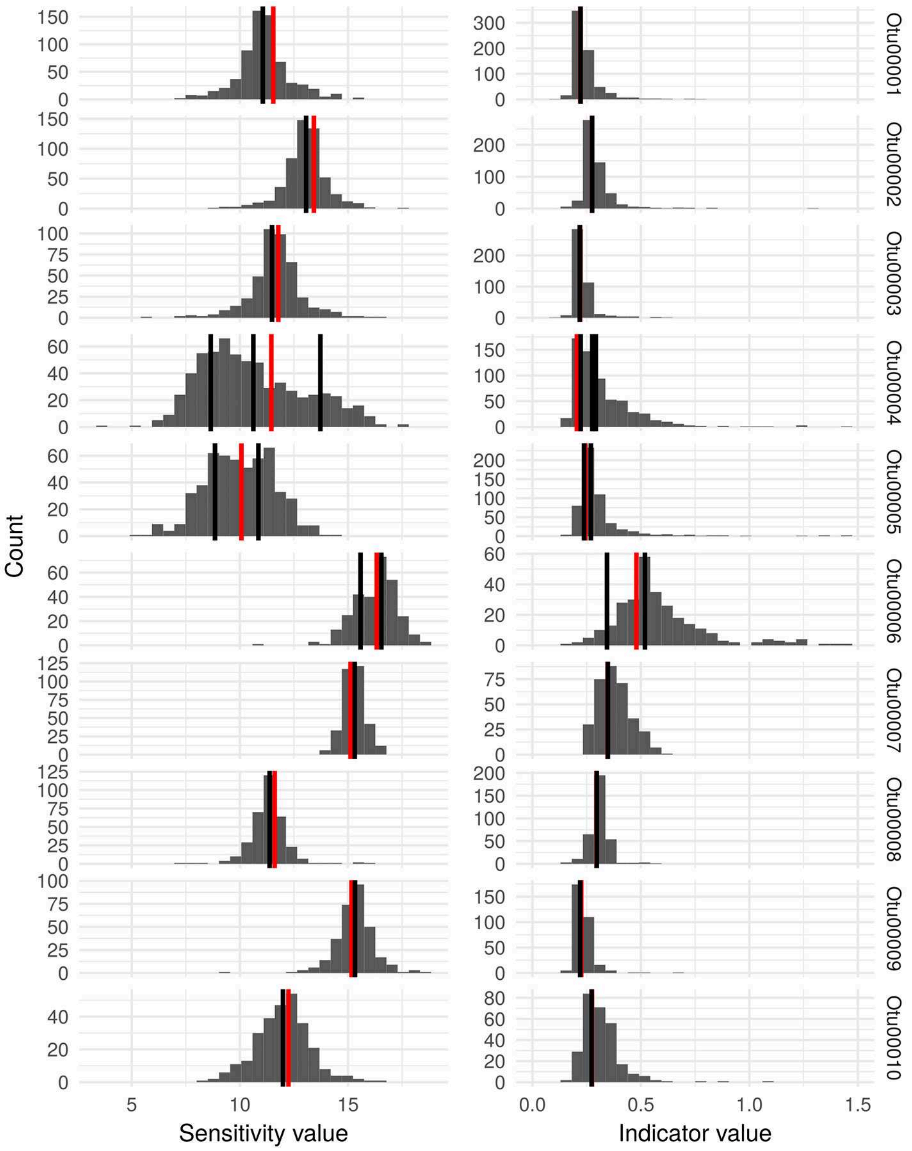


FIGURE 4 | Histograms representing the distribution of the estimated sensitivity values and indicator values in the Idx_ISU for the ISU that were clustered into the 10 most abundant OTU. For each OTU, the vertical black lines indicate the ecological values of the most abundant ISU (relative abundance in the entire dataset >0.1%) and the vertical red line shows the ecological values estimated for the complete OTU (Idx_OTU).

inventories but also studied the differences within molecular methods, between OTU-, ISU-, and ESV- based indices.

Last decade(s) has seen a tremendous evolvement in implementing molecular-based methods in biomonitoring with the purpose to improve it in terms of standardization, cost- and time-efficiency, accuracy, etc. (Leese et al., 2016). The first step of this process was to imitate biomonitoring approaches already used with microscopic data by substituting morphospecies inventories with the ones obtained via metabarcoding. Numerous studies revealed characteristic features in which OTU taxonomic inventories perform differently than morphospecies, mainly regarding taxonomic coverage issues or the quantification of the biological signal (Zimmermann et al., 2015; Vasselon et al., 2017b, 2018). Molecular data was also used to create inventories for already existing diatom indices based on morphospecies with the common drawback of uncomplete reference libraries (Kermarrec et al., 2014; Visco et al., 2015; Pawlowski et al., 2016; Rivera et al., 2018). Recent studies have started to develop OTU-based, so-called taxonomy-free indices in order to test the possibility of using such approaches in diatom-based quality assessment (Apothéloz-Perret-Gentil et al., 2017; Tapolczai et al., 2019), with promising results.

It is worth to note that the literature makes a clear distinction between taxonomy-free indices and machine learning based

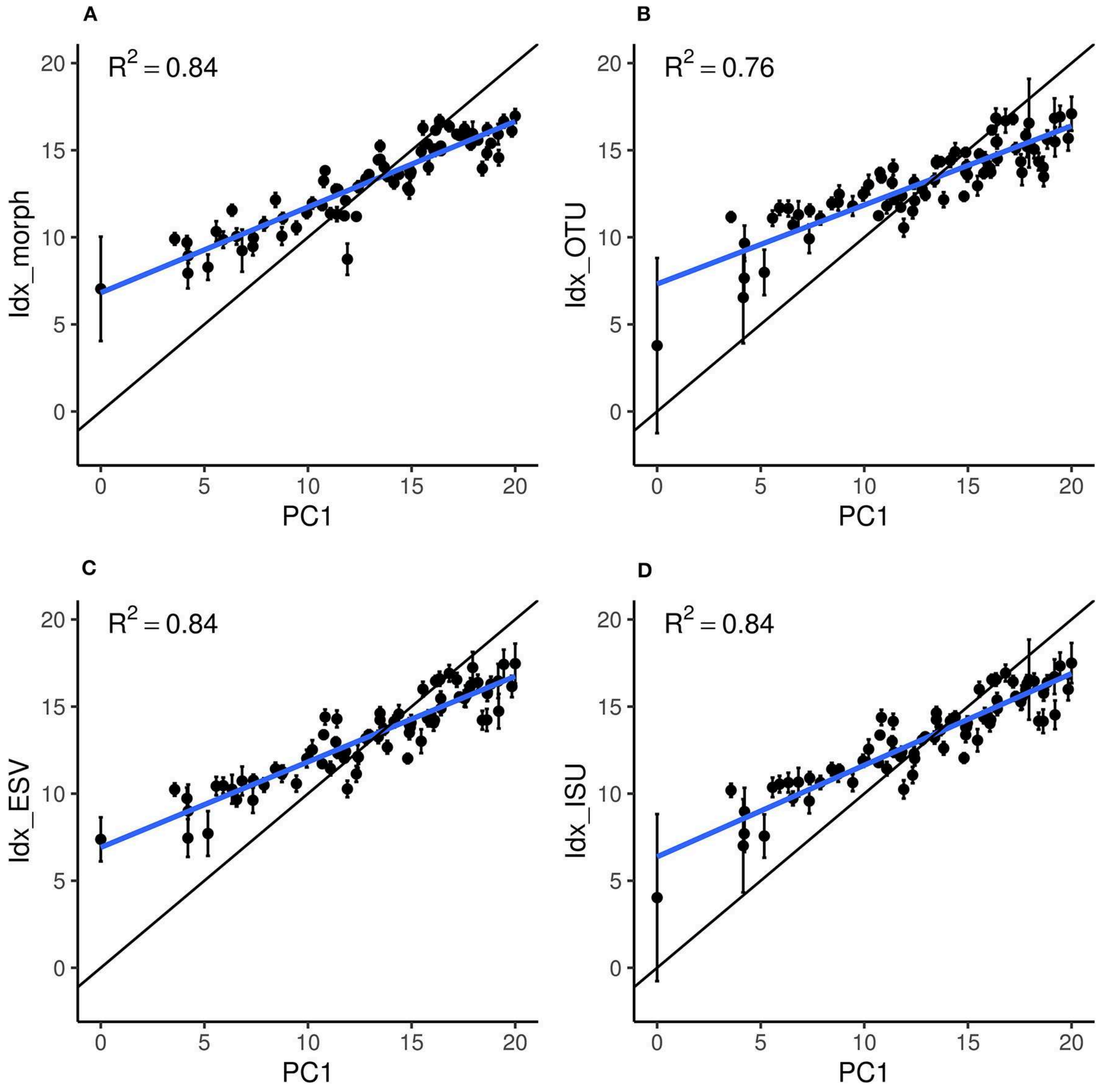


FIGURE 5 | Relation between the site scores on the reference gradient (PC1) and the scores estimated by each index: **(A)** Idx_morph, **(B)** Idx_OTU, **(C)** Idx_ESV, and **(D)** Idx_ISU. Black dots and error bars represent the average and standard deviations, respectively, over the 100 training repetitions. The thick blue line represents the estimated linear regression between PC1 and the index values. The black line materialize the perfect equivalence between PC1 and the indices (i.e., the optimal 1:1 line).

indices (Cordier et al., 2017, 2018). For consistency and clarity we advocate that this dichotomy is not relevant and the term machine learning can be employed to refer to both approaches. Indeed, machine learning is a generic term for a very broad statistical approach (basically consisting in training predictive functions and testing their performance) rather than the application of a reduced set of learning algorithms. Although derived from a simple function, taxonomy-free indices based on Zelinka and Marvan equation are obtained by optimizing morphospecies/OTU/ISU/ESV weights with a training set or through cross validation. This procedure is typical of supervised machine learning.

In this study we take a step forward and assess the performance of *de novo* developed molecular diatom indices for the first time. We do not only assessed the performance of molecular methods compared to microscopic one but we reconsidered the already existing molecular methods too. We showed that beside being a step toward a more standardized biomonitoring, the Idx_ISU unveiled hidden ecological differences between ISUs that are otherwise grouped together into the same OTU due to their high genetic similarity, masking the bioindication signal. Thus, the construction of *de novo* indices enabled a fair comparison of different approaches for the improvement in bioassessment.

TABLE 2 | Summary table of the performances of the four indices.

	Idx_morph	Idx_OTU	Idx_ESV	Idx_ISU
Linear regression slope (m)	0.49	0.45	0.49	0.53
R^2	0.84^a	0.76^b	0.84^a	0.84^a
Prediction performance MSE (Wilcoxon-test)	6.85 ($SD = 10.44$)^{ac}	8.73 ($SD = 13.61$)^b	6.98 ($SD = 10.52$)^c	6.75 ($SD = 11.79$)^a
Prediction instability/mean standard deviation from CV (Wilcoxon-test)	0.40 ($SD = 0.35$)^a	0.58 ($SD = 0.66$)^b	0.51 ($SD = 0.26$)^b	0.58 ($SD = 0.63$)^b

Superscript letters indicate significant pairwise differences detected by Wilcoxon tests.

Currently used diatom indices, as the Trophic Diatom Index (TDI; Kelly and Whitton, 1995), the Biological Diatom Index (Jean Prygiel, 2002) or the Specific Pollution sensitivity Index (Coste, 1982), were developed using the ecological profile of species along particular physical and chemical parameters related to eutrophication, organic pollution, etc. Following the strategy of previous studies of the authors (Tapolczai et al., 2017, 2019), this study used another approach by applying the first principal component of a PCA, carried out on our dataset, as the reference gradient. It is a way to integrate the effect of the several environmental parameters affecting the position of samples on this gradient. We observed that all variables measured, except pH, correlated well with the defined reference gradient. This approach avoids completely the use of already existing index values based on morphology as reference (Apothéloz-Perret-Gentil et al., 2017) and serves perfectly the comparison of the effect of different biological inventories on a newly developed quality index. One technical disadvantage of this strategy is that the gradient, together with the taxa's ecological values are specific to our data and cannot be directly used in other studies. However, they can be always linked to values of environmental parameters via their correlation with PC1. It is important to note that the ecological validity of the use of a reference based on solely physical and chemical parameters to assess ecological quality is often contested (Kelly et al., 2009; Schneider et al., 2016). The main critic is that although the WFD introduced the new fundamental concept of the ecological quality defined by the status of the biota instead of physical and chemical parameters, the methods adopted are the already existing metrics based on old concepts.

To define the ecological optimum of species, the weighted average method was used. Even though it is sometimes criticized by the literature, we used this method due to its simplicity and the fact that the majority of the diatom indices are still based on this calculation. Since the weighted average assesses species optima the best where abundance distribution of species is symmetric and unimodal, it usually overestimates the quality note of poor quality sites and underestimates the quality of high quality sites where species distributions are strongly right- and left- skewed, respectively as already shown by Tapolczai et al. (2017). Potapova et al. (2004) proposed different strategies to improve the calculation of the optima including generalized linear models or giving multiple indicator values for species based on the probability that it can be found in the different quality classes, based on the "smoothed" distribution along the reference gradient.

Comparison of the Performance of the Four Indices

As highlighted in the previous section, the *de novo* morphological (Idx_morph) and molecular (Idx_OTU, Idx_ISU, Idx_ESV) diatom indices were all relevant to predict correctly the ecological status of the study sites using machine learning approach. Despite the relation between the site scores on the reference gradient and the scores estimated with the four indices are highly similar, the Idx_morph, Idx_ISU, and Idx_ESV performed equally and outperformed the Idx_OTU.

The biological information used to compute the four indices were based on diatom morphospecies (Idx_morph), OTU (Idx_OTU), ISU (Idx_ISU), and ESV (Idx_ESV) tables. Despite the methodological and biological biases introduced by molecular and morphological approaches applied to obtain those tables (Pawlowski et al., 2018), they were all derived from the same environmental diatom community. Thus, as we expected, the community structures revealed by the four matrices were highly correlated, as shown by the procrustean analyses and already observed in previous diatom metabarcoding studies (Vasselon et al., 2017b; Rivera et al., 2018). The highest correlation was observed between ESV and ISU structures, as they are based on the same metabarcoding data, and both correlated better with morphospecies than OTU. However, we would expect OTUs to be more related to morphospecies as OTUs are supposed to be proxies for species (Porter and Hajibabaei, 2018). This may be explained partially by (i) our OTU definition, determined by the choice of the OTU clustering algorithm and the genetic distance similarity threshold applied, which may not reflect properly the morphological diatom species concept (Hugerth and Andersson, 2017; Tapolczai et al., 2019); (ii) the bioinformatics biases introduced at the different steps used to proceed raw DNA reads into OTUs, like the alignment of DNA reads, the chimera detection or the OTU clustering algorithm (Mysara et al., 2015; Edgar, 2018; Hardge et al., 2018); (iii) the consistency of OTU, as genetically close taxa may be grouped within the same OTU, reducing the final resolution of the OTUs in comparison to ISUs and ESVs (Callahan et al., 2017).

In our study, the consistency of OTU is more likely to affect the efficiency of the Idx_OTU in comparison to the Idx_ISU and Idx_ESV. By confronting the distribution of the sensitivity and indicator values of each ISU (estimated with the Idx_ISU index) to the values of their corresponding OTU (estimated with the Idx_OTU index), we observed two patterns: (i) OTU composition is consistent: the OTU is dominated by one abundant ISU and both shared similar ecological preferences; (ii) OTU composition is not consistent: the OTU is dominated by several abundant ISUs which may have various ecological preferences, the ecological preferences of the OTU corresponding to an average of the dominant ISUs values. Among the 10 most dominant OTUs observed, 3 of them appeared to be inconsistent as they were composed by several abundant ISUs with different ecological preferences (e.g., OTU00004). As the calculation method used to create the quality index gives more weight to dominant taxa (Bigler et al., 2010), the misestimation of dominant OTU ecological preferences, due to their inconsistency, reduces the efficiency of the Idx_OTU in comparison of the Idx_ISU where estimation of ISU ecological preferences is more realistic.

OTU consistency is mainly affected by methodological biases introduced during the bioinformatics steps applied to create OTUs, like the choice of the clustering method (Schmidt et al., 2014). In our study we used the furthest neighbor method as implemented in Mothur, which is known to create numerous OTUs in comparison to recently developed clustering algorithm like Opticlust (Westcott and Schloss, 2017) or Swarm (Mahé et al., 2015). However, hierarchical complete linkage method, like furthest neighbor, enables to create more consistent OTUs with ecologically consistent partitions (Schmidt et al., 2014). The sequence similarity threshold applied to define OTUs can also affect their consistency, the smaller the threshold, the greater the risk of merging genetically and ecologically diverse taxa. As we used a 95% similarity threshold, this risk is increased, however a previous study shown that the use of a threshold between 95 and 99%, using furthest neighbor clustering method, has a limited effect on the efficiency of the computed OTU index (Tapolczai et al., 2019). Furthermore, we observed that the dominant ISUs belonging to the same OTU (e.g., OTU0004) were genetically distant of only 2 or 3 nucleotides, corresponding to 1–2% of differences. So even if we had applied the 97% similarity threshold, the problem would have remained. There are some clustering algorithms though, with strategies avoiding the use of a global similarity threshold. These methods, e.g., Swarm (Mahé et al., 2015) with a d parameter equal to one would potentially separate this ecological signal. Similarly, OTU clustering based on sequence distribution among samples (Preheim et al., 2013) or the application of post-clustering curation procedure to denoise OTUs (Frøslev et al., 2017) are attempts to handle the bias of using sequence similarity threshold. However, in comparison to OTUs, ISUs and ESVs are able to take into account intra-species and intra-population variability which provide relevant ecological information for freshwater biomonitoring.

Finally, even if the Idx_ISU and Idx_ESV outperformed the Idx_OTU, it provided similar predictive power than the Idx_morph with a higher correlation slope between the expected gradient and the estimated index values, but appeared to be significantly less stable. The highest prediction instability was observed for sites corresponding to the extreme situation on the physico-chemical reference gradient characterized by few sites (particularly on polluted sites). It was already described that the instability of the index development is related to the cross validation process, used for defining training and test datasets, which is sensitive to the size of the dataset and the presence of outliers (Tapolczai et al., 2019). Even if this bias also occurred in the Idx_morph, the highest instability was observed for the Idx_ISU certainly due to the high number of ISUs obtained which fragmented the ecological signal. Furthermore, highly impacted sites are usually characterized by lower diatom richness and can contribute to increase the instability of indices based on molecular data (Tapolczai et al., 2019). This problem should be mitigated by increasing the size of the dataset.

The Place of Molecular Metabarcoding Approaches Within Actual Freshwater Biomonitoring

In the context of freshwater biomonitoring and WFD, we need transferable tools. We have shown that all indices produced (Idx_morph, Idx_OTU, Idx_ISU, Idx_ESV) are suitable to evaluate the ecological status of rivers using diatoms. However, they do not perform equally in terms of routine monitoring applicability. We already introduced the limitations related to morphological approaches (time-consuming, limiting spatio-temporal surveys, high expertise required), justifying the development of molecular biomonitoring approaches. However, these new molecular tools are not yet straightforward for stakeholders and water managers. In this study we showed that ISUs, after applying bioinformatic limited filtering steps, provide enough resolution for monitoring and offer several transferable advantages in comparison to OTUs or ESVs: (i) analysis are more reproducible as ISUs correspond to the basic untransformed unit produced with metabarcoding, without affecting their composition with algorithm (e.g., chimera detection, denoiser, clustering); (ii) they are consistent from one study to another as their identifier is the DNA sequence itself (unlike OTU); (iii) with less bioinformatics steps they are faster to analyze, require less computing power and thus tools are more easily transferable; (iv) like ESVs, they allow a higher resolution as they include intraspecific/intrapopulation level.

A further advantage of new molecular approaches is the detection of rare biosphere which might be of interest for freshwater biomonitoring. In this context, if new molecular indices are developed based on this rare biosphere, efficiency of filtered ISUs should be validated as sequencing errors may bias the ecological assessment. However, as discussed by Elbrecht et al. (2018), the increasing number of metabarcoding data obtained from freshwater sampling sites mitigates sequencing errors and the need of denoising algorithm. With this data deluge, machine learning methods combined with molecular approaches like metabarcoding will change our way to perform biomonitoring (Bohan et al., 2017).

Molecular approaches offer the possibility to increase spatial and temporal survey of freshwater monitoring networks. On the other hand, morphological approaches offer the possibility to work with ecologically meaningful information relevant for biomonitoring and not achieved by molecular ones, like morphological features observed at different life-stages of organisms, the detection of teratologic forms, as well as traits or ecosystem functions. The final objective is to improve our ability to survey and protect freshwater ecosystems, which can not be achieved with molecular based approaches alone for now. Stability of those methods is still scarce due to permanent technological and methodological evolution, meaning that molecular and morphological approaches must be used in a complementary way.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the https://doi.org/10.5281/zenodo.3244156, https://doi.org/10.5281/zenodo.3463043.

AUTHOR CONTRIBUTIONS

KT, FK, MK, FR, AB, and VV contributed to the study design and the construction of the article. KT, FK, and VV performed analysis and wrote the paper.

FUNDING

We thank COST Action DNAqua-Net (CA15219), supported by the COST (European Cooperation in Science and Technology) program which supported this work, especially through a Short-Term Scientific Mission grant. We would like to thank the Agence Française pour la Biodiversité (AFB) which founded the studies realized on the French rivers and the Institut National pour la Recherche Agronomique (INRA) which leaded the project research. We further thank the Premium Postdoctoral Research Program of the Hungarian Academy of Sciences (PPD2018-026/2018) for partly funding the study.

ACKNOWLEDGMENTS

We acknowledge all the people and institutions in charge of the French freshwater monitoring survey (governmental agencies, private consultancies) who collaborate actively to the project:

- DREALs: Aquitaine (Delphine Sagnet), Auvergne (Franck Véry), Bourgogne (Valérie Peeters), Bretagne (Gael Gicquiaud), Centre (Simon Saadat, Chafika Karabaghli), Corse (Isabelle Boulier), Franche-Comté (Eric Parmentier), Ile-de-France (Odile Cortial), Limousin (Jean Marc Vouters), Lorraine (David Heudre), Midi-Pyrénées (Eléonore Seigneur), Pays-de-la-Loire (Didier Guillard), Rhône-Alpes (Rémy Chavaux), Normandie (Frédéric Pételo), PACA (Véronique Vassal), Picardie (David Fouré), Nord-Pas-de-Calais (Nathalie Zydek).
- Water governmental agencies: Artois-Picardie (Christophe Lesniak), Rhône-Méditerranée et Corse (Loïc Imbert, Franck Repellini), Adour-Garonne (Majlis Durand, Jean-Pierre Rebillard, Margaux Saut), Rhin-Meuse (Jean-Luc Matte, Guillaume Demortier), Loire-Bretagne (Jacky Durocher), Seine-Normandie (Marie Berdoulay).
- Private consultancies: AQUABIO (Rémy Marcel, Bruno Fontan), AQUASCOP (Jessica Vizinet), ECOMA (Lénaïg Kermarrec), SAGE (Anne Rolland, Jean-Philippe Vulliet, Carole Geret), GREBE (Philippe Prompt), Eurofins Expertises Environnementales (Léa Feret).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2019.00409/full#supplementary-material

Table S1 | Description of the 15 physico-chemical parameters used to compute the principal component analysis (PCA) and to define the reference gradient.

Table S2 | *rbc*L primers, PCR1 Mix, and thermal cycler condition used for diatom DNA metabarcoding.

Table S3 | Morphospecies list of the 76 samples used as input to compute the Idx_morph. Proportions correspond to relative abundance of diatom valves counted under microscope.

Table S4 | OTU list of the 76 samples used as input to compute the Idx_OTU. Proportions correspond to relative abundance of DNA reads.

Table S5 | ISU list of the 76 samples used as input to compute the Idx_ISU. Proportions correspond to relative abundance of DNA reads.

Table S6 | ESV list of the 76 samples used as input to compute the Idx_ESV. Proportions correspond to relative abundance of DNA reads.

Table S7 | Effect of the different bioinformatics steps applied to produce OTU and ISU list on the DNA reads number.

Table S8 | Ecological preferences of each individual morphospecies estimated during the training test phase of the Idx_morph, represented by the sensitivity and the indicator value. Each value corresponds to an average of the values obtained during the 100 times cross validation procedure.

Table S9 | Ecological preferences of each individual OTU estimated during the training test phase of the Idx_OTU, represented by the sensitivity and the indicator value. Each value corresponds to an average of the values obtained during the 100 times cross validation procedure.

Table S10 | Ecological preferences of each individual ISU estimated during the training test phase of the Idx_ISU, represented by the sensitivity and the indicator value. Each value corresponds to an average of the values obtained during the 100 times cross validation procedure.

Table S11 | Ecological preferences of each individual ESV estimated during the training test phase of the Idx_ESV, represented by the sensitivity and the indicator value. Each value corresponds to an average of the values obtained during the 100 times cross validation procedure.

REFERENCES

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

Bailet, B., Bouchez, A., Franc, A., Frigerio, J.-M., Keck, F., Karjalainen, S.-M., et al. (2019). Molecular versus morphological data for benthic diatoms biomonitoring in Northern Europe freshwater and consequences for ecological status. *Metabarcod. Metagenom.* 3:e34002. doi: 10.3897/mbmg.3.34002

Baird, D. J., and Hajibabaei, M. (2012). Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Mol. Ecol.* 21, 2039–2044. doi: 10.1111/j.1365-294X.2012.05519.x

Bere, T., Mangadze, T., and Mwedzi, T. (2014). The application and testing of diatom-based indices of stream water quality in Chinhoyi Town, Zimbabwe. *Water SA* 40:503. doi: 10.4314/wsa.v40i3.14

Bigler, C., Gälman, V., and Renberg, I. (2010). Numerical simulations suggest that counting sums and taxonomic resolution of diatom analyses to determine IPS pollution and ACID acidity indices can be reduced. *J. Appl. Phycol.* 22, 541–548. doi: 10.1007/s10811-009-9490-1

Bohan, D. A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A. J., and Woodward, G. (2017). Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. *Trends Ecol. Evol.* 32, 477–487. doi: 10.1016/j.tree.2017.03.001

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi: 10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Chambers, J. M., and Hastie, T. J. (eds.). (1992). "Linear models," in *Statistical Models* (Wadsworth & Brooks/Cole Advanced Books & Software), 608.

Chen, W., Zhang, C. K., Cheng, Y., Zhang, S., and Zhao, H. (2013). A comparison of methods for clustering 16S rRNA sequences into OTUs. *PLoS ONE* 8:e70837. doi: 10.1371/journal.pone.0070837

Chonova, T., Kurmayer, R., Rimet, F., Labanowski, J., Vasselon, V., Keck, F., et al. (2019). Benthic diatom communities in an alpine river impacted by waste water treatment effluents as revealed using DNA metabarcoding. *Front. Microbiol.* 10:653. doi: 10.3389/fmicb.2019.00653

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., et al. (2017). Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning. *Environ. Sci. Technol.* 51, 9118–9126. doi: 10.1021/acs.est.7b01518

Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., and Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Mol. Ecol. Resour.* 18, 1381–1391. doi: 10.1111/1755-0998.12926

Coste, M. (1982). *Etude des méthodes biologiques quantitatives d'appréciation de la qualité des eaux.* Rapport Division Qualité des Eaux Lyon. Agence financiè de Bassin Rhone-Méditerarée Corse^ ePierre-Bénite Pierre-Bénite

Diedenhofen, B., and Musch, J. (2015). cocor: a comprehensive solution for the statistical comparison of correlations. *PLoS ONE* 10:e0121945. doi: 10.1371/journal.pone.0121945

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* 081257. doi: 10.1101/081257

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* 34, 2371–2375. doi: 10.1093/bioinformatics/bty113

Elbrecht, V., Vamos, E. E., Steinke, D., and Leese, F. (2018). Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ* 6:e4644. doi: 10.7717/peerj.4644

European Commission (2000). Directive 2000/60/EC of the European parliament and of the Council of 23 October 2000 establishing a framework for community action in the field of water policy. *Off. J. Eur. Commun.* 327, 1–73.

European Committee for Standardization (2014). *Water Quality - Guidance Standard for the Identification, Enumeration and Interpretation of Benthic Diatom Samples from Running Waters*. Brussels.

European Committee for Standardization (2016). *Water Quality - Guidance Standard for the Routine Sampling and Pretreatment of Benthic Diatoms from Rivers*. Brussels.

Forster, D., Lentendu, G., Filker, S., Dubois, E., Wilding, T. A., and Stoeck, T. (2019). Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants. *Environ. Microbiol*. doi: 10.1111/1462-2920.14764. [Epub ahead of print].

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., et al. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nat. Commun.* 8:1188. doi: 10.1038/s41467-017-01312-x

Hardge, K., Neuhaus, S., Kilias, E. S., Wolf, C., Metfies, K., and Frickenhaus, S. (2018). Impact of sequence processing and taxonomic classification approaches on eukaryotic community structure from environmental samples with emphasis on diatoms. *Mol. Ecol. Resour*. 18, 204–216. doi: 10.1111/1755-0998.12726

Hebert, P. D., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. B Biol. Sci.* 270, 313–321. doi: 10.1098/rspb.2002.2218

Hollander, M., and Wolfe, D. A. (1973). *Nonparametric Statistical Methods. 2nd Edn*. New York, NY: Wiley.

Hugerth, L. W., and Andersson, A. F. (2017). Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. *Front. Microbiol.* 8:1561. doi: 10.3389/fmicb.2017.01561

Jean Prygiel, P. C. (2002). Determination of the biological diatom index (IBD NF T 90–354): results of an intercomparison exercise. *J. Appl. Phycol.* 14, 27–39. doi: 10.1023/A:1015277207328

Kahlert, M., Kelly, M., Albert, R.-L., Almeida, S. F. P., Bešta, T., Blanco, S., et al. (2012). Identification versus counting protocols as sources of uncertainty in diatom-based ecological status assessments. *Hydrobiologia* 695, 109–124. doi: 10.1007/s10750-012-1115-z

Keck, F., Vasselon, V., Tapolczai, K., Rimet, F., and Bouchez, A. (2017). Freshwater biomonitoring in the information age. *Front. Ecol. Environ.* 15, 66–274. doi: 10.1002/fee.1490

Kelly, M., King, L., and, Ní Chatháin, B. (2009). The conceptual basis of ecological-status assessments using diatoms. *Biol. Environ. Proc. R. Ir. Acad.* 109, 175–189. doi: 10.3318/BIOE.2009.109.3.175

Kelly, M. G., and Whitton, B. A. (1995). The trophic diatom index: a new index for monitoring eutrophication in rivers. *J. Appl. Phycol.* 7, 433–444

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Frigerio, J.-M., Humbert, J.-F., et al. (2014). A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshw. Sci.* 33, 349–363. doi: 10.1086/675079

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Humbert, J. F., and Bouchez, A. (2013). Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. *Mol. Ecol. Resour.* 13, 607–619. doi: 10.1111/1755-0998.12105

Leese, F., Altermatt, F., Bouchez, A., Ekrem, T., Hering, D., Meissner, K., et al. (2016). DNAqua-net: developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. *Res. Ideas Outcomes* 2:e11321. doi: 10.3897/rio.2.e11321

Lindeque, P. K., Parry, H. E., Harmer, R. A., Somerfield, P. J., and Atkinson, A. (2013). Next generation sequencing reveals the hidden diversity of zooplankton assemblages. *PLoS ONE* 8:e81327. doi: 10.1371/journal.pone.0081327

Linhart, C., and Shamir, R. (2002). The degenerate primer design problem. *Bioinformatics* 18, S172–S181. doi: 10.1093/bioinformatics/18.suppl_1.S172

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., et al. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* 30, 434–439. doi: 10.1038/nbt.2198

Mahé, F., Rognes, T., Quince, C., Vargas, C., de, and Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* 3:e1420. doi: 10.7717/peerj.1420

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Mortágua, A., Vasselon, V., Oliveira, R., Elias, C. L., Chardon, C., Bouchez, A., et al. (2019). Applicability of DNA metabarcoding approach in the bioassessment of Portuguese rivers using diatoms. *Ecol. Indic.* 106: 105470. doi: 10.1016/j.ecolind.2019.105470

Mysara, M., Saeys, Y., Leys, N., Raes, J., and Monsieurs, P. (2015). CATCh, an ensemble classifier for chimera detection in 16S rRNA sequencing studies. *Appl. Environ. Microbiol.* 81, 1573–1584. doi: 10.1128/AEM.02896-14

Nearing, J. T., Douglas, G. M., Comeau, A. M., and Langille, M. G. I. (2018). Denoising the denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ* 6:e5364. doi: 10.7717/peerj.5364

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2016). *vegan: Community Ecology Package*. Available online at: https://CRAN.R-project.org/package=vegan (accessed September 27, 2019).

Pardo, I., Gómez-Rodríguez, C., Wasson, J.-G., Owen, R., van de Bund, W., Kelly, M., et al. (2012). The European reference condition concept: a scientific and technical approach to identify minimally-impacted river ecosystems. *Sci. Total Environ.* 420, 33–42. doi: 10.1016/j.scitotenv.2012.01.026

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., et al. (2018). The future of biotic indices in the ecogenomic era: integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Sci. Total Environ.* 637–638, 1295–1310. doi: 10.1016/j.scitotenv.2018.05.002

Pawlowski, J., Lejzerowicz, F., Apotheloz-Perret-Gentil, L., Visco, J., and Esling, P. (2016). Protist metabarcoding and environmental biomonitoring: time for change. *Eur. J. Protistol.* 55, 12–25. doi: 10.1016/j.ejop.2016.02.003

Peres-Neto, P. R., and Jackson, D. A. (2001). How well do multivariate data sets match? *The advantages of a procrustean superimposition approach over the mantel test. Oecologia* 129, 169–178. doi: 10.1007/s004420100720

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

Potapova, M. G., Charles, D. F., Ponader, K. C., and Winter, D. M. (2004). Quantifying species indicator values for trophic diatom indices: a comparison of approaches. *Hydrobiologia* 517, 25–41. doi: 10.1023/B:HYDR.0000027335.73651.ea

Preheim, S. P., Perrotta, A. R., Martin-Platero, A. M., Gupta, A., and Alm, E. J. (2013). Distribution-based clustering: using ecology to refine the operational taxonomic unit. *Appl. Environ. Microbiol.* 79, 6593–6603. doi: 10.1128/AEM.00342-13

R Development Core Team (2008). *R: A language and Environment for Statistical Computing*. Vienna: R Fondation for Statistical Computing. Available online at: http://www.r-project.org (accessed September 27, 2019).

Rivera, S. F., Vasselon, V., Jacquet, S., Bouchez, A., Ariztegui, D., and Rimet, F. (2018). Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment. *Hydrobiologia* 807, 37–51. doi: 10.1007/s10750-017-3381-2

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Schmidt, T. S. B., Rodrigues, J. F. M., and Mering, C. v. (2014). Ecological consistency of SSU rRNA-based operational taxonomic units at a global scale. *PLoS Comput. Biol.* 10:e1003594. doi: 10.1371/journal.pcbi.1003594

Schneider, S. C., Hilt, S., Vermaat, J. E., and Kelly, M. (2016). *The "Forgotten" Ecology Behind Ecological Status Evaluation: Re-Assessing the Roles of Aquatic Plants and Benthic Algae in Ecosystem Functioning*. Available online at: http://link.springer.com/10.1007/124_2016_7 (accessed June 24, 2016).

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

Stackebrandt, E., and Goebel, B. M. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int. J. Syst. Evol. Microbiol.* 44, 846–849. doi: 10.1099/00207713-44-4-846

Stenger-Kovács, C., Buczk,ó, K., Hajnal, É., and Padisák, J. (2007). Epiphytic, littoral diatoms as bioindicators of shallow lake trophic status: Trophic Diatom Index for Lakes (TDIL) developed in Hungary. *Hydrobiologia* 589, 141–154. doi: 10.1007/s10750-007-0729-z

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

Tapolczai, K., Bouchez, A., Stenger-Kovács, C., Padisák, J., and Rimet, F. (2017). Taxonomy- or trait-based ecological assessment for tropical rivers? Case study on benthic diatoms in Mayotte island (France, Indian Ocean). *Sci. Total Environ.* 607–608, 1293–1303. doi: 10.1016/j.scitotenv.2017.07.093

Tapolczai, K., Vasselon, V., Bouchez, A., Stenger-Kovács, C., Padisák, J., and Rimet, F. (2019). The impact of OTU sequence similarity threshold on diatom-based bioassessment: a case study of the rivers of Mayotte (France, Indian Ocean). *Ecol. Evol.* 9, 166–179. doi: 10.1002/ece3.4701

Tikhonov, M., Leach, R. W., and Wingreen, N. S. (2015). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J.* 9, 68–80. doi: 10.1038/ismej.2014.117

U.S. Congress (1972). *Federal Water Pollution Control Act Amendments.*

Vasselon, V., Bouchez, A., Rimet, F., Jacquet, S., Trobajo, R., Corniquel, M., et al. (2018). Avoiding quantification bias in metabarcoding: application of a cell biovolume correction factor in diatom molecular biomonitoring. *Methods Ecol. Evol.* 9, 1060–1069. doi: 10.1111/2041-210X.12960

Vasselon, V., Domaizon, I., Rimet, F., Kahlert, M., and Bouchez, A. (2017a). Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: do DNA extraction methods matter? *Freshw. Sci.* 36, 162–177. doi: 10.1086/690649

Vasselon, V., Rimet, F., Tapolczai, K., and Bouchez, A. (2017b). Assessing ecological status with diatoms DNA metabarcoding: scaling-up on a WFD monitoring network (Mayotte island, France). *Ecol. Indic.* 82, 1–12. doi: 10.1016/j.ecolind.2017.06.024

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S. 4th Edn.* New York, NY: Springer. Available online at: http://www.stats.ox.ac.uk/pub/MASS4 (accessed September 27, 2019).

Visco, J. A., Apothéloz-Perret-Gentil, L., Cordonier, A., Esling, P., Pillet, L., and Pawlowski, J. (2015). Environmental monitoring: inferring the diatom index from next-generation sequencing data. *Environ. Sci. Technol.* 49, 7597–7605. doi: 10.1021/es506158m

Westcott, S. L., and Schloss, P. D. (2017). OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. *mSphere* 2:e00073-17. doi: 10.1128/mSphereDirect.00073-17

Zelinka, M., and Marvan, P. (1961). Zur präzisierung der biologischen klassifikation der reinheit flies sender gewässer. *Arch. Hydrobiol.* 57, 389–407.

Zimmermann, J., Glöckner, G., Jahn, R., Enke, N., and Gemeinholzer, B. (2015). Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Mol. Ecol. Resour.* 15, 526–542. doi: 10.1111/1755-0998.12336

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

HYPOTHESIS AND THEORY
published: 29 October 2019
doi: 10.3389/fevo.2019.00402

Artificial Intelligence for Ecological and Evolutionary Synthesis

Philippe Desjardins-Proulx[1,2]*, *Timothée Poisot*[2] *and Dominique Gravel*[1]

[1] *Université de Sherbrooke, Sherbrooke, QC, Canada,* [2] *Université de Montréal, Montreal, QC, Canada*

The grand ambition of theorists studying ecology and evolution is to discover the logical and mathematical rules driving the world's biodiversity at every level from genetic diversity within species to differences between populations, communities, and ecosystems. This ambition has been difficult to realize in great part because of the complexity of biodiversity. Theoretical work has led to a complex web of theories, each having non-obvious consequences for other theories. Case in point, the recent realization that genetic diversity involves a great deal of temporal and spatial stochasticity forces theoretical population genetics to consider abiotic and biotic factors generally reserved to ecosystem ecology. This interconnectedness may require theoretical scientists to adopt new techniques adapted to reason about large sets of theories. Mathematicians have solved this problem by using formal languages based on logic to manage theorems. However, theories in ecology and evolution are not mathematical theorems, they involve uncertainty. Recent work in Artificial Intelligence in bridging logic and probability theory offers the opportunity to build rich knowledge bases that combine logic's ability to represent complex mathematical ideas with probability theory's ability to model uncertainty. We describe these hybrid languages and explore how they could be used to build a unified knowledge base of theories for ecology and evolution.

Keywords: artificial intelligence, theoretical biology, theoretical ecology, evolution, theoretical population genetics, machine learning, knowledge representation

OPEN ACCESS

Edited by:
David Andrew Bohan,
INRA Centre Dijon Bourgogne
Franche-Comté, France

Reviewed by:
Battle Karimi,
INRA Centre Dijon Bourgogne
Franche-Comté, France
Zacchaeus Greg Compson,
University of New Brunswick
Fredericton, Canada

Correspondence:
Philippe Desjardins-Proulx
philippe.d.proulx@gmail.com

Specialty section:
This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Ecology and Evolution

Received: *27 May 2019*
Accepted: *08 October 2019*
Published: *29 October 2019*

Citation:
Desjardins-Proulx P, Poisot T and
Gravel D (2019) Artificial Intelligence
for Ecological and Evolutionary
Synthesis. Front. Ecol. Evol. 7:402.
doi: 10.3389/fevo.2019.00402

1. INTRODUCTION

Almost four decades ago, Ralph W. Lewis argued for the formalization of evolutionary theory and the recognition of evolution as a system of theories. In his words, "when theories are partially formalized [...] the intra- and interworkings of theories become more clearly visible, and the total structure of the discipline becomes more evident" (Lewis, 1980). Supporting Lewis' point, Queller recently showed how Fisher's fundamental theorem of natural selection, Price's theorem, the Breeder equation of quantitative genetics, and other key formulas in ecology and evolution were related (Queller, 2017). In the same vein, Rice formulated an axiomatic theory of evolution based on a stochastic version of Price's theorem (Rice and Papadopoulos, 2009). These projects fall under the scope of automated theorem proving, one of the oldest and most mature branches of Artificial Intelligence (Harrison, 2009a). Theories can be written in some formal language, such as first-order logic or type theory, and then algorithms are used to ensure the theories can be derived from a knowledge base of axioms and existing results. In the last few decades, mathematicians have built knowledge bases with millions of helper theorems to assist the discovery of new ideas (Kaliszyk and Urban, 2015). For example, the Mizar Mathematical Library is a growing library of theorems, which are added after new candidate theorems are approved by the proof checker and peer-reviewed for style. Such libraries help mathematicians juggle with a growing body of knowledge and offers a concrete answer to the issue of knowledge synthesis. Mizar uses a language powerful enough for the formalization of evolutionary theories envisioned by Lewis and the result of Queller on Price's theorem and its relationship to other theories. It is also expressive enough to build a knowledge base out of Rice's axiomatic theory of evolution. Doing so would force us to think more clearly about the theoretical structure of evolution, with theoretical ecology facing a similar state of disorganization (Lewis, 1980). Case in point: theoretical community ecologists have been criticized for focusing on a single prediction for theories capable of making several (McGill et al., 2007). An example of this is Hubbell's neutral theory of biodiversity (Hubbell, 2001), which uses an unrealistic point-mutation model that does not fit with our knowledge of speciation, leading to odd predictions (Etienne and Haegeman, 2011; Desjardins-Proulx and Gravel, 2012a,b). In logic-based (also called symbolic) systems like Mizar, all formulas involving speciation would be implicitly linked together. Storing ecological theories in such knowledge base would automatically prevent inconsistencies and highlight the consequences of theories.

Despite the importance of formalization, it remains somewhat divorced from an essential aspect of theories in ecology and evolution: their probabilistic and fuzzy nature. As a few examples: a surprisingly common idea found in ecological theories is that predators are generally larger than their prey, a key assumption of the food web model of Williams and Martinez (2000); deviations from the Hardy-Weinberg principle are not only common but tend to give important information on selective pressures; and nobody expects the Rosenzweig-MacArthur predator-prey model to be exactly right. In short, important ideas in ecology and evolution do not fit the true/false epistemological framework of systems like Mizar, and ideas do not need to be derived from axiomatic principles to be useful. We are often less concerned by whether a formula can be derived from axioms than in how it fits a particular dataset. In the 1980s, Artificial Intelligence experts developed probabilistic graphical models to handle large probabilistic systems (Pearl, 1988). While probabilistic graphical models are capable of answering probabilistic queries for large systems of variables, they cannot represent or reason with sophisticated mathematical formulas. Alone, neither logic nor probability theory is enough to elucidate the structure of theories in ecology and evolution.

For decades, researchers have tried to unify probability theory with rich logics to build knowledge bases both capable of the sophisticated mathematical reasoning found in automated theorem provers and the probabilistic reasoning of graphical models. Recent advances moved us closer to that goal (Richardson and Domingos, 2006; Getoor et al., 2007; Wang and Domingos, 2008; Nath and Domingos, 2015; Hu et al., 2016; Staton et al., 2016; Bach et al., 2017). Using these systems, it is possible to check if a mathematical formula can be derived from existing results and also possible to ask probabilistic queries about theories and data. The probabilistic nature of these representations is a good fit to learn complex logical and mathematical formulas from data (Kok and Domingos, 2009). Within this framework, there is no longer a sharp distinction between *theory* and *data*, since the knowledge base defines a probability distribution over all objects, including logical relationships and mathematical formulas.

For this article, we introduce key ideas on methods at the frontier of logic and probability, beginning with a short survey of knowledge representations based on logic and probability. First-order logic is described, along with how it can be used in a probabilistic setting with Markov logic networks (Richardson and Domingos, 2006). We detail how theories in ecology and evolution can be represented with Markov logic networks, as well as highlighting some limitations. We present a case study involving a tritrophic system to demonstrate the strengths and weaknesses of Markov logic networks. Synthesis in ecology and evolution has been made difficult by the sheer number of theories involved and their complex relationships (Poisot et al., 2018). Practical representations to unify logic and probability are relatively new, but we argue they could be used to achieve greater synthesis by allowing the construction of large, flexible knowledge bases with a mix of mathematical and scientific knowledge.

2. KNOWLEDGE REPRESENTATIONS

Traditional scientific theories and models are mathematical, or logic-based. Einstein's $e = mc^2$ established a relationship between energy e, mass m, and the speed of light c. This mathematical knowledge can be reused: in any equation with energy, we could replace e with mc^2. This ability of mathematical theories to establish precise relationships between concepts, which can then be used as foundations for other theories, is fundamental to how science grows and forms an interconnected corpus of knowledge. The formula is implicitly connected to other formulas involving the same symbol, such that if we were to establish a different but equivalent way to represent the speed of light c, it could automatically substitute c in $e = mc^2$.

Artificial Intelligence researchers have long been interested in expert systems capable of scientific discoveries, or simply capable of storing scientific and medical knowledge in a single coherent system. *Dendral*, arguably the first expert system, could form hypotheses to help identify new molecules using its knowledge of chemistry (Lindsay et al., 1993). In the 1980s, *Mycin* was used to diagnose blood infections (and did so more accurately than professionals) (Buchanan and Shortliffe, 1984). Both systems were based on logic, with *Mycin* adding a "confidence factor" to its rules to model uncertainty. These expert systems generally relied on a simple logic system not powerful enough to handle uncertainty. With few exceptions, the rules were hand-crafted by human experts. After the experts established the logic formulas, the systems acted as static knowledge bases and unable to discover new rules. Algorithms have been developed to learn new logic rules from data (Muggleton and Feng, 1990; Muggleton and de Raedt, 1994), but the non-probabilistic nature of the resulting knowledge base makes it difficult to handle real-world uncertainty. In addition to expert systems, logic systems are used to store mathematical knowledge and perform automatic theorem proving (Harrison, 2009a). Pure logic has rarely been used in ecology and evolution, but recent studies have shown its ability to reconstruct food webs from data (Bohan et al., 2011; Tamaddoni-Nezhad et al., 2013).

There are many different logics for expert systems and automatic theorem proving (Harrison, 2009a; Program, 2013; Nederpelt and Geuvers, 2014). We will focus on first-order logic, the most commonly used logic in efforts to unify logic with probability. A major reason for adopting rich logics, whether first-order or higher-order, is to allow for the complex relationships found in ecology and evolution to be expressed in concise formulas. Stuart Russell noted that "the rules of chess occupy 10^0 pages in first-order logic, 10^5 pages in propositional logic, and 10^{38} pages in the language of finite automata" (Russell, 2015). Similarly, first-order logic allows us to directly express complex ecological ideas in a simple but formal language.

In mathematics, a function f maps terms $\mathbf{X}$ (its domain) to other terms $\mathbf{Y}$ (its codomain) $f:\mathbf{X} \rightarrow \mathbf{Y}$. The number of arguments of a function, $|\mathbf{X}|$, is called its arity. The atomic element of first-order logic is the **predicate**: a function that maps 0 or more terms to a truth value: false or true. In first-order logic, terms are either variables, constants, or functions. A **variable** ranges over a domain, for example x could range over integers, p over a set of species, and $city$ over a set of cities. **Constants** represent values such as 42, *Manila*, π. Lastly, **functions** map terms to other terms such as multiplication, integration, *sin*, *CapitalOf* (mapping a country to its capital). Variables have to be **quantified** either universally with ∀ (forall), existentially with ∃ (exists), or uniquely with ∃!. $\forall x : p(x)$ means $p(x)$ must hold true for all possible values of x. $\exists x : p(x)$ means it must hold for at least one value of x while $\exists! x : p(x)$ means it must hold for exactly one value of x. Using this formal notation, we can write the relationship between the basal metabolic rate (BMR) and body mass (*Mass*) for mammals (Ahlborn, 2004):

$$\forall m \in Mammal : BMR(m) = 4.1 \times Mass(m)^{0.75}. \tag{1}$$

This formula has one variable m which is universally quantified: $\forall m \in Mammal$ reads "for all m in the set *Mammal*." It has two constants: the numbers 4.1 and 0.75, along with four functions (*BMR*, *Mass*, multiplication, exponentiation). The equal sign $=$ is the sole predicate.

A first-order logic **formula** is either a lone predicate or a complex formula formed by linking formulas using the unary connective ¬ (negation) or binary connectives (*and* ∧, *or* ∨, *implication* ⇒, see **Table 1**). For example, $PreyOn(s_x, s_y)$ is a predicate that maps two species to a truth value, in this case whether the first species preys on the second species, and $IsParasite(s)$ is a predicate that is true if species s is a parasite. We could also have a function $Mass(s_x)$ mapping a species to its body mass. We can build more complex formulas from there, for example:

$$\forall s_x : \neg PreyOn(s_x, s_x). \tag{2a}$$

$$\forall s_x, s_y : PreyOn(s_x, s_y) \Rightarrow Mass(s_x) > Mass(s_y). \tag{2b}$$

$$\forall s_x, s_y : PreyOn(s_x, s_y) \wedge \neg IsParasite(s_x) \Rightarrow Mass(s_x) > Mass(s_y). \tag{2c}$$

The first formula says that species don't prey on themselves. The second formula says that predators are larger than their prey (> is a shorthand for the *greater than* predicate). The third formula refines the second one by adding that predators are larger than their prey unless the predator is a parasite. None of these rules are expected to be true all the time, which is where mixing probability with logic will come in handy. The Rosenzweig-MacArthur equation can also easily be expressed with first-order logic:

$$\forall x, y : \dot{x} = r_0\left(1 - \frac{x}{K}\right) - \frac{Cxy}{D+x} \wedge \dot{y} = X\frac{Cxy}{D+x} - \delta_0 y. \tag{3}$$

This formula has four functions: the time differential $\dot{x} \equiv dx/dt$, multiplication, addition, and subtraction. Prey x and predator y are universally quantified variables while $r_0, K, C, D, X, \delta_0$ are constants. The formula has only one predicate, $=$, and both sides of the formula are connected by ∧, the symbol for conjunction ("and").

A **knowledge base** $\mathcal{K}$ in first-order logic is a set of formulas $\mathcal{K} = \{f_0, f_1, ..., |f_{|\mathcal{K}|-1}\}$. First-order logic is expressive enough to represent and manipulate complex logic and mathematical ideas. It can be used for simple ideas such that predators are generally larger than their prey (Equation 2b), mathematical formulas for predator-prey systems equation (Equation 3), and also to establish the logical relationship between various predicates. We may want a *PreyOn* predicate to tell us whether s_x preys on s_y, but also a narrower $PreyOnAt(s_x, s_y, l)$ predicate to model whether s_x preys on s_y at a specific location l. In this case, it would be a good idea to have the formula $\forall s_x, s_y, l : PreyOnAt(s_x, s_y, l) \Rightarrow PreyOn(s_x, s_y)$. Given this formula and the data point $PreyOnAt(Wolverine, Rabbit, Quebec)$, we do not need $PreyOn(Wolverine, Rabbit)$ to be explicitly stated, ensuring the larger metaweb (Poisot et al., 2016) is always consistent with information from local food webs.

An **interpretation** defines which object, predicate, or function is represented by which symbol, e.g., it says *PreyOnAt* is a predicate with three arguments, two species and one location. The process of replacing variables with constants is called **grounding**, and we talk of ground terms / predicates / formulas when no variables are present. Together with an interpretation, a **possible world** assigns truth values to each possible ground predicate, which can then be used to assign truth values to a knowledge base's formulas. $PreyOn(s_x, s_y)$ can be neither true nor false until we assign constants to the variables s_x and s_y. Constants are typed, so a set of constants $\mathcal{C}$ may include two species $\{Gulo\ gulo, Orcinus\ orca\}$ and three locations $\{Quebec, Fukuoka, Arrakis\}$. The constants $\mathcal{C}$ yield $2^2 \times 3$ possible ground predicates for $PreyOnAt(s_x, s_y, l)$:

$$\begin{aligned}
&PreyOnAt(Gulo\ gulo, Gulo\ gulo, Quebec)\\
&PreyOnAt(Gulo\ gulo, Orcinus\ orca, Quebec)\\
&PreyOnAt(Orcinus\ orca, Orcinus\ orca, Quebec)\\
&PreyOnAt(Orcinus\ orca, Gulo\ gulo, Quebec)
\end{aligned}$$

$$PreyOnAt(Gulo\ gulo, Gulo\ gulo, Fukuoka)$$
$$\ldots$$

and only two possible ground predicates for *IsParasite*:

$$IsParasite(Gulo\ gulo)$$
$$IsParasite(Orcinus\ orca)$$

We say a possible world **satisfies** a knowledge base (or a single formula) if all the formulas are true given the ground predicates. A basic question in first-order logic is to determine whether a knowledge base $\mathcal{K}$ **entails** a formula f, or $\mathcal{K} \models f$. Formally, the entailment $\mathcal{K} \models f$ means that for all possible worlds in which all formulas in $\mathcal{K}$ are true, f is also true. More intuitively, it can be read as the formula *following from* the knowledge base (Russell and Norvig, 2009). A **proof** in first-order logic can be derived using **inference rules** such as **Modus Ponens**:

$$\frac{\alpha \Rightarrow \beta \quad \alpha}{\beta}. \tag{6}$$

This notation reads: infer β if $\alpha \Rightarrow \beta$ is true and α is true. See Harrison (2009a) for a detailed look at inference rules in first-order logic.

Probabilistic graphical models, which combine graph theory with probability theory to represent complex probability distributions, can provide an alternative to logic-based representations (Koller and Friedman, 2009; Barber, 2012). There are primarily two motivations behind probabilistic graphical models. First, even for binary random variables, we need to learn $2^n - 1$ parameters for a distribution of n variables. This is unmanageable on many levels: it is computationally difficult to do inference with so many parameters, requires a large amount of memory, and makes it difficult to learn parameters without an unreasonable volume of data (Koller and Friedman, 2009). Second, probabilistic graphical models provide important information about independences and the overall structure of the distribution. Probabilistic graphical models were also used as expert systems: *Munin* had a network of more than 1,000 nodes to analyze electromyographic data (Andreassen et al., 1996), while *PathFinder* assisted medical professionals for the diagnostic of lymph-node pathologies (Heckerman and Nathwani, 1992) (**Figure 1**).

The two key inference problems in probabilistic machine learning are finding the most probable joint state of the unobserved variables (maximum a posteriori, or MAP) and computing conditional probabilities (conditional inference). In a simple presence/absence model for 10 species ($s_0, s_1, ..., s_9$), given that we know the state of species $s_0 = Present, s_1 = Absent, s_2 = Absent$, MAP inference would tell us the most likely state for species $s_3, ..., s_9$, while conditional inference could answer queries such as $P(s_4 = Absent|s_0 = Present)$.

TABLE 1 | Common binary connectives.

			Truth table			
Name	**Common**	**Symbol**	**T × T**	**T × F**	**F × T**	**F × F**
Conjunction	and	$\wedge$	T	F	F	F
Disjunction	or	$\vee$	T	T	T	F
Implication	implies	$\Rightarrow$	T	F	T	T
Material equivalence	iff	$\Leftrightarrow$	T	F	F	T
Exclusive disjunction	xor	$\veebar$	F	T	T	F

The table shows the resulting truth value (T: True, F: False) for all possible combinations. iff *is read* if and only if. *Implication is one of the most common connective and may have surprising behavior. In particular, it will always return true when the left-side is false. While this may seem odd, it allows us to make statements such as* $\forall x \in \mathbb{R} : x \geq 0 \Rightarrow \sqrt{x^2} = x$. *This formula holds for all real numbers, including negative ones, since with* $x = -1$, $x \geq 0$ *is false and F* $\Rightarrow$ *F returns true.*

3. MARKOV LOGIC

At this point we have first-order logic, which is capable of manipulating complex logic and mathematical formulas but cannot handle uncertainty, and probabilistic graphical models, which cannot be used to represent mathematical formulas (and thus theories in ecology and evolution) but can handle uncertainty. The limit of first-order logic can be illustrated with our previous example: predators generally have a larger body weight (*Mass*) than their prey, which we expressed in predicate logic as $\forall s_x, s_y : PreyOn(s_x, s_y) \Rightarrow Mass(s_x) > Mass(s_y)$, but this is obviously false for some assignments such as $s_x : grey\ wolf$ and $s_y : moose$. However, it is still useful knowledge that underpins many ecological theories (Williams and Martinez, 2000). When our domain involves a great number of variables, we should expect useful rules and formulas that are not always true.

A core idea behind many efforts to unify rich logics with probability theory is that formulas can be weighted, with higher values meaning we have greater certainty in the formula. In pure logic, it is impossible to violate a single formula. With weighted formulas, an assignment of concrete values to variables is only *less likely* if it violates formulas, and how much less likely will depend on the weight assigned to the violated formula. The higher the weight of the formula violated, the less likely the assignment is. It is conjectured that all perfect numbers are even ($\forall x : Perfect(x) \Rightarrow Even(x)$), thus, if we were to find a single odd perfect number, that formula would be refuted. It makes sense for mathematics but for many disciplines, such as biology, important principles are only expected to be true *most* of the time. If we were to find a single predator smaller than its prey, it would definitely not make our rule useless.

The idea of weighted formulas is not new. Markov logic networks (or just Markov logic), invented a decade ago, allows for logic formulas to be weighted (Richardson and Domingos, 2006; Domingos and Lowd, 2009). Similar efforts also use weighted formulas (Hu et al., 2016; Bach et al., 2017). Markov logic supports algorithms to add weights to existing formulas given a dataset, learn new formulas or revise existing ones, and answer probabilistic queries (MAP or conditional). As a case study, Yoshikawa et al. used Markov logic to understand how events in a document were time-related (Yoshikawa et al., 2009). Their research is a good case study of interaction between traditional theory-making and artificial intelligence. The formulas they used as a starting point were well-established logic rules to understand

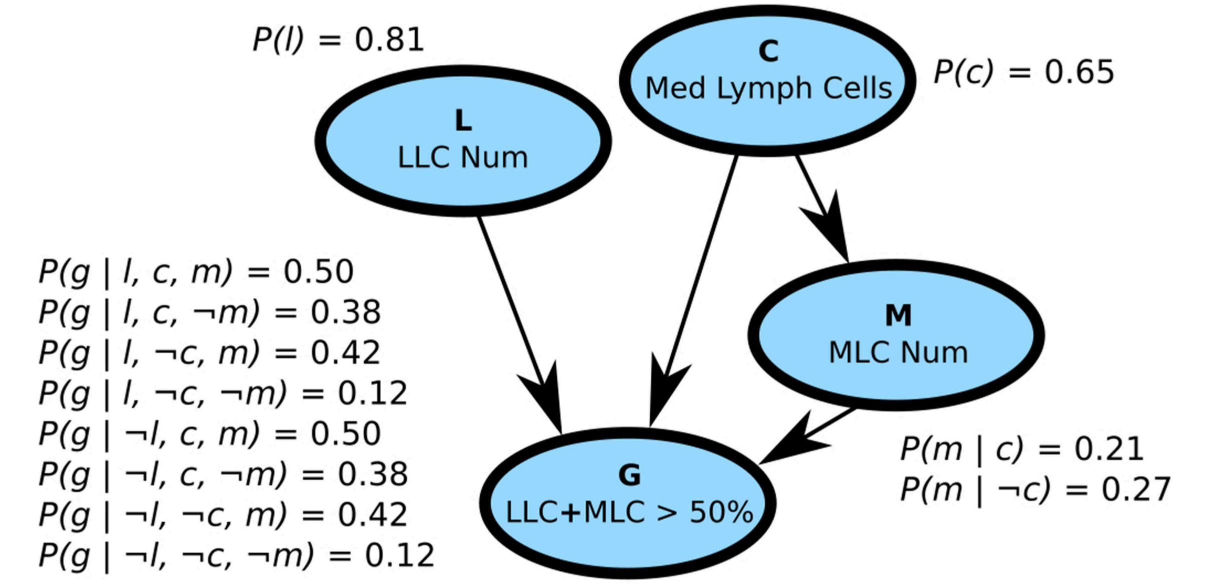


FIGURE 1 | A Bayesian network with four binary variables (the vertices) and possible conditional probability tables. Bayesian networks encode the distribution as directed acyclic graphs such that $P(\mathbf{X} = \mathbf{x}) = \prod_i P(x_i|Pa(x_i))$, where $Pa(x_i)$ is the set of parents of variable x_i. Because no cycles are allowed, the variables form an ordering so the set $Pa(x_i)$ can only involve variables already seen on the left of x_i. Thus, $P(a)P(b|a)P(c)$ is a valid Bayesian networks but not $P(a)P(b|c)P(c|b)$. The four vertices represented here were extracted from *PathFinder*, a Bayesian network with more than 1,000 vertices used to help diagnose blood infections (Heckerman and Nathwani, 1992). The vertices represent four variables related to blood cells and are denoted by a single character (in bold in the figure): C, M, L, G. We denote a positive value with a lowercase letter and a negative value with ¬ (e.g.,: $C = c$, $M = \neg m$). Since $P(\neg x|\mathbf{y}) = 1 - P(x|\mathbf{y})$, we need only $2^{|Pa(x)|}$ parameters per vertex, with $|Pa(x)|$ being the number of parents of vertex x. The structure of Bayesian networks highlights the conditional independence assumptions of the distribution and reduces the number of parameters for learning and inference. As a example query: $P(l, \neg c, m, \neg g) = P(l)P(\neg c)P(m|\neg c)P(\neg g|l, \neg c, m) = 0.81 \times (1 - 0.65) \times 0.27 \times (1 - 0.42) = 0.044$. See Darwiche (2009) for a detailed treatment of Bayesian networks and Koller and Friedman (2009) for a more general reference on probabilistic graphical models.

temporal expressions. From there, they used Markov logic to weight the rules, adding enough flexibility to their system to beat the best approach of the time. Brouard et al. (2013) used Markov logic to understand gene regulatory networks, noting how the resulting model provided clear insights, in contrast to more traditional machine learning techniques. Markov logic greatly simplifies the process of growing a base of knowledge: two research labs with different knowledge bases can simply put all their formulas in a single knowledge base. The only steps required to merge two knowledge bases is to put all the formulas in a single knowledge base and reevaluate the weights.

In a nutshell, a knowledge base in Markov logic $\mathcal{M}$ is a set of formulas $f_0, f_1, f_2, ...$ along with their weights $w_0, w_1, w_2, \ldots$:

$$\mathcal{M} = \{(f_0, w_0), (f_1, w_1), ..., (f_{|\mathcal{M}|-1}, w_{|\mathcal{M}|-1})\}. \tag{7}$$

Given constants $\mathcal{C} = \{c_0, c_1, \ldots, c_{|\mathcal{C}|-1}\}$, $\mathcal{M}$ defines a Markov network (an undirected probabilistic graphical model) which can be used to answer probabilistic queries. **Weights** are real numbers in the $-\infty$, ∞ range. The intuition is: the higher the weight associated with a formula, the greater the penalty for violating it (or alternatively: the less likely a possible world is). The **cost** of an assignment is the sum of the weights of the unsatisfied formulas (those that are false). The higher the cost, the less likely the assignment is. Thus, if a variable assignment violates 12 times a formula with a weight of 0.1 and once a formula with a weight of 1.1, while another variable assignment violates a single formula with a weight of 5, the first assignment has a higher likelihood (cost of 2.3 vs. 5). Formulas with an infinite weight act like formulas in pure logic: they cannot be violated without setting the probabilities to 0. In short, a knowledge base in pure first-order logic is exactly the same as a knowledge base in Markov logic where all the weights are infinite. In practice, it means mathematical ideas and axioms can easily be added to Markov logic as formulas with an infinite weight. Formulas with weights close to 0 have little effect on the probabilities and the cost of violating them is small. A formula with a negative weight is expected to be false. It is often assumed that all weights are positive real numbers without loss of generality since $(f, -w) \equiv (\neg f, w)$. See Jain (2011) for a detailed treatment of knowledge engineering with Markov logic. Markov logic can answer queries of complex formulas of the form:

$$P(f_0|f_1, \mathcal{M}, \mathcal{C}) = \frac{P(f_0 \wedge f_1|\mathcal{M}, \mathcal{C})}{P(f_1|\mathcal{M}, \mathcal{C})}, \tag{8}$$

where f_0 and f_1 are first-order logic formulas while $\mathcal{M}$ is a weighted knowledge base and $\mathcal{C}$ a set of constants. It's important to note that neither f_0 nor f_1 need to be in $\mathcal{M}$. Logical entailment $\mathcal{M} \models f$ is equivalent to finding $P(f|\mathcal{M}) = 1$ (Domingos and Lowd, 2009).

We build a small knowledge base for an established ecological theory: the niche model of trophic interactions (Williams and Martinez, 2000). The first iteration of the niche model posits that all species are described by a niche position N (their body size for instance) in the $[0, 1]$ interval, a diet D in the $[0, N]$ interval, and a range R such that a species preys on all species with a niche in the $[D - R/2, D + R/2]$ interval. We can represent these ideas with three formulas:

$$\forall x, y : \neg PreyOn(x, y), \tag{9a}$$

$$\forall x : D(x) < N(x), \tag{9b}$$

$$\forall x, y : PreyOn(x, y) \Leftrightarrow D(x) - R(x)/2 < N(y) \wedge N(y) < D(x) + R(x)/2, \tag{9c}$$

As pure logic, this knowledge base makes little sense. Formula 9a is obviously not true all the time but often is since most pairs of species do not interact. In Markov logic, it is common to have a formula for each lone predicate, painting a rough picture of its marginal probability (Domingos and Lowd, 2009; Jain, 2011). We could also add that cannibalism is rare $\forall x : \neg PreyOn(x, x)$ and that predator-prey relationships are generally asymmetrical $\forall x, y : PreyOn(x, y) \Rightarrow \neg PreyOn(y, x)$ (although this formula is redundant with the idea that predators are generally larger than their prey). Formulas that are often wrong are assigned a lower weight but can still provide useful information about the system. The second formula says that the diet is smaller than the niche value. The last formula is the niche model: species x preys on y if and only if species y's niche is within the diet interval of x. Using Markov logic and a dataset, we can learn a weight for each formula in the knowledge base. This step alone is useful and provides insights into which formulas hold best in the data. With the resulting weighted knowledge base, we can make probabilistic queries and even attempt to revise the theory automatically. We could find, for example, that the second rule does not apply to parasites or some group and get a revised rule such as $\forall x : \neg IsParasite(x) \Rightarrow D(x) < N(x)$.

4. FUZZINESS

First-order logic provides a formal language for expressing mathematical and logical ideas while probability theory provides a framework for reasoning about uncertainty. A third dimension often found in discussions on unifying logic with probability is fuzziness. A struggle with applying logic to ecology is that all predicates are either true or false. Even probabilistic logics like Markov logic define a distribution over *binary* predicates. Going back to Rosenzweig-MacArthur (Equation 3), this formula's weight in Markov logic is almost certainly going to be zero since it's never *exactly* right. If the Rosenzweig-MacArthur equation predicts a population size of 94 and we observe 93, the formula is false. Weighted formulas help us understand *how often a formula is true*, but in the end the formula has to give a binary truth value: true or false, there is no place for nuance. Logicians studied more flexible logics where truth is a real number in the $[0, 1]$ range. These logics are said to be "infinitely many-valued" or "fuzzy." In this setting: 0 is false, 1 is true, and everything in-between is used to denote nuances of truth (Zadeh, 1965; Behounek et al., 2011). Predicates returning truth values in the $[0, 1]$ range are called **fuzzy predicates**, while standard predicates returning *false*, *true* are said to be bivalent. To show fuzziness in action, let's look at a simple formula that says that small populations experience exponential growth:

$$\forall s, l, t : SmallPopSize(s, l, t) \Rightarrow N(s, l, t + 1) = R(s) \times N(s, l, t). \tag{10}$$

Variables s, l, t, respectively, denote a species, a location, and time. Function N returns the population size of a species at a specific location and time while function R returns the growth rate of the species. The predicates *SmallPopSize* and $=$ are both problematic from a bivalent perspective. Equality poses problem for the same reason it did with the Rosenzweig-MacArthur example: we do not expect this formula to be exactly right. The notion of a small population size should also be flexible, yet logic forces us to determine a strict threshold where *SmallPopSize* will change from true to false. Using truth values in the $[0, 1]$ range makes it possible to have a wide range of nuances for both *SmallPopSize* and equality. See **Table 2** for the definitions of fuzzy logic connectives.

TABLE 2 | Definitions of logic connectives for the three main fuzzy logics.

	Logic		
Connective	**Lukasiewicz**	**Gödel-Dummett**	**Product**
$x \wedge y$	$max(0, x + y - 1)$	$min(x, y)$	$x \times y$
$x \vee y$	$min(1, x + y)$	$max(x, y)$	$x + y - x \times y$
$x \Rightarrow y$	$min(1, 1 - x + y)$	1 if $x \leq y$, y otherwise	1 if $x \leq y$, y/x otherwise
$\neg x$	$1 - x$	0 if $x > 0$, 1 otherwise	0 if $x > 0$, 1 otherwise

These three logics are said to be normal, meaning they behave exactly like classical logic when restricted to truth values of 0 (false) and 1 (true). When truth values are between 0 and 1, these logics will often behave differently than classical logic. For example, in both classical and Lukasiewicz logics, $\neg\neg x \equiv x$, but it is not the case for Gödel-Dummett and Product logics (unless $x \in \{0, 1\}$). Another example is that conjunction and disjunction are idempotent in classical and Gödel-Dummett logics, meaning $x \wedge x \equiv x$ and $x \vee x \equiv x$, but it is not the case for Lukasiewicz and Product logics. See Behounek et al. (2011) for a detailed explanation of how the connectives are defined.

Fuzzy logic is not a replacement for probability theory. The most interesting aspect of fuzzy logic is how it interacts with probability theory to form truly flexible languages. For examples, fuzzy predicates are used in both probabilistic soft logic (Kimmig et al., 2012; Bach et al., 2017) and deep learning approaches to predicate logic (Hu et al., 2016; Zahavy et al., 2016). Hybrid Markov logic (Wang and Domingos, 2008; Domingos and Lowd, 2009) extends Markov logic by allowing not only weighted formulas but terms like soft equality, which applies a Gaussian penalty to deviations from equality. While not exactly a full integration of fuzzy logic into Markov logic, soft equality behaves in a similar matter and is a good fit for formulas like the Rosenzweig-MacArthur system or our previous example with exponential growth. Hybrid Markov logic is not as well-developed as standard Markov logic, for example there are no algorithms to learn new formulas from data. On the other hand, Hybrid Markov logic solves many of the problems

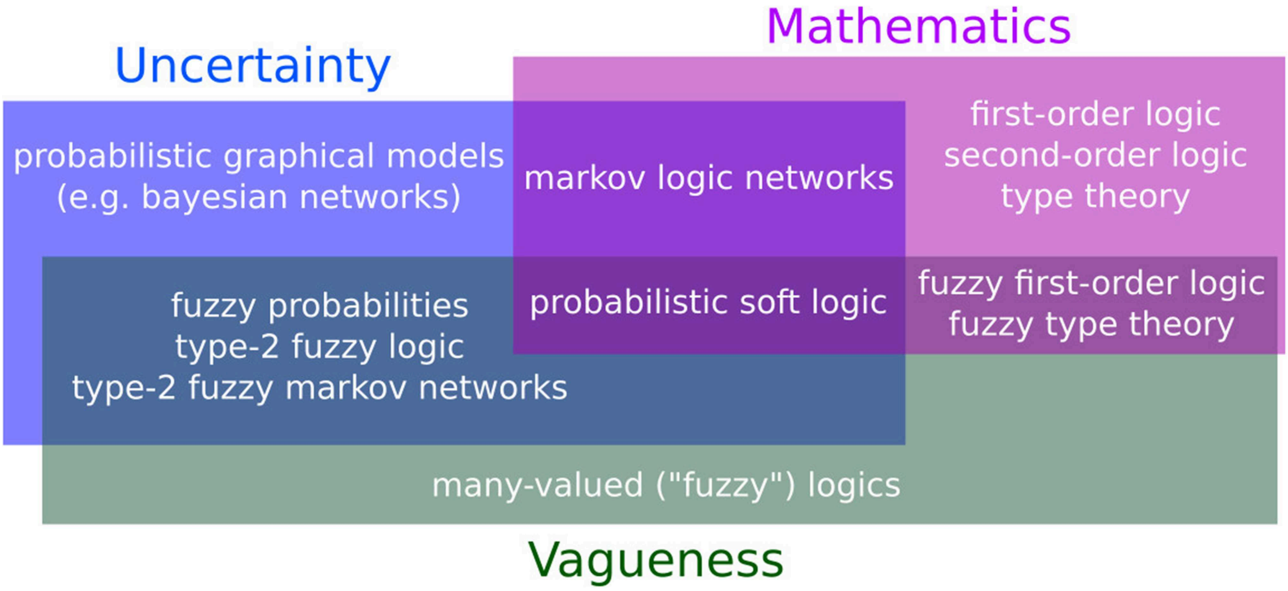


FIGURE 2 | Various languages and their ability to model uncertainty, vagueness, and mathematics (the size of the rectangles has no meaning). In the blue rectangle: languages capable of handling uncertainty. Probabilistic graphical models combine probability theory with graph theory to represent complex distributions (Koller and Friedman, 2009). Alternatives to probability theory for reasoning about uncertainty include possibility theory and Dempster-Shafer belief functions, see Halpern (2003) for an extended discussion. In the green rectangle: Fuzzy logic extends standard logic by allowing truth values to be anywhere in the [0, 1] interval. Fuzziness models vagueness and is particularly popular in linguistics, engineering, and bioinformatics, where complex concepts and measures tend to be vague by nature. See Kosko (1990) for a detailed comparison of probability and fuzziness. In the purple rectangle: languages capable of modeling mathematical formulas. It is important to note that while first-order logic is expressive enough to express a large class of mathematical ideas, many languages rely on a restricted from of first-order logic without functions. Alone, these languages are not powerful enough to express scientific ideas, we must thus focus on what lies at their intersection. Type-2 Fuzzy Logic is a fast-expanding (Sadeghian et al., 2014; Mendel, 2017) extension to fuzzy logic, which, in a nutshell, models uncertainty by considering the truth value itself to be fuzzy (Mendel and Bob John, 2002; Zeng and Liu, 2008). Markov logic networks (Richardson and Domingos, 2006; Domingos and Lowd, 2009) extends predicate logic with weights to unify probability theory with logic. Probabilistic soft logic (Kimmig et al., 2012; Bach et al., 2015) also has formulas with weights, but allows the predicates to be fuzzy, i.e., have truth values in the [0, 1] interval. Some recent deep learning studies also combine all three aspects (Garnelo et al., 2016; Hu et al., 2016).

caused by bivalent predicates while retaining the ability to answer conditional queries. In the next section we'll explore hybrid Markov logic and its application to an ecological dataset. Several languages for reasoning have combined fuzziness with probability or logic (**Figure 2**). It has been argued that, in the context of Bayesian reasoning, fuzziness plays an important role in bridging logic with probability (Jacobs and Zanasi, 2018; Nedbal and Serafini, 2018). However, how to effectively combine rich logics with probability theory remains an open question, as is the role of fuzziness.

5. MARKOV LOGIC AND THE SALIX TRITROPHIC SYSTEM

The primary goal of unifying logic and probability is to be able to grow knowledge bases of formulas in a clear, precise language. For Markov logic, it means a set of formulas in first-order logic. For this example, we used Markov logic to build a knowledge base for ecological interactions around the Salix dataset (Kopelke et al., 2017). The Salix dataset has 126 parasites, 96 species of gallers (insects), and 52 species of salix, forming a tritrophic ecological network (*Parasite* → *Galler* → *Salix*). Furthermore, we have partial phylogenetic information for the species, their presence/absence in 374 locations, interactions, and some environmental information on the locations. To fully illustrate the strengths and limits of Markov logic in this setting, we will not limit ourselves to the data available for this particular dataset (e.g., we do not have body mass for all species).

TABLE 3 | A sample of three tables for the Salix dataset (Kopelke et al., 2017).

PreyOnAt			***IsParasitoid***	***HighTemperature***
Amorri	Ovesic	Site060	Ppecti	Site006
Chalci	Halien	Site116	Psoemi	Site311
Ireuni	Hpolit	Site291	Tspone	Site296
Eacicu	Ovimin	Site121	Tsptwo	Site183
…	…	…	…	…

*Species are denoted by the first six letters of their names while sites are numbered from 1 to 374. Data in first-order logic is often organized in tables with one table per predicate and where entries represent true values while absent combinations are assumed to be false. For example, given this sample, HighTemp(Site006) is true while HighTemp(Site001) would be false. The full data formatted for Alchemy-2 (Richardson and Domingos, 2006) is provided as **Supplementary Material**.*

Data in first-order logic can be organized as a set of tables (one for each predicate). For our example, we have a table named *PreyOnAt* with three columns (its arguments) and a table named *IsParasitoid* with only one column. This format implies the closed-world assumption: if an entry is not found, it is false (see **Table 3** for an example). For this problem we defined several functions and predicates to describe everything from predator-prey relationships, whether pairs of species often co-occurred, along with information on locations such as humidity, precipitation, and temperature (see **Table 4**). We ran the basic learning algorithm from Alchemy-2 (Richardson and Domingos, 2006), which is used both to learn new formulas and weight them. The weights are listed at the end of each formula. We use the "?" character at the end of the formula involving data that were unavailable for this dataset (and thus, we could not learn the weight). Here's a sample of a knowledge base where the first three formulas were learned directly from our dataset and the last two serve as example for Hybrid Markov logic:

$$\forall s_0, s_1 : \mathit{IsGaller}(s_0) \wedge \mathit{PreyOn}(s_0, s_1) \;\Rightarrow\; \mathit{IsSalix}(s_1), 4.15. \tag{11a}$$

$$\forall s_0, s_1 : \mathit{IsParasitoid}(s_0) \wedge \mathit{PreyOn}(s_0, s_1) \;\Rightarrow\; \mathit{IsGaller}(s_1), 3.49. \tag{11b}$$

$$\forall s_0, s_1 : \mathit{PreyOn}(s_0, s_1) \;\Rightarrow\; \mathit{HighCooccurrence}(s_0, s_1), 1.57. \tag{11c}$$

$$\forall s_0, s_1, \exists \alpha : \mathit{PPreyOn}(s_0, s_1) \;\approx\; \alpha \exp\left(-2(N(s_1) - C(s_0))^2 / R(s_0)\right)? \tag{11d}$$

$$\forall s_0, s_1 : \mathit{CloselyRelated}(s_0, s_1) \wedge T(\mathit{Occ}(s_0)) > T(\mathit{Occ}(s_1)) \;\Rightarrow\; \mathit{Mass}(s_0) > \mathit{Mass}(s_1)? \tag{11e}$$

The first two formulas correctly define the tritrophic relationship between parasites, galler and salix, while the third shows a solid, but not as strong, relationship between predation and co-occurence. Formula (9c) would require hybrid Markov logic and a fuzzy predicate $\approx$.

Integration of macroecology and food web ecology may rely on a better understanding of macroecological rules (Baiser et al.). These rules are easy to express with first-order logic, for example Equation (11e) is a formulation of Bergmann's rule. We also used the learning algorithm to test whether closely related species had similar prey, but the weight attributed to the formula was almost zero, telling us the formula was right as often as it was wrong:

$$\forall s_0, s_1 : \mathit{CloselyRelated}(s_0, s_1) \wedge \mathit{PreyOn}(s_0, s_2) \Rightarrow \mathit{PreyOn}(s_1, s_2), 0.00. \tag{12}$$

This example shows both the promise and the current issues with hybrid logic-probabilistic techniques. Many of the predicates would benefit from being fuzzy, for example, *PreyOn* should take different values depending on how often predation occurs. We also had to use arbitrary cut-offs for predicates like *CloselyRelated* and *HighTemperature*. Fortunately, many recent approaches integrate logic with both fuzziness and probability theory (Adams and Jacobs, 2015; Hu et al., 2016; Bach et al., 2017). Weights are useful to understand which relationship is strong in the data, and this example shows the beginning of a knowledge base for food web ecology. The next step would be to discover new formulas, whether manually or using machine learning algorithms, and add data to revise the weights. If a formula involves a predicate operating on food webs and we want to apply our knowledge base to a dataset without food webs, this formula will simply be ignored (because it won't have grounded predicates to evaluate it; see section 2). This is a strong advantage of this knowledge representation: our little knowledge base here can be used as a basis for any other ecological datasets even if they quite different. With time, it's possible to grow an increasingly connected knowledge base, linking various ideas from different fields together.

6. BAYESIAN HIGHER-ORDER PROBABILISTIC PROGRAMMING

Artificial Intelligence has a long history with first-order logic (Russell and Norvig, 2009) but type theory (or higher-order logic), a more expressive logic, is currently more popular both as a tool to formalize mathematics and as foundation for programming languages. We explored hybrid approaches based on first-order logic and, for this section, we'll briefly discuss Bayesian Higher-Order Probabilistic Programming (BHOPP) along with its relationship with type theory. Probabilistic programming languages are programming languages built to describe probabilistic models and simplify the inference process. Stan Carpenter et al. (2017) and BUGS Lunn et al. (2012) are two popular examples of probabilistic programming languages used for Bayesian inference, but even more flexible languages for Bayesian probabilistic programming have recently emerged. These languages, like Church Goodman et al. (2008) and Anglican Wood et al. (2014), accept higher-order constructs (that is: functions accepting other functions as arguments). The ambition is that "ultimately we would like simply to be able to do probabilistic programming using any existing programming language as the modeling language" (van de Meent et al., 2018).

First-order logic allowed us to model intricate theories but, in practice, almost all modern systems used to formalize mathematics are based on type theory (higher-order logic) (Nederpelt and Geuvers, 2014). The "first" in first-order logic refers to the limitation that quantification can only be done on individual elements of a set, but not on higher-order structures like sets, predicates, or functions. As a consequence, several important concepts in mathematics cannot be formalized directly with first-order logic. Since type theory supports higher-order quantification, it is used as a foundation to reason about mathematics. Coq, HOL, HOL Light, and Microsoft's LEAN are all popular languages for automated theorem proving based on different forms of type theory (The Coq Development Team, 2004; Harrison, 2009b; de Moura et al., 2015). Programming languages in general, not just those targeted at mathematicians, tend to also rely on type theory as foundation (Pierce, 2002). See Farmer (2008)

TABLE 4 | Predicates and functions used for the Salix example.

Functions	Meaning
$PPreyOn : species \times species \mapsto [0, 1]$	Probability that a species preys on another
$PreyOn : species \times species \mapsto bool$	Predator-prey relationship
$PreyOnAt : species \times species \times location \mapsto bool$	Predator-prey relationship at a given location
$PresenceAt : species \times location \mapsto bool$	Presence of a species at a location
$IsParasite : species \mapsto bool$	Whether the species is a parasite
$IsGaller : species \mapsto bool$	Whether the species is a galler
$IsSalix : species \mapsto bool$	Whether the species is a salix
$CloselyRelated : species \times species \mapsto bool$	Whether two species are closely related
$Occ : species \mapsto \{location\}$	Set of locations where a species is found
$Cooccurrence : species \times species \mapsto \mathbb{R}^+$	Proportion of locations where the species co-occur
$HighCooccurrence : species \times species \mapsto bool$	Pair of species with high co-occurence
$HighTemperature : location \mapsto bool$	Location with above-average temperature
$T : \{location\} \mapsto \mathbb{R}$	Mean temperature for a set of locations
$Mass : species \mapsto \mathbb{R}^+$	Mean adult body mass for a species
$FoodWeb : location \mapsto Graph$	Food web at a given location
$Connectance : Graph \mapsto \mathbb{R}^+$	$Edges/Vertices^2$
$SpeciesRichness : Graph \mapsto \mathbb{N}$	Number of species in the food web
$N : species \mapsto \mathbb{R}^+$	Niche of species per Williams et al. (2010)
$C : species \mapsto \mathbb{R}^+$	Diet of the species per Williams et al. (2010)
$R : species \mapsto \mathbb{R}^+$	Range of species' diet per Williams et al. (2010)

A predicate is simply a function mapping to a boolean value (false or true, denoted bool). $\mathbb{N}$ stands for natural numbers (0, 1, 2,...) while $\mathbb{R}$ stands for real numbers, and [0, 1] is a shorthand for a real number in the [0, 1] range. We must often force continuous values into boolean values. For example, HighTemperature and CloselyRelated both require arbitrary cutoffs, often the line between true and false is set at the mean. Recent languages push for greater integration with fuzziness, which would allow predicates to take any values in the [0, 1] range.

and Nederpelt and Geuvers (2014) for an introduction to type theory. Here is where it gets confusing: the *higher* in higher-order logic has a different meaning than in *higher-order probabilistic programming* and yet, Bayesian higher-order probabilistic programming languages (BHOPPL) may hold the key to sound inference mixed with type theory. In BHOPPL, *higher-order* means functions can take functions as arguments, a common capability of modern programming languages. This is necessary for higher-order logic but not sufficient. Where it gets exciting is that a lot of progress is being made in framing BHOPPL in the language of type theory (Borgström et al., 2016). In effect, it would bring Bayesian and higher-order logic reasoning together.

Furthermore, software-wise, BHOPPLs are well ahead of the approaches described in previous sections such as hybrid Markov logic networks. Current higher-order probabilistic programming languages operate on variants of well-known languages: Anglican is based on Clojure (Wood et al., 2014), Pyro is based on Python (Bingham et al., 2019), Turing.jl uses Julia (Ge et al., 2018). Many BHOPPLs have been designed to exploit the high-performance architecture developed for deep learning such as distributed systems of GPUs (graphics cards). GPUs have been important in the development of fast learning and inference in deep learning (Goodfellow et al., 2016). Pyro (Bingham et al., 2019) is a BHOPPL built on top of PyTorch, one of the most popular frameworks for deep learning, allowing computation to be distributed on systems of GPUs. In contrast, there are no open-source implementations of Markov logic networks running on GPUs. The main downside of BHOPPLs is that, while in theory they may support the richer logics used to formalize modern mathematics, in practice higher-order probability theory is itself not well understood. This is an active research topic (van de Meent et al., 2018) but formalization faces serious issues. For one, there are incompatibilities with the standard measure-theoretic foundation of probability theory, which may require rethinking how probability theory is formulated (Borgströ et al., 2011; Staton et al., 2016; Heunen et al., 2017; Staton, 2017; Ścibior et al., 2018). First-order logic is among the most studied formal languages, making it easy to use a first-order knowledge base with various software. The current informal nature of BHOPPLs make them hard to recommend for the synthesis of knowledge in ecology and evolution, even though they may very well hold the most potential.

7. WHERE'S OUR UNREASONABLY EFFECTIVE PARADIGM?

Legitimate abstractions can often obfuscate how much various subfields are related. Natural selection is a good example. Many formulas in population genetics rely on fitness. Nobody disputes the usefulness of this abstraction, it allows us to think about changes in populations without worrying whether selection is caused by predation or climate change. On the other hand, fitness has also allowed the development of theoretical population genetics to evolve almost independently of ecology. There is a realization that much of the complexity of evolution is related to how selection varies in time and space, which puts evolution in ecology's backyard (Bell, 2010). Achieving Lewis' goal of formalization would not prevent the use of fitness, but having formulas with fitness cohabiting with formulas explaining the components of fitness would implicitly link ecology and evolution. This goes in both directions: what are the consequences of new discoveries on speciation and adaptive radiations on the formation of metacommunities? How can community dynamics explain the extinction and persistence of new species? If there isn't a single theory of biodiversity, the imperative is to understand biodiversity as a system of theories. Given the scope of ecology and evolution and the vast number of theories involved, it seems difficult to achieve a holistic understanding without some sort of formal system to see how the pieces of the puzzle fit together. Connolly et al. noted how theories for metacommunities were divided between those derived from first principles and those based on statistical methods (Connolly et al., 2017). In systems unifying rich logics with a probabilistic representation, this distinction does not exist, theories are fully realized as symbolic and statistical entities. Efforts to bring theories in ecology and evolution into a formal setting should be primarily seen as an attempt to put them in context, to force us to be explicit about our assumptions and see how our ideas interact (Suppes, 1968).

Despite recent progress at the frontier of logic and probability, there are still practical and theoretical issues to overcome to make a large database of knowledge for ecology and evolution possible. Inference can be difficult in rich knowledge representations, not all methods have robust open-source implementations, and some approaches such as Bayesian higher-order probabilistic programming are themselves not well understood. Plus, while mathematicians benefit from decades of experience making large databases of theorems, there have been no such efforts for ecology and evolution. Lewis' case for the formalization is worth repeating: "when theories are partially formalized [...] the intra- and interworkings of theories become more clearly visible, and the total structure of the discipline becomes more evident" (Lewis, 1980). This vision might soon become reality thanks to increased access to data in evolution and evolution and recent advances at the frontier of logic and probability. Given the pressing need to understand a declining biodiversity, ecologists and evolutionary biologists should be at the forefront of the efforts to organize theories in unified knowledge bases.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

PD-P wrote the manuscript, designed, and performed the experiments. PD-P, TP, and DG edited the manuscript.

FUNDING

PD-P has been funded by an Alexander Graham Bell Graduate Scholarship from the National Sciences and Engineering Research Council of Canada, an Azure for Research award from Microsoft, and benefited from the Hardware Donation Program from NVIDIA. DG was funded by the Canada Research Chair program and NSERC Discovery grant. TP was funded by an NSERC Discovery grant and an FQRNT Nouveau Chercheur grant.

ACKNOWLEDGMENTS

We thank three reviewers for their helpful comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2019.00402/full#supplementary-material

REFERENCES

The Coq development team (2004). *The Coq Proof Assistant Reference Manual*. LogiCal Project. Version 8.0.

Adams, R., and Jacobs, B. (2015). A type theory for probabilistic and bayesian reasoning. *CoRR* abs/1511.09230.

Ahlborn, B. (2004). *Zoological Physics: Quantitative Models of Body Design, Actions, and Physical Limitations of Animals*. Berlin; Heidelberg: Springer.

Andreassen, S., Rosenfalck, A., Falck, B., Olesen, K. G., and Andersen, S. K. (1996). Evaluation of the diagnostic performance of the expert EMG assistant MUNIN. *Electroencephalogr. Clin. Neurophysiol.* 101, 129–144. doi: 10.1016/0924-980X(95)00252-G

Bach, S., Broecheler, M., Huang, B., and Getoor, L. (2015). Hinge-loss markov random fields and probabilistic soft logic. *arXiv: 1505.04406*.

Bach, S., Broecheler, M., Huang, B., and Getoor, L. (2017). Hinge-loss markov random fields and probabilistic soft logic. *J. Mach. Learn. Res.* 18, 1–67.

Baiser, B., Gravel, D., Cirtwill, A., Dunne, J., Fahimipour, A., Gilarranz, L., et al. (2009). Ecogeographical rules and the macroecology of food webs. *Glob. Ecol. Biogeogr.* 28, 1204–1218. doi: 10.1111/geb.12925

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge: Cambridge University Press.

Behounek, L., Cintula, P., and Hájek, P. (2011). "Introduction to mathematical fuzzy logic," in *Handbook of Mathematical Fuzzy Logic volume 1*, Chap. 1, eds P. Cintula, P. Hájek, and C. Noguera (London: College Publications), 1–101.

Bell, G. (2010). Fluctuating selection: the perpetual renewal of adaptation in variable environments. *Phil. Trans. R. Soc. B* 365, 87–97. doi: 10.1098/rstb.2009.0150

Bingham, E., Chen, J., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., et al. (2019). Pyro: deep universal probabilistic programming. *J. Mach. Learn. Res.* 20, 973–978.

Bohan, D. A., Caron-Lormier, G., Muggleton, S., Raybould, A., and Tamaddoni-Nezhad, A. (2011). Automated discovery of food webs from ecological data using logic-based machine learning. *PLoS ONE* 6:e29028. doi: 10.1371/journal.pone.0029028

Borgströ, J., Gordon, A., Greenberg, M., Margetson, J., and Gael, J. V. (2011). "Measure transformer semantics for bayesian machine learning," in *Programming Languages and Systems*, ed G. Barthe (Berlin; Heidelberg: Springer), 77–96.

Borgström, J., Dal Lago, U., Gordon, A., and Szymczak, M. (2016). "A lambda-calculus foundation for universal probabilistic programming," in *Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming, ICFP 2016* (Nara), 33–46.

Brouard, C., Vrain, C., Dubois, J., Castel, D., D, M., and d'Alche Buc, F. (2013). Learning a markov logic network for supervised gene regulatory network inference. *BMC Bioinformatics* 14:273. doi: 10.1186/1471-2105-14-273

Buchanan, B., and Shortliffe, E. (1984). *Rule-based Expert Systems: The Mycin experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: a probabilistic programming language. *J. Stat. Softw.* 76, 1–32. doi: 10.18637/jss.v076.i01

Connolly, S. R., Keith, S., Colwell, R., and Rahbek, C. (2017). Process, mechanism, and modeling in macroecology. *Trends Ecol. Evol.* 32, 835–844. doi: 10.1016/j.tree.2017.08.011

Darwiche, A. (2009). *Modeling and Reasoning with Bayesian Networks*. Cambridge: Cambridge University Press.

de Moura, L., Kong, S., Avigad, J., van Doorn, F., and von Raumer, J. (2015). "The lean theorem prover," in *25th International Conference on Automated Deduction (CADE-25)* (Berlin).

Desjardins-Proulx, P., and Gravel, D. (2012a). A complex speciation-richness relationship in a simple neutral model. *Ecol. Evol.* 2, 1781–1790. doi: 10.1002/ece3.292

Desjardins-Proulx, P., and Gravel, D. (2012b). How likely is speciation in neutral ecology? *Am. Nat.* 179, 137–144. doi: 10.1086/663196

Domingos, P., and Lowd, D. (2009). *Markov Logic: An Interface Layer for Artificial Intelligence*. San Rafael, CA: Morgan & Claypool Publishers.

Etienne, R., and Haegeman, B. (2011). The neutral theory of biodiversity with random fission speciation. *Theor. Ecol.* 4, 87–109. doi: 10.1007/s12080-010-0076-y

Farmer, W. (2008). The seven virtues of simple type theory. *J. Appl. Logic* 6, 267–286. doi: 10.1016/j.jal.2007.11.001

Garnelo, M., Arulkumaran, K., and Shanahan, M. (2016). Towards deep symbolic reinforcement learning. *arXiv:1609.05518v2*.

Ge, H., Xu, K., and Ghahramani, Z. (2018). "Turing: a language for flexible probabilistic inference," in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, *Vol.* 84, eds A. Storkey and F. Perez-Cruz (Playa Blanca), 1682–1690.

Getoor, L., Friedman, N., Koller, D., Pfeffer, A., and Taskar, B. (2007). "Probabilistic relational models," in *Introduction to Statistical Relational Learning*, eds L. Getoor and B. Taskar (Cambridge: MIT Press).

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge: MIT Press.

Goodman, N., Mansinghka, V., Roy, D., Bonawitz, K., and Tenenbaum, J. (2008). Church: a language for generative models. *Uncertainty Artific. Intell.* arXiv:1206.3255

Halpern, J. (2003). *Reasoning About Uncertainty*. Quebec City, QC: The MIT Press.

Harrison, J. (2009a). *Handbook of Practical Logic and Automated Reasoning*. Cambridge: Cambridge University Press.

Harrison, J. (2009b). "Hol light: an overview," in *Theorem Proving in Higher Order Logics*, eds S. Berghofer, T. Nipkow, C. Urban, and M. Wenzel (Berlin; Heidelberg: Springer Berlin Heidelberg), 60–66.

Heckerman, D. E., and Nathwani, B. N. (1992). An evaluation of the diagnostic accuracy of Pathfinder. *Comput. Biomed. Res.* 25, 56–74. doi: 10.1016/0010-4809(92)90035-9

Heunen, C., Kammar, O., Staton, S., and Yang, H. (2017). "A convenient category for higher-order probability theory," in *Proceedings of 32nd Annual ACM/IEEE Symposium on Logic in Computer Science* (LICS). doi: 10.1109/LICS.2017.8005137

Hu, Z., Ma, X., Liu, Z., Hovy, E., and Xing, E. (2016). Harnessing deep neural networks with logic rules. *arXiv:1603.06318*. doi: 10.18653/v1/P16-1228

Hubbell, S. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography, Volume 32 of Monographs in Population Biology*. Princeton, NJ: Princeton University Press.

Jacobs, B., and Zanasi, F. (2018). The logical essentials of bayesian reasoning. *CoRR* abs/1804.01193.

Jain, D. (2011). "Knowledge engineering with markov logic networks: a review," in *DKB 2011: Proceedings of the Third Workshop on Dynamics of Knowledge and Belief* (Hershey, PA).

Kaliszyk, C., and Urban, J. (2015). Learning-assisted theorem proving with millions of lemmas. *J. Symb. Comput.* 69, 109–128. doi: 10.1016/j.jsc.2014.09.032

Kimmig, A., Bach, S., Broecheler, M., Huang, B., and Getoor, L. (2012). "A short introduction to probabilistic soft logic," in *Proceedings of the NIPS Workshop on Probabilistic Programming* (Barcelona).

Kok, S., and Domingos, P. (2009). "Learning markov logic network structure via hypergraph lifting," in *Proceedings of the 26nd international conference on Machine learning* (Montreal, QC).

Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models*. Cambridge, MA: The MIT Press.

Kopelke, J. P., Nyman, T., Cazelles, K., Gravel, D., Vissault, S., and Roslin, T. (2017). Food-web structure of willow-galling sawflies and their natural enemies across europe. *Ecology* 98:1730. doi: 10.1002/ecy.1832

Kosko, B. (1990). Fuzziness vs probability. *Int. J. Gen. Syst.* 17, 211–240. doi: 10.1080/03081079008935108

Lewis, R. (1980). Evolution: a system of theories. *Perspect. Biol. Med.* 23, 551–572. doi: 10.1353/pbm.1980.0053

Lindsay, R., Buchanan, B., Feigenbaum, E., and Lederberg, J. (1993). Dendral: a case study of the first expert system for scientific hypothesis formation. *Artif. Intell.* 61, 209–261. doi: 10.1016/0004-3702(93)90068-M

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). *The BUGS Book – A Practical Introduction to Bayesian Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

McGill, B. J., Etienne, R., Gray, J., Alonso, D., Anderson, M., Benecha, H., et al. (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol. Lett.* 10, 995–1015. doi: 10.1111/j.1461-0248.2007.01094.x

Mendel, J. (2017). *Uncertain Rule-Based Fuzzy Systems, 2nd Edn*. Springer.

Mendel, J., and Bob John, R. (2002). Type-2 fuzzy sets made simple. *IEEE Trans. Fuzzy Syst.* 10, 117–127. doi: 10.1109/91.995115

Muggleton, S., and de Raedt, L. (1994). Inductive logic programming: theory and methods. *J. Logic Programm.* 19–20, 629–679. doi: 10.1016/0743-1066(94)90035-3

Muggleton, S., and Feng, C. (1990). "Efficient induction of logic programs," in *New Generation Computing* (Academic Press).

Nath, A., and Domingos, P. (2015). "Learning relational sum-product networks," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, TX), 2878–2886.

Nedbal, R., and Serafini, L. (2018). "Bayesian markov logic networks," in *AI*IA 2018 – Advances in Artificial Intelligence* (Trento), 348–361.

Nederpelt, R., and Geuvers, H. (2014). *Type Theory and Formal Proof: An Introduction*. New York, NY: Cambridge University Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, MA: Morgan Kaufmann.

Pierce, B. (2002). *Types and Programming Languages*. MIT Press.

Poisot, T., Labrie, R., Larson, E., and Rahlin, A. (2018). Data-based, synthesis-driven: setting the agenda for computational ecology. *bioRxiv* 150128. doi: 10.1101/150128

Poisot, T., Stouffer, D., and Kéfi, S. (2016). Describe, understand and predict: why do we need networks in ecology? *Funct. Ecol.* 30, 1878–1882. doi: 10.1111/1365-2435.12799

Program, T. U. F. (2013). *Homotopy Type Theory: Univalent Foundations of Mathematics*. Princeton, NJ: Institute for Advanced Study.

Queller, D. (2017). Fundamental theorems of evolution. *Am. Nat.* 189, 345–353. doi: 10.1086/690937

Rice, S. H., and Papadopoulos, A. (2009). Evolution with stochastic fitness and stochastic migration. *PLoS ONE* 4:e7130. doi: 10.1371/journal.pone.0007130

Richardson, M., and Domingos, P. (2006). Markov logic networks. *Mach. Learn.* 62, 107–136. doi: 10.1007/s10994-006-5833-1

Russell, S. (2015). Unifying logic and probability. *Commun. ACM* 58, 88–97. doi: 10.1145/2699411

Russell, S., and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach, 3rd Edn*. Upper Saddle River, NJ: Prentice Hall.

Sadeghian, A., Mendel, J., and Tahayori, H. (2014). *Advances in Type-2 Fuzzy Sets and Systems*. Springer.

Ścibior, A., Kammar, O., Vákár, M., Staton, S., Hongseok, Y., Cai, Y., et al. (2018). Denotational validation of higher-order bayesian inference. *Proc. ACM Progr. Lang.* 2, 60:1–60:29. doi: 10.1145/3158148

Staton, S. (2017). "Commutative semantics for probabilistic programming," in *Programming Languages and Systems*, ed H. Yang (Berlin, Heidelberg: Springer), 855–879.

Staton, S., Yang, H.,Wood, F., Heunen, C., and Kammar, O. (2016). Semantics for probabilistic programming: higher-order functions, continuous distributions, and soft constraints. *Logic Comp. Sci.* 525–534. doi: 10.1145/2933575.2935313

Suppes, P. (1968). The desirability of formalization in science. *J. Philos.* 65, 651–664. doi: 10.2307/2024318

Tamaddoni-Nezhad, A., Milani, G., Raybould, A., Muggleton, S., and Bohan, D. (2013). Construction and validation of food webs using logic-based machine learning and text mining. *Adv. Ecol. Res.* 49, 225–289. doi: 10.1016/B978-0-12-420002-9.00004-4

van de Meent, J.-W., Paige, B., Yang, H., and Wood, F. (2018). An introduction to probabilistic programming. *ArXiv:1809.10756*

Wang, J., and Domingos, P. (2008). "Hybrid markov logic networks," in *AAAI'08 Proceedings of the 23rd National Conference on Artificial Intelligenc* (Menlo Park, CA), Vol. 2, 1106–1111.

Williams, R., Anandanadesan, A., and Martinez, N. (2010). The probabilistic niche model reveals the niche structure and role of body size in a complex food web. *PLoS ONE* 5:e12092. doi: 10.1371/journal.pone.0012092

Williams, R. J., and Martinez, N. D. (2000). Simple rules yield complex food webs. *Nature* 404, 180–183. doi: 10.1038/35004572

Wood, F., van de Meent, J., and Mansinghka, V. (2014). "A new approach to probabilistic programming inference," in *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics* (Reykjavik), 1024–1032.

Yoshikawa, K., Riedel, S., Asahara, M., and Matsumoto, Y. (2009). "Jointly identifying temporal relations with Markov Logic," in *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (Singapore). doi: 10.3115/1687878.1687936

Zadeh, L. (1965). Fuzzy sets. *Inform. Control* 8, 338–353. doi: 10.1016/S0019-9958(65)90241-X

Zahavy, T., Ben-Zrihem, N., and Mannor, S. (2016). Graying the black box: understanding DQNS. *arXiv: 1602.02658.*

Zeng, J., and Liu, Z.-Q. (2008). Type-2 fuzzy markov random fields and their application to handwritten chinese character recognition. *IEEE Trans. Fuzzy Syst.* 16, 747–760. doi: 10.1109/TFUZZ.2007.905916

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, ZC, declared a past co-authorship with one of the authors, DG, to the handling Editor.

POLICY AND PRACTICE REVIEWS
published: 08 November 2019
doi: 10.3389/fevo.2019.00434

Studying Ecosystems With DNA Metabarcoding: Lessons From Biomonitoring of Aquatic Macroinvertebrates

Alex Bush[1]*, *Zacchaeus G. Compson*[1], *Wendy A. Monk*[1,2], *Teresita M. Porter*[3,4], *Royce Steeves*[5], *Erik Emilson*[3], *Nellie Gagne*[5], *Mehrdad Hajibabaei*[4], *Mélanie Roy*[5] and *Donald J. Baird*[1]

[1] Department of Biology, Environment and Climate Change Canada, Canadian Rivers Institute, University of New Brunswick, Fredericton, NB, Canada, [2] Faculty of Forestry and Environmental Management, University of New Brunswick, Fredericton, NB, Canada, [3] Great Lakes Forestry Centre, Natural Resources Canada, Marie, ON, Canada, [4] Centre for Biodiversity Genomics and Department of Integrative Biology, University of Guelph, Guelph, ON, Canada, [5] Department for Fisheries and Oceans, Gulf Fisheries Centre, Moncton, NB, Canada

An ongoing challenge for ecological studies has been the collection of data with high precision and accuracy at a suitable scale to detect and manage critical global change processes. A major hurdle has been the time-consuming and challenging process of sorting and identification of organisms, but the rapid development of DNA metabarcoding as a biodiversity observation tool provides a potential solution. As high-throughput sequencing becomes more rapid and cost-effective, a "big data" revolution is anticipated, based on higher and more accurate taxonomic resolution, more efficient detection, and greater sample processing capacity. These advances have the potential to amplify the power of ecological studies to detect change and diagnose its cause, through a methodology termed "Biomonitoring 2.0." Despite its promise, the unfamiliar terminology and pace of development in high-throughput sequencing technologies has contributed to a growing concern that an unproven technology is supplanting tried and tested approaches, lowering trust among potential users, and reducing uptake by ecologists and environmental management practitioners. While it is reasonable to exercise caution, we argue that any criticism of new methods must also acknowledge the shortcomings and lower capacity of current observation methods. Broader understanding of the statistical properties of metabarcoding data will help ecologists to design, test and review evidence for new hypotheses. We highlight the uncertainties and challenges underlying DNA metabarcoding and traditional methods for compositional analysis, specifically comparing the interpretation of otherwise identical bulk-community samples of freshwater benthic invertebrates. We explore how taxonomic resolution, sample similarity, taxon misidentification, and taxon abundance affect the statistical properties of these samples, but recognize these issues are relevant to applications across all ecosystem types. In conclusion, metabarcoding has the capacity to improve the quality and utility of ecological data, and consequently the quality of new research and efficacy of management responses.

Keywords: biodiversity observation, high-throughput sequencing, taxonomic resolution, community ecology, environmental genomics, freshwater, benthic macroinvertebrate

OPEN ACCESS

Edited by:
Alireza Tamaddoni-Nezhad, Imperial College London, United Kingdom

Reviewed by:
Christian Henri Nozais, Université du Québec à Rimouski, Canada
Rosetta Charlotte Blackman, Swiss Federal Institute of Aquatic Science and Technology, Switzerland

****Correspondence:***
Alex Bush
alexalbush@gmail.com

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Ecology and Evolution

Received: *29 July 2019*
Accepted: *23 October 2019*
Published: *08 November 2019*

Citation:
Bush A, Compson ZG, Monk WA, Porter TM, Steeves R, Emilson E, Gagne N, Hajibabaei M, Roy M and Baird DJ (2019) Studying Ecosystems With DNA Metabarcoding: Lessons From Biomonitoring of Aquatic Macroinvertebrates. Front. Ecol. Evol. 7:434. doi: 10.3389/fevo.2019.00434

INTRODUCTION

Biodiversity loss and the risks it poses to ecosystem functions and services remain a major societal concern (Cardinale et al., 2012), but due to a lack of consistently-observed data, there is no consensus regarding the speed or severity of this decline (Vellend et al., 2013; Newbold et al., 2015). There are very few ecosystems in which we can quantify the magnitude of degradation, nor can we discriminate among multiple stressors, both key goals for environmental monitoring programs (Bonada et al., 2006). The power to detect change in ecological communities has been hampered by sampling costs predominantly associated with skilled human labor and travel. As a result, ecosystem monitoring programs must manage a trade-off between the scope of a study, including the phylogenetic breadth of taxon coverage and the resolution to which taxa are described (our universe of observation), and its spatial and temporal coverage (e.g., tropical forests Gardner et al., 2008; marine sediments Musco et al., 2009). A history of such trade-offs has led to entrenched practices relying on observation of a narrow range of taxa, which aim to provide a surrogate for the full biodiversity complement, yet whose taxonomic, spatial, or temporal relationships are largely undefined (Lindenmayer and Likens, 2011). Landscapes are under increasing stress from multiple drivers, and yet the troubling reality is that management decisions are informed by very limited and potentially biased information, generated by approaches that no longer reflect our understanding of how ecosystems and species interact (Woodward et al., 2013).

Fortunately, technological advances offer the opportunity to generate high-quality biodiversity data in a consistent manner, increasingly automating processing pipelines, and radically expanding the scope of ecosystem monitoring (Turner, 2014; Bush et al., 2017). One of the most promising of these is the technique of DNA metabarcoding, which supports the massively-parallelized, and hence high-throughput, taxonomic identification of organism assemblages within a biological sample. While single-specimen DNA barcoding uses short genetic sequences to identify individual taxa, often at the species-level, metabarcoding supports simultaneous identification of entire assemblages via high-throughput sequencing (Taberlet et al., 2012; Yu et al., 2012). The application of metabarcoding for ecosystem monitoring has been termed "Biomonitoring 2.0" (Baird and Hajibabaei, 2012) because it could provide a universal platform to identify any, and potentially all, phylogenetic groups occurring within an ecosystem, including many taxa currently not identifiable by expert taxonomists (e.g., streams: Sweeney et al., 2011; rainforest: Brehm et al., 2016; marine zooplankton: Zhang et al., 2018). As DNA sequencing capacity continues to increase, there is a growing interest from ecological researchers and environmental managers for guidance in how to apply these new tools, and to provide clear evidence of their value relative to existing microscopy-based methods. However, it is important to emphasize that comparisons between traditional morphological identifications and DNA sequences are far from straightforward. For example, while metabarcoding can observe the occurrence of DNA sequences within a specified environmental matrix (e.g., soil sample), it does not discriminate between intact, living organisms, and their presence as parts, ingested, or extraneous tissue. While some may see this as a challenge to be overcome, to retrofit a new method to an old system of observation, we view this as an opportunity to expand our universe of interest, and gain new insight into metacommunity assembly and structure (Bohan et al., 2017). We draw on recent research into metabarcoding of freshwater macroinvertebrates to illustrate these issues, the most widely applied non-microbial applications of DNA metabarcoding to date, but many of the analytical concepts we discuss will be common to other ecosystems and assemblages.

Aquatic researchers have long recognized the challenges of taxonomic identification and resulting limitations it imposes on the scale and scope of observational, experimental and monitoring studies (Jones, 2008). Freshwater monitoring programs rely upon a subset of taxa, primarily aquatic macroinvertebrates, fish, or algae, with little consistency across environmental agencies or regions (Friberg et al., 2011), although we acknowledge efforts in Europe to rectify these divides (Birk et al., 2012). Sparse spatial and temporal coverage and limited taxonomic resolution (e.g., Orlofske and Baird, 2013) ultimately constrains outcomes to "pass/fail" (impacted/non-impacted; Clarke et al., 2006; Strachan and Reynoldson, 2014), with causes of degradation inferred rather than supported by direct evidence. After decades of research, our ability to disentangle the influence of even the most basic drivers that impact the state of freshwater ecosystems is still limited (Woodward et al., 2013). Given the challenges faced by aquatic ecologists it is not surprising that within a decade of the first preliminary studies (Hajibabaei et al., 2011), attention is now focused on how to overcome the barriers to full-scale implementation (e.g., technological and regulatory Keck et al., 2017; Hering et al., 2018; Leese et al., 2018; Porter and Hajibabaei, 2018a). It is therefore timely to highlight how the interpretation of metabarcoding and traditional morphological identification differ, their sources of error, and sources of uncertainty.

OUR UNIT AND UNIVERSE OF OBSERVATION

The science of aquatic biomonitoring is based on the principle that site-level observations of biological assemblages integrate responses to prevailing environmental conditions over space and time, reducing the intensity of sampling required to detect stressor-related changes in the environment, and providing an immediate signal of "ecosystem health" (Friberg et al., 2011). However, consistently observing more than a narrow range of taxa within an ecological community has proved costly and impractical, with accuracy of identification often unrecorded or difficult to quantify, and varying across taxa. The observation universe is further constrained by sampling methods (e.g., mesh-size of collection nets), rather than common phylogenetic or ecological characteristics, with further downgrading or exclusion of groups that are difficult to identify (e.g., Vlek et al., 2006). Even with the best taxonomic expertise available, it is practically impossible to identify all specimens to species-level, since many early life-stages lack necessary diagnostic features (Orlofske and Baird, 2013). Species are subsequently aggregated at higher taxonomic ranks, obscuring species-level responses, constraining our knowledge of whether species' environmental preferences are conserved or variable (Macher et al., 2016; Beermann et al., 2018). In our view, the level of observation provided by direct morphological identification of biological specimens in a sample is highly variable (typically referred to as "lowest taxonomic level"), disconnected from ecological theory, and contains an unknown, yet potentially significant degree of bias (Jones, 2008; Nakov et al., 2018).

DNA metabarcoding offers the potential to reduce many of the costs involved in routine morphological identification (Ji et al., 2013), and can also generate a richer list of taxa (Sweeney et al., 2011; Gibson et al., 2015). Taxonomic assignment is continually improving as DNA-barcode reference libraries expand (e.g., Curry et al., 2018; Weigand et al., 2019), and in contrast to morphological approaches, a universe of observation defined by the DNA region and primers (see below) is less ambiguous. The opportunity this represents has triggered a wide range of metabarcoding studies in aquatic ecosystems (e.g., rivers Hajibabaei et al., 2011; wetlands Gibson et al., 2015; lakes Bista et al., 2017), and applied to describe community composition in a wide variety of taxa (e.g., worms Vivien et al., 2015; insects Emilson et al., 2017; diatoms Vasselon et al., 2017).

THE UNIVERSE OF OBSERVATION FOR MONITORING WITH METABARCODING IN FRESHWATERS

While metabarcoding offers the potential to observe a greater diversity of taxa, a crucial step for any metabarcoding study is the selection of primers used to amplify specific DNA sequence marker regions, as they determine the taxonomic groups under study and resolution of assignment (Hajibabaei et al., 2012; Gibson et al., 2014). In order to expand taxonomic coverage, it is necessary to employ a range of primers, and marker sequences (see Figure 3 in Gibson et al., 2014). The cost of sequencing additional primers can therefore limit the number of sites surveyed, but these costs may rapidly decline as automated processing becomes available. Refining primers for different taxonomic groups or species has taken considerable effort, but primers with broad coverage for invertebrates have now been established (e.g., Hajibabaei et al., 2012; Elbrecht and Leese, 2017). However, amplification bias due to variable affinity among sequence variants for amplification can distort the relationship between sample biomass and the number of sequence reads (Elbrecht and Leese, 2015; Zhang et al., 2018). Metabarcoding can therefore support a taxonomically broad universe of observation, but outputs should be treated as occurrences and do not support reliable estimation of organism biomass or abundance.

Another key issue is the distinction between bulk-community sampling and environmental DNA (eDNA). eDNA samples focus on a signal derived predominantly from traces of intracellular and extracellular DNA without attempting to isolate organisms (e.g., from water or soil; Deiner et al., 2017; Cristescu and Hebert, 2018), whereas bulk-community samples include eDNA, but target the collection of whole organisms. eDNA can be effective in detecting biological signal from the environment, but the significant spatial and temporal uncertainty of that signal clouds its application in observational studies. In addition, the ease with which trace amounts of DNA can be transported makes cross-contamination a critical issue for eDNA studies (i.e., the addition of false-positives Ficetola et al., 2015), whereas the high concentrations of template material in bulk samples mean this is less of a concern (Majaneva et al., 2018). As a result, our examples of metabarcoding below focus entirely on observations derived from unsorted bulk-community samples that are otherwise identical to traditional monitoring surveys.

INTERPRETATION

The statistical power and precision of any ecological assessment that is based on sample assemblage composition depends upon how results are aggregated and analyzed, how misidentification (i.e., false-presences and false-absences) can obscure expectations when setting the baseline composition, limiting our ability to detect deviations from this baseline and infer that change has occurred (e.g., Clarke et al., 2002; Clarke, 2009). Although many sources of uncertainty affect our ability to infer regional and landscape-level trends from site-level observations, these are difficult to address with traditional approaches (Clarke, 2009; Carstensen and Lindegarth, 2016). To illustrate this problem, and whether metabarcoding can alleviate it, we focus on how four sources of error involved in describing freshwater biodiversity differ between morphological and metabarcoding workflows: (a) taxonomic resolution, (b) replicate similarity, (c) taxonomic misidentification, and (d) quantitative measures like abundance.

TAXONOMIC RESOLUTION

Biomonitoring 2.0 (Baird and Hajibabaei, 2012) employs metabarcoding to overcome the taxonomic bottleneck of sample processing, removing a critical trade-off between sample taxonomic resolution and the number of samples that can be studied (Jones, 2008). Moreover, sample metrics derived from higher taxonomic categories, such as family- or genus-level, make a tacit assumption that species within those higher categories share similar environmental responses, and possess similar ecological functions. However, when studies are able to differentiate taxa at the species level, they may reject this assumption (e.g., nutrient and sediment sensitivity; Macher et al., 2016; Beermann et al., 2018), and this can significantly influence study outcomes (Hawkins et al., 2000; Schmidt-Kloiber and Nijboer, 2004; Sweeney et al., 2011).

Observing taxonomic assemblages at genus- or family-level masks turnover in composition, reducing our power to detect subtle changes among communities over space and time. As each species is less common than its parent taxonomic group, there will be fewer observations with which to establish reliable associations, and their inclusion could add noise to statistical models, echoing the long-running debate about the value of rare taxa in biomonitoring (Nijboer and Schmidt-Kloiber, 2004; Lavoie et al., 2009). This "noise" is not only due to the stochastic occurrence of uncommon species, but also sampling error, which can be quantified using hierarchical occupancy models (Clarke, 2009; Guillera-Arroita, 2017). We should therefore be particularly cautious about concluding how taxonomic resolution affects the strength of statistical relationships (Arscott et al., 2006; Martin et al., 2016). Instead, our current challenge is understanding when these subtle changes, previously invisible to traditional monitoring, are related to natural environmental factors or anthropogenic disturbance.

One criticism of DNA metabarcoding is that high taxonomic resolution is not valuable if those taxa cannot be linked to a binomial taxonomic name, a limitation that emerges when barcode reference libraries are incomplete (Curry et al., 2018). However, many methods of ecological assessment evaluate community level characteristics such as alpha- and beta-diversity, and therefore do not retain taxon identity, particularly at the species-level (Birk et al., 2012). For this reason, interest in taxonomy-free approaches is increasing among those studying poorly-known assemblages whose morphological identification is challenging (e.g., meiofauna or diatoms; Vasselon et al., 2017). Clearly defining the unit and universe of observation (i.e., taxonomic breadth and resolution) is fundamental to comparing such characteristics (Cordier et al., 2018; Pawlowski et al., 2018), but doing so could also improve compatibility between biogeographically separated programs (Turak et al., 2017; Bailet et al., 2019). Nonetheless, to tie DNA-based monitoring to historic surveys, and to assign ancillary information such as traits, it is still a requirement to assign taxonomic names to identified sequences (e.g., Compson et al., 2018). Based on the wealth of ecological information available that could complement DNA-based ecological studies, and the considerable body of legacy data generated by historical studies, including regulatory monitoring, increasing reference library coverage should be a priority for management agencies transitioning to DNA-based surveys (Rimet et al., 2018; Stokstad, 2018; Weigand et al., 2019).

REPLICATE SIMILARITY

Depending on the scale of observation, species are rarely distributed randomly or uniformly in nature (e.g., Soininen et al., 2016). For example, the distribution of macroinvertebrate taxa in streams is notoriously dynamic, as species adjust to changes in both abiotic (e.g., flow velocity, substratum size) and biotic (e.g., fish predation, mussel aggregation) factors (Downes et al., 1993; Vaughn and Spooner, 2006). Heterogeneity may also result from stochastic processes such as dispersal and colonization (Fonseca and Hart, 2001), ephemeral resources (Lancaster and Downes, 2014), or disturbance regimes at multiple scales (Effenberger et al., 2006). Indeed, heterogeneity is so pervasive that a shift toward greater homogeneity within aquatic communities could indicate human modification of the landscape (Petsch, 2016). Given such heterogeneity, the challenge for ecological studies or biomonitoring is to detect a sufficient proportion of the community, whilst also minimizing processing costs, so that further detections are unlikely to alter the interpretation of subsequent analyses. Counting all individuals in a sample can have value, but it is prohibitive for routine observational studies, and not cost-effective for biomonitoring purposes (e.g., Vlek et al., 2006). Most studies therefore employ subsampling (i.e., identifying a subset of individuals collected from the field) to reduce the time, effort, and cost of processing macroinvertebrate samples. However, reducing the effort per sampling unit can significantly underestimate the richness per sample (Doberstein et al., 2000; Buss et al., 2014) and although subsampling is standardized by volume, time, weight, or number of individuals, it is often difficult to compare among survey methods and biomonitoring schemes (Buss et al., 2014). Although sensitivity to subsampling depends on the metric employed, subsampling can substantially increase the misclassification of site status (Clarke et al., 2006; Petkovska and Urbanič, 2010), and exaggerate the perceived rarity of many taxa, whose exclusion from analyses may further bias interpretations of condition (Schmidt-Kloiber and Nijboer, 2004).

Regardless of the sub-sampling approach, a single sample only recovers a subset of the community, particularly in heterogeneous environments. As sampling effort increases, either by area or time, more taxa are recovered until the rate of new discoveries declines (Vlek et al., 2006). The rate of accumulation depends on taxon abundance distributions, their dispersion, and ease of collection, including the effects of environment on collection efficiency (Guillera-Arroita, 2017). For example, a typical 3-min kick-sample recovered only 50% of the macroinvertebrates species, and 60% of the families, found in total from six replicate kick-samples (Furse et al., 1981). **Figure 1** illustrates a similar degree of turnover also occurs among replicate samples from the same location for other standardized protocols that study aquatic benthic invertebrates.

Metabarcoding can, in principle, substantially reduces detection error by identifying damaged and juvenile specimens, and because aliquots from homogenized bulk community samples are likely to be more representative than morphological subsamples. Nonetheless, successfully detecting all taxa is still conditional on which primers were selected, on the sequencing platform (Singer et al., 2019), the sequencing "effort" (checked by rarefaction of taxon richness and sequencing depth), and, particularly with bulk biological samples, the representativeness of each extraction (checked by analyzing extractions from multiple DNA aliquots). Although low-biomass, low abundance taxa are more likely to be missed (Hajibabaei et al., 2012; Elbrecht et al., 2017a), metabarcoding can detect a higher proportion of the target assemblage compared to morphologically-identified samples (i.e., faster rate of accumulation: **Figure 2**), thereby increasing the power of monitoring programs to detect change. **Figure 2** compares the accumulation curves of macroinvertebrate families collected in the Peace-Athabasca Delta between 2011 and 2016 (updated from surveys published in Gibson et al., 2015). Note that to compare the efficiency of sampling, the metabarcoding data in **Figure 2** were aggregated to an equivalent family-level taxonomy of the morphologically-identified samples, but the complete metabarcoding dataset

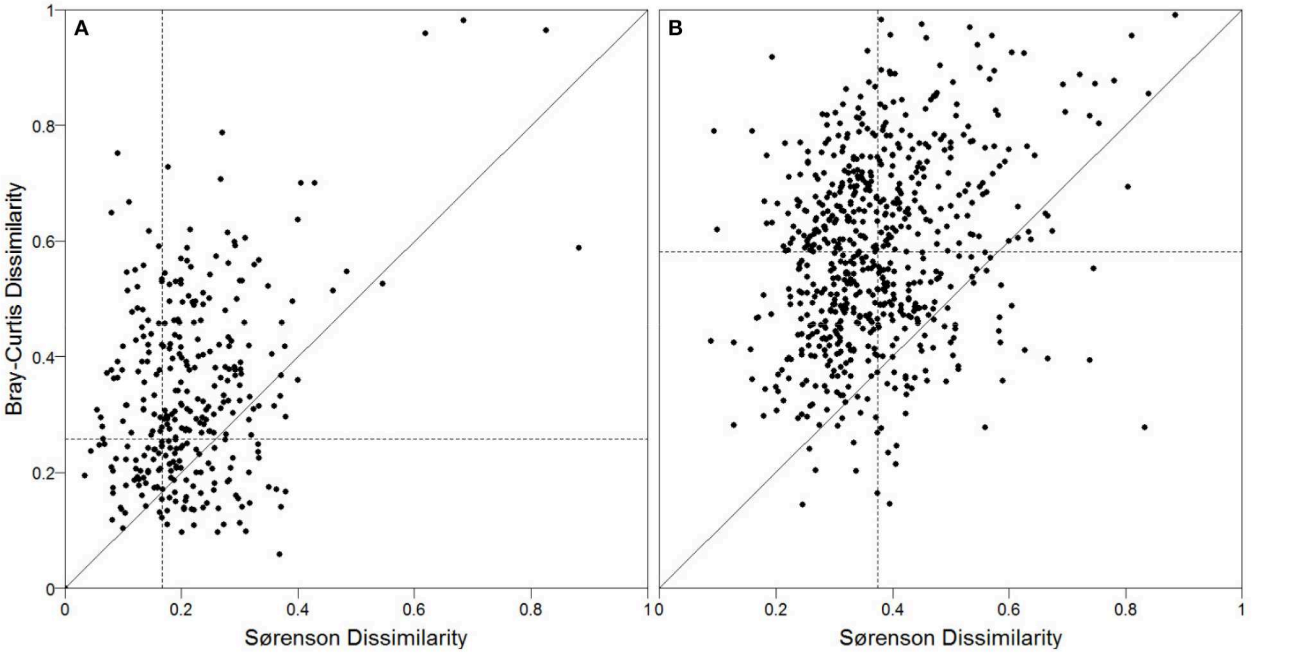


FIGURE 1 | Dissimilarity between replicate samples (same location and time) based on presence/absence data (Sørensen), and count data (Bray-Curtis) of morphologically identified macroinvertebrate families from **(A)** 417 CABIN (Canadian Aquatic Biomonitoring Network; ECCC, 2018) surveys (total *n* = 1,656, mean richness = 16 ± 4.8), and **(B)** 787 surveys from the STAR-AQEM dataset (total *n* = 1,673) from 14 European countries (mean richness = 51 ± 18.4; Furse et al., 2006; Schmidt-Kloiber et al., 2014).

actually observed 109 families, 263 genera, and thousands of unique sequences.

MISIDENTIFICATION

Morphological identification of diverse taxonomic groups, such as invertebrates, is challenging, as demonstrated by a lack of reliable species-level data generated by routine biomonitoring programs. The probability of misidentifying an individual depends on the quality of the specimen (e.g., is the specimen partial or complete? Is it mature or immature?), the availability and completeness of identification keys, and the taxonomist's experience. Early audits of the RIVPACS program showed that 8.3% of family occurrences were missed, and approximately one false presence was added in every four samples (Clarke, 2009). Similarly, an audit of a range of European programs by Haase et al. (2006) found that after accounting for misidentifications and sorting errors, samples were on average 40% dissimilar to their initial composition (based on lowest taxonomic level). Though procedures for quality control and assessment in biomonitoring programs have reduced the likelihood of misidentification (Haase et al., 2010), false positives and negatives are still common, identification errors compound the loss of taxa during sub-sampling, and misidentifications remain difficult to predict.

A major advantage of metabarcoding over traditional morphological identification is the ability to generate more accurate identifications in a consistent manner (Orlofske and Baird, 2013; Jackson et al., 2014). However, if organisms are misidentified at the time of sequence deposition, reference library sequences become associated with an incorrect taxonomic name. To minimize this challenge, the Barcode of Life Database (BOLD) stores information on voucher specimens, supporting linkage of sequences to material in curated reference specimen collections. Overall, database coverage for animals is expanding rapidly (Porter and Hajibabaei, 2018b) and is already relatively high for freshwater invertebrates (Leese et al., 2018; Weigand et al., 2019). For example, sequences exist for 95% of the genera observed in >1% of samples collected by the Canadian national biomonitoring program (Curry et al., 2018). Currently, the BOLD reference library is better suited to identifying macroinvertebrate families routinely observed in Canada, reflecting the greater effort on DNA barcode library development in that country when compared to Australia and the UK (**Figure 3**, **Supplementary Material S1**). At the time of writing, a routine Bayesian classifier (Porter and Hajibabaei, 2018c) is expected to misidentify 4.4, 6.1, and 7.7% of families within CABIN, RIVPACS, and AUSRIVAS programs, respectively. It cannot be overstated that this is a significant improvement on the documented ability of current best-available morphological identification, and is accompanied by an ability to drill down to species-level, which will only improve as DNA libraries become more complete.

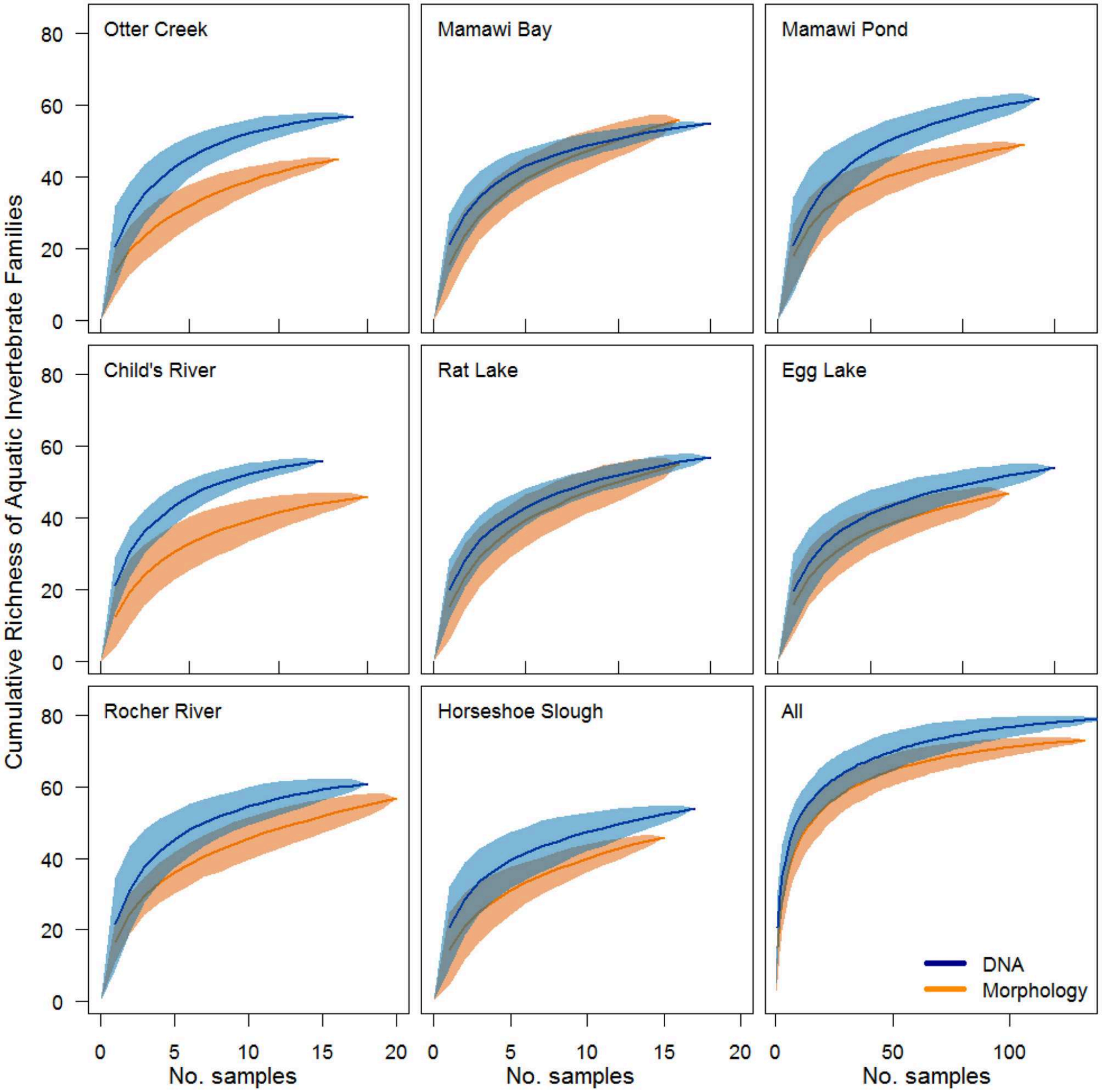


FIGURE 2 | Accumulated richness (mean ± 95% confidence interval) of aquatic invertebrate families from 8 wetland sites in the Peace-Athabasca Delta, and for all samples combined (note different scale) using DNA metabarcoding and morphological identification. Metabarcoded sequences were aggregated and restricted to the same taxa as observed in the morphological dataset for the entire delta.

QUANTITATIVE MEASURES OF BIODIVERSITY

As stated above, DNA metabarcoding results do not currently produce a reliable signal of abundance or biomass (Elbrecht and Leese, 2015), although at the same time a bias in organism biomass can reduce the detectability of rare taxa (Elbrecht et al., 2017a). Nonetheless, it is equally misleading to suggest that current biomonitoring practices are themselves able to effectively detect differences in macroinvertebrate abundance without substantial effort. The difficulty of processing samples, coupled with species' patchy distributions, means few studies can claim to have truly quantified patterns of abundance for multispecies invertebrate assemblages (e.g., Hawkins et al., 2000).

Reliable estimates of taxon abundance or biomass can support studies of many key ecological processes, and are fundamental to

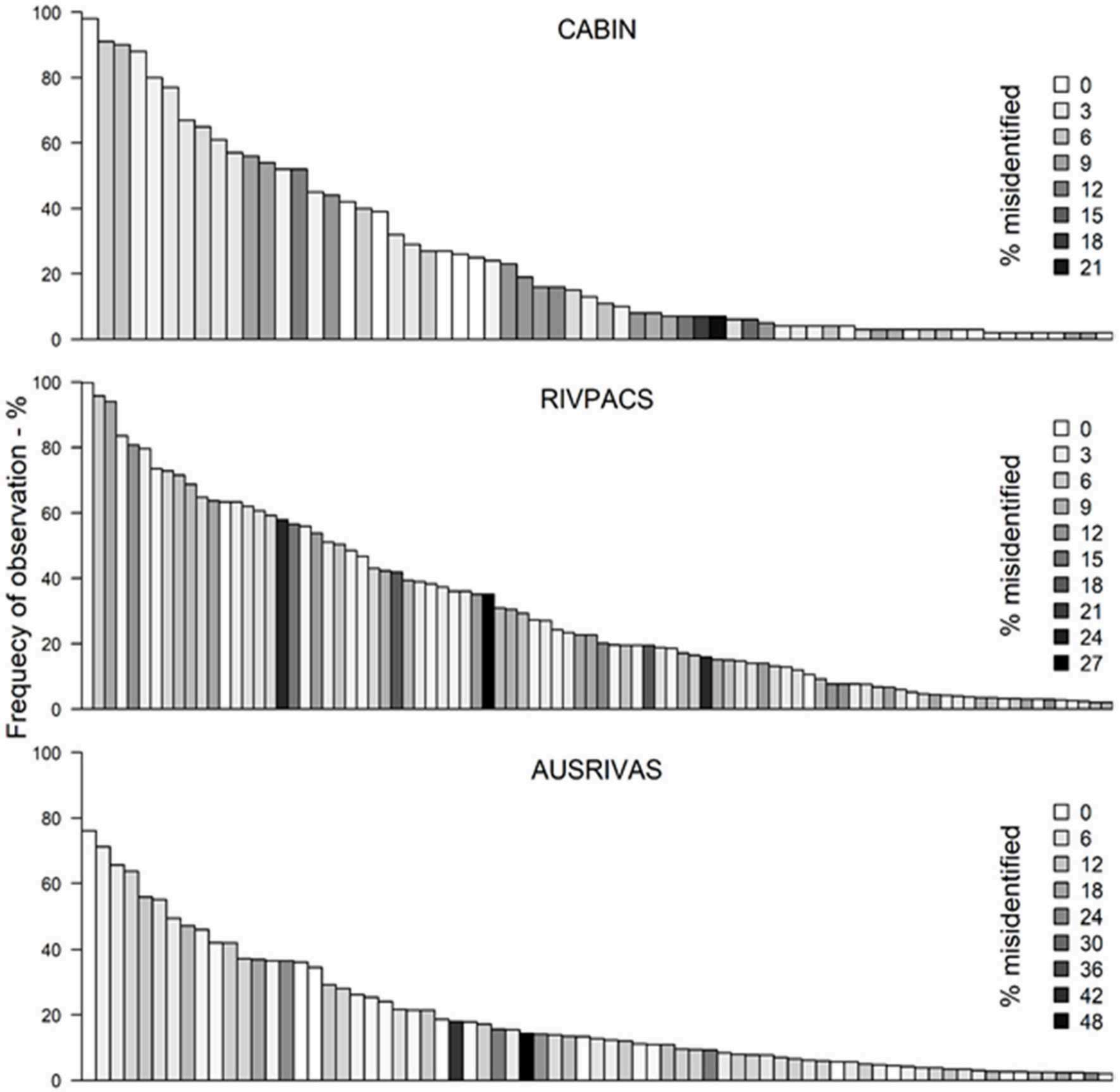


FIGURE 3 | Families ordered by frequency of occurrence within three biomonitoring programs: the CABIN ($n = 540$), the UK River Invertebrate Prediction and Classification System (RIVPACS, $n = 2,504$), and the Australian River Assessment System (AUSRIVAS $n = 1,516$) from Victoria. Shading reflects the likelihood taxa could be misidentified using the CO1 RDP classifier v.3 (see **Supplementary Material S1** for further details).

detecting shifts in species dominance that are not associated with changes in composition. This is particularly true in depauperate systems, if species are pooled at higher taxonomic levels, or rare taxa are discarded (Reynoldson et al., 1997). Nonetheless, differences in the composition of diverse assemblages are often sufficient to discriminate among sites, even at relatively coarse taxonomic resolution (Thorne et al., 1999; Hawkins et al., 2000). Thus, the challenge has always been the reliable identification of those taxa. While count or relative abundance information may provide another axis for discrimination, their inherent variability exaggerates the dissimilarity among replicate samples (**Figure 1**), rendering baseline conditions more variable, thus reducing statistical power to detect change. These limitations are well illustrated by studies that have replaced quantitative count data with qualitative categories or occurrence data (e.g., Wright et al., 1984; Armanini et al., 2013). These approaches have proved acceptable to practitioners precisely because count data provide little or no incremental improvement to detecting differences among sites. Moreover, approaches based on occurrence data illustrate a direct pathway to implement DNA metabarcoding in routine biomonitoring programs (Beentjes et al., 2018).

PERFORMANCE

Study design and interpretation should acknowledge the sources of uncertainty in both morphological and metabarcoding approaches to deliver specific goals. As they are driven by regulatory needs, most monitoring programs focus on relatively simple outcomes (e.g., local deviation from baseline; categorical quality assessment), and thus can greatly benefit from increased precision and statistical power. Recent freshwater ecosystem studies have demonstrated that metabarcoding data can support detection of ecological change at a greater level of discrimination than traditional approaches (Gibson et al., 2015; Elbrecht et al., 2017b; Emilson et al., 2017). Although regulators have thus far remained hesitant to transition to monitoring with metabarcoding, these early studies have highlighted a lack of precision and consistency in the application of existing morphological approaches, shortcomings that are too often overlooked (but see Giupponi, 2007; Clarke, 2009; Birk et al., 2012; Voulvoulis et al., 2017).

Our intent has been to explore the ability of DNA metabarcoding as an observational tool that provides

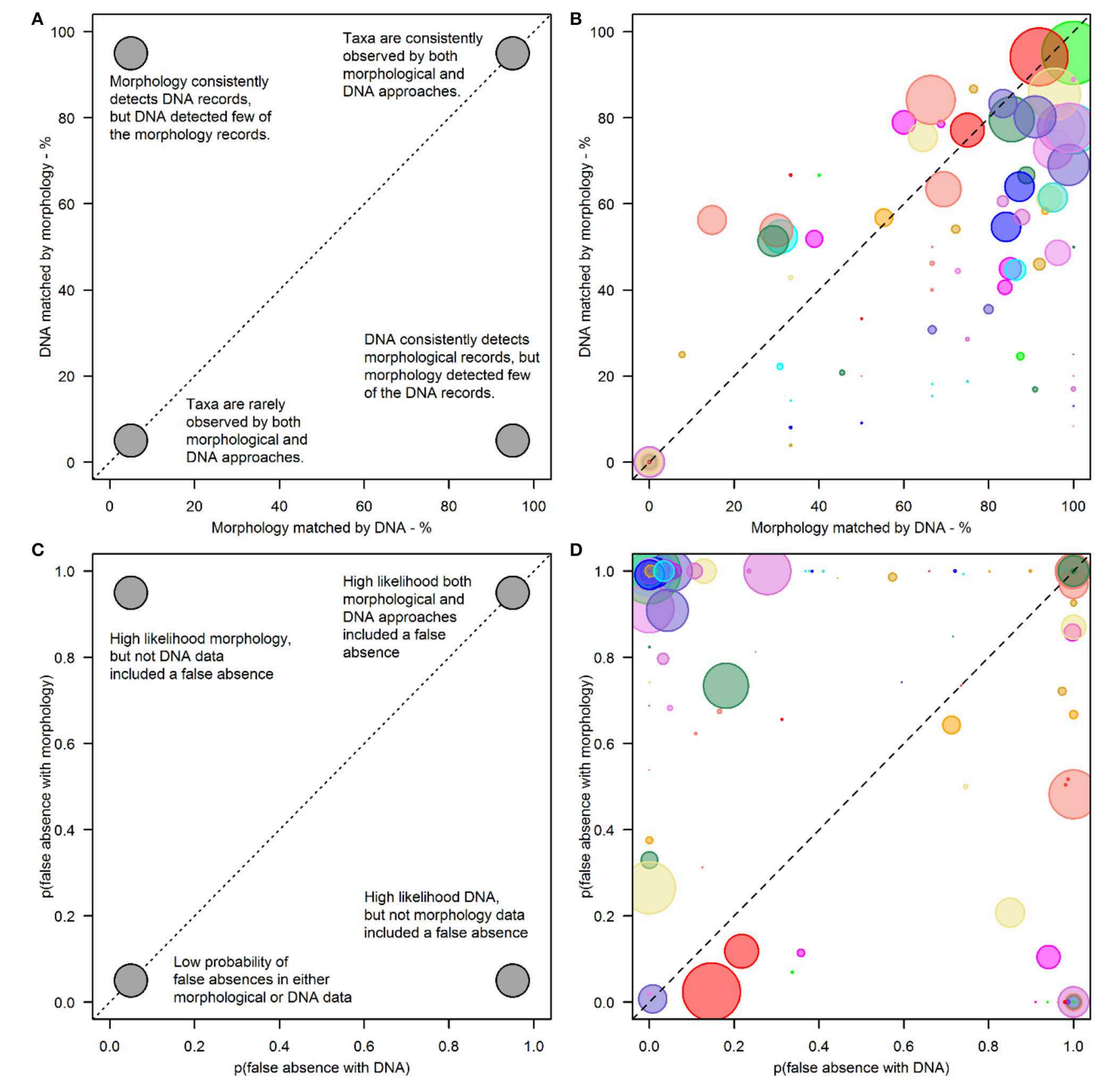


FIGURE 4 | Comparison of macroinvertebrate families (*n* = 114) observed in pairs of standard 3-min river benthos kick samples (*n* = 141 sites). **(A,B)** Shows the correspondence between observations of each taxonomic family using either morphological identification or DNA metabarcoding. Points are scaled relative to the number of morphological observations. **(C,D)** Shows the probability that each method included at least one false absence for each taxon (see **Supplementary Material S2** for code and raw data).

consistently-observed information to answer routine questions posed by managers (e.g., Is biological composition at a site significantly different from expectations, and if so, is there evidence of impact?). Comparisons between metabarcoding and morphology-based methods have involved sorting and identification of a sample using existing taxonomic keys, followed by the reassembly of the sample for metabarcoding (but see also Hajibabaei et al., 2012; Gibson et al., 2015). These approaches have demonstrated that DNA metabarcoding recovered ∼90% of the taxa identified by morphology, and all false-absences were from taxa that represented <1% of individuals. Most recently, we have also evaluated the similarity of taxa recovered by metabarcoding using paired samples (**Figure 4**; GRDI-Ecobiomics, 2017). DNA was extracted from unsorted bulk samples and, as for **Figure 2**, the data are aggregated to family-level for comparison with the resolution of routine monitoring in Canada (ECCC, 2018). The average similarity of morphological and metabarcoded samples was 73%, within the range of variation expected for replicate samples (see **Figure 1**; Clarke et al., 2002). Of the families observed by both methods, DNA observed 79% of the observations made by morphology, whereas morphology only matched 61% of those made by DNA. Some families also appear to be consistently under-represented or absent from this DNA dataset (**Figures 4A,B**, bottom-left), most likely due to a combination of gaps in the reference library (aquatic mites and oligochaetes in particular) and primer bias (Gibson et al., 2014; Elbrecht et al., 2017b). Beyond mere overlap, a better estimate of performance could be the likelihood each family was missed based on their detectability in replicate samples (**Figure 4B**). Both methods are likely to have missed many families at least once, but the mean and likelihood of multiple false absences was lower among metabarcoding samples than for samples identified by morphology (**Supplementary Material S2**). Metabarcoding therefore represents a major advance in how consistently we observe the taxonomic structure of aquatic invertebrate communities.

CONCLUSIONS

Biomonitoring 2.0 (Baird and Hajibabaei, 2012) envisaged the use of DNA metabarcoding to generate consistently-observed biodiversity data to detect environmental change efficiently and rapidly. This can be done with only minor modification of existing sample collection methods, ensuring backwards compatibility with legacy data. Finer taxonomic resolution, more efficient detection (**Figure 2**), and the capacity to increase spatiotemporal coverage can all increase the statistical power to detect change and diagnose its cause (Bonada et al., 2006). Finer taxonomic resolution and more samples with metabarcoding would improve the estimation of detection errors(e.g., Davis et al., 2018), and once standard operating procedures emerge, many tasks can be automated, further reducing the risk of handling errors and the costs of sequencing (Porter and Hajibabaei, 2018a). Currently, the cost of processing an invertebrate community sample (from DNA-extraction to sequencing) is approximately half the cost of morphological identification by taxonomists, but as we have stressed, the divergent properties of each approach make it misleading to base comparisons on costs alone.

We can only manage what we can measure, and at present the unknown magnitude and consequences of global biodiversity loss emphasizes the value of metabarcoding as a technique to support improved ecological observation in all field studies of multispecies assemblages. We expect the increasing numbers of metabarcoding studies, and sequences in reference libraries, will help refine the uncertainties associated with observations, and accelerate the large-scale implementation of metabarcoding (e.g., Leese et al., 2018). Metabarcoding is also being used for increasingly novel applications, such as the study of trophic interactions, either through direct analyses of gut contents, or via the reconstruction of networks of multi-trophic assemblages (Bohan et al., 2017). Other fields of research such as meta-community theory (Miller et al., 2018), and ecosystem function relationships (Vamosi et al., 2017) also benefit where previously the statistical relationships were obscured by coarse taxonomic resolution. These applications could generate substantial added value to existing or future biomonitoring programs (Compson et al., 2018).

In conclusion, ecologists in all ecosystems should be aware of the shortcomings in their data, and acknowledge it publicly if the uncertainty could alter their conclusions. Metabarcoding is now an established technique, with the capacity to improve the quality and utility of ecological data, and understanding its statistical properties will help ecologists to design, test and review evidence for new hypotheses.

AUTHOR CONTRIBUTIONS

AB wrote the first draft and all coauthors contributed to the final edit. Data was collected by AB, ZC, WM, DB, and EE and analyzed by RS, TP, MR, NG, and MH.

FUNDING

This work was supported by the Ontario Genomics Institute and Genome Canada, NSERC, Environment and Climate Change Canada program funds, and the Canadian federal Genomics Research & Development Initiative.

ACKNOWLEDGMENTS

We thank Guy Woodward, Richard Marchant, Astrid Schmidt-Kloiber, and Daniel Hering for providing data from monitoring programs in the UK, Australia, and EU.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2019.00434/full#supplementary-material

REFERENCES

Armanini, D. G., Monk, W. A., Carter, L., Cote, D., and Baird, D. J. (2013). Towards generalised reference condition models for environmental assessment: a case study on rivers in Atlantic Canada. *Environ. Monit. Assess.* 185, 6247–6259. doi: 10.1007/s10661-012-3021-2

Arscott, D. B., Jackson, J. K., and Kratzer, E. B. (2006). Role of rarity and taxonomic resolution in a regional and spatial analysis of stream macroinvertebrates. *J. North Am. Benthol. Soc.* 25, 977–997. doi: 10.1899/0887-3593(2006)025[0977:RORATR]2.0.CO;2

Bailet, B., Bouchez, A., Franc, A., Frigerio, J.-M., Keck, F., Karjalainen, S.-M., et al. (2019). Molecular versus morphological data for benthic diatoms biomonitoring in Northern Europe freshwater and consequences for ecological status. *Metabarcod. Metagenom.* 3:e34002. doi: 10.3897/mbmg.3.34002

Baird, D., and Hajibabaei, M. (2012). Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Mol. Ecol.* 21, 2039–2044. doi: 10.1111/j.1365-294X.2012.05519.x

Beentjes, K. K., Speksnijder, C. L., Schilthuizen, M., Schaub, B. E. M., and van der Hoorn, B. B. (2018). The influence of macroinvertebrate abundance on the assessment of freshwater quality in The Netherlands. *Metabarcod. Metagenom.* 2:e26744. doi: 10.3897/mbmg.2.26744

Beermann, A. J., Zizka, V. M. A., Elbrecht, V., Baranov, V., and Leese, F. (2018). DNA metabarcoding reveals the complex and hidden responses of chironomids to multiple stressors. *Environ. Sci. Eur.* 30:26. doi: 10.1186/s12302-018-0157-x

Birk, S., Bonne, W., Borja, A., Brucet, S., Courrat, A., Poikane, S., et al. (2012). Three hundred ways to assess Europe's surface waters: an almost complete overview of biological methods to implement the Water Framework Directive. *Ecol. Indic.* 18, 31–41. doi: 10.1016/j.ecolind.2011.10.009

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

Bohan, D. A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A. J., and Woodward, G. (2017). Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. *Trends Ecol. Evol.* 32, 477–487. doi: 10.1016/j.tree.2017.03.001

Bonada, N., Prat, N., Resh, V. H., and Statzner, B. (2006). Developments in aquatic insect biomonitoring: a comparative analysis of recent approaches. *Annu. Rev. Entomol.* 51, 495–523. doi: 10.1146/annurev.ento.51.110104.151124

Brehm, G., Hebert, P. D. N., Colwell, R. K., Adams, M.-O., Bodner, F., Friedemann, K., et al. (2016). Turning up the heat on a hotspot: DNA barcodes reveal 80% more species of geometrid moths along an andean elevational gradient. *PLoS ONE* 11:e0150327. doi: 10.1371/journal.pone.0150327

Bush, A., Sollmann, R., Wilting, A., Bohmann, K., Cole, B., Balzter, H., et al. (2017). Connecting earth observation to high-throughput biodiversity data. *Nat. Ecol. Evol.* 1:0176. doi: 10.1038/s41559-017-0176

Buss, D. F., Carlisle, D. M., Chon, T.-S., Culp, J., Harding, J. S., Keizer-Vlek, H. E., et al. (2014). Stream biomonitoring using macroinvertebrates around the globe: a comparison of large-scale programs. *Environ. Monit. Assess.* 187:4132. doi: 10.1007/s10661-014-4132-8

Cardinale, B. J., Duffy, J. E., Gonzalez, A., and Hooper, D. U. (2012). Biodiversity loss and its impact on humanity. *Nature* 486, 59–67. doi: 10.1038/nature11148

Carstensen, J., and Lindegarth, M. (2016). Confidence in ecological indicators: a framework for quantifying uncertainty components from monitoring data. *Ecol. Indic.* 67, 306–317. doi: 10.1016/j.ecolind.2016.03.002

Clarke, R. (2009). *Uncertainty in WFD Assessments for Rivers Based on Macroinvertebrates and RIVPACS.* Integrated Catchment Science Programme Science Report: SC060044/SR4, Bristol, UK, 1–87.

Clarke, R. T., Furse, M. T., Gunn, R. J. M., Winder, J. M., and Wright, J. F. (2002). Sampling variation in macroinvertebrate data and implications for river quality indices. *Freshw. Biol.* 47, 1735–1751. doi: 10.1046/j.1365-2427.2002.00885.x

Clarke, R. T., Lorenz, A., Sandin, L., Schmidt-Kloiber, A., Strackbein, J., Kneebone, N. T., et al. (2006). Effects of sampling and sub-sampling variation using the STAR-AQEM sampling protocol on the precision of macroinvertebrate metrics. *Hydrobiologia* 566, 441–459. doi: 10.1007/s10750-006-0078-3

Compson, Z. G., Monk, W. A., Curry, C. J., Gravel, D., Bush, A., C., et al. (2018). Linking DNA metabarcoding and text mining to create network-based biomonitoring tools: a case study on boreal wetland macroinvertebrate communities. *Adv. Ecol. Res.* 59, 33–74. doi: 10.1016/bs.aecr.2018.09.001

Cordier, T., Forster, D., Dufresne, Y. C., Martins, I. M., Stoeck, T., et al. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Mol. Ecol. Resour.* 18, 1381–1391. doi: 10.1111/1755-0998.12926

Cristescu, M. E., and Hebert, P. D. N. (2018). Uses and misuses of environmental DNA in biodiversity science and conservation. *Annu. Rev. Ecol. Evol. Syst.* 49, 209–230. doi: 10.1146/annurev-ecolsys-110617-062306

Curry, C. J., Gibson, J. F., Shokralla, S., Hajibabaei, M., and Baird, D. J. (2018). Identifying North American freshwater invertebrates using DNA barcodes: are existing COI sequence libraries fit for purpose? *Freshw. Sci.* 37, 178–189. doi: 10.1086/696613

Davis, A. J., Williams, K. E., Snow, N. P., Pepin, K. M., and Piaggio, A. J. (2018). Accounting for observation processes across multiple levels of uncertainty improves inference of species distributions and guides adaptive sampling of environmental DNA. *Ecol. Evol.* 8, 10879–10892. doi: 10.1002/ece3.4552

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., et al. (2017). Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol. Ecol.* 26, 5872–5895. doi: 10.1111/mec.14350

Doberstein, C. P., Karr, J. R., and Conquest, L. L. (2000). The effect of fixed-count subsampling on macroinvertebrate biomonitoring in small streams. *Freshw. Biol.* 44, 355–371. doi: 10.1046/j.1365-2427.2000.00575.x

Downes, B. J., Lake, P. S., and Schreiber, E. S. G. (1993). Spatial variation in the distribution of stream invertebrates: implications of patchiness for models of community organization. *Freshw. Biol.* 30, 119–132. doi: 10.1111/j.1365-2427.1993.tb00793.x

ECCC (2018). *CABIN Canadian Aquatic Biomonitoring Network, Environment and Climate Change Canada*. Available online at: https://open.canada.ca/data/en/dataset/13564ca4-e330-40a5-9521-bfb1be767147

Effenberger, M., Sailer, G., Townsend, C. R., and Matthaei, C. D. (2006). Local disturbance history and habitat parameters influence the microdistribution of stream invertebrates. *Freshw. Biol.* 51, 312–332. doi: 10.1111/j.1365-2427.2005.01502.x

Elbrecht, V., and 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. doi: 10.1371/journal.pone.0130324

Elbrecht, V., and Leese, F. (2017). Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. *Front. Environ. Sci.* 5:38. doi: 10.3389/fenvs.2017.00038

Elbrecht, V., Peinert, B., and Leese, F. (2017a). Sorting things out: assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecol. Evol.* 7, 6918–6926. doi: 10.1002/ece3.3192

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

Emilson, C. E., Thompson, D. G., Venier, L. A., Porter, T. M., Swystun, T., Chartrand, D., et al. (2017). DNA metabarcoding and morphological macroinvertebrate metrics reveal the same changes in boreal watersheds across an environmental gradient. *Sci. Rep.* 7:12777. doi: 10.1038/s41598-017-13157-x

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., et al. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Mol. Ecol. Resour.* 15, 543–556. doi: 10.1111/1755-0998.12338

Fonseca, D. M., and Hart, D. D. (2001). Colonization history masks habitat preferences in local distributions of stream insects. *Ecology* 82, 2897–2910. doi: 10.1890/0012-9658(2001)082[2897:CHMHPI]2.0.CO;2

Friberg, N., Bonada, N., Bradley, D. C., Dunbar, M. J., Edwards, F. K., Grey, J., et al. (2011). "Biomonitoring of human impacts in freshwater ecosystems: the good, the bad and the ugly," in *Advances in Ecological Research, Vol. 44*, ed G. Woodward (Amsterdam), 1–68.

Furse, M., Hering, D., Moog, O., Verdonschot, P., Johnson, R. K., Brabec, K., et al. (2006). "The STAR project: context, objectives and approaches," in *The Ecological Status of European Rivers: Evaluation and Intercalibration of*

Assessment Methods, eds M. T. Furse, D. Hering, K. Brabec, A. Buffagni, L. Sandin, and P. F. M. Verdonschot (Dordrecht: Springer Netherlands), 3–29.

Furse, M. T., Wright, J. F., Armitage, P. D., and Moss, D. (1981). An appraisal of pond-net samples for biological monitoring of lotic macro-invertebrates. *Water Res*. 15, 679–689. doi: 10.1016/0043-1354(81)90160-3

Gardner, T. A., Barlow, J., Araujo, I. S., Ávila-Pires, T. C., Bonaldo, A. B., Costa, J. E., et al. (2008). The cost-effectiveness of biodiversity surveys in tropical forests. *Ecol. Lett.* 11, 139–150. doi: 10.1111/j.1461-0248.2007.01133.x

Gibson, J., Shokralla, S., Porter, T. M., King, I., van Konynenburg, S., Janzen, D. H., et al. (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proc. Natl. Acad. Sci. U.S.A*. 111, 8007–8012. doi: 10.1073/pnas.1406468111

Gibson, J. F., Shokralla, S., Curry, C., Baird, D. J., Monk, W. A., King, I., et al. (2015). Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. *PLoS ONE* 10:e0138432. doi: 10.1371/journal.pone.0138432

Giupponi, C. (2007). Decision support systems for implementing the european water framework directive: the MULINO approach. *Environ. Model. Softw*. 22, 248–258. doi: 10.1016/j.envsoft.2005.07.024

GRDI-Ecobiomics (2017). *Ecobiomics: Metagenomics Based Ecosystem Biomonitoring Project, Government of Canada, Genomics R&D Initiative, Year-End Performance Report for Shared Priority Projects (2017–2018)*. GRDI-Ecobiomics.

Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography* 40, 281–295. doi: 10.1111/ecog.02445

Haase, P., Murray-Bligh, J., Lohse, S., Pauls, S., Sundermann, A., Gunn, R., et al. (2006). Assessing the impact of errors in sorting and identifying macroinvertebrate samples. *Hydrobiologia* 566, 505–521. doi: 10.1007/s10750-006-0075-6

Haase, P., Pauls, S. U., Schindehütte, K., and Sundermann, A. (2010). First audit of macroinvertebrate samples from an EU Water Framework Directive monitoring program: human error greatly lowers precision of assessment results. 29, 1279–1291, 13. doi: 10.1899/09-183.1

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A. C., and Baird, D. J. (2011). Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE* 6:e17497. doi: 10.1371/journal.pone.0017497

Hajibabaei, M., Spall, J. L., Shokralla, S., and van Konynenburg, S. (2012). Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecol.* 12:28. doi: 10.1186/1472-6785-12-28

Hawkins, C. P., Norris, R. H., Hogue, J. N., and Feminella, J. W. (2000). Development and evaluation of predictive models for measuring the biological integrity of streams. *Ecol. Appl*. 10, 1456–1477. doi: 10.1890/1051-0761(2000)010[1456:DAEOPM]2.0.CO;2

Hering, D., Borja, A., Jones, J. I., Pont, D., Boets, P., Bouchez, A., et al. (2018). Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. *Water Res.* 138, 192–205. doi: 10.1016/j.watres.2018.03.003

Jackson, J. K., Battle, J. M., White, B. P., Pilgrim, E. M., Stein, E. D., Miller, P. E., et al. (2014). Cryptic biodiversity in streams: a comparison of macroinvertebrate communities based on morphological and DNA barcode identifications. *Freshw. Sci*. 33, 312–324. doi: 10.1086/675225

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., et al. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecol. Lett.* 16, 1245–1257. doi: 10.1111/ele.12162

Jones, F. C. (2008). Taxonomic sufficiency: The influence of taxonomic resolution on freshwater bioassessments using benthic macroinvertebrates. *Environ. Rev*. 16, 45–69. doi: 10.1139/A07-010

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

Lancaster, J., and Downes, B. J. (2014). Population densities and density–area relationships in a community with advective dispersal and variable mosaics of resource patches. *Oecologia* 176, 985–996. doi: 10.1007/s00442-014-3062-z

Lavoie, I., Dillon, P. J., and Campeau, S. (2009). The effect of excluding diatom taxa and reducing taxonomic resolution on multivariate analyses and stream bioassessment. *Ecol. Indic.* 9, 213–225. doi: 10.1016/j.ecolind.2008.04.003

Leese, F., Bouchez, A., Abarenkov, K., Altermatt, F., Borja, Á., Bruce, K., et al. (2018). "Chapter 2: Why we need sustainable networks bridging countries, disciplines, cultures and generations for aquatic biomonitoring 2.0: a perspective derived from the DNAqua-net COST action," in *Advances in Ecological R*esearch, eds D. A. Bohan, A. J. Dumbrell, G. Woodward, and M. Jackson (Amsterdam: Academic Press), 63–99. doi: 10.1016/bs.aecr.2018.01.001

Lindenmayer, D. B., and Likens, G. E. (2011). Direct measurement versus surrogate indicator species for evaluating environmental change and biodiversity loss. *Ecosystems* 14, 47–59. doi: 10.1007/s10021-010-9394-6

Macher, J. N., Salis, R. K., Blakemore, K. S., Tollrian, R., Matthaei, C. D., and Leese, F. (2016). Multiple-stressor effects on stream invertebrates: DNA barcoding reveals contrasting responses of cryptic mayfly species. *Ecol. Indic.* 61, 159–169. doi: 10.1016/j.ecolind.2015.08.024

Majaneva, M., Diserud, O. H. S.,Eagle, H. C., Hajibabaei, M., and Ekrem, T. (2018). Choice of DNA extraction method affects DNA metabarcoding of unsorted invertebrate bulk samples. *Metabarcod. Metagenom.* 2:e26664. doi: 10.3897/mbmg.2.26664

Martin, G. K., Adamowicz, S. J., and Cottenie, K. (2016). Taxonomic resolution based on DNA barcoding affects environmental signal in metacommunity structure. *Freshw. Sci*. 35, 701–711. doi: 10.1086/686260

Miller, E. T., Svanbäck, R., and Bohannan, B. J. M. (2018). Microbiomes as metacommunities: understanding host-associated microbes through metacommunity ecology. *Trends Ecol. Evol.* 33, 926–935. doi: 10.1016/j.tree.2018.09.002

Musco, L., Terlizzi, A., Licciano, M., and Giangrande, A. (2009). Taxonomic structure and the effectiveness of surrogates in environmental monitoring: a lesson from polychaetes. *Mar. Ecol. Prog. Ser.* 383, 199–210. doi: 10.3354/meps07989

Nakov, T., Beaulieu, J. M., and Alverson, A. J. (2018). Insights into global planktonic diatom diversity: the importance of comparisons between phylogenetically equivalent units that account for time. *ISME J.* 12, 2807–2810. doi: 10.1038/s41396-018-0221-y

Newbold, T., Hudson, L. N., Hill, S. L., Contu, S., Lysenko, I., and Senior, R. A. (2015). Global effects of land use on local terrestrial biodiversity. *Nature* 520, 45–50. doi: 10.1038/nature14324

Nijboer, R. C., and Schmidt-Kloiber, A. (2004). The effect of excluding taxa with low abundances or taxa with small distribution ranges on ecological assessment. *Hydrobiologia* 516, 347–363. doi: 10.1023/B:HYDR.0000025275.49062.55

Orlofske, J. M., and Baird, D. J. (2013). The tiny mayfly in the room: implications of size-dependent invertebrate taxonomic identification for biomonitoring data properties. *Aquat. Ecol.* 47, 481–494. doi: 10.1007/s10452-013-9460-1

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., et al. (2018). The future of biotic indices in the ecogenomic era: integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Sci. Total Environ.* 637–638, 1295–1310. doi: 10.1016/j.scitotenv.2018.05.002

Petkovska, V., and Urbanič, G. (2010). Effect of fixed-fraction subsampling on macroinvertebrate bioassessment of rivers. *Environ. Monitor. Assess.* 169, 179–201. doi: 10.1007/s10661-009-1161-9

Petsch, D. K. (2016). Causes and consequences of biotic homogenization in freshwater ecosystems. *Int. Rev. Hydrobiol.* 101, 113–122. doi: 10.1002/iroh.201601850

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

Porter, T. M., and Hajibabaei, M. (2018b). Over 2.5 million COI sequences in GenBank and *growing*. *PLoS ONE* 13:e0200177. doi: 10.1371/journal.pone.0200177

Porter, T. M., and Hajibabaei, M. (2018c). Automated high throughput animal CO1 metabarcode classification. *Sci. Rep.* 8:4226. doi: 10.1038/s41598-018-22505-4

Reynoldson, T. B., Norris, R. H., Resh, V. H., Day, K. E., and Rosenberg, D. M. (1997). The reference condition: a comparison of multimetric and multivariate approaches to assess water-quality impairment using benthic macroinvertebrates. *J. North Am. Benthol. Soc.* 16, 833–852. doi: 10.2307/1468175

Rimet, F., Abarca, N., Bouchez, A., Kusber, W.-H., Jahn, R., Kahlert, M., et al. (2018). The potential of High-Throughput Sequencing (HTS) of natural samples as a source of primary taxonomic information for reference libraries of diatom barcodes. *Fottea* 18, 37–54. doi: 10.5507/fot.2017.013

Schmidt-Kloiber, A., and Nijboer, R. C. (2004). The effect of taxonomic resolution on the assessment of ecological water quality classes. *Hydrobiologia* 516, 269–283. doi: 10.1023/B:HYDR.0000025270.10807.10

Schmidt-Kloiber, A., Strackbein, J., Vogl, R., Furse, M. T., and Hering, D. (2014). Description of the AQEM/STAR invertebrate database. *Freshw. Metadata J.* 2, 1–8. doi: 10.15504/fmj.2014.2

Singer, G. A. C., Fahner, N. A., Barnes, J. G., McCarthy, A., and Hajibabaei, M. (2019). Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. *Sci. Rep.* 9:5991. doi: 10.1038/s41598-019-42455-9

Soininen, J., Jamoneau, A., Rosebery, J., and Passy, S. I. (2016). Global patterns and drivers of species and trait composition in diatoms. *Glob. Ecol. Biogeogr.* 25, 940–950. doi: 10.1111/geb.12452

Stokstad, E. (2018). Researchers launch plan to sequence 66,000 species in the United Kingdom. *Science* 366. doi: 10.1126/science.aav9295.

Strachan, S. A., and Reynoldson, T. B. (2014). Performance of the standard CABIN method: comparison of BEAST models and error rates to detect simulated degradation from multiple data sets. *Freshw. Sci.* 33, 1225–1237. doi: 10.1086/678948

Sweeney, B. W., Battle, J. M., Jackson, J. K., and Dapkey, T. (2011). Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? *J. North Am. Benthol. Soc.* 30, 195–216. doi: 10.1899/10-016.1

Taberlet, P., Coissac, E., Hajibabaei, M., and Rieseberg, L. H. (2012). Environmental DNA. *Mol. Ecol.* 21, 1789–1793. doi: 10.1111/j.1365-294X.2012.05542.x

Thorne, R. S. J., Williams, W. P., and Cao, Y. (1999). The influence of data transformations on biological monitoring studies using macroinvertebrates. *Water Res.* 33, 343–350. doi: 10.1016/S0043-1354(98)00247-4

Turak, E., Harrison, I., Dudgeon, D., Abell, R., Bush, A., Darwall, W., et al. (2017). Essential biodiversity variables for measuring change in global freshwater biodiversity. *Biol. Conserv.* 213, 272–279. doi: 10.1016/j.biocon.2016.09.005

Turner, W. (2014). Sensing biodiversity. *Science* 346, 301–302. doi: 10.1126/science.1256014

Vamosi, J. C., Gong, Y.-B., Adamowicz, S. J., and Packer, L. (2017). Forecasting pollination declines through DNA barcoding: the potential contributions of macroecological and macroevolutionary scales of inquiry. *N. Phytol.* 214, 11–18. doi: 10.1111/nph.14356

Vasselon, V., Rimet, F., Tapolczai, K., and Bouchez, A. (2017). Assessing ecological status with diatoms DNA metabarcoding: scaling-up on a WFD monitoring network (Mayotte island, France). *Ecol. Indic.* 82, 1–12. doi: 10.1016/j.ecolind.2017.06.024

Vaughn, C. C., and Spooner, D. E. (2006). Unionid mussels influence macroinvertebrate assemblage structure in streams. *J. North Am. Benthol. Soc.* 25, 691–700. doi: 10.1899/0887-3593(2006)25[691:UMIMAS]2.0.CO;2

Vellend, M., Baeten, L., Myers-Smith, I. H., Elmendorf, S. C., Beauséjour, R., Brown, C. D., et al. (2013). Global meta-analysis reveals no net change in local-scale plant biodiversity over time. *Proc. Natl. Acad. Sci. U.S.A.* 110, 19456–19459. doi: 10.1073/pnas.1312779110

Vivien, R., Wyler, S., Lafont, M., and Pawlowski, J. (2015). Molecular barcoding of aquatic oligochaetes: implications for biomonitoring. *PLoS ONE* 10:e0125485. doi: 10.1371/journal.pone.0125485

Vlek, H. E., Šporka, F., and Krno, I. J. (2006). Influence of macroinvertebrate sample size on bioassessment of streams. *Hydrobiologia* 566, 523–542. doi: 10.1007/s10750-006-0074-7

Voulvoulis, N., Arpon, K. D., and Giakoumis, T. (2017). The EU water framework directive: from great expectations to problems with implementation. *Sci. Total Environ.* 575, 358–366. doi: 10.1016/j.scitotenv.2016.09.228

Weigand, H., Beermann, A. J., Ciampor, F., Costa, F. O., Csabai, Z., Duarte, S., et al. (2019). DNA barcode reference libraries for the monitoring of aquatic biota in Europe: gap-analysis and recommendations for future work. *Sci. Total Environ.* 678, 499–524. doi: 10.1016/j.scitotenv.2019.04.247

Woodward, G., Gray, C., and Baird, D. J. (2013). Biomonitoring for the 21 st Century: new perspectives in an age of globalisation and emerging environmental threats. *Limnetica* 29, 159–174. doi: 10.23818/limn.32.14

Wright, J. F., Moss, D., Armitage, P. D., and Furse, M. T. (1984). A preliminary classification of running-water sites in Great Britain based on macro-invertebrate species and the prediction of community type using environmental data. *Freshw. Biol.* 14, 221–256. doi: 10.1111/j.1365-2427.1984.tb00039.x

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., et al. (2012). Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods Ecol. Evol.* 3, 613–623. doi: 10.1111/j.2041-210X.2012.00198.x

Zhang, G. K., Chain, F. J. J., Abbott, C. L., and Cristescu, M. E. (2018). Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities. *Evol. Appl.* 11, 1901–1914. doi: 10.1111/eva.12694

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

ORIGINAL RESEARCH
published: 25 November 2019
doi: 10.3389/fevo.2019.00395

Network-Based Biomonitoring: Exploring Freshwater Food Webs With Stable Isotope Analysis and DNA Metabarcoding

Zacchaeus G. Compson[1,2,3], Wendy A. Monk[1,4], Brian Hayden[3], Alex Bush[1,3,5], Zoë O'Malley[3], Mehrdad Hajibabaei[2,6], Teresita M. Porter[6,7], Michael T. G. Wright[6], Christopher J. O. Baker[8,9], Mohammad Sadnan Al Manir[8,9], R. Allen Curry[3,4] and Donald J. Baird[1,3]*

[1] Environment and Climate Change Canada, University of New Brunswick, Fredericton, NB, Canada, [2] Centre for Environmental Genomics Applications, St. John's, NL, Canada, [3] Canadian Rivers Institute, Department of Biology, University of New Brunswick, Fredericton, NB, Canada, [4] Faculty of Forestry and Environmental Management, University of New Brunswick, Fredericton, NB, Canada, [5] Lancaster Environment Centre, Lancaster University, Lancaster, United Kingdom, [6] Department of Integrative Biology, Centre for Biodiversity Genomics, University of Guelph, Guelph, ON, Canada, [7] Great Lakes Forestry Centre, Natural Resources Canada, Sault Ste. Marie, ON, Canada, [8] Department of Computer Science, University of New Brunswick, Saint John, NB, Canada, [9] IPSNP Computing Inc., Saint John, NB, Canada

OPEN ACCESS

Edited by:
*Alireza Tamaddoni-Nezhad,
Imperial College London,
United Kingdom*

Reviewed by:
*Michelle Jackson,
Imperial College London,
United Kingdom
Lingxin Chen,
Yantai Institute of Coastal Zone
Research (CAS), China*

***Correspondence:**
*Zacchaeus G. Compson
zacchaeus.greg.compson@gmail.com*

Specialty section:
*This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Ecology and Evolution*

Received: *22 June 2019*
Accepted: *04 October 2019*
Published: *25 November 2019*

Citation:
*Compson ZG, Monk WA, Hayden B,
Bush A, O'Malley Z, Hajibabaei M,
Porter TM, Wright MTG, Baker CJO,
Al Manir MS, Curry RA and Baird DJ
(2019) Network-Based Biomonitoring:
Exploring Freshwater Food Webs With
Stable Isotope Analysis and DNA
Metabarcoding.
Front. Ecol. Evol. 7:395.
doi: 10.3389/fevo.2019.00395*

Threatened freshwater ecosystems urgently require improved tools for effective management. Food web analysis is currently under-utilized, yet can be used to generate metrics to support biomonitoring assessments by measuring the stability and robustness of ecosystems. Using a previously developed analysis pipeline, we combined taxonomic outputs from DNA metabarcoding with a text-mining routine to extract trait information directly from the literature. This pipeline allowed us to generate heuristic food webs for sites within the lower Saint John/Wolastoq River and the Grand Lake Meadows (hereafter called the "GLM complex"), Atlantic Canada's largest freshwater wetland. While these food webs are derived from empirical traits and their structure has been shown to discriminate sites both spatially and temporally, the accuracy of their properties have not been assessed against other methods of trophic analysis. We explored two approaches to validate the utility of heuristic food webs. First, we qualitatively compared how well-trophic position derived from heuristic food webs recovered spatial and temporal differences across the GLM complex in comparison to traditional stable isotope approaches. Second, we explored how the trophic position of invertebrates, derived from heuristic food webs, predicted trophic position measured from δ^{15}N values. In general, both heuristic food webs and stable isotopes were able to detect seasonal changes in maximum trophic position in the GLM complex. Samples from the entire GLM complex demonstrated that prey-averaged trophic position measured from heuristic food webs strongly predicted trophic position inferred from stable isotopes ($R^2 = 0.60$), and even stronger relationships were observed for some individual models ($R^2 = 0.78$ for best model). Beyond their areas of congruence, heuristic food web and stable isotope analyses also appear to complement one another, suggesting a surprising degree of independence between community trophic niche width (assessed from stable isotopes) and food web size and complexity (assessed from heuristic food webs). Collectively, these analyses indicate that trait-based networks have properties that correspond to those of actual food webs, supporting the routine adoption of food web metrics for ecosystem biomonitoring.

Keywords: DNA metabarcoding, ecological network, trait, food web, stable isotope, trophic position, Bayesian mixing model

INTRODUCTION

Freshwater ecosystems, which house a disproportionate amount of Earth's biodiversity (Dudgeon et al., 2006), face multiple threats (Cazzolla Gatti, 2016; Hu et al., 2017). Freshwater availability (Rodell et al., 2018) and habitat extent (e.g., Dixon et al., 2016) are in decline in most parts of the world. Yet, the very structural and ecological complexity that gives freshwater systems their capacity for biodiversity and ecosystem services also makes them very difficult to study. This is particularly true for the planet's wetlands, which are generally viewed as hard to define, seasonally variable, and often inaccessible.

While widely studied by ecologists, food web networks are under-utilized in bioassessment, even though they provide a wide range of information–from taxon-specific, node-to-node information to higher order information aggregated across the community–and are a tool for visualizing the dense information in complex systems, such as wetlands. Metrics from food web analysis can be used to infer their stability (May, 1972), robustness to biodiversity loss (Estrada, 2007; Gilbert, 2009), and different assembly or interaction mechanisms (Vázquez, 2005; Williams, 2011). Food web networks explicitly show biodiversity, species interactions, and structural and functional relationships of ecosystems (Dunne et al., 2002b; Thompson et al., 2012), and are an intuitive communication tool for environmental managers, particularly when presenting to lay audiences. However, constructing food webs is laborious (Thompson et al., 2012), underscoring the need for new tools that can facilitate this process to support wider implementation in bioassessment (Bohan et al., 2017).

DNA metabarcoding has the potential to revolutionize biomonitoring and bioassessment by providing a fast way of consistently observing biodiversity in high-resolution detail (Baird and Hajibabaei, 2012). Further, DNA metabarcoding can be both more cost-effective and more efficient than traditional biomonitoring (Aylagas et al., 2018). The rapid adoption of DNA metabarcoding can be seen in the exponential rise in the number of papers published about biomonitoring with DNA metabarcoding in the last decade[1]. While much of the early literature focused on assessing how well DNA metabarcoding technologies could reproduce biodiversity data collected by traditional means (e.g., Gibson et al., 2015; Emilson et al., 2017), more recent efforts have expanded the application of this approach, exploring possibilities for leveraging genomic data in novel ways (e.g., Gray et al., 2014; Bohan et al., 2017; Derocles et al., 2018; Deagle et al., 2019).

Concurrently, efforts have sought to make use of organism traits to leverage existing biodiversity knowledge for bioassessment. Traits-based approaches assume that environmental filtering selects species with suites of traits that allow them to coexist under similar environmental conditions (Poff, 1997). Since many species share the same traits, traits-based approaches are taxon-free measures of biodiversity (sensu Damuth et al., 1992; Doledec and Statzner, 2008; Andrews and Hixson, 2014). Body size, for example, is a trait that aggregates information across taxonomic groups, and has been invoked as a powerful, trait-based indicator of community responses to disturbance (Liu et al., 2015). While growing interest in traits-based approaches has prompted some of its key advocates to call it a "bandwagon" (McGill, 2015; Didham et al., 2016), we argue trait-based approaches have failed to realize their full potential, focusing primarily upon phenomenological case studies and relying on re-application of traits approaches to traditional analyses (e.g., ordination approaches). This approach has led to vacuous generalizations about traits approaches, lacking in mechanistic evidence (Didham et al., 2016). Recently there has been growing interest in using key traits to construct ecological networks by linking them to the rich taxonomic lists generated by DNA metabarcoding. Bohan et al. (2017) advocated for such an approach to improve biomonitoring, and tools for the construction of heuristic food webs from biological community data have been developed (Gray et al., 2015; Compson et al., 2018). Nonetheless, while these studies have demonstrated a proof-of-concept for heuristic food web construction from ecological traits, the scale of these applications has been limited, and their connection to real food webs remains unknown, as are their relationships to ecosystem functions.

Stable isotope analysis is one of the primary ways of assessing food webs, yet it is an imperfect approach, often elucidating only part of the food web, and requires (1) information for all consumer food sources, failing when isotopic signatures are too similar (Birkhofer et al., 2017), (2) an understanding of how different sources fractionate for different isotopes, tissues, and life stages of the consumer (Post, 2002; McCutchan et al., 2003), and (3) an appropriate number of isotopes (i.e., n – 1 per food source, Fry, 2006), as the usefulness of mixing models declines when the number of sources exceeds the number of isotopes (Lerner et al., 2018). Thus, despite the complementarity of DNA metabarcoding and stabile isotope information (Kartzinel et al., 2015), the promise of merging this information to provide new ecological insights has not been fully realized.

[1] Makiola, A., Compson, Z. G., Baird, D. J., Barnes, M. A., Boerlijst, S. P., Bouchez, A., et al. (2019). Key questions for the next-generation of biomonitoring. *Front. Ecol. Evol.*

Here, we assess the spatial and temporal variability of invertebrate food webs of a large wetland complex, and examine how trait-based and stable isotope approaches can be used to assess these complex systems. Specifically, this study explores how stable isotope information can be used to both validate and improve the inference from heuristic food webs that are themselves constructed from DNA metabarcoding data integrated with trait information. We first explore the performance of heuristic food web properties (e.g., trophic links, omnivory, trophic position) at resolving spatial and temporal differences in a large wetland complex. Second, we qualitatively explore how heuristic food web and stable isotope approaches compare at resolving spatial and temporal patterns of trophic position. Third, we compare how strongly the trophic position of invertebrates inferred using heuristic food webs predicts trophic position as measured by $\delta^{15}N$ values. Fourth, we examine how the unprecedented detail provided by DNA-derived heuristic food webs provides complementary information to stable isotope analysis of trophic niche width. We conclude with an exploration of the utility of heuristic food web analysis as a management tool for rapid bioassessment.

MATERIALS AND METHODS

Study Sites and Sample Collection

Our study area encompassed the lower Saint John/Wolastoq River (SJWR) and the connected Grand Lake Meadows (GLM), Atlantic Canada's largest freshwater wetland (**Figure 1**); hereafter, we refer to the SJWR and GLM collectively as the "GLM complex." We examined three regions within this vast wetland complex: a region within the mainstem SJWR ("mainstem"), a region within the Portobello National Wildlife Refuge area in the heart of the GLM ("wetland"), and a region in the Jemseg River ("transition"), which is a low-flow, intermediate system connecting the GLM to the SJWR. Within each of these three regions, sites ($n = 6$ sites per region) were chosen to capture the range of habitat and flow variability across these regions, and to provide a wide range of trophic variability by which to explore relationships between metrics from heuristic food web and stable isotope analyses.

Sites were sampled in early June, early September, and mid-December, 2016. We chose these time points because they represented the beginning, middle, and end of the active, ice-free season in our system. In the spring, ice break up occurs, creating large ice jams and massive spring flooding in the SJRW and GLM systems; consequently, June was the earliest point that we could sample with conditions returning to base flow. Peak productivity occurred in late-August into early September, and the active, ice-free season ended in mid-December during our final sampling event. Because major ice-flows scoured our system in spring of 2016, very little aquatic insect biomass was observed post-flood in June. Because of this, stable isotope samples were only collected in September and December, when there was enough biomass for sampling. Because early September (the peak of biological activity) and mid-December (when ice began reforming in our system) represented extremes in the biological activity in our system, the time points we selected for stable isotope sampling were expected to represent extremes in associated food webs and trophic dynamics.

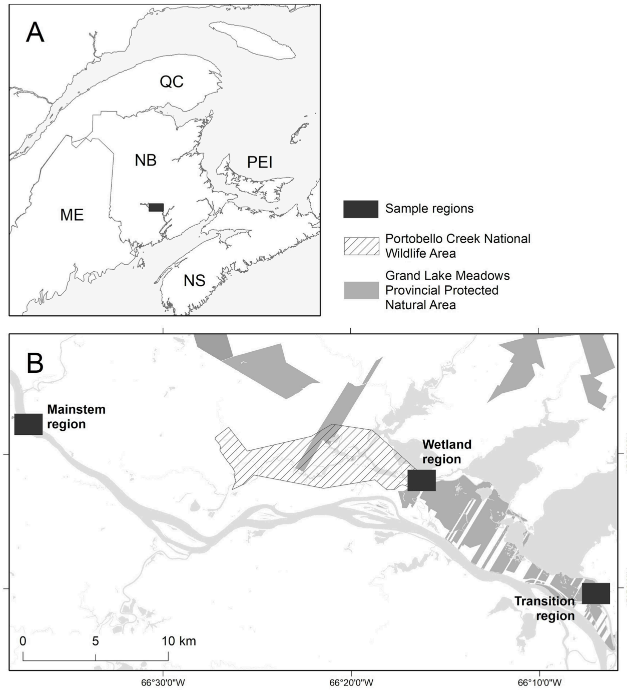


FIGURE 1 | Map of the study area. **(A)** The Grand Lake Meadows complex, located in southern New Brunswick, is Atlantic Canada's largest freshwater wetland. **(B)** This complex is protected both nationally (diagonal lines) and provincially (light gray). Our study area consisted of three distinct regions (dark gray): the wetland (Wetland), the mainstem Saint John/Wolastoq River (Mainstem), and the Jemseg River (Transition), which connects the wetland to the mainstem region. Within each region, six sites were sampled three times (early June, early September, and mid-December).

At each site ($n = 18$, each sampled at three time points), paired benthic kick-net samples were collected using the standard protocol from the Canadian Aquatic Biomonitoring Network (CABIN; https://www.canada.ca/en/environment-climate-change/services/canadian-aquatic-biomonitoring-network.html): one sample for bulk sequencing and DNA metabarcoding of invertebrates, and a second sample for assessment of morphological identification of these taxa, organism body size, abundance, and analysis of stable isotopes ($\delta^{13}C$ and $\delta^{15}N$). For bulk DNA samples, sterile technique was used and nets were sterilized between samples in a 1% bleach solution. DNA benthic samples ("DNA") were preserved in 95% ethanol and stored at −80°C, but samples for microscopy and stable isotope analysis ("morphological") were directly frozen (−20°C) on return to the lab, to avoid altering their isotopic composition (Arrington and Winemiller, 2002; Barrow et al., 2008). Additionally, samples for producer baselines were taken at each site, including detritus, biofilm, and dominant macrophytes; these samples were placed in paper bags and dried at 60°C in the lab for at least 72 h.

Laboratory Processing

Morphological samples were thawed and sorted, with invertebrates identified to the lowest taxonomic level (usually genus) using standard morphological keys (e.g., Merritt et al., 2008). Voucher samples and high-resolution digital images of key taxa are stored at the Environment and Climate Change Canada lab at the University of New Brunswick, Fredericton. Additionally, individuals were measured (total length, mm), and dried at 60°C. For stable isotope samples, all dominant taxa (based on biomass) were selected from samples covering all major functional feeding groups. Because of the mass requirements for stable isotope samples, we only included samples for taxonomic groups that had a dry biomass of $\geq$0.6 mg. This usually translated to each stable isotope sample including many individuals (>20); however, many predator groups were assessed from fewer individuals (often <3) because of the rarity of these taxa. Dried macroinvertebrate and food web base (i.e., biofilm, macrophytes, and leaf litter) samples were homogenized, weighed on a Sartorius MC21S microbalance (3.0 ± 0.1 mg for biofilm and plant tissue and 1.0 ± 0.1 mg for macroinvertebrate tissue), enclosed in 4 × 6 mm tin cups (Costech Analytical Technologies Inc., Valencia, California, USA), and delivered to the Stable Isotopes in Nature Laboratory (SINLAB) at the University of New Brunswick (http://www.unb.ca/research/institutes/cri/sinlab/) for stable isotope analysis.

Stable Isotope Analysis

Natural abundance stable isotopes of C and N were assessed for aquatic invertebrates and the food web base. Invertebrate and food web base ^{13}C, ^{15}N, C, and N content were measured using a Carlo Erba NC 2500 Elemental Analyzer (CE Instruments, Milan, Italy) with a Thermo-Finnigan Delta Plus XP (Thermo-Electron Corp., Bremen, Germany) isotope ratio mass spectrometer at SINLAB. Macroinvertebrate and food web base δ^{13}C and δ^{15}N isotope compositions were expressed in parts per thousand (‰) relative to Vienna PeeDee Belemnite for C and air for N, as follows:

$$\delta = ([R_{\text{sample}}/R_{\text{standard}}] - 1)\times 1,000 \tag{1}$$

where R is the ratio ^{13}C/^{12}C or ^{15}N/^{14}N. Instrumental error—measured as the standard deviation of repeated measurements of working laboratory standards (i.e., caffeine (δ^{13}C = −35.05‰, δ^{15}N = −2.87‰), bovine liver (δ^{13}C = −18.8‰, δ^{15}N = 7.2‰) and muskellunge liver (δ^{13}C = −22.3‰, δ^{15}N = 14‰))—was <0.1 ‰ for both δ^{13}C and δ^{15}N.

DNA Extraction and Sequencing

Benthic samples for DNA metabarcoding were packed on ice and shipped to the Biodiversity Institute of Ontario at the University of Guelph for DNA extraction, PCR amplification, and high throughput sequencing (HTS). Briefly, samples were homogenized in sterile blenders and the slurry was subsampled into 50 mL conical tubes. Samples were centrifuged, excess preservative ethanol was removed, and residual ethanol was evaporated at 65°C. Once dry, the homogenate was subsampled into 2 mL lysing matrix tubes (MP Biomedicals, Solon, Ohio) and further homogenized using a MP FastPrep-24 Classic tissue homogenizer (MP Biomedicals). Samples were then extracted using a NucleoSpin Tissue Kit (Machery-Nagel, Düren, German) according to the manufacturer's protocol, eluting with 30 uL molecular grade water. Samples were extracted in batches of 12–18 with a negative control (no sample added) for each batch.

Two COI fragments were amplified using the primer sets BR5 (B_F 5′ CCIGAYATRGCITTYCCICG, R5_R 5′ GTRATIGCICCIGCIARIACIGG—314 bp) and F230R (Folmer_F 5′ GGTCAACAAATCATAAAGATATTGG 230R_R 5′ CTTATRTTRTTTATICGIGGRAAIGC—230 bp) in a two-step PCR following the protocol outlined in Gibson et al. (2015), with the exception of having a 35 cycle regime in the first PCR. For both primer sets, the annealing temperature (Ta) was 46°C for 1 min. The melting temperatures (Tm) for these primers are as follows: BR5 (forward = 61.4°C, reverse = 56.4°C) and F230 (forward = 50.5°C, reverse = 56.7°C). For further information about these primers, which were designed to target a wide range of arthropod orders, see Gibson et al. (2015). A negative control was included for each batch of PCR, which was carried through each of the two PCR steps. Amplification success was confirmed visually using a 1.5% agarose gel. Amplicons were purified using a MinElute DNA purification system (Qiagen) and quantified using a Quant-iT (Invitrogen, Waltham Massachusetts, United States) PicoGreen dsDNA assay on a TBS-380 Mini-Fluorometer (Turner Biosystems Sunnyvale California, United States). All samples were normalized to the same concentration, and the two amplified fragments were pooled for each sample prior to dual-indexing using the Nextera XT Index Kit (Illumina, San Diego, California) (FC-131-1002). Indexed samples were pooled into one tube, purified through magnetic bead purification, and quantified using the PicoGreen dsDNA assay. Average fragment length was determined on an Agilent Bioanalyzer 2100 (Santa Clara, California, United States) before sequencing the library on an Illumina MiSeq using the V3 sequencing chemistry kit (2 × 300) (MS-102-3003). A 10% spike-in of PhiX was used as a control.

Bioinformatic Methods

Raw Illumina MiSeq paired-end reads were processed using the SCVUC v2.1 COI metabarcode pipeline, available on GitHub at https://github.com/Hajibabaei-Lab/SCVUC_COI_metabarcode_pipeline. At each step, read and ESV statistics were calculated (**Supplementary Table S1**) using custom scripts (available at the above link). Briefly, forward and reverse raw reads were paired using SEQPREP (available from https://github.com/jstjohn/SeqPrep) with a Phred quality score cutoff of 20 and an overlap of at least 25 bp (St. John, 2016). For each marker, forward and reverse primers were trimmed using CUTADAPT v1.14, ensuring trimmed reads were at least 150 bp long, allowing no more than 3 N's, and ensuring a minimum Phred quality score of 20 at the ends (Martin, 2011). A global ESV analysis was conducted by pooling all the data together, dereplicating the reads using VSEARCH v2.4.2 with the "derep_fulllength" command, and denoising with USEARCH v10.0.240 using the unoise3 algorithm (Edgar, 2016; Rognes et al., 2016). The denoising step removes sequences with predicted sequence errors, any PhiX carryover from MiSeq sequencing, putative chimeric sequences, and rare clusters. We defined rare clusters as exact sequence variant (ESV) clusters including only one or two reads (singletons and doubletons) (Callahan et al., 2017). A sample ESV matrix was generated using VSEARCH with the "usearch_global" command with an identity of 1.0 (100% exact sequence mapping, including matching of exact substrings). The denoised ESVs were taxonomically assigned using the COI classifier v3.2 (Porter and Hajibabaei, 2018; https://github.com/terrimporter/CO1Classifier).

Heuristic Food Web Construction

Heuristic food webs were constructed using a previously published pipeline (Compson et al., 2018). Briefly, this pipeline takes presence-absence taxonomic lists generated by DNA metabarcoding and pairs it with a customized interaction database covering the taxa found in our system. Our database of pairwise trophic interactions was created using a) information gathered from existing trophic databases (e.g., Database of Trophic Interactions; Brose et al., 2005), b) information from a secondary text-mining pipeline, and c) information manually gathered from systematic literature searches. Specifically, we used an updated trophic linkage database from Compson et al. (2018), which we updated to include novel taxa found in the GLM complex. Information gaps on species linkages, caused by missing species in our trophic interaction database, were inferred using a series of trait filters based on other information, including functional feeding group, body size, and phylogenetic relatedness. From the complete set of possible pairwise interactions, linkages were first reduced based on the known functional feeding group of each taxa, and then further reduced based on the average body size of each taxa. When functional feeding group or body size traits were not available, we obtained these traits from the next closest related species. The updated trophic linkage database includes 50,975 pairwise interactions and covers 965 invertebrate genera. Using this updated database, we created adjacency matrices by constraining interactions to only taxa present in a sample (for individual food webs) or region (for metawebs). We then used the *cheddar* R package (version 0.1-633; Hudson et al., 2013) to create food webs for each replicate sample and extract relevant food web metrics used for subsequent analyses.

Mixing Models for Trophic Position

Analysis of trophic position was done in two ways. First, we created a two-source, two-isotope Bayesian mixing model using the *tRophicPosition* position package (version 0.7.7; Quezada-Romegialli et al., 2018) in R to summarize trophic position of consumers in each of the dominant functional feeding groups in our system (i.e., predators, collectors, grazers, omnivores, and shredders) (Model 1):

$$\delta^{15}N_c = \Delta N(TP - \lambda) + \alpha(\delta^{15}N_{b1} + \delta^{15}N_{b2}) - \delta^{15}N_{b2}, \quad (2)$$

where $\delta^{15}N_c$ is the nitrogen isotopic ratio of the consumer, $\delta^{15}N_{b1}$ is the nitrogen isotopic ratio of the first baseline (biofilm), $\delta^{15}N_{b2}$ is the nitrogen isotopic ratio of the second baseline (leaf litter), ΔN is the trophic enrichment factor (TEF) for nitrogen, *TP* is the trophic position of the target consumer, and λ is the trophic position of the baseline. Additionally, this model uses a secondary mixing model to calculate α, which accounts for fraction in $\delta^{13}C$ and estimates the relative contribution of each source to the consumer's trophic position:

$$\alpha = ([\delta^{13}C_{b2} - (\delta^{13}C_c + \Delta C)]/(TP - \lambda))/(\delta^{13}C_{b2} + \delta^{13}C_{b1}), \quad (3)$$

where $\delta^{13}C_c$ is the carbon isotopic ratio of the consumer for which we want to estimate trophic position, $\delta^{13}C_{b1}$ is the carbon isotopic ratio of the first baseline (biofilm), $\delta^{13}C_{b2}$ is the carbon isotopic ratio of the second baseline (leaf litter), and ΔC is the TEF for carbon. The Bayesian approach allows Equations (2) and (3), which both include *TP* and α, to be solved iteratively, with δ^{13}C and δ^{15}N values and TEFs for both consumers and baselines modeled as random variables with vague prior normal distributions of their means [$dnorm(0,\tau)$, $\tau = 1/\text{SD}^2$] and vague prior uniform distributions of their standard deviations [*dunif*(1,100)]; *TP* and α are treated as random parameters with uniform and Beta prior distributions, respectively (Quezada-Romegialli et al., 2018). We used the function "multiSpeciesTP" to define and initialize the Bayesian model, and to sample the posterior distribution of trophic position. The Bayesian model ran 10,000 iterations used for the parameters "n.adapt," "n.iter," and "burnin" and used five parallel Markov Chain Monte Carlo (MCMC) simulations using the JAGS (version 4.3.0) Gibbs sampler (Plummer, 2003).

The second approach we employed for calculating trophic position (*TP*) was a more conventional model (Model 2; Post, 2002):

$$TP = \lambda + (\delta^{15}N_c - [\delta^{15}N_{b1} \times \alpha + \delta^{15}N_{b2} \times (1 - \alpha)])/\Delta N. \quad (4)$$

Here, α is defined as,

$$\alpha = (\delta^{13}C_c - \delta^{13}C_{b2})/(\delta^{13}C_{b1} - \delta^{13}C_{b2}). \quad (5)$$

This model allowed us to calculate individual trophic position values for all consumers in our system, including those taxa represented by few individuals; Bayesian models, which require replicate observations for each consumer estimated, could only provide group-level estimates and confidence intervals for well-represented individuals and functional feeding groups across food webs. Consequently, trophic position values obtained from Model 2 were used for all linear regression analyses. For both Model 1 and Model 2, TEFs were based on values for whole organisms reported in McCutchan et al. (2003). Additionally, for predators, we explored models using consumers as the baseline, which involved changing the baseline trophic position from $\lambda = 1$ (for biofilm and detritus) to $\lambda = 2$ for consumers feeding primarily on biofilm (i.e., grazers) and terrestrial leaf litter (i.e., shredders). Finally, from trophic position (*TP*) estimated from Model 2, we also calculated adjusted trophic position (*ATP*), which used known information about the habits of each organism or functional feeding group to constrain trophic position to not go below these trophic levels (i.e., $\lambda = 2$ for consumers and $\lambda = 3$ for predators).

Bayesian Estimates of Isotopic Food Web Size

To compare how community-wide trophic niche breadth varied spatially and temporally, we used the *SIBER* R package (version 2.1.4; Jackson et al., 2011). Specifically, we examined the Bayesian posterior estimate of the convex hull area of each community, which encompasses all species in δ^{13}C-δ^{15}N bi-plot space and is a measure of the total amount of niche space occupied by the community (Layman et al., 2007). The Bayesian model utilized a JAGS Gibbs sampler with five MCMC chains. This model fit initial multivariate normal distributions to each group in the dataset with *rjags* (version 4-8), using the recommended default *SIBER* parameters and priors (Jackson et al., 2011). The default priors included an inverse Wishart prior for fitting ellipses and a vague normal prior for the means; vague normal priors are recommended for fitting the means because *SIBER* internally *z*-score standardizes the data before model fitting, aiding in the JAGS fitting process (Jackson et al., 2011). Because of spatial and temporal variation in our producer baselines (**Supplementary Figure S1**, **Supplementary Table S2**), prior to running Bayesian models we converted our δ^{13}C data to autochthonous reliance values (0–1) and our δ^{15}N data to trophic position values (based on Model 2).

Statistical Analyses

All statistical analyses were conducted in R (version 3.6.0; R Core Team, 2013). To assess the spatial and temporal variability in heuristic food web metrics, including estimates of trophic position, we created a linear mixed effects model using the *lme4* package (version 1.1-21; Bates et al., 2019), where Season and Region were fixed effects, and Site was a random effect nested in Region. Contrasts were set up *a priori* to maximize comparisons across the different levels of the individual terms and interactions. Model terms and interactions were assessed using both *t* values and calculated statistical significance using Satterthwaite's method for approximating the degrees of freedom using the *lmerTest* package (version 3.1-0, Kuznetsova et al., 2017).

Linear regression analyses were used to assess how well-trophic position measured from heuristic food webs predicted trophic position measured from stable isotope values. First, we extracted trophic position values from all food webs, including metawebs, using the *TrophicLevels* function in the *cheddar* package (version 0.1-633; Hudson et al., 2013). This function provides multiple estimates of trophic position for each node in the food web, including prey averaged trophic position (PATP) and chain-averaged trophic position (CATP) (Levine, 1980; Cohen et al., 2003; Williams and Martinez, 2004; Jonsson et al., 2005). We then paired these values with trophic position estimates from stable isotope values (Model 2, Equations 4 and 5) according to taxa (at the genus level) and conducted separate analyses with different predictor (i.e., PATP and CATP) and response (i.e., TP and ATP) variables. Model significance was assessed using *p*-values ($\alpha = 0.05$), and models were compared qualitatively using R^2 statistics.

RESULTS

Heuristic Food Webs

Heuristic food webs constructed from DNA and paired trait information elucidated both spatial and temporal patterns in the GLM complex. Metawebs (i.e., food webs aggregated across samples) from the wetland region of the complex were relatively larger (i.e., more nodes), denser (i.e., higher connectance), and had a higher maximum trophic position (due in part because of more predators) than metawebs from the transition and mainstem regions of the complex; metawebs from the transition region were generally the smallest and sparsest, with lower numbers of nodes, links, and maximum trophic positions compared to metawebs of the other regions (**Figure 2**). Metawebs generally constricted through time, such that they got smaller moving from early June, to early September, to mid-December; however, metawebs from the transition region of the GLM complex were the smallest, least complex food webs overall, and varied little over the study period (**Figure 2**).

Assessment of food web properties across individual heuristic food webs revealed seasonal but little spatial variation. The strongest patterns appeared to occur between June and December, but differences between other months were also apparent (**Supplementary Table S3**, **Figure 3**), with September exhibiting the most variation in food web metrics compared to the other months (**Figure 3**). Spatial patterns were weaker, with no significant Region terms for any of the metrics and Site variation explaining only a small amount of the residual variation via the random effects, but there were Season*Region interactions for the number of trophic links and chain-averaged trophic level (**Supplementary Table S3**).

Stable Isotope Analysis of Trophic Position and Community Niche Width

Bayesian estimation of trophic position using stable isotopes (Model 1, Equations 2 and 3) revealed no significant seasonal differences in maximum trophic position (i.e., the trophic position of predators) between September and December (**Supplementary Table S4**). However, when the model was run with primary consumers as the baseline instead of the food web base (i.e., leaf litter and biofilm), trophic position of predators was significantly lower in the wetland region in December compared to September (**Supplementary Table S4**). This, however, was the only significant pattern revealed by Bayesian analysis of trophic position (**Supplementary Table S4**). Because Bayesian mixing models are very different from linear mixed effects models and stable isotope data were only collected in September and December, we also qualitatively assessed maximum trophic position assessed from stable isotopes (Model 2, Equations 4 and 5) with maximum trophic position assessed from heuristic food webs using a series of reduced linear mixed effects models (**Supplementary Table S5**). These results indicated that the stable isotope approach showed differences in maximum trophic position in September compared to December, while heuristic food webs did not (**Supplementary Table S5**). Similar to Bayesian and full linear mixed effects models, these reduced

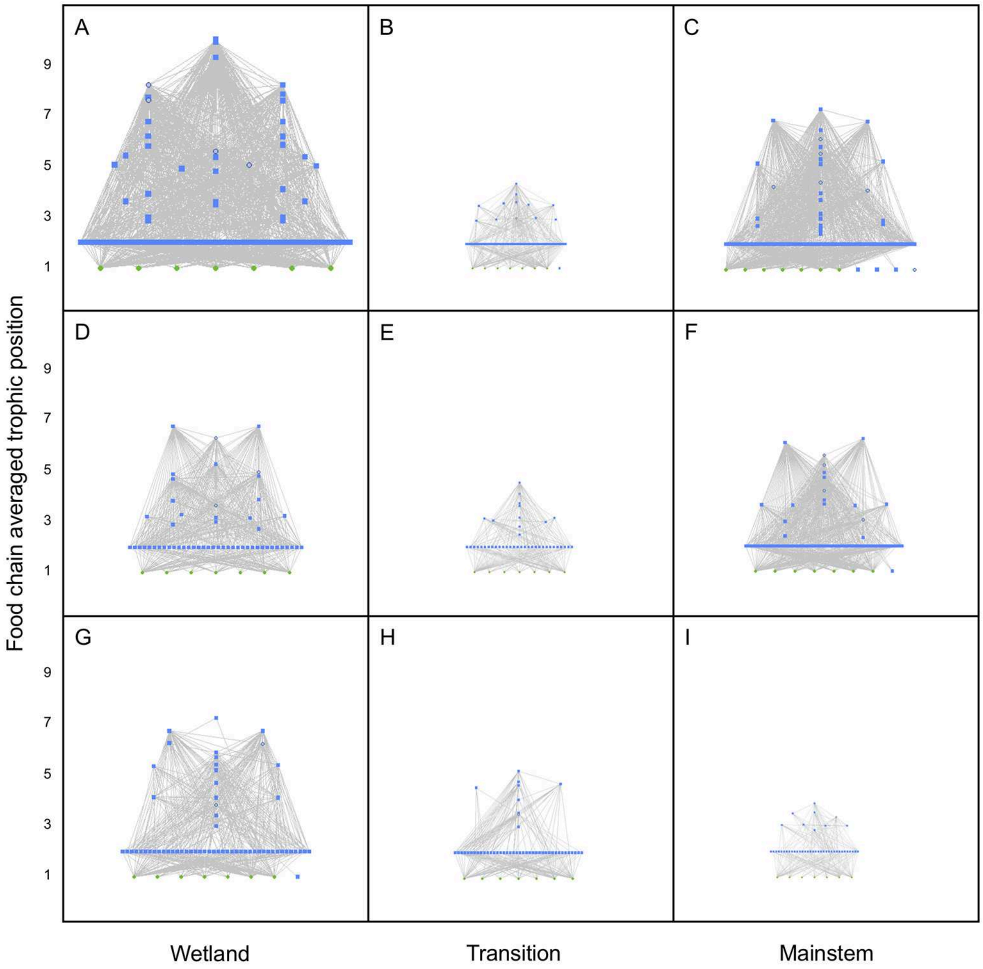


FIGURE 2 | Metawebs for the Portobello Creek wetland complex (Wetland), the Jemseg River connecting the wetland to the mainstem (Transition) and the Saint John/Wolastoq River (Mainstem) in **(A–C)** June, **(D–F)** September, and **(G–I)** December of 2016. Trophic position for each node was calculated as the food chain-averaged value for that consumer or resource. For nodes, green squares depict producers, blue squares depict consumers, and open blue circles depict cannibalistic consumers. The size of nodes and thickness of links are scaled to the maximum trophic position for each food web.

models failed to show any significant effects among regions (**Supplementary Table S5**).

Invertebrate community trophic niche widths varied little spatially and temporally (**Figure 4**). In general, trophic niche widths were slightly larger for all regions in December—when trophic positions were also higher for all regions (**Figure 4A**)—compared to September, but these differences were not strong as the 95% confidence intervals for all regions and seasons overlapped (**Figure 4B**). While the trophic niche widths did not change spatially or temporally, energy pathways did change among regions through time (**Figure 4A**). Communities in the wetland and transition regions, for example, both increased in their autochthonous reliance moving from September to December; communities in the mainstem region, however, decreased in their autochthonous reliance moving from September to December (**Figure 4A**). In general, communities in the mainstem region were fueled by autochthonous resources in September, but shifted toward a mixture of autochthonous and allochthonous resources in December, while communities in the wetland and transition regions relied on both autochthonous and allochthonous resources in September and increased in their reliance on autochthonous resources in December (**Figure 4A**).

The Relationship Between Estimated and Measured Trophic Position

Trophic position estimated from heuristic food webs was generally a strong predictor of trophic position estimated from stable isotope values (**Supplementary Table S6**). Prey-averaged trophic position (PATP) was consistently the strongest predictor of trophic position estimated from stable isotope values, exhibiting the highest R^2 values across models (seasonal models, R^2 range: 0.51–0.78; global models, R^2 range: 0.48–0.60). The strength of the relationships generally increased for predictions of adjusted trophic position (ATP) (**Supplementary Table S6**), as this variable constrained trophic position estimates based

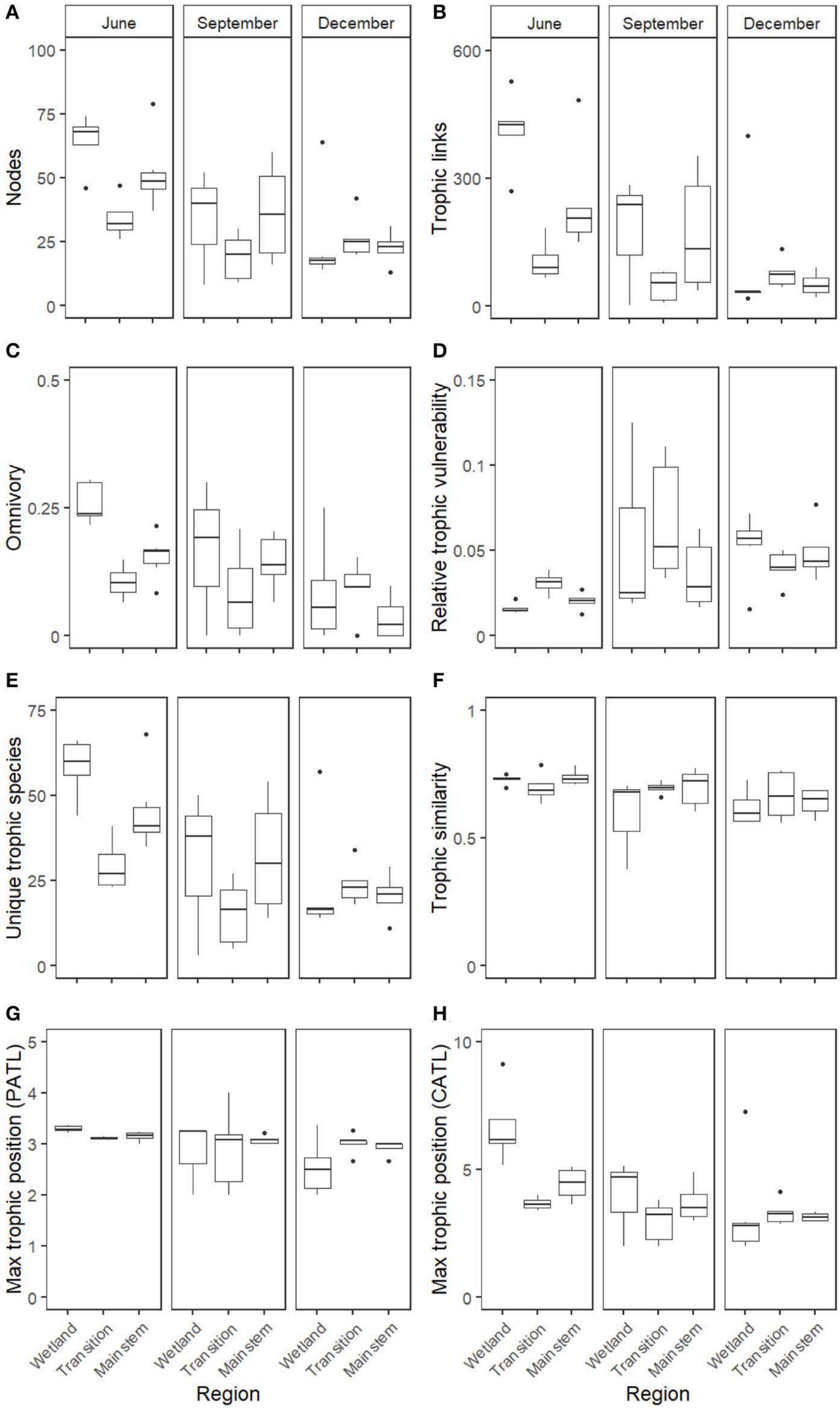


FIGURE 3 | Boxplots of the eight measured heuristic food web properties, expressed across the three regions of the GLM complex (Wetland, Transition, and Mainstem) and across the three seasons of the study (June, September, and December). Food web metrics included the following whole-network properties: **(A)** the number of nodes, **(B)** the number of trophic linkages among nodes, **(C)** the proportion of omnivory, **(D)** relative trophic vulnerability, or the average vulnerability across all nodes, standardized by the number of trophic linkages, **(E)** the number of unique trophic species, **(F)** the trophic similarity measured across all nodes, **(G)** the maximum prey-averaged trophic position, and **(H)** the maximum chain-averaged trophic position.

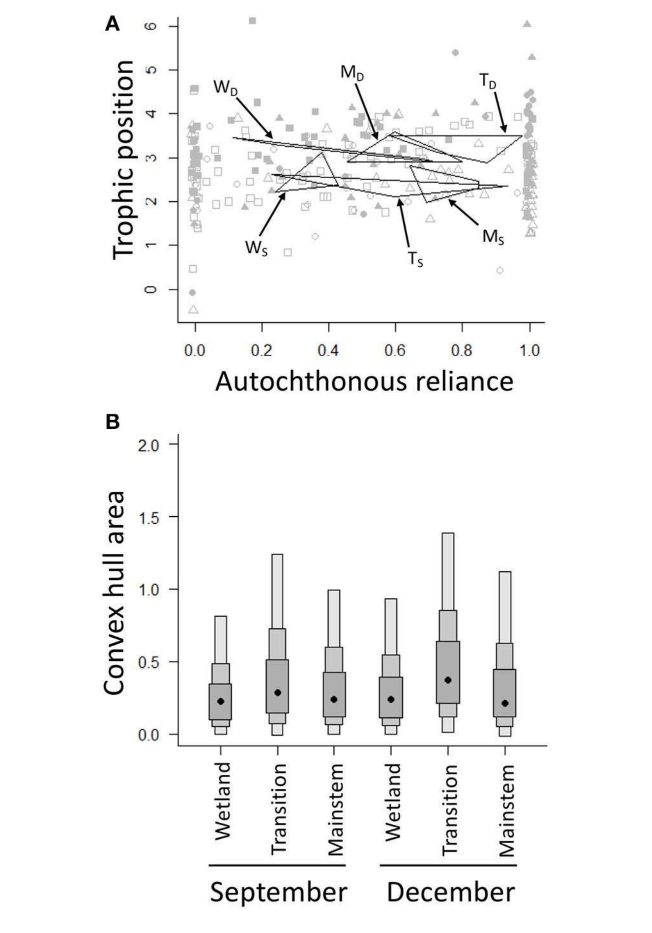


FIGURE 4 | (A) Convex hulls for the three regions of the GLM complex in September [open squares = wetland (W_S), open circles = transition (T_S), open triangles = mainstem (M_S)] and December [filled squares = wetland (W_D), filled circles = transition (T_D), filled triangles = mainstem (M_D)]. Convex hulls are based on the centers of each of the functional feeding groups that make up each community. **(B)** Density plots of the convex hull area are based on posterior estimates from Bayesian mixing models analyzed using the R package *SIBER*. Black dots depict group modes, and the shaded boxes represent the 50% (dark gray), 75% (gray), and 95% (light gray) confidence intervals.

on knowledge of the minimum possible trophic position of a consumer. The global model predicting ATP (including all paired samples across all regions and seasons) was significant for both maximum PATP ($R^2 = 0.60$, $p < 1.00$ e-15; **Figure 5A**) and maximum chain-averaged trophic position (CATP; $R^2 = 0.38$, $p < 1.00$ e-15; **Figure 5C**), and these relationships were generally stronger for specific regions at specific time periods (PATP R^2 range: 0.58–0.78; CATP R^2 range: 0.34–0.62). For these models, the deviation in predicted and measured trophic position varied based on functional feeding group. For example, the best predictor (i.e., PATP) tended to underestimate the trophic position inferred from stable isotopes for filter-feeders, but predictions closely matched for collector-gatherers, shredders, and predators, the latter which exhibited the lowest amount of variation in trophic deviation (**Figure 5B**). Trophic deviation patterns changed when CATP was used as the predictor, which tended to better predict the trophic position of functional feeding groups closer to the base of the food web and over predict the trophic position of functional feeding groups at higher trophic levels (**Figure 5D**). Finally, across all individual food webs, models that only included the maximum trophic position of each food web revealed that heuristic food web estimates significantly predicted the maximum trophic position estimated by stable isotope analysis (**Supplementary Table S6**, producer baseline models; all $p < 0.05$), but that these patterns were relatively weak (R^2 range: 0.13–0.17); these patterns, however, did not hold when consumer baselines were used in trophic position models instead of producer baselines (**Supplementary Table S6**, consumer baseline models; all $p > 0.05$).

DISCUSSION

The Predictive and Discriminatory Power of Heuristic Food Webs

We have demonstrated that heuristic food webs can provide a reliable and powerful tool for the characterization of invertebrate community structure and assessment of spatial and temporal differences among the wetland, transitional, and mainstem regions of the GLM complex. Metawebs of the three regions indicated that all food webs became less connected moving from June to September to December, and that the largest spatial differences across regions were in June and December, where the wetland trophic network was clearly larger and denser than transition and mainstem networks. Analysis of individual food webs indicated that temporal patterns were more pronounced than spatial patterns. All the food web properties we examined showed significant variation seasonally, whereas none showed significant spatial variation, though there were significant Region*Season terms for both trophic links and chain-averaged trophic position (**Supplementary Table S3**). These results support those found in another study examining DNA-based heuristic food webs in a different wetland complex, the Peace-Athabasca Delta, in northern Alberta (Compson et al., 2018). Given that extensive flooding in both the Peace-Athabasca Delta and GLM complex are regular events, connecting the wetlands to the main river channels, perhaps it is not surprising that these aquatic habitats can appear structurally homogeneous (Thomaz et al., 2007). However, the apparent contradiction between our metawebs, which showed clear structural differences among regions, and analysis of individual food webs suggests that it is more likely that spatial variability among sites within these regions was high, and this was certainly the case at all sites in September (**Figure 3**). This highlights one of the key advantages of metawebs as a visualization tool for biodiversity and community structure: while they do not convey the site-level variation within a region, because they are an aggregator of all detected biodiversity in a system, they give an overview of how these taxa are structured and interact trophically, providing scientists with a tool for making predictions about how a system might respond to perturbations (e.g., species extirpations or invasions, changes in resource availability, anthropogenic

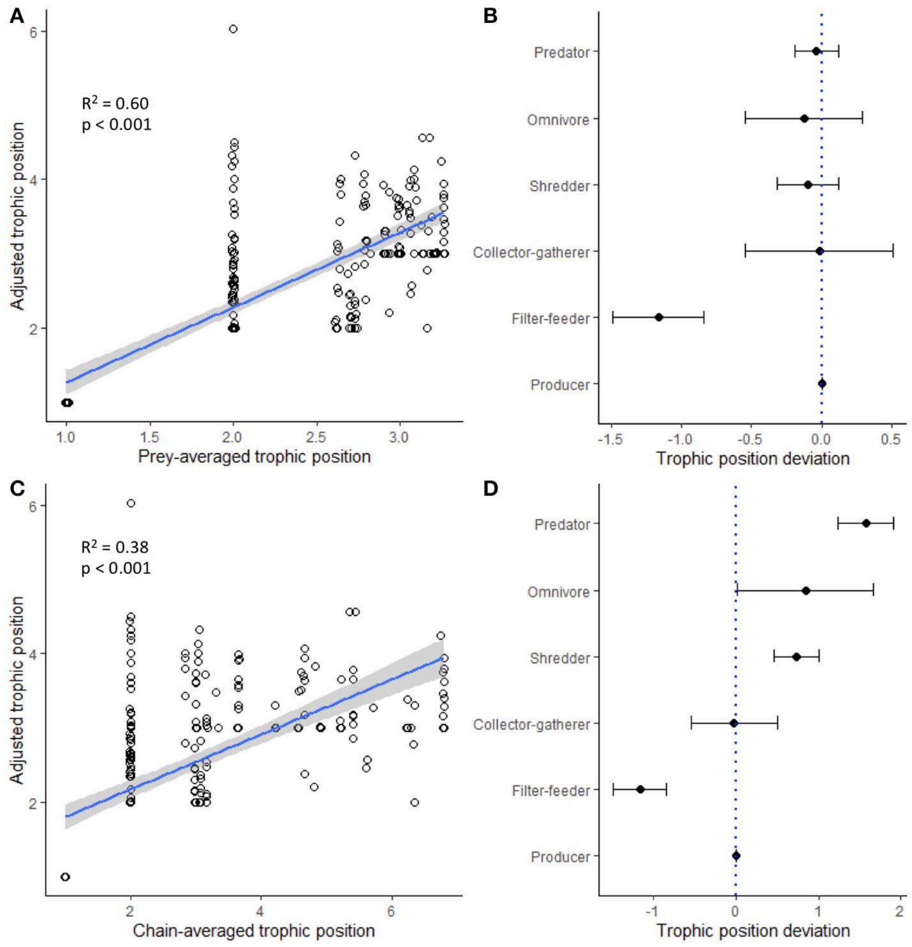


FIGURE 5 | Global linear regression models illustrating how well two heuristic food web estimates of trophic position [**(A)** prey-averaged trophic position and **(C)** chain-averaged trophic position] predicted trophic position inferred from stable isotope analysis. Points have been jittered for visualization purposes. **(B,D)** Deviation of different functional feeding groups are indicated by black dots with 95% confidence intervals; negative values indicate when heuristic food web analysis underestimated stable isotope trophic position, and positive values indicate when heuristic food web analysis overestimated stable isotope trophic position.

impacts) and how ecosystems function (reviewed in Thompson et al., 2012).

Stable isotope analysis generally confirmed the spatial and temporal patterns revealed by our heuristic food web analysis. For example, our Bayesian mixing models (Model 1, Equations 2 and 3) demonstrated that predators had a significantly higher trophic position in September compared to December at sites in the wetland region, but that there were no spatial differences in the trophic position of predators or any other consumers. Given that we only collected stable isotope samples in September and December, a more direct comparison of how well-heuristic food webs and stable isotopes resolved spatial and temporal patterns is a qualitative assessment of the reduced linear mixed effects models (**Supplementary Table S5**). Again, analysis of trophic position calculated from stable isotope values (Model 2, Equation 4 and 5) indicated that maximum trophic position only varied seasonally, and that there were no spatial differences. Interestingly, the two food web metrics we examined (i.e., prey-averaged trophic position and chain-averaged trophic position) differed neither spatially nor temporally using the reduced linear mixed effects models (i.e., using only September and December data) (**Supplementary Table S5**). Here, the discrepancy of the two approaches could have arisen because the maximum trophic position measured using stable isotopes is based on the single, highest value of all taxa examined in the community, whereas both heuristic food web metrics for maximum trophic position are integrated estimates of all the possible links to the top predator, meaning that the heuristic approach is a more integrated estimate across the entire food web. Additionally, these differences could have arisen because of our sampling design, since we targeted functional feeding groups with sufficient biomass to support stable isotope analysis. While we assessed all predators in our samples in order to get the best estimate of maximum trophic position, this approach means that in September, when we found many more predators than in December, we increased our chances of finding a single predator with a high trophic position value, potentially exacerbating differences between September and December.

This illustrates why Bayesian mixing models, which integrate variation in trophic position across all predators (or other functional feeding groups), are a more robust approach for measuring trophic position compared to point estimates from simpler mixing models. Similarly, heuristic food webs, which are created from DNA-metabarcoding data of the entire community in the sample, are likely to be a more robust estimate of maximum trophic position, as these estimates integrate all members of the community and are less biased to subsampling of larger individuals.

One of the most promising results emerging from this study was the finding that trophic position measured from heuristic food webs predicted trophic position inferred from stable isotope analysis. Significant variation was explained when measured across all paired samples (i.e., global models, R^2 range: 0.31–0.60; all models $p < 0.05$), which improved further (R^2 range: 0.34–0.78, all models $p < 0.05$) when the analysis was constrained within the metaweb of a specific region and season (**Supplementary Table S6**, **Figure 5**). One caveat of these models is that the cluster of points at the baselines has a strong leveraging effect on the linear patterns; when we examined the models with the baselines removed, patterns were generally weaker (data not shown). Further, when we explored models for individual functional feeding groups or maximum trophic position, patterns were much weaker and often not significant (**Supplementary Table S6**). Consequently, it is likely that heuristic food webs will do a better job at predicting the trophic structure of an entire community and not necessarily the specific trophic position of individual consumers or functional feeding groups. However, it is important to emphasize that deviation from these linear patterns was predictable, especially for some functional feeding groups. For example, models using prey-averaged trophic position consistently underestimated the trophic position of filter-feeders, while other groups were much more consistently predicted (**Figure 5B**). These findings suggest the possibility of calibrating heuristic food webs using stable isotope data. While it is impractical—if not impossible—to collect stable isotope data for all members of a community, our findings suggest that collecting samples from a few key functional feeding groups could allow the trophic position of some groups to be better predicted by heuristic food web analysis. Importantly, despite some groups (e.g., collector-gatherers, omnivores) exhibiting a wide range of variation in their deviation from these linear predictions, estimates of trophic position for invertebrate predators exhibited the least deviation, perhaps because they obtain biomass from many different chains in a food web.

Given that our heuristic food web and stable isotope analyses did not assess fish and other vertebrate predators feeding higher in the food web, these taxa represent important groups for future case studies linking ecological network and stable isotope approaches. Based on our findings, which indicate that heuristic food webs best predict the trophic position of invertebrate predators in models covering a wide range of functional feeding groups, we hypothesize that including vertebrate predators in these food webs will (a) improve whole-food web regression models of trophic position, and (b) lead to more accurate estimates of trophic position of these vertebrate top predators, with less trophic deviation, compared to invertebrate predators and consumers. These hypotheses are contingent upon the scale of the study, the hydrological connectivity of the system, seasonal flow dynamics, and the disbursal and trophic specialization of the taxa studied. For example, in hydrologically distinct systems where vertebrate predators are disbursal limited or have narrow trophic niches, trophic position estimates of these predators will likely be the most accurate, while they should be relatively weaker in hydrologically interconnected systems with mobile, generalist predators. A larger-scale study—covering a wider range of spatially and hydrologically distinct systems—would likely be needed to assess patterns of mobile predators like fish. However, given the hydrological interconnection of the GLM complex and the extreme seasonal dynamics of this system, the strong trophic position patterns we demonstrate for invertebrate predators shows the promise of the heuristic food web approach, even when ideal conditions are not met. If the trophic deviation patterns we demonstrate hold in different systems, heuristic food webs might live up to the promise of being a rapid indicator of both trophic structure and trophic dynamics, which would be especially useful in biodiverse systems that are difficult to study.

Merging DNA Metabarcoding and Ecological Network Analysis

Measuring food webs poses a great challenge. Constructing a food web requires the ability to sample and identify every species in a system and then to determine, or at least infer, all the trophic interactions among these species, which requires further information about species traits (reviewed in Thompson et al., 2012). These challenges illustrate why so few quality food webs have been described in the literature (Dunne et al., 2002a). Here, we have demonstrated the utility of employing a food-web generating pipeline based on DNA-derived biodiversity knowledge (Compson et al., 2018). Food webs can be generated in this manner in a fraction of the time that would otherwise be needed to quantify a trophic network, especially those as complex as wetland food webs (Halls, 1997; Millennium Ecosystem Assessment, 2005); yet, the quality and coverage of trait information available for the breadth of biodiversity in trait databases remains heterogeneous and incomplete (reviewed in Schneider et al., 2019). Our pipeline was therefore based on many assumptions about species interactions (detailed in Compson et al., 2018). Nonetheless, the generated heuristic food webs performed well as predictors of trophic position derived from stable isotope analysis, and exhibited similar spatial and temporal patterns in trophic position compared to those revealed by stable isotope analysis.

Exploring the composition of biological communities based on their DNA signature permits rapid acquisition of sequence-based occurrence data and thus orders of magnitude more taxonomic information when compared with traditional microscope-based taxonomy (Gibson et al., 2015). When this high-resolution biodiversity information is organized into ecological networks, it yields even more information on connections among organisms and how this structures the food web; understanding the variation in this connectivity can reveal complex ecological relationships (Winemiller, 1989; Dunne et al., 2002b; Poisot and Gravel, 2014). Indeed, while ecological networks have long been proposed as inexpensive tools for assessment of biostructure (McCann, 2007), the added resolution DNA-based networks provide can improve their use as a tool, and even radically change the inferences we make (Wirta et al., 2014). What is more, using DNA metabarcoding to assess communities does not always require direct observation of interactions, as gut contents, blood meal, or feces, for example, can be sequenced and interactions inferred directly from DNA metabarcoding results, circumventing the need for laborious field observations, rearing experiments, or gut content analysis (Clare et al., 2019). For this reason, genomics approaches are particularly useful for resolving difficult trophic situations, such as those involving hard to identify taxa, relationships involving cryptic species, or interactions with fluid feeders. These potential advantages of DNA-based ecological networks are opening a new frontier in ecosystem monitoring, permitting exploration of how networks change through space and time in other ecosystems and, importantly, across stronger gradients of environmental change. Our study demonstrated that strong seasonal gradients dominate in the GLM complex, but this system is relatively unimpaired and is hydrologically connected, so it is not surprising that stronger spatial patterns were not observed among regions. Further, while our study explored one important food web metric—trophic position—it is unclear how this and other food web metrics will relate to ecosystem function. Certainly, network metrics provide a promising opportunity to develop novel indicators (sensu Kissling et al., 2018) of ecosystem change[1].

Stable Isotope Analysis and DNA-Based Ecological Network Analysis: Complimentary Approaches

One of the more interesting results that emerged from this study was how the unique information from heuristic food web analysis and stable isotope analysis provided surprising, yet complementary results, illuminating the complexity of the food webs in the GLM complex. While heuristic food web analysis provided both visual and quantitative data on the relative structure, size, and complexity of the food webs in the three regions of our study, stable isotope analysis illuminated the trophic niche widths and energy pathways of communities in these regions. For example, while trophic niche widths (based on stable isotope analysis) differed neither seasonally nor spatially in our system (**Figure 4B**), metawebs (based on heuristic food web analysis) were clearly larger in the wetland region, and across all regions, became generally smaller later in the year (**Figure 2**). One of the explanations for these findings is that heuristic food webs measure all of the organisms DNA can detect in a system, whereas stable isotope analysis in our study considered only dominant taxa (i.e., taxa with enough abundance or biomass to constitute a composite isotope sample), and this could mean that while a lot of the rare or non-dominant taxa were reduced (at least below the levels of DNA detectability) later in the season, the core food web backbone (sensu Serrano et al., 2009; Lu et al., 2016) was more resilient to seasonal change in our system, an idea that is beyond the scope of this study but that warrants more attention. Future studies might be able to use network principles (e.g., the friendship paradox; Pires et al., 2017) to identify highly connected species critical to food webs prior to sampling the entire network, which could aid in project development, enabling researchers to identify key community members of a food web to sample for stable isotope analysis.

Another example of the complementary information stable isotope and heuristic food web analyses provide is the finding that—despite the lack of differences in trophic niche widths across space and time—stable isotopes revealed a shift in autochthonous reliance from September to December: the wetland and transition regions generally increased in autochthonous reliance moving later into the year, while the mainstem region decreased in autochthonous reliance (**Figure 4A**). It is possible that these differing patterns in resource use could reflect the different flow and productivity dynamics in the three regions of the GLM complex. In the highly productive wetland and transition regions, where allochthonous litter subsidies are probably exhausted or buried in sediments later in the year, leaf litter is likely less important in the winter; however, in the mainstem region, where flows are much higher and ice cover takes longer to establish, tributaries of the SJWR likely deliver a relatively high allochthonous subsidy later in the year. Consequently, while food webs were getting relatively smaller across the regions of the GLM complex throughout the year, the dominant energy pathways of the food webs changed in different regions and in different ways, indicating that seasonality, and potentially other disturbances that reduce food web size (e.g., Lu et al., 2016), could impact the structure and function of these ecosystems differently. It should be noted that while our study used aggregate samples (i.e., many individuals of a particular taxon made up an isotope sample), one advantage of stable isotope analysis compared to heuristic food web analysis is that, when a single sample is taken for each individual, it is possible to measure the variation among individuals in a population, enabling the elucidation of intraspecific energy flow pathways, especially for larger bodied consumers, like fish; in studies where a more nuanced energy flow assessment is the aim, the stable isotope and heuristic food web approaches will provide even more complementary information, with heuristic food webs providing a broad picture of how all of the organisms in a food web are connected, and stable isotope analysis elucidating specific pathways of interest.

Collectively, the complementary information gleaned from stable isotope and heuristic food web analyses may indicate important ways communities in the GLM complex function and utilize resources. Intra- and interspecific competition, ecological opportunity, and predation all govern among-individual niche variation, which likely both affects and is affected by community dynamics (Araújo et al., 2011). Because these mechanisms can be affected by seasonality in wetlands, where the flood regime and seasonal drying can exert strong pressures on organisms and communities (Costa-Pereira et al., 2017), they also likely influence community niche width, which is linked to ecosystem function (Salles et al., 2009). In the GLM complex, which is subject to late-season drying and early winter ice formation, these seasonal processes can act to both decrease ecological opportunity by reducing habitat connectivity and increase competition by reducing resource availability, two processes that would have opposite effects on community niche width. This supposition is congruent with our findings that trophic niche width differed among regions in neither September (when habitat connectivity was reduced due to late season drying) nor December (when habitat connectivity was further reduced by ice formation), and likely the reduced habitat connectivity by these events limited any increase in trophic niche width that could have arisen from increased competition. Our findings differed from those of another study in the Pantanal wetland, where habitat constriction in the dry season led to reduced niche width in a tetra fish population despite increased competition (Costa-Pereira et al., 2017). The differing patterns found in our study could be attributed to the fact that habitat constriction likely impacts the niche width of fish, which are relatively mobile, more than invertebrates, especially at the scales examined in our study.

Our results illustrate the complementary nature of DNA-based network analysis and stable isotope analysis. Stable isotope analysis provides a longer-term picture of energy flow patterns of key or dominant taxa, while DNA-based heuristic food web analysis provides a high-resolution snapshot in time of the entire community of interest. These two approaches will likely be synergistic in cases where (1) multiple pressures drive biodiversity and trophic patterns differently, (2) direct observation of trophic interactions cannot be made, (3) a community has a lot of cryptic species that are in competition or could undergo niche differentiation, (4) general energy flow pathways can be established with stable isotope mixing models, but more resolution is required to elucidate the players responsible for these patterns (e.g., DNA metabarcoding the gut contents of fish or riparian predators to better resolve aquatic-terrestrial linkages), and (5) researchers exploring heuristic food web analysis require additional evidence about interaction strengths among linkages. Of these potential synergies, the latter is probably the most challenging, especially in complex food webs, because while heuristic food webs can accommodate an unlimited number of basal food web resources, even the most sophisticated isotopic mixing model is mathematically constrained by the number of isotopic tracers in the system, which must also exhibit isotopically distinct signatures (Fry, 2006). Even in cases where food webs are very complex, however, stable isotopes could elucidate the food web backbone (sensu Serrano et al., 2009), such that dominant energy pathways of a food web are quantified. Certainly, interaction strengths among nodes of heuristic food webs could be quantified in other ways, including through added abundance or biomass information (Thompson et al., 2012), mathematical occupancy modeling with replicate DNA metabarcoding samples (Doi et al., 2019), probabilistic models of interaction (Morales-Castilla et al., 2015), or even using relative read abundances (Deagle et al., 2019). How much this added information will improve heuristic food web predictions of ecosystem structure and function remains to be seen and will likely vary based on ecosystem type and spatial and temporal scales, but this question is at the forefront of the field of ecological network analysis.

Overcoming the Limitations of DNA-Based Heuristic Food Webs as a Rapid Bioassessment Tool

Ecological network analysis has long been argued to be a tool that could provide inexpensive analysis of biostructure (McCann, 2007), and with the advent of next-generation sequencing approaches, this tool has the potential to be part of an analytical pipeline for rapid bioassessment (Gray et al., 2014; Bohan et al., 2017). At the time of writing, we were unaware of any international jurisdiction which is actively employing DNA metabarcoding for biomonitoring purposes or ecological network analysis. Heuristic food webs—which take ecological co-occurrence networks and build upon them by integrating known or measured trait information, such as information about feeding habits, species interactions, or stable isotopes—present challenges for use as a rapid bioassessment tool, despite the clear advantages they provide over simple co-occurrence networks (e.g., calculation of food web metrics, such as trophic position or relative network vulnerability).

We have identified five key advances that will overcome many of the limitations preventing widespread adoption of DNA-based heuristic food web analysis as a tool for rapid bioassessment. (1) A more widespread adoption of genomics tools is needed, particularly among groups in charge of biomonitoring programs. As standardized field sampling methods are established for environmental genomics sampling (e.g., see CABIN and National Ecological Observatory Network protocols), DNA sequencing technologies are advanced (Singer et al., 2019), genomics laboratory procedures are refined, and primers are optimized (sensu Hajibabaei et al., 2019), the cost of implementing genomics approaches will come down and public adoption should increase, but technological advancements are not often readily adopted by resource managers and policy makers (Darling and Mahon, 2011). Consequently, more needs to be done to improve biomonitoring of aquatic ecosystems by bringing stakeholders together, such as GEO BON (www.geobon.org), GEOSS (www.earthobservations.org), COST action DNAqua-Net (www.dnaqua.net), and SYNAQUA (www.interreg-francesuisse.eu) (Hering et al., 2018; Leese et al., 2018; Lefrançois et al., 2018; Pawlowski et al., 2018). (2) Bioinformatics pipelines need to be developed, reviewed (sensu Mangul et al., 2019), and made publicly available via open-source archival services, like GitHub or SourceForge, and through package managers, like Bioconda (Grüning et al., 2018). (3) Open-source databases for both genomic (e.g., BOLD, GenBank) and trait data (e.g., GloBI, EPA's Freshwater Biological Traits Database) need to be improved. Currently, the coverage of these databases is lacking, especially for understudied systems (Compson et al., 2018; Curry et al., 2018), but efforts to develop and integrate databases for ecological network analysis are underway (e.g., Poisot et al., 2016; Vissault et al., 2019). (4) We require more case studies demonstrating the utility of DNA-based network and food web analyses and the meaning of their derived network metrics. Testing these tools will be important in novel ecosystems, across extreme environmental gradients, and across large spatial and temporal scales, especially in cases where we can pair these assessments with measured estimates of ecosystem function. (5) To facilitate these efforts and to house and curate the massive amount of data next-generation biomonitoring will generate (sensu Hey and Trefethen, 2003; Bell et al., 2009), an international biomonitoring consortium needs to emerge, with federated centers for data aggregation[1]. Promisingly, advancements in any one of these areas will improve the utility and adoption of DNA-based network approaches, as progress in these areas will be linked but not necessarily limited by uneven advancement. Collective advancements made on these five fronts will enable heuristic food webs to steadily improve in their resolution, utility, and predictive power.

DATA AVAILABILITY STATEMENT

Raw data, R scripts, metadata, and supplementary material supporting the conclusions of this manuscript can be found in the project GitHub repository: https://github.com/zacchaeus-compson/Biomonitoring-with-DNA-based-food-webs. Additionally, the NCBI SRA BioProject ID is PRJNA555584, and the data will be released upon publication of the manuscript.

AUTHOR CONTRIBUTIONS

ZC, WM, BH, ZO'M, and DB conceived and designed the experiment. ZC and ZO'M conducted the field and lab work. MH and MW performed the DNA extraction, amplification, and sequencing. TP performed all the bioinformatics related to genomics data. The lab of BH processed the stable isotope samples. ZC and WM performed all the statistical analyses. ZC wrote the first draft of the manuscript, and all authors contributed to subsequent revisions.

FUNDING

DB wishes to acknowledge funding support for this work from Environment and Climate Change Canada's Strategic Application of Genomics in the Environment (STAGE) program. CB acknowledges a joint IPSNP/NSERC Engage Grant (EGP 486592-15), which provided funding for database development and an associated text mining pipeline, as well as financial support to MA from this program and from Environment & Climate Change Canada. Research support was also provided by a Natural Sciences and Engineering Research Council of Canada Collaborative Research and Development grant (NSERC CRD CRDPJ 462708-13) awarded to DB, RC, and others, Grand Lake Meadows Endowment Fund awarded to WM, ZC, and others, and New Brunswick Environmental Trust Fund and New Brunswick Wildlife Trust Fund awarded to WM, ZC, and BH. ZC, TP, and AB were funded by the Government of Canada through the Genomics Research and Development Initiative (GRDI) Ecobiomics project.

ACKNOWLEDGMENTS

We thank members of the Canadian Rivers Institute and Mactaquac Aquatic Ecosystem Study who provided field and laboratory assistance, including Colin DeCoste, Liam MacNeil, and Laura Hunt.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2019.00395/full#supplementary-material

REFERENCES

Andrews, P., and Hixson, S. (2014). Taxon-free methods of palaeoecology. *Ann. Zool. Fennici* 51, 269–285. doi: 10.5735/086.051.0225

Araújo, M. S., Bolnick, D. I., and Layman, C. A. (2011). The ecological causes of individual specialisation. *Ecol. Lett.* 14, 948–958. doi: 10.1111/j.1461-0248.2011.01662.x

Arrington, D. A., and Winemiller, K. O. (2002). Preservation effects on stable isotope analysis of fish muscle. *Trans. Am. Fish. Soc.* 131, 337–342. doi: 10.1577/1548-8659(2002)131<0337:PEOSIA>2.0.CO;2

Aylagas, E., Borja, Á., Muxika, I., and Rodríguez-Ezpeleta, N. (2018). Adapting metabarcoding-based benthic biomonitoring into routine marine ecological status assessment networks. *Ecol. Indic.* 95, 194–202. doi: 10.1016/j.ecolind.2018.07.044

Baird, D. J., and Hajibabaei, M. (2012). Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Mol. Ecol.* 21, 2039–2044. doi: 10.1111/j.1365-294X.2012.05519.x

Barrow, L. M., Bjorndal, K. A., and Reich, K. J. (2008). Effects of preservation method on stable carbon and nitrogen isotope values. *Physio. Biochem. Zool.* 81, 688–693. doi: 10.1086/588172

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., et al. (2019). *Package 'lme4'. Linear Mixed-Effects Models Using S4 Classes*. R package version, 1.1-21. Available online at: https://github.com/lme4/lme4/

Bell, G., Hey, T., and Szalay, A. (2009). Beyond the data deluge. *Science* 323, 1297–1298. doi: 10.1126/science.1170411

Birkhofer, K., Bylund, H., Dalin, P., Ferlian, O., Gagic, V., Hambäck, P. A., et al. (2017). Methods to identify the prey of invertebrate predators in terrestrial field studies. *Ecol. Evol.* 7, 1942–1953. doi: 10.1002/ece3.2791

Bohan, D. A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A. J., and Woodward, G. (2017). Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. *Trends Ecol. Evol.* 32, 477–487. doi: 10.1016/j.tree.2017.03.001

Brose, U., Cushing, L., Berlow, E. L., Jonsson, T., Banasek-Richter, C., Bersier, L. F., et al. (2005). Body sizes of consumers and their resources. *Ecology* 86, 2545–2545. doi: 10.1890/05-0379

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi: 10.1038/ismej.2017.119

Cazzolla Gatti, R. (2016). Freshwater biodiversity: a review of local and global threats. *Int. J. Environ. Stud.* 73, 887–904. doi: 10.1080/00207233.2016.1204133

Clare, E. L., Fazekas, A. J., Ivanova, N. V., Floyd, R. M., Hebert, P. D., Adams, A. M., et al. (2019). Approaches to integrating genetic data into ecological networks. *Mol. Ecol.* 28, 503–519. doi: 10.1111/mec.14941

Cohen, J. E., Jonsson, T., and Carpenter, S. R. (2003). Ecological community description using the food web, species abundance, and body size. *Proc. Nat. Acad. Sci. U.S.A.* 100, 1781–1786. doi: 10.1073/pnas.232715699

Compson, Z. G., Monk, W. A., Curry, C. J., Gravel, D., Bush, A., Baker, C. J., et al. (2018). Linking DNA metabarcoding and text mining to create network-based biomonitoring tools: a case study on boreal wetland macroinvertebrate communities. *Adv. Ecol. Res.* 59, 33–74 doi: 10.1016/bs.aecr.2018.09.001

Costa-Pereira, R., Tavares, L. E., de Camargo, P. B., and Araújo, M. S. (2017). Seasonal population and individual niche dynamics in a tetra fish in the Pantanal wetlands. *Biotropica* 49, 531–538. doi: 10.1111/btp.12434

Curry, C. J., Gibson, J. F., Shokralla, S., Hajibabaei, M., and Baird, D. J. (2018). Identifying North American freshwater invertebrates using DNA barcodes: are existing COI sequence libraries fit for purpose? *Freshw. Sci.* 37, 178–189. doi: 10.1086/696613

Damuth, J. D., Jablonski, D., Harris, J. A., Potts, R., Stucky, R. K., Sues, H. D., et al. (1992). "Taxon-free characterization of animal communities," in *Terrestrial Ecosystems Through Time: Evolutionary Paleoecology of Terrestrial Plants and Animals*, eds A. K. Behrensmeyer, J. D. Damuth, W. A. DiMichele, R. Potts, H. Sues, and S. L. Wing (Chicago, IL: University of Chicago Press), 183–203.

Darling, J. A., and Mahon, A. R. (2011). From molecules to management: adopting DNA-based methods for monitoring biological invasions in aquatic environments. *Env. Res.* 111, 978–988. doi: 10.1016/j.envres.2011.02.001

Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., et al. (2019). Counting with DNA in metabarcoding studies: how should we convert sequence reads to dietary data? *Mol. Ecol.* 28, 391–406. doi: 10.1111/mec.14734

Derocles, S. A., Bohan, D. A., Dumbrell, A. J., Kitson, J. J., Massol, F., Pauvert, C., et al. (2018). Biomonitoring for the 21st century: integrating next-generation sequencing into ecological network analysis. *Adv. Ecol. Res.* 58, 1–62. doi: 10.1016/bs.aecr.2017.12.001

Didham, R. K., Leather, and, S. R., and Basset, Y. (2016). Circle the bandwagons–challenges mount against the theoretical foundations of applied functional trait and ecosystem service research. *Insect Conserv. Divers.* 9, 1–3. doi: 10.1111/icad.12150

Dixon, M. J. R., Loh, J., Davidson, N. C., Beltrame, C., Freeman, R., and Walpole, M. (2016). Tracking global change in ecosystem area: the wetland extent trends index. *Biol. Conserv.* 193, 27–35. doi: 10.1016/j.biocon.2015.10.023

Doi, H., Fukaya, K., Oka, S. I., Sato, K., Kondoh, M., and Miya, M. (2019). Evaluation of detection probabilities at the water-filtering and initial PCR steps in environmental DNA metabarcoding using a multispecies site occupancy model. *Sci. Rep.* 9:3581. doi: 10.1038/s41598-019-40233-1

Doledec, S., and Statzner, B. (2008). Invertebrate traits for the biomonitoring of large European rivers: an assessment of specific types of human impact. *Freshw. Biol.* 53, 617–634. doi: 10.1111/j.1365-2427.2007.01924.x

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z. I., Knowler, D. J., Lévêque, C., et al. (2006). Freshwater biodiversity: importance, threats, status and conservation challenges. *Biol. Rev.* 81, 163–182. doi: 10.1017/S1464793105006950

Dunne, J. A., Williams, R. J., and Martinez, N. D. (2002a). Food-web structure and network theory: the role of connectance and size. *Proc. Nat. Acad. Sci. U.S.A.* 99, 12917–12922. doi: 10.1073/pnas.192407699

Dunne, J. A., Williams, R. J., and Martinez, N. D. (2002b). Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecol. Lett.* 5, 558–567. doi: 10.1046/j.1461-0248.2002.00354.x

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv [Preprint]*. doi: 10.1101/081257

Emilson, C. E., Thompson, D. G., Venier, L. A., Porter, T. M., Swystun, T., Chartrand, D., et al. (2017). DNA metabarcoding and morphological macroinvertebrate metrics reveal the same changes in boreal watersheds across an environmental gradient. *Sci. Rep.* 7:12777. doi: 10.1038/s41598-017-13157-x

Estrada, E. (2007). Food webs robustness to biodiversity loss: the roles of connectance, expansibility and degree distribution. *J. Theor. Biol.* 244, 296–307. doi: 10.1016/j.jtbi.2006.08.002

Fry, B. (2006). *Stable Isotope Ecology*. New York, NY: Springer.

Gibson, J. F., Shokralla, S., Curry, C., Baird, D. J., Monk, W. A., King, I., et al. (2015). Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. *PLoS ONE* 10:e0138432. doi: 10.1371/journal.pone.0138432

Gilbert, A. J. (2009). Connectance indicates the robustness of food webs when subjected to species loss. *Ecol. Indic.* 9, 72–80. doi: 10.1016/j.ecolind.2008.01.010

Gray, C., Baird, D. J., Baumgartner, S., Jacob, U., Jenkins, G. B., O'Gorman, E. J., et al. (2014). Ecological networks: the missing links in biomonitoring science. *J. Appl. Ecol.* 51, 1444–1449. doi: 10.1111/1365-2664.12300

Gray, C., Figueroa, D. H., Hudson, L. N., Ma, A., Perkins, D., and Woodward, G. (2015). Joining the dots: an automated method for constructing food webs from compendia of published interactions. *Food Webs* 5, 11–20. doi: 10.1016/j.fooweb.2015.09.001

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., et al. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* 15, 475–476. doi: 10.1038/s41592-018-0046-7

Hajibabaei, M., Porter, T. M., Wright, M., and Rudar, J. (2019). COI metabarcoding primer choice affects richness and recovery of indicator taxa in freshwater systems. *PLoS ONE* 14:e0220953. doi: 10.1371/journal.pone.0220953

Halls, A. (1997). *Wetlands, Biodiversity and the Ramsar Convention: The Role of the Convention on Wetlands in the Conservation and Wise Use of Biodiversity*. Gland: Ramsar Convention Bureau.

Hering, D., Borja, A., Jones, J. I., Pont, D., Boets, P., Bouchez, A., et al. (2018). Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. *Water Res.* 138, 192–205. doi: 10.1016/j.watres.2018.03.003

Hey, A. J., and Trefethen, A. E. (2003). "The data deluge: an e-science perspective," in *Grid Computing: Making the Global Infrastructure a Reality*, eds F. Berman, G. Fox, and T. Hey (West Sussex, England: John Wiley & Sons Ltd.), 809–824. doi: 10.1002/0470867167.ch36

Hu, S., Niu, Z., Chen, Y., Li, L., and Zhang, H. (2017). Global wetlands: potential distribution, wetland loss, and status. *Sci. Total Environ.* 586, 319–327. doi: 10.1016/j.scitotenv.2017.02.001

Hudson, L. N., Emerson, R., Jenkins, G. B., Layer, K., Ledger, M. E., Pichler, D. E., et al. (2013). Cheddar: analysis and visualisation of ecological communities in R. *Methods Ecol. Evol.* 4, 99–104. doi: 10.1111/2041-210X.12005

Jackson, A. L., Inger, R., Parnell, A. C., and Bearhop, S. (2011). Comparing isotopic niche widths among and within communities: SIBER–Stable Isotope Bayesian Ellipses in R. *J. Anim. Ecol.* 80, 595–602. doi: 10.1111/j.1365-2656.2011.01806.x

Jonsson, T., Cohen, J. E., and Carpenter, S. R. (2005). Food webs, body size, and species abundance in ecological community description. *Adv. Ecol. Res.* 36, 1–84. doi: 10.1016/S0065-2504(05)36001-6

Kartzinel, T. R., Chen, P. A., Coverdale, T. C., Erickson, D. L., Kress, W. J., Kuzmina, M. L., et al. (2015). DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proc. Nat. Acad. Sci. U.S.A.* 112, 8019–8024. doi: 10.1073/pnas.1503283112

Kissling, W. D., Walls, R., Bowser, A., Jones, M. O., Kattge, J., Agosti, D., et al. (2018). Towards global data products of essential biodiversity variables on species traits. *Nat. Ecol. Evol.* 2, 1531–1540. doi: 10.1038/s41559-018-0667-3

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13

Layman, C. A., Arrington, D. A., Montaña, C. G., and Post, D. M. (2007). Can stable isotope ratios provide for community-wide measures of trophic structure? *Ecology* 88, 42–48. doi: 10.1890/0012-9658(2007)88[42:CSIRPF]2.0.CO;2

Leese, F., Bouchez, A., Abarenkov, K., Altermatt, F., Borja, Á., Bruce, K., et al. (2018). Why we need sustainable networks bridging countries, disciplines, cultures and generations for aquatic biomonitoring 2.0: a perspective derived from the DNAqua-Net COST action. *Adv. Ecol. Res.* 58, 63–99. doi: 10.1016/bs.aecr.2018.01.001

Lefrançois, E., Apothéloz-Perret-Gentil, L., Blancher, P., Botreau, S., Chardon, C., Crepin, L., et al. (2018). Development and implementation of eco-genomic tools for aquatic ecosystem biomonitoring: the SYNAQUA French-Swiss program. *Environ. Sci. Pollut. Res.* 25, 33858–33866. doi: 10.1007/s11356-018-2172-2

Lerner, J. E., Ono, K., Hernandez, K. M., Runstadler, J. A., Puryear, W. B., and Polito, M. J. (2018). Evaluating the use of stable isotope analysis to infer the feeding ecology of a growing US gray seal (*Halichoerus grypus*) population. *PLoS ONE* 13:e0192241. doi: 10.1371/journal.pone.0192241

Levine, S. (1980). Several measures of trophic structure applicable to complex food webs. *J. Theor. Biol.* 83, 195–207. doi: 10.1016/0022-5193(80)90288-X

Liu, T., Guo, R., Ran, W., Whalen, J. K., and Li, H. (2015). Body size is a sensitive trait-based indicator of soil nematode community response to fertilization in rice and wheat agroecosystems. *Soil Biol. Biochem.* 88, 275–281. doi: 10.1016/j.soilbio.2015.05.027

Lu, X., Gray, C., Brown, L. E., Ledger, M. E., Milner, A. M., Mondragón, R. J., et al. (2016). Drought rewires the cores of food webs. *Nat. Clim. Change* 6:875. doi: 10.1038/nclimate3002

Mangul, S., Martin, L. S., Eskin, E., and Blekhman, R. (2019). Improving the usability and archival stability of bioinformatics software. *Genome Biol.* 20:47. doi: 10.1186/s13059-019-1649-8

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12. doi: 10.14806/ej.17.1.200

May, R. M. (1972). Will a large complex system be stable? *Nature* 238:413. doi: 10.1038/238413a0

McCann, K. (2007). Protecting biostructure. *Nature* 446:29. doi: 10.1038/446029a

McCutchan, J. H. Jr., Lewis, W. M. Jr., Kendall, C., and McGrath, C. C. (2003). Variation in trophic shift for stable isotope ratios of carbon, nitrogen, and sulfur. *Oikos* 102, 378–390. doi: 10.1034/j.1600-0706.2003.12098.x

McGill, B. J. (2015). *Steering the Trait Bandwagon*. Dynamic Ecology. Available online at: https://dynamicecology.wordpress.com/2015/07/01/steering-the-trait-bandwagon/

Merritt, R. W., Cummins, K. W., and Berg, M. B. (2008). *An Introduction to the Aquatic Insects of North America, 4th Edn*. Dubuque, IA: Kendall-Hunt.

Millennium Ecosystem Assessment (2005). *Ecosystems and Human Well Being: Wetlands and Water Synthesis.* Millennium Ecosystem Assessment Series. Washington, DC: World Resources Institute.

Morales-Castilla, I., Matias, M. G., Gravel, D., and Araujo, M. B. (2015). Inferring biotic interactions from proxies. *Trends Ecol. Evol.* 30, 347–356. doi: 10.1016/j.tree.2015.03.014

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., et al. (2018). The future of biotic indices in the ecogenomic era: integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Sci. Total Environ.* 637, 1295–1310. doi: 10.1016/j.scitotenv.2018.05.002

Pires, M. M., Marquitti, F. M., and Guimarães, P. R. Jr. (2017). The friendship paradox in species-rich ecological networks: implications for conservation and monitoring. *Biol. Conserv.* 209, 245–252. doi: 10.1016/j.biocon.2017.02.026

Plummer, M. (2003). "JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling," in *Proceedings of the 3rd International Workshop Distributed Statistical Computing* (Vienna: Technische Universität Wien), 1–8.

Poff, N. L. (1997). Landscape filters and species traits: towards mechanistic understanding and prediction in stream ecology. *J. N. Am. Benthol. Soc.* 16, 391–409. doi: 10.2307/1468026

Poisot, T., Baiser, B., Dunne, J. A., Kéfi, S., Massol, F., Mouquet, N., et al. (2016). Mangal–making ecological network analysis simple. *Ecography* 39, 384–390. doi: 10.1111/ecog.00976

Poisot, T., and Gravel, D. (2014). When is an ecological network complex? Connectance drives degree distribution and emerging network properties. *PeerJ* 2:e251. doi: 10.7717/peerj.251

Porter, T. M., and Hajibabaei, M. (2018). Automated high throughput animal CO1 metabarcode classification. *Sci. Rep.* 8:4226. doi: 10.1038/s41598-018-22505-4

Post, D. M. (2002). Using stable isotopes to estimate trophic position: models, methods, and assumptions. *Ecol.* 83, 703–718. doi: 10.1890/0012-9658(2002)083[0703:USITET]2.0.CO;2

Quezada-Romegialli, C., Jackson, A. L., Hayden, B., Kahilainen, K. K., Lopes, C., and Harrod, C. (2018). tRophicPosition, an R package for the Bayesian estimation of trophic position from consumer stableisotope ratios. *Methods Ecol. Evol.* 9, 1592–1599. doi: 10.1111/2041-210X.13009

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Rodell, M., Famiglietti, J. S., Wiese, D. N., Reager, J. T., Beaudoing, H. K., Landerer, F. W., et al. (2018). Emerging trends in global freshwater availability. *Nature* 557, 651–659. doi: 10.1038/s41586-018-0123-1

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584. doi: 10.7717/peerj.2584

Salles, J. F., Poly, F., Schmid, B., and Roux, X. L. (2009). Community niche predicts the functioning of denitrifying bacterial assemblages. *Ecology* 90, 3324–3332. doi: 10.1890/09-0188.1

Schneider, F. D., Fichtmuller, D., Gossner, M. M., Güntsch, A., Jochum, M., König-Ries, B., et al. (2019). Towards an ecological trait-data standard. *Methods Ecol. Evol.* doi: 10.1111/2041-210X.13288. [Epub ahead of print].

Serrano, M. Á., Boguná, M., and Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proc. Nat. Acad. Sci. U.S.A.* 106, 6483–6488. doi: 10.1073/pnas.0808904106

Singer, G. A. C., Fahner, N. A., Barnes, J. G., McCarthy, A., and Hajibabaei, M. (2019). Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. *Sci. Rep.* 9:5991. doi: 10.1038/s41598-019-42455-9

St. John, J. (2016). *SeqPrep*. Available online at: https://github.com/jstjohn/SeqPrep/releases

Thomaz, S. M., Bini, L. M., and Bozelli, R. L. (2007). Floods increase similarity among aquatic habitats in river-floodplain systems. *Hydrobiologia* 579, 1–13. doi: 10.1007/s10750-006-0285-y

Thompson, R. M., Brose, U., Dunne, J. A., Hall, R. O. Jr., Hladyz, S., Kitching, R. L., et al. (2012). Food webs: reconciling the structure and function of biodiversity. *Trends Ecol. Evol.* 27, 689–697. doi: 10.1016/j.tree.2012.08.005

Vázquez, D. P. (2005). Degree distribution in plant–animal mutualistic networks: forbidden links or random interactions? *Oikos* 108, 421–426. doi: 10.1111/j.0030-1299.2005.13619.x

Vissault, S., Gravel, D., and Poisot, T. (2019). Mangal: an open infrastructure for ecological interactions. *Biodivers. Info. Sci. Stand.* 3:e37037. doi: 10.3897/biss.3.37037

Williams, R. J. (2011). Biology, methodology or chance? The degree distributions of bipartite ecological networks. *PLoS ONE* 6:e17645. doi: 10.1371/journal.pone.0017645

Williams, R. J., and Martinez, N. D. (2004). Limits to trophic levels and omnivory in complex food webs: theory and data. *Am. Nat.* 163, 458–468. doi: 10.1086/381964

Winemiller, K. O. (1989). Must connectance decrease with species richness? *Am. Nat.* 134, 960–968. doi: 10.1086/285024

Wirta, H. K., Hebert, P. D., Kaartinen, R., Prosser, S. W., Várkonyi, G., and Roslin, T. (2014). Complementary molecular information changes our perception of food web structure. *Proc. Nat. Acad. Sci. U.S.A.* 111, 1885–1890. doi: 10.1073/pnas.1316990111

Conflict of Interest: MA employed by company IPSNP Computing Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Environmental Science

REVIEW
published: 09 January 2020
doi: 10.3389/fenvs.2019.00197

Key Questions for Next-Generation Biomonitoring

Andreas Makiola[1], Zacchaeus G. Compson[2,3], Donald J. Baird[2], Matthew A. Barnes[4], Sam P. Boerlijst[5,6], Agnès Bouchez[7], Georgina Brennan[8], Alex Bush[2,9], Elsa Canard[10], Tristan Cordier[11], Simon Creer[8], R. Allen Curry[12], Patrice David[13], Alex J. Dumbrell[14], Dominique Gravel[15], Mehrdad Hajibabaei[16], Brian Hayden[17], Berry van der Hoorn[5], Philippe Jarne[13], J. Iwan Jones[18], Battle Karimi[1], Francois Keck[7], Martyn Kelly[19], Ineke E. Knot[20], Louie Krol[5,6], Francois Massol[21,22], Wendy A. Monk[2,23], John Murphy[18], Jan Pawlowski[11], Timothée Poisot[24], Teresita M. Porter[16,25], Kate C. Randall[14], Emma Ransome[26], Virginie Ravigné[27], Alan Raybould[28,29,30], Stephane Robin[31], Maarten Schrama[5,6], Bertrand Schatz[13], Alireza Tamaddoni-Nezhad[32], Krijn B. Trimbos[6], Corinne Vacher[33], Valentin Vasselon[34], Susie Wood[35], Guy Woodward[26] and David A. Bohan[1]*

[1] Agroécologie, AgroSup Dijon, INRA, Université Bourgogne, Université Bourgogne Franche-Comté, Dijon, France, [2] Environment and Climate Change Canada @ Canadian Rivers Institute, Department of Biology, University of New Brunswick, NB, Canada, [3] Centre for Environmental Genomics Applications, St. John's, NL, Canada, [4] Department of Natural Resources Management, Texas Tech University, Lubbock, TX, United States, [5] Naturalis Biodiversity Center, Leiden, Netherlands, [6] Institute of Environmental Sciences, Leiden University, Leiden, Netherlands, [7] CARRTEL, USMB, INRA, Thonon-les-Bains, France, [8] School of Natural Sciences, Bangor University, Bangor, United Kingdom, [9] Lancaster Environment Centre, Lancaster University, Lancaster, United Kingdom, [10] UMR 1349 IGEPP, INRA, Université de Rennes 1, Agrocampus Ouest Rennes, Domaine de la Motte, Le Rheu, France, [11] Department of Genetics and Evolution, University of Geneva, Science III, Geneva, Switzerland, [12] Canadian Rivers Institute, Biology, Forestry and Environmental Management, University of New Brunswick, Fredericton, NB, Canada, [13] CEFE UMR 5175, CNRS—Université de Montpellier - Université Paul-Valéry Montpellier–IRD—EPHE, Montpellier, France, [14] School of Biological Sciences, University of Essex, Colchester, United Kingdom, [15] Département de Biologie, Université de Sherbrooke, Sherbrooke, Canada, [16] Centre for Biodiversity Genomics and Department of Integrative Biology, University of Guelph, Guelph, ON, Canada, [17] Stable Isotopes in Nature Laboratory, Canadian Rivers Institute, University of New Brunswick, Fredericton, NB, Canada, [18] School of Biological and Chemical Sciences, Queen Mary University of London, London, United Kingdom, [19] Bowburn Consultancy, Durham, United Kingdom, [20] Institute of Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, Netherlands, [21] Univ. Lille, CNRS, UMR 8198—Evo-Eco-Paleo, SPICI Group, Lille, France, [22] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, U1019–UMR 8204–CIIL–Center for Infection and Immunity of Lille, Lille, France, [23] Faculty of Forestry and Environmental Management, University of New Brunswick, Fredericton, NB, Canada, [24] Département de sciences biologiques, Université de Montréal, Montreal, QC, Canada, [25] Great Lakes Forestry Centre, Natural Resources Canada, Sault Ste. Marie, ON, Canada, [26] Department of Life Sciences, Silwood Park Campus, Imperial College London, London, United Kingdom, [27] CIRAD, UMR PVBMT, Saint-Pierre, France, [28] Syngenta Crop Protection AG, Basel, Switzerland, [29] School of Social and Political Science, The University of Edinburgh, Edinburgh, United Kingdom, [30] The Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Global Academy of Agriculture and Food Security, Edinburgh, United Kingdom, [31] Laboratoire MMIP—UMR INRA 518/AgroParisTech, Paris, France, [32] Department of Computer Science, University of Surrey, Guildford, United Kingdom, [33] BIOGECO, INRA, Univ. Bordeaux, Pessac, France, [34] Agence Française pour la Biodiversité, Pôle R&D ECLA, Évian-les-Bains, France, [35] Cawthron Institute, Nelson, New Zealand

OPEN ACCESS

Edited by:
Marco Casazza,
Università degli Studi di Napoli
Parthenope, Italy

Reviewed by:
Erin Grey,
Governors State University,
United States
Shuisen Chen,
Guangzhou Institute of
Geography, China

***Correspondence:**
David A. Bohan
david.bohan@inra.fr

Specialty section:
This article was submitted to
Environmental Informatics,
a section of the journal
Frontiers in Environmental Science

Received: 06 June 2019
Accepted: 04 December 2019
Published: 09 January 2020

Citation:
Makiola A, Compson ZG, Baird DJ, Barnes MA, Boerlijst SP, Bouchez A, Brennan G, Bush A, Canard E, Cordier T, Creer S, Curry RA, David P, Dumbrell AJ, Gravel D, Hajibabaei M, Hayden B, van der Hoorn B, Jarne P, Jones JI, Karimi B, Keck F, Kelly M, Knot IE, Krol L, Massol F, Monk WA, Murphy J, Pawlowski J, Poisot T, Porter TM, Randall KC, Ransome E, Ravigné V, Raybould A, Robin S, Schrama M, Schatz B, Tamaddoni-Nezhad A, Trimbos KB, Vacher C, Vasselon V, Wood S, Woodward G and Bohan DA (2020) Key Questions for Next-Generation Biomonitoring. Front. Environ. Sci. 7:197. doi: 10.3389/fenvs.2019.00197

Classical biomonitoring techniques have focused primarily on measures linked to various biodiversity metrics and indicator species. Next-generation biomonitoring (NGB) describes a suite of tools and approaches that allow the examination of a broader spectrum of organizational levels—from genes to entire ecosystems. Here, we frame 10 key questions that we envisage will drive the field of NGB over the next decade. While not exhaustive, this list covers most of the key challenges facing NGB, and provides the basis of the next steps for research and implementation in this field. These questions have been grouped into current- and outlook-related categories, corresponding to the organization of this paper.

Keywords: eDNA, metabarcoding, biodiversity assessment, artificial intelligence, ecological networks

INTRODUCTION

Classical biomonitoring techniques (**Table 1**) have focused primarily on measures linked to various biodiversity metrics (e.g., species richness, beta diversity; Li et al., 2010; Gutiérrez-Cánovas et al., 2019) and indicator species (but see Vandewalle et al., 2010; Culhane et al., 2014; Saito et al., 2015 for other approaches). Next-generation biomonitoring (NGB) describes a suite of tools and approaches that allow the examination of a broader spectrum of organizational levels—from genes to entire ecosystems. A more holistic vision of evaluating ecological structure and change has long been a goal of ecology, but only recently have the tools emerged to bring it toward fruition. In this issue of Frontiers in Ecology & Evolution, which explores the research topic, "A Next Generation of Biomonitoring to Detect Global Ecosystem Change," we explore this complementary suite of new tools that could be forged into a global approach to biomonitoring. In this overview paper, we attempt to synthesize opinion on the key issues that are necessary to address en route to this next generation of biomonitoring tools. We focus on a key subset of these tools—those based on DNA metabarcoding as a new standard methodology for multiple taxonomic identifications—for which the number of papers published has increased exponentially since 2010 (**Figure 1**).

DNA metabarcoding generates massive amounts of data on taxonomic units (e.g., operational taxonomic units, OTUs, or exact sequence variants, ESVs; Callahan et al., 2017) rapidly, and these can be linked increasingly to functional attributes (Douglas et al., 2018; Makiola et al., 2019). DNA metabarcoding is highly complementary to whole metagenomic and metatranscriptomic sequencing (Knight et al., 2018), existing sources of ecological information (Cordier et al., 2018; Derocles et al., 2018) and classical biomonitoring approaches (Deiner et al., 2017); in all cases, adding genomic and/or ecological information to the rich taxonomic lists afforded by DNA metabarcoding would allow deeper exploration of ecological or biodiversity patterns. This would move biomonitoring closer to being able to extract both structural and functional attributes from the same multispecies sample (Keck et al., 2017; Cordier et al., 2019). By merging DNA metabarcoding with ecological information and machine learning approaches, NGB extends modern analytical potential beyond the classical morphological identification of bioindicator species. For instance, taxonomic lists from DNA metabarcodes can identify anthropogenic drivers behind community change and infer networks of possible ecological interactions and associated ecosystem properties (Bohan et al., 2017; Compson et al., 2018). While challenges to constructing these networks from NGB data remain (e.g., Barner et al., 2018; Freilich et al., 2018; Deagle et al., 2019), this overview paper discusses some promising ways of overcoming these limitations, including using trait filters developed from published literature and methods of inferring interactions (e.g., machine learning), and these ideas are developed in more depth in the associated manuscripts of this special issue. Indeed, the ultimate aim of NGB is to deliver this more integrated view of natural ecosystems at a fraction of the time and cost of classical approaches (Baird and Hajibabaei, 2012; Keck et al., 2017; Leese et al., 2018; Cordier et al., 2019). Building this large-scale monitoring poses many challenging questions, from the practical and logistical to the political and philosophical.

Here, we frame and describe the interplay of ten key questions that we envisage will drive the field over the next decade (**Figure 2**). Questions 1–7 address issues that are of current importance, and pertain to the scope of NGB. Questions 8–10 are questions of outlook and opportunity, exploring where the field might be going. This list emerged as an overview of the current Frontiers special issue on the research topic: "A Next Generation of Biomonitoring to Detect Global Ecosystem Change." While not exhaustive it covers most of the key challenges facing NGB, and provides the basis of the next steps for research and implementation in this field.

Current Questions

How Can the Benefits of NGB Be Most Successfully Communicated to Citizens, Scientists, and Policymakers?

Managing issues of human health, food production and security, and the intertwined environmental issues of biodiversity and ecosystem services necessitates biomonitoring (Bush et al., 2019a; Schmidt-Traub et al., 2019). Information about the status of these issues, such as changes in the frequency of human (Jones et al., 2008) and crop diseases (Savary et al., 2019), insect declines (Hallmann et al., 2017), and losses of species of flowering plants (Carvell et al., 2006) are expected to lead to profound changes in human behavior and appreciation of the environment (Schröter et al., 2017). However, the vision of a broader scale evaluation of ecosystem change, and the benefits this will bring to citizens, scientists, and policymakers, needs to be clearly communicated if wide adoption of NGB approaches is to be realized.

There are three clear benefits of NGB. First, as is argued across the papers of this Issue, NGB has the potential to provide a more holistic method of assessment than classical biomonitoring, affording improved decision-making and management of issues that affect citizens' quality of life. Second, while NGB will provide methods for detailing the complexity of ecosystems, it will also use methods,

TABLE 1 | Glossary of terms as used in this paper.

Term	Definition
Bioindicator	An organism used as an indicator of the qualitative status of the environment or an ecosystem
Classical biomonitoring	The methodologies of observing and assessing the state and ongoing changes in ecosystems, components of biodiversity, and landscape, including the types of natural habitats, populations and species
Community science	Public participation in scientific research (citizen science)
DNA barcoding	A method of taxonomic identification using a section of DNA from a specific gene or genes (genetic marker)
DNA metabarcoding	A method for taxonomic identification of multiple organisms out of a mixed DNA sample. Usually amplifies genetic markers with universal primers and uses next generation sequencing technologies
Ecological network	A representation of biotic interactions in an ecosystem, in which species (nodes) are connected by pairwise interactions (links). Links can be used to represent any type of ecological interaction, including antagonistic interactions, such as those of competition and predation (trophic), or mutualistic, such as pollination
Environmental DNA (eDNA)	DNA that can be sampled from environments such as water, soil or feces, without the isolation of organisms
Explainable artificial intelligence	The set of artificial intelligence methods and techniques producing solutions and results that can be understood by humans
Hierarchical modeling	A statistical model where quantities (observations) are sorted in a hierarchy. The key idea is that inferences made about one observation affects inferences about the others in the hierarchy. This contrasts with linear-based methods, where observations are independent
Heuristic food web	Synthetic ecological network constructed from species lists where interactions are inferred from traits (e.g., published consumer–resource linkages), mathematical rules of interaction, or a combination of both
Machine-learning	The study and use of algorithms and statistical models that perform specific tasks without explicit instructions, using instead inference of data patterns
Metacommunity	A set of otherwise distinct communities that interact or are linked by the dispersal of species
Metagenomic sequencing	A comprehensive sequencing approach where all genes from all organisms present in a sample or community are processed
Meta-interpretive learning	An inductive logic program that infers (learns) logic programs (rules) from a combination of background knowledge and examples (observations)
Metatranscriptomic sequencing	The sequencing of the total genes expressed (transcribed) from a community of organisms
Network construction	One of any number of approaches for inferring taxonomic linkages in a community in order to generate a visual representation of co-occurrence patterns
Network inference	The process of hypothesizing and predicting network structure and topology.
Next-generation biomonitoring	The suite of emerging tools and approaches that allow the observation and assessment of the state and ongoing change in ecological systems across a broader spectrum of organizational levels—from genes to entire ecosystems
Occupancy modeling	A type of hierarchical modeling used to infer probabilities of species presence or absence in sample data where there is imperfect detection of organisms
Essential Biodiversity Variable (EBV)	Basic ecological quantities used to assess local to global change in biodiversity as part of monitoring progress toward policy goals and the effects of management
Operational Taxonomic Unit (OTU)	A pragmatic, operational classification of taxa with closely related DNA sequences into groups
Exact Sequence Variant (ESV)	Taxonomic classification where the exact DNA sequence is used for identification as opposed to clustering related sequences into taxonomic units (i.e., as for OTUs)
Functional traits	Key characteristics of individual organisms, whether morphological, structural, biochemical, physiological, phenological or behavioral, which influence performance and fitness

such as ecological networks, which render this complexity comprehensible, communicating to citizens the richness of their local ecosystem and responses to change (Pocock et al., 2016). Third, NGB can foster citizen participation and buy-in to biomonitoring if it underpins evidence-based decision-making (Hodgetts et al., 2018), and projects with high public participation or strong community science components can produce tangible change in management (Schröter et al., 2017). Portable DNA sequencing instruments allow individuals with relatively little training to generate data; for example, Quick et al. (2016) used this approach to develop a tool to monitor the 2015 Ebola outbreak in Central Africa with a 24 h response time. Similar kits are being developed for use by members of the public to monitor local plant and human disease prevalence and the status of pests in agricultural fields and waterways.

For policy, NGB will not only achieve what classical biomonitoring currently does, such as by reporting on agreed classic indicator species or assemblages, but will also allow the inference and prediction of higher level ecosystem properties (Evans et al., 2016; Compson et al., 2019). In principle, NGB could facilitate remedial decision-making, allowing its accompanying management to be trialed before it is implemented. NGB has the potential to enable monitoring of changes in ecosystem structure and function in something close to real time, because large elements of biomonitoring

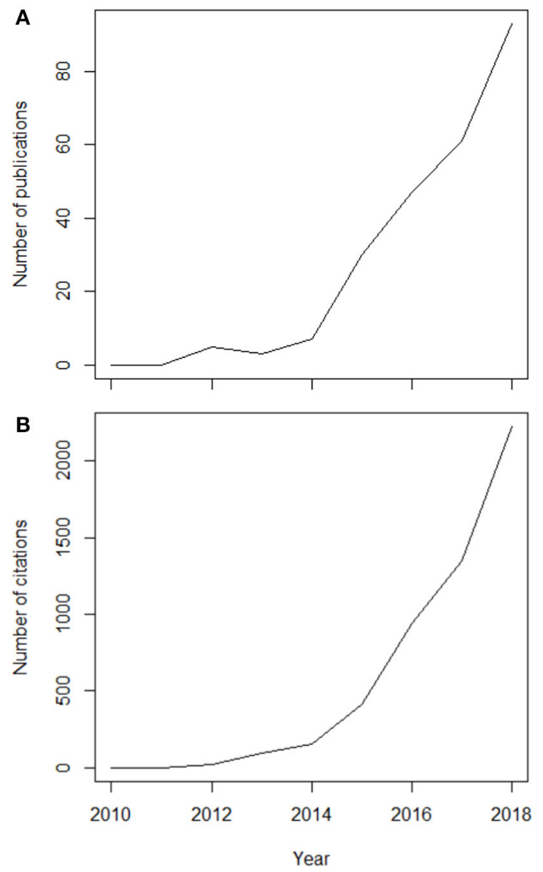


FIGURE 1 | Exponential rise in the number of published, peer-reviewed articles **(A)** and the number of citations of these articles **(B)** about next-generation biomonitoring. Figures depict data obtained through a systematic query of the Web of Science database using the Boolean search: “*monitoring” AND “*DNA” AND “metabarcoding”.

can be automated, reducing the latencies and biases in human-dependent biomonitoring (Quick et al., 2016; Bohan et al., 2017), bringing science one step closer to the vision of biomonitoring any ecosystem in any biome of the globe. Large coverage would also help to avoid some of the “shocks” associated with the loss or the sharp decline of keystone species and major ecosystem processes long after a tipping point has occurred (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, IPBES[1]). Finally, the generality of NGB enables fusing of multiple areas of biomonitoring that are currently distinct and managed separately. Monitoring of disease, invasions, climate, and land-use change could be undertaken simultaneously, greatly reducing the cost of NGB by pooling resources and sharing expenses. This would, in turn, increase the amount of biomonitoring that might be done, increasing its efficiency.

Despite these potential benefits, which have become apparent, adoption of the latest methodologies into management and decision-making processes has been slow, often hindered by miscommunication between research/scientists and management/policy partners (Darling and Mahon, 2011). Nevertheless, the number of initiatives for a global-scale biomonitoring of biodiversity that maximize cooperation and communication between scientists, policymakers and citizens is increasing. These include the development of new indicators, such as Essential Biodiversity Variables (EBVs, Kissling et al., 2018), ontologies for global biomonitoring (Global Infrastructures for Supporting Biodiversity, GLOBIS-B), storage and linking of data-sets (Global Biodiversity Information Facility, GBIF[2]), and routes into global scale policy (Group on Earth Observations Biodiversity Observation Network, GEO BON[3]; Global Earth Observation System of Services, GEOSS[4]). Scientists working on NGB should participate actively in these efforts. For example, the EU Co-Operation in Science and Technology (COST) action DNAqua-Net[5] gathers scientists in order to improve biomonitoring of aquatic ecosystems (Leese et al., 2018; Pawlowski et al., 2018), and has a working group dedicated to the discussion of regulatory and policy frameworks where scientists and stakeholders work collaboratively (Hering et al., 2018). The Interreg European Regional Development Fund project Synergie transfrontalière pour la bio-surveillance et la préservation des écosystèmes Aquatiques (SYNAQUA[6]) shares a similar aim to gather panels of stakeholders to design scenarios for future NGB implementation for freshwater ecosystem biomonitoring in France and Switzerland (Lefrançois et al., 2018). The benefits of NGB should, in turn, leverage new policy, providing a better fit into the current regulatory and policy frameworks for the more “complex” metrics and indicators of ecosystem structure, function, and services. The role of science and scientists should be to critically appraise the development of the NGB approach and, in so doing, advocate for the benefits of NGB and establish a dialogue between relevant biomonitoring scientists, citizens, industry end-users, and policymakers.

What Is the Appropriate Spatio-Temporal Scale for NGB?

A recurrent message of the papers in this issue is that the scales of biomonitoring, both in terms of spatial extent and temporal frequency of sampling, need to be greatly enlarged if we are to appropriately monitor and assess the risks to ecosystems (see Ovaskainen et al., 2019), identify and evaluate the core drivers of ecosystem dynamics and stability, and make decisions for their management. This will require solutions to some of the practical framework problems that limit the scales of the current generation of biomonitoring approaches, including socio-economic, political, and local management issues.

[1] www.ipbes.net/assessment-reports (accessed May 30, 2019).

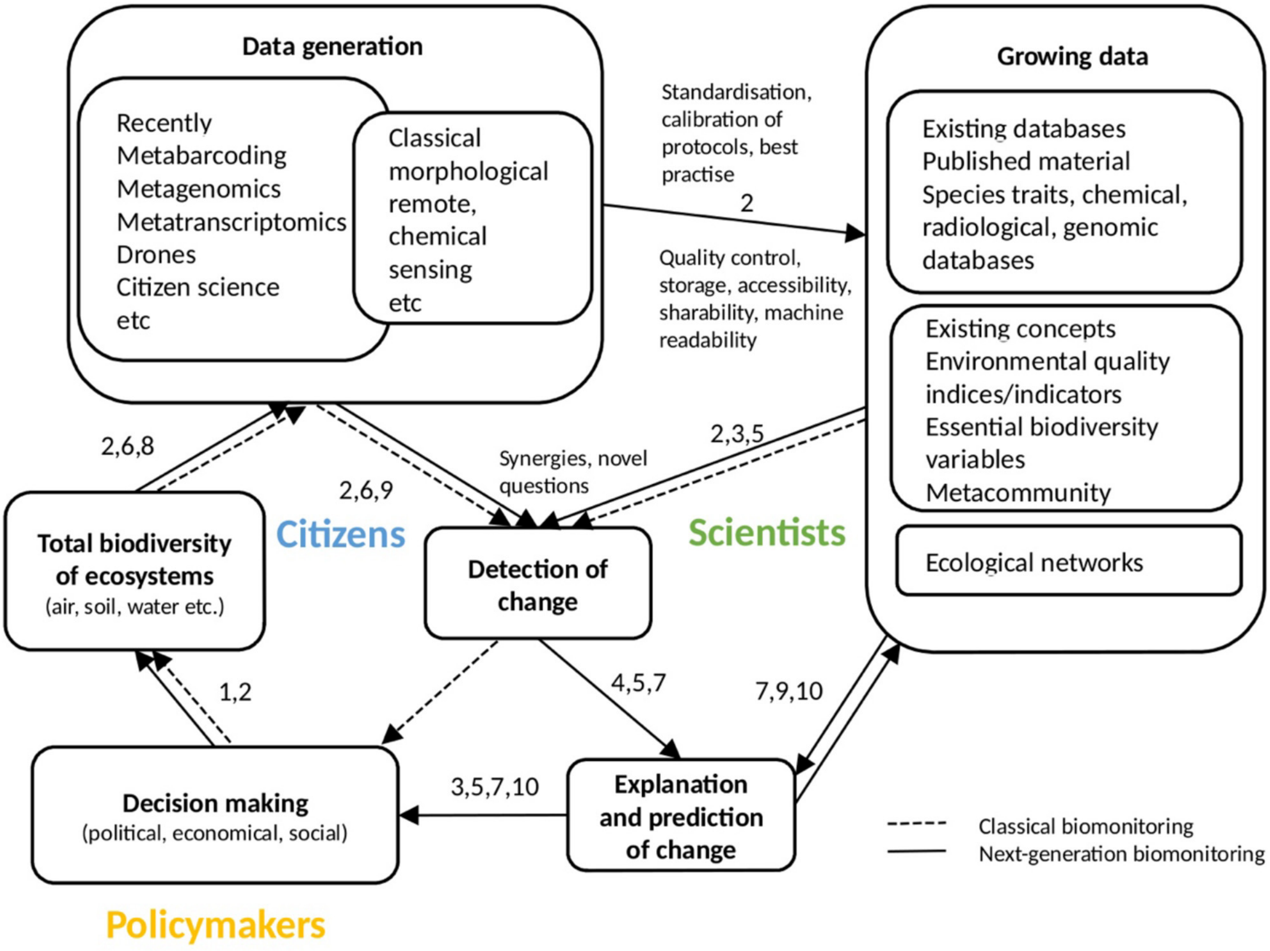
[2] www.gbif.org (accessed May 30, 2019).

[3] www.geobon.org (accessed May 30, 2019).

[4] http://www.earthobservations.org/ (accessed May 30, 2019).

[5] www.dnaqua.net (accessed May 30, 2019)

[6] www.interreg-francesuisse.eu/beneficiaire/synaqua-synergie-transfrontaliere-pour-la-bio-surveillance-et-la-preservation-des-ecosystemes-aquatiques (accessed May 30, 2019)


Data generation
Recently
Metabarcoding
Metagenomics
Metatranscriptomics
Drones
Citizen science
etc
Classical
morphological
remote,
chemical
sensing
etc
Standardisation, calibration of protocols, best practise
2
Quality control, storage, accessibility, sharability, machine readability
Growing data
Existing databases
Published material
Species traits, chemical, radiological, genomic databases
Existing concepts
Environmental quality indices/indicators
Essential biodiversity variables
Metacommunity
Ecological networks
2,6,8
2,6,9
Synergies, novel questions
2,3,5
Total biodiversity of ecosystems
(air, soil, water etc.)
Citizens
Detection of change
Scientists
1,2
4,5,7
7,9,10
Decision making
(political, economical, social)
3,5,7,10
Explanation and prediction of change
Classical biomonitoring
Next-generation biomonitoring
Policymakers

1. How can the benefits of NGB be most successfully communicated to citizens, scientists, and policymakers?
2. What is the appropriate spatio-temporal scale for NGB?
3. What is the most productive balance between case-specific and generic NGB methodologies?
4. What are the appropriate indicators of change?
5. How will NGB benefit from machine-learning approaches?
6. What are the key technical challenges to the advancement of NGB?
7. How can NGB be applied to risk management?
8. What are the most promising future advancements in genomics tools?
9. What are the most promising future advancements in computing and bioinformatics?
10. What are the most promising future advancements in modeling for addressing targeted questions?

FIGURE 2 | Diagrammatic representation of the interplay between the *Key questions for next-generation biomonitoring* presented in this paper. Next-generation biomonitoring (NGB) is based on a holistic view of ecosystems through integrating new technologies and exploring synergies with existing data sources. For its realization, it will be necessary to both automate many bioassessment processes and separate the steps of biodiversity detection and explanation of ecosystem change.

Current biomonitoring is heavily skewed toward terrestrial Europe, North America, Australia, and New Zealand (Cavallo et al., 2019; McGee et al., 2019). This is due, in part, to a lack of expertise in biomonitoring and interpretation in many countries, a global shortage of finance, as well as a limited acceptance of conventional methods. One avenue that might contribute to a solution, besides better communication of NGB (see Question 1), is to simplify the biomonitoring process into component steps. NGB would consist of two essential steps: (1) sample collection and the detection of ecosystem change; and then, only where change is detected, (2) explanation and prediction. Such separation would greatly reduce the need for expertise in all parts of the globe. Automated and high-throughput sampling and detection of change would take place at large temporal and spatial scales, including parts of the globe with poor coverage at present (using field technicians, citizen scientists, or drones), with the expertise to explain any detected change being outsourced to regional centers of excellence, much as already exists for the World Health Organization Regional Offices and the networks of experts they support (WHO[7]). The two-stage process would also lower costs for a given scale of coverage, thus making better use of the available finance. A challenging framework question will be what the definition of "change" is, which may vary between different countries and regions. Having the necessary, near-real-time assessments of change is something that is currently only achievable using the NGB approach.

While it is clear that scalability and reusability of global biomonitoring data are necessary to answer large-scale ecological management questions, this can only be achieved where the steps of sampling and detection of change are automated and standardized, making data machine-readable so that information from different systems is comparable and shareable, and can be integrated with other, existing sources of environmental and ecological information (Poisot et al., 2016, 2019). Automating the process of taxa identification, network construction and inference, and comparison to reference states will require considerable technological development (Bohan et al., 2017; Lausch et al., 2018).

Environmental DNA (eDNA) describes genomic materials shed from organisms into their environment that represent the "template" for NGB analysis. eDNA data quality can be influenced by almost every step in the taxa identification workflow (Zinger et al., 2019), from sample collection (Dickie et al., 2018), DNA extraction (Lear et al., 2018), choice of gene or target region, selection of Taq polymerase, polymerase chain reaction (PCR) cycling protocol, primer, choice of sequencing platform, bioinformatic pipelines (Deiner et al., 2017; Makiola et al., 2018; Bush et al., 2019a; Pauvert et al., 2019), and taxonomic reference databases utilized (Porter and Hajibabaei, 2018). These potential challenges compound with the myriad context-specific influences on the ecology of eDNA, such as abiotic and biotic influences on eDNA production, degradation, and transport in the environment (Barnes and Turner, 2016). Standardization or calibration of sampling protocols and other methods in the workflow can improve reproducibility by allowing compilation and comparison of data from across studies (Dickie et al., 2018). Such standardization can be attractive for the majority of users, being both cheap and efficient, even where their research needs differ, as has been successfully demonstrated in The Earth Microbiome Project[8] (Thompson et al., 2017) and the Global ARMS (Autonomous Reef Monitoring Structures) Program[9] (Ransome et al., 2017).

To tap the full potential of biomonitoring data, it will be necessary to improve curation and access to the rich reference datasets that have already been generated. Due in part to specific institutional regulations, there is a lot of genetic reference material that is only available to researchers within certain institutions. Since molecular-based identifications are heavily dependent on the quality and completeness of the reference databases, this research field will collectively benefit from incentives to curate and upload reference sequences to publicly available databases. Ensuring that these datasets are available in a usable format to interested researchers across the globe represents a major challenge to the field, but one which must be met in order to address global changes in biodiversity and species distribution (Poisot et al., 2016, 2019; Desjardins-Proulx et al., 2019). The definition of the ontologies that will allow NGB data to be machine-read and automated, assuring quality control and the integration of metadata from biomonitoring and associated disciplines, has begun but requires large-scale adoption across fields to be useful.

Knowledge from existing sources (e.g., remote sensing, chemical screening, trait databases) could be integrated into NGB via machine-readable ontologies to generate data synergies and explore novel ecological questions (Bohan et al., 2017; Lausch et al., 2018). For example, this approach could be used to supplement DNA taxa lists with functional trait information for the development of more advanced, predictive heuristic network models (sensu, Compson et al., 2018), while simultaneously creating new—and supplementing existing—databases of taxonomic traits, such as organismal body size or trophic linkages (Kissling et al., 2018). Since the integration of multiple traits and bioindicators holds one of the biggest potential synergies, a possible answer to this question could be working with other initiatives, such as GLOBIS-B, GEO BON, GBIF (Canhos et al., 2015), and the Aquatic eDNA Atlas Project[10], as noted in Question 1, toward a common, decentralized, global biodiversity data platform.

What Is the Most Productive Balance Between Case-Specific and Generic NGB Methodologies?

One promise of NGB is to provide general biomonitoring methodologies and comparisons across potentially any ecosystem, including those currently poorly studied or unknown. The search for rationalized, common approaches has begun in certain disciplines, including in aquatic environments (Goldberg et al., 2016), but as the field matures more general guides or approaches may be achievable. Ecosystems are ecologically distinct, but each has unique scales of operation that should be reflected in the spatial scales, frequencies, and replication of sampling. The scales of application of biomonitoring are currently constrained by the methodology used, with most survey methods designed to assess local taxonomic groups of interest. This leads to methodological heterogeneity across regions (Borja et al., 2009; Birk et al., 2012), encumbering efforts to scale up to regional or national levels (Voulvoulis et al., 2017).

[7] www.who.int/ (accessed May 30, 2019).

[8] www.earthmicrobiome.org (accessed May 30, 2019)

[9] www.oceanarms.org (accessed May 30, 2019)

[10] www.fs.fed.us/rm/boise/AWAE/projects/the-aquatic-eDNAtlas-project.html (accessed May 30, 2019).

Two approaches might be adopted to standardize NGB methodologies. The first would be to sample at the finest spatial resolution possible—at high frequency, in any or all ecosystems across the globe—to store copious amounts of data and to invest in the computational hardware and bioinformatics to detect, forecast, and monitor change. This approach would produce datasets that are both close to complete and an invaluable monitoring and ecological resource, with as yet unforeseen benefits, but the data would come at the cost of collection and curation that may not warrant the increase in efficacy, especially where the detection of system change or changing processes does not require such high-resolution data. However, with plummeting costs this approach will likely be increasingly feasible in the future.

The alternative approach would build upon generic expectations of the rate and temporal dynamics of change in order to identify the required frequency of sampling. The spatial scales of sample independence and representation might then be identified across examples of the ecosystem, indicating appropriate levels of replication to assure detection, with an appropriate power, of given levels of acceptable change. Ma et al. (2018a) described generic, multiscale approaches adopted from the theory of networks to examine temporal and spatial variation. These approaches treat network structure as essentially being independent of the taxa involved in the networks, and use network profiling, null models, and multilayer networks to make statements about the expected level of change that *is* and *is not* acceptable in pure network structural terms. This information can then be fed into ecological modeling and robust forecasting studies.

A standardized but general methodology for sampling would maximize scalability, interpretability, and impact of NGB. It is unlikely, however, that the specification of sampling would conveniently lead to a common set of results for all ecosystems to be examined. Rather, any generality that might exist would likely be limited to some combination of the biome being sampled (i.e., air, soil, water), and the organizational (i.e., regional and local networks, communities, species, populations, individuals, or genes) and taxonomic levels. Generality may only be delivered by an ecological understanding of ecosystem structure, probably facilitated by network approaches.

What Are the Appropriate Indicators of Change?

To move biomonitoring forward, science and policy need to explore how: (1) NGB information could lead to new indicators for metacommunities; (2) novel indicators build upon and contribute to existing indicators and frameworks (e.g., Tapolczai et al., 2019); and (3) spatio-temporal metacommunity scales influence the interpretation of these novel indicators. The indicator concept proposes that the ecological state of an ecosystem can be evaluated by observing a particular taxon or taxonomic group or function (De Cáceres and Legendre, 2009). Taxon-free indicator metrics, such as Indices of Biotic Integrity (IBI), are appealing to environmental practitioners and policymakers because they distill a lot of information down to a simple metric that, in principle, can be compared across systems. However, their simplicity is likely the reason why such metrics may be misused in practice (Seegert, 2000). Further, while indicator species or IBIs might be useful at local spatial scales, they are not applicable across the many habitats, ecosystems, or biomes (Angermeier et al., 2000) that can be monitored using next-generation methods. Pairing molecular-based approaches with machine learning for NGB can potentially recover orders of magnitude more information in biomonitoring data, thus eliminating many of the constraints that hindered the development of biomonitoring indicators we use today. For example, building ecological networks from this recovered data might be used to analyze whole-network properties with ecosystem functions and services (Evans et al., 2016), providing a mechanistic link between network structural change and ecological functions. There certainly is a lot of work to be done to explore and develop these higher-level, network indicators, as well as to determine which network properties will be useful for predicting ecosystem consequences to environmental change. Once developed, however, these tools should provide immediate added-value to the taxonomic lists generated by NGB, as well as to the classical, biomonitoring approaches, especially considering the cost effectiveness of routine, open-source pipelines for the rapid calculation of such (e.g., ecological network) indicators.

Scaling up from a local- to a large-scale approach should furthermore incorporate recent advances in metacommunity ecology into biomonitoring, in order to make sense of the connections that exist among communities across landscapes. Leibold and Chase (2017) expounded the compelling argument that we should combine previously competing concepts of community assembly, such as neutral theory, species sorting, patch dynamics, and mass effects into a single, overarching theory. Ecosystem biomonitoring is strongly rooted in local observation and a normative interpretation, yet it often fails to take into account spatio-temporal variability and connections among sampled localities, arguably leading to over-interpretation of local-scale deviations from a putative "normal state" (Baattrup-Pedersen et al., 2017). We may also underestimate the influence of metacommunity effects on the drivers of local dynamics and, consequently, biomonitoring observations. The scale-limited spatio-temporal scope of biomonitoring studies also carries a serious risk of missing large-scale phenomena that could have potentially devastating impacts, such as biological invasions (Kamenova et al., 2017) or global declines in insects (Hallmann et al., 2017) that went largely unnoticed in policy for nearly 30 years (IPBES[1]). DNA-based approaches offer a potential avenue to address this challenge, and we should seize the opportunity to both develop NGB methods by further refinement and testing and promote these methods to policymakers, citing their many benefits.

How Will NGB Benefit From Machine-Learning Approaches?

Statistical methods for extracting information from data represent some of the basic tools that ecologists wield. Standard statistics are used to explore the covariation between dependent and independent variables and to test hypotheses of interaction. Machine-learning approaches work analogously, exploring the probabilistic or logical correlations across matrices of species data. Machine learning of networks has been successfully applied to classical, macro-ecological sample data (e.g., Bohan et al., 2011) and to evaluate ecosystem responses to changed management (Ma et al., 2018b). In contrast, the reconstruction of microbial networks or the inference of networks and trophic links from DNA data has proven to be more difficult (Barner et al., 2018; Freilich et al., 2018; Deagle et al., 2019), with results that appear to depend upon a combination of the machine learning technique and the data used. No one algorithm will work best for every problem, mirroring the "no free lunch" theorem of Wolpert and Macready (1997). The rhetorical question, "How will NGB benefit from machine-learning approaches?," is one that we can answer only by continual work to further develop and integrate ever better learning approaches into ecology and biomonitoring.

Because NGB represents an emerging field, it is useful to look at examples where machine learning and metabarcoding have been successfully combined. Naïve Bayesian and random forest classifiers have been used to make taxonomic assignments from metabarcodes, produce statistical measures of confidence, and reduce rates of false positive identifications (Wang et al., 2007). Supervised machine learning has been used to classify environmental samples in a meta-analysis of microbial community samples collected by hundreds of researchers for the Earth Microbiome Project (Thompson et al., 2017). Recently, eDNA datasets have been analyzed using supervised machine learning to predict the status of aquatic ecosystems (Cordier et al., 2018). The combination of taxonomy-free molecular data and machine-learning techniques outperformed biomonitoring methods based on the screening of known indicator species by classic metabarcoding (Cordier et al., 2018).

Moving toward the reconstruction of networks of explicit interactions is a logical next step that would afford an ecological explanation of change. Such ecological network reconstruction would require the incorporation of background knowledge or information, for example, about species traits or existing interactions (Tamaddoni-Nezhad et al., 2013, 2015). Taxon interaction knowledge can be text-mined from direct observations recorded in the literature, or inferred from published trait information, and, when used to reconstruct interaction networks such as food webs, offer the potential to generate new biomonitoring metrics derived from network properties (Compson et al., 2018). Recent results suggest that, in the absence of background information, model-free inference of network structure is also feasible using information from the overall network structure and those interactions that are known (Stock et al., 2017). Hypotheses or explicit models for how species interact can also be incorporated into machine learning as background knowledge (Tamaddoni-Nezhad et al., 2013, 2015). As symbolic representations of interactions, these hypotheses and models have the benefit of rendering the machine-learning output human-comprehensible and explainable for decision-making and prediction (Muggleton et al., 2018). The challenge for this model-based approach is that we have relatively few symbolic descriptions of species interactions for organisms, especially in understudied biomes. While there are rules for trophic interactions between macro-organisms, for example, based upon body- or gape-size (Jonsson et al., 2018), there are few such rules for microorganisms. The generation of hypotheses for potentially new mechanisms of interaction in understudied systems could also be supported by artificial intelligence: first, using text mining to recover information about taxa and functions that is not readily accessible from reference databases like Global Biotic Interactions (GloBI) or the United States Geological Survey (USGS) traits database; and then by employing machine learning, such as Meta-Interpretive Learning (Tamaddoni-Nezhad et al., 2015), to hypothesize interaction rules that explain the text-mined information and metabarcoding data.

Considerable amounts of this kind of information exist in literature databases such as Google Scholar, Academic Search Premier, and Web of Science. Unfortunately, the publishing rights to these data are often difficult for scientists to disentangle, and the various text-mining exercises that have been conducted have been treated as hacking attacks, which are resisted. Until these publishing rights are relaxed, such as is proposed in Europe (Enserink, 2018), populating many ecosystems with biological and functional information will remain a limitation.

What Are the Key Technical Challenges to the Advancement of NGB?

NGB aims to detect and explain changes in the total biodiversity of ecosystems to understand and predict the ecological structure of ecosystems. This requires that NGB methods generate accurate data for the presence, absence, and abundance of taxa. Uncertainty in the detection of a taxon, as false negatives or positives, can lead to erroneous conclusions with consequences that could impair biomonitoring and decision making. As noted in Question 3, detection uncertainty can arise from multiple sources, such as sampling, laboratory, and bioinformatics, and these have been extensively reviewed elsewhere (e.g., Deiner et al., 2017; Knight et al., 2018; Larsson et al., 2018; Lear et al., 2018; Porter and Hajibabaei, 2018; Zinger et al., 2019). Work to reduce rates of false negatives and positives in DNA metabarcoding data is an active field of research, and progress has been made through using occupancy modeling (Ficetola et al., 2015, 2016) and probability distribution modeling for tag jumping and contamination issues (Larsson et al., 2018).

The next logical step is to ask whether DNA concentrations in the environment relate to organismal abundance or biomass. The question is intuitive, in the sense that a greater abundance or biomass of organisms should, in principle, produce a higher concentration of DNA, but as with detection uncertainty DNA concentration is determined by many other factors. Studies have demonstrated that the *relative* abundance of an organism between samples can relate to eDNA concentrations (Takahara et al., 2012; Thomas et al., 2016; Piñol et al., 2019). However, the leap from *relative* abundance to *absolute* abundance (or anything close) has been confounded by multiple effects, including an inability to distinguish between live and dead biomass, the observation that different age classes of the same organism release DNA at different rates into the environment (Maruyama et al., 2014), and an increased awareness of the complex environmental interactions of eDNA, relating to its origin, state, transport, and fate (Cristescu and Hebert, 2018). How to treat read count data is critical now that microbiome datasets are understood to be compositional in nature and sensitive to library size and several other biases (Gloor et al., 2017). For NGB it is clear that we need to establish how DNA technologies relate to absolute organismal abundance and how we can minimize methodological biases through best practices (e.g., Knight et al., 2018). However, the debate about the confidence to be invested in metabarcoding data will likely continue until we attain technical advances, such as PCR-free sequencing systems, curated and complete reference databases, and modeling that can explain and correct for errors.

How Can NGB Be Applied to Risk Management?

With further development of NGB, multiple lines of evidence and data will need to be combined in real time to provide managers with cost-effective tools needed to make robust decisions and mitigate impacts on the natural environment. To incorporate these multiple sources of information and move beyond purely descriptive models of ecosystem structure and change, such as eDNA-derived lists of taxa and co-occurrence networks, it will be necessary both to develop explanatory and predictive models of ecosystem function and services, and to test, explore, and understand these models, possibly using developments in text-mining (Compson et al., 2018) and Explainable Artificial Intelligence (Miller, 2019; Rudin, 2019).

As the "universe of observation" (Bush et al., 2019b) expands toward a more integrative ecosystem approach, driven by the growing capacity of molecular and analytical methods, it remains unclear what amount of information will be needed to make good management decisions. For example, how much do we benefit if we incorporate all possible data, or do we just add noise? The application of DNA-isolation from bulk environmental samples or mixed communities coupled with high throughput sequencing and automated taxonomic assignment removes many of the taxonomic constraints currently hindering biomonitoring, particularly for multiple trophic groups and otherwise cryptic groups of organisms (Hug et al., 2016). Increasing taxonomic resolution and greater sampling intensity expands the number of observed biological units. This greater volume of information will also require a parallel expansion of our abilities to interpret biodiversity changes.

Artificial intelligence, in the form of machine learning algorithms such as Meta-interpretive Learning, can help process these large amounts of information and aid in hypothesizing explanatory models of interaction that humans can comprehend and machines can read symbolically (Tamaddoni-Nezhad et al., 2015). The explanations used in biomonitoring will evolve from existing concepts of ecosystem indicators and indices that do not attempt to explain the reason for changes in ecosystems (Derocles et al., 2018) toward models that provide a holistic view of ecological change, such as EBVs (Jetz et al., 2019); models that provide an understanding of the underlying mechanisms behind ecosystem functions; and models that recognize the complex and dynamic nature of ecosystems, including all trophic levels and their interactions. This evolution of biomonitoring, moving from a descriptive toward a predictive risk management tool, based on new hypotheses and models, will have the greatest impact on decision and policy making, which will in turn feed-back to biomonitoring.

Outlook Questions

To this point, the questions posed have focused on contemporary issues about the framework of NGB, as well as technical and conceptual challenges to implementing NGB (**Figure 2**). We also foresee rapid advancement in this field beyond what is needed to establish NGB as a biomonitoring approach, facilitating exploration of new frontiers of science and providing solutions to some of the problems we have outlined in this article. These are related, in large part, to rapid developments in computing and genomics. Specifically, we believe that three areas of advancement in biodiversity assessment and analytical capacity will drastically improve NGB: (1) advances in genomics tools that will lead to greater sequencing capacity, providing unprecedented recovery of information from DNA (Question 8); (2) advances in computing, bioinformatics, and open-source pipelines (Question 9); and, (3) improved models that will allow for more targeted use by practitioners interested in adopting NGB approaches (Question 10).

What Are the Most Promising Future Advancements in Genomics Tools?

Many widely used, next-generation sequencing technologies have attained greater sequencing depth (i.e., the product of the number of reads and the read length standardized to the genome length) despite using shorter read lengths by exponentially increasing the amount of sequences generated (Sims et al., 2014). We anticipate a *next*-next-generation revolution that will achieve whole genome sequencing for entire communities, with enough sequencing depth to provide information about individual sequence variation necessary to begin exploration of evolutionary and functional questions in conjunction with NGB. Already, technologies are emerging that provide orders of magnitude more sequencing depth than current platforms. For example, a single flow cell of Illumina's Novaseq platform can generate ∼700 times greater sequencing depth than is typically available, allowing for the detection of dramatically more diversity, even at coarse taxonomic levels; standardizing sequencing depth using patterned flow cells further improves sequencing performance by preventing the merging of neighboring sequences (Singer et al., 2019). Eventually, as such platforms advance, shotgun sequencing will become the norm, and the need for PCR will be circumvented, eliminating many of the issues currently associated with sequencing and subsequent data processing. Such advances in sequencing capacity and error reduction will translate to higher detection probabilities, greater coverage of species, and better assessments of abundance and rare or endangered species in all systems, including those that are remote and difficult to access or under-studied. Additionally, we foresee three new frontiers of science that the added information from new sequencing technologies will enable us to explore.

First, greater sequencing depth across a larger complement of the community will make it possible to construct robust phylogenetic trees for entire communities, which will help advance NGB method development by providing better phylogenetic information for improving ecological information and prediction. The practice of metaphylogenetics is currently limited by short sequence read lengths (i.e., normally ∼150–300 base pairs). Furthermore, PCR and primer choice can greatly influence the resultant community (Hajibabaei et al., 2019), leading to coarse and unresolved phylogenetic trees. With greater sequencing depth, these limitations will become a thing of the past, allowing for more robust phylogenetic analysis. Community assembly can only be understood in the context of species' evolutionary histories, and such an advance in phylogenetic community ecology will not only be crucial for advancement of ecological theory, but also improve the current standards of biodiversity assessment, allowing for a more holistic exploration of rare or unknown taxa in hyperdiverse, poorly studied ecosystems (Papadopoulou et al., 2015).

Second, future platforms will improve sequencing depth per individual such that it will soon be possible to assess intraspecific genetic variation in an assemblage. To date, studies of population genetics have been limited by primer development for target organisms, focusing on no more than a few taxa at a time in order to answer very specific questions. For example, mitochondrial metagenomics approaches that combine shotgun sequencing and DNA metabarcoding allow for read mapping that may provide the quantitative information on intraspecific genetic variation needed to assess population genetic structure (reviewed in Crampton-Platt et al., 2016). In combination with DNA metabarcoding (sensu, Elbrecht et al., 2018), these approaches would then make it possible to assess the genetic structure of any taxa of interest in the community, and enable practitioners to ask questions about the entire metacommunity and test macroecological theory (e.g., species-genetic diversity correlations).

Third, enhanced sequencing depth will allow for a wider exploration of functional genes in environmental samples. This would make it possible to map functional genes to taxa for entire communities of organisms, linking communities and networks with broad-scale ecosystem assessment of function. Recent efforts have attempted to utilize machine learning to link genus-level predictions of function in microbial communities, for example by using Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt), for inferential assessments of function and hypothesis generation (Douglas et al., 2018). With more sequence data and better inferential methodologies, machine learning in biomonitoring will progress. Concurrent efforts to expand and annotate functional gene databases (e.g., Kyoto Encyclopedia of Genes and Genomes, KEGG[11]) are facilitating the mapping of genes to function across a wide range of biodiversity, bringing incredible added value to projects using the greater sequencing depth afforded by newer sequencing platforms. As these efforts advance, not only will metacommunity and ecosystem theory be advanced by linking structure to function at multiple scales of observation, but potentially transformative changes in biomonitoring and biodiversity assessment will occur, as functional profiles could have greater discriminatory power for detecting change compared to taxonomic profiles, especially in cases where taxonomic profiles are highly variable.

What Are the Most Promising Future Advancements in Computing and Bioinformatics?

With unprecedented data generation, NGB practitioners will be confronted with the enormous task of dealing with an overwhelming amount of information (Keck et al., 2017). Advances in computing and bioinformatics are required to maximize the use of this biodiversity information. Much work still needs to be done to test for and correct errors that inherently emerge from bioinformatics approaches (reviewed in Olson et al., 2017). One solution is to quantitatively assess genome assembly by incorporating evolutionary expectations of gene content, using single copy orthologs (Seppey et al., 2019). These problems of genome assembly and amplification bias will eventually be eliminated as whole-genome sequencing approaches are adopted, but this will, in turn, require even more sophisticated bioinformatics tools (e.g., NanoPack, De Coster et al., 2018).

Another area that will benefit greatly from advances in computing and bioinformatics is database generation, maintenance, and expansion. Existing taxonomic, trait, and functional gene databases (e.g., GenBank, GloBI, KEGG) are incomplete, and the task of updating and expanding these databases is daunting. Artificial intelligence could also be used to advance data discovery (Gonzalez et al., 2016; Compson et al., 2018). Text-mining pipelines, for example, currently make use of open-source, artificial intelligence tools (e.g., OrganismTagger: Naderi et al., 2011). The consequent improvements that these tools will make to taxonomic and functional databases will lead to further advancements of biomonitoring tools, such as cloud-based, rapid ecological network and food web construction, driving a virtuous cycle where more robust datasets lead to improved models.

The promise of these advancements will only be met, however, via improvements in data accessibility, data discoverability, and development of data standards. These will likely emerge from consortiums developing ontologies for genomics and other data (reviewed by Levy and Myers, 2016), as noted in Questions 1 and 3. More work needs to be done, in particular, to develop, peer-review, and publish open-source tools for bioinformatics pipelines (Mangul et al., 2019). Without parallel improvements in tool archival and version control, the improvements that should follow will be inconsistent, reducing their utility and widespread adoption. This work would likely be facilitated by open-source archival services (e.g., GitHub or SourceForge) or package managers (e.g., Bioconda, Grüning et al., 2018).

What Are the Most Promising Future Advancements in Modeling for Addressing Targeted Questions?

While genomic and technological advancements will affect the field of biodiversity assessment, advances in modeling will specifically help end-users, including regulators and resource managers, using NGB approaches. For example, as the costs of sample and bioinformatic processing reduce, more sophisticated hierarchical occupancy models could be applied to repeated sampling data to quantify detection

[11] www.genome.jp/kegg (accessed May 30, 2019).

probabilities and inform practitioners about the sampling effort required to answer system-specific questions. These models, which can account for multiple categorical factors influencing a response variable, can accommodate samples of repeated presence-absence data to provide estimates of occurrence and detection probabilities of species and communities, enabling to account for false negatives due to imperfect detection (Campos-Cerqueira and Aide, 2016; Steenweg et al., 2016), a limitation that is seldom considered in bioassessment studies (McClenaghan et al., 2019). Occupancy modeling could also provide a way past the critical limitation of current DNA metabarcoding—that of obtaining absolute abundance information. Applied hierarchical occupancy modeling has been used to address questions related to the detection and abundance of species (Kery and Andrew Royle, 2015), and future genomic and technical advancements will broaden the application of these models via the generation of larger datasets covering wider ranges and along more gradients of environmental change. Hierarchical occupancy models will enable further leveraging of these more robust datasets by incorporating variation in the pathway from sample collection to sequencing and bioinformatics. Detection probabilities, for example, can be built into Bayesian hierarchical models to detect probabilities associated with different primers, sequencing approaches, and other steps along the sampling-to-sequencing pathway (Doi et al., 2019), providing NGB practitioners with better experiments that make more efficient use of resources (Lugg et al., 2018).

As the field of NGB evolves, we foresee synergistic advancements from merging occupancy-modeling and machine-learning approaches with additional layers of information coded in DNA, recovered by improved sequencing technologies and greater sequencing depth. Incorporating relative read abundance information into occupancy models could be used to assess the abundance of functional gene classes in environmental samples. Shotgun sequencing will also remove the constraints and biases of PCR amplification of DNA, leading to better estimates of sample abundance and biomass (Bista et al., 2018). Much of this information could be incorporated into ecological networks and heuristic food webs to estimate interaction strengths and calculate probabilities of interaction (Morales-Castilla et al., 2015). Finally, with increases in occupancy and food web model sophistication, and as more data are generated that capitalize on these approaches, there will be increasing volumes of high-quality information to feed into machine learning algorithms, leading to more predictive modeling of diverse ecosystems and an unprecedented opportunity for NGB practitioners to anticipate change and prevent ecosystem impairment in real time.

AUTHOR CONTRIBUTIONS

AM conceived, contributed and led the writing of the paper. DAB conceived, contributed and wrote the paper. ZC conceived, contributed and wrote the paper. All other authors contributed and wrote the paper.

ACKNOWLEDGMENTS

ABo and JP would like to acknowledge funding from the FEDER and Swiss Confederation through the SYNAQUA project. ABo, FK, VV, MK, JP, and TC would like to acknowledge fruitful discussions in the framework of the COST action DNAqua-Net (CA15219).

AD, KR, AT-N, and GW would like to acknowledge funding support from the UK Natural Environment Research Council (NERC-NE/M02086X/1 and NE/M020843/1).

DAB, AM, EC, PD, FM, SR, VR, and CV would like to acknowledge the financial support of the French Agence Nationale de la Recherche project NGB (ANR-17-CE32-011).

DAB, AM, SR, and CV would like to acknowledge the support of the Consortium Biocontrole, which provides funding for the BCMicrobiome project.

TMP would like to acknowledge funding from the Canadian government through the Genomics Research and Development Initiative, Ecobiomics project.

REFERENCES

Angermeier, P. L., Smogor, R. A., and Stauffer, J. R. (2000). Regional frameworks and candidate metrics for assessing biotic integrity in mid-Atlantic highland streams. *Trans. Am. Fish. Soc.* 129, 962–981. doi: 10.1577/1548-8659(2000)129<0962:RFACMF>2.3.CO;2

Baattrup-Pedersen, A., Emma Göthe, Riis, T., Andersen, D. K., and Larsen, S. E. (2017). A new paradigm for biomonitoring: an example building on the danish stream plant index. *Methods Ecol. Evol.* 8, 297–307. doi: 10.1111/2041-210X.12676

Baird, D. J., and Hajibabaei, M. (2012). Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Mol. Ecol.* 21, 2039–2044. doi: 10.1111/j.1365-294X.2012.05519.x

Barner, A., Coblentz, K., Hacker, S., and Menge, B. (2018). Fundamental contradictions among observational and experimental estimates of non-trophic species interactions. *Ecology* 99, 557–566. doi: 10.1002/ecy.2133

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

Birk, S., Bonne, W., Borja, A., Brucet, S., Courrat, A., Poikane, S., et al. (2012). Three hundred ways to assess Europe's surface waters: an almost complete overview of biological methods to implement the water framework directive. *Ecol. Indic.* 18, 31–41. doi: 10.1016/j.ecolind.2011.10.009

Bista, I., Carvalho, G. R., Tang, M., Walsh, K., Zhou, X., Hajibabaei, M., et al. (2018). Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Mol. Ecol. Resour.* 18, 1020–1034. doi: 10.1111/1755-0998.12888

Bohan, D. A., Caron-Lormier, G., Muggleton, S., Raybould, A., and Tamaddoni-Nezhad, A. (2011). Automated discovery of food webs from ecological data using logic-based machine learning. *PLoS ONE* 6:e29028. doi: 10.1371/journal.pone.0029028

Bohan, D. A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A. J., and Woodward, G. (2017). Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. *Trends Ecol. Evol.* 32, 477–487. doi: 10.1016/j.tree.2017.03.001

Borja, A., Miles, A., Occhipinti-Ambrogi, A., and Berg, T. (2009). Current status of macroinvertebrate methods used for assessing the quality of European marine waters: implementing the water framework directive. *Hydrobiologia* 633, 181–196. doi: 10.1007/s10750-009-9881-y

Bush, A., Compson, Z., Monk, W., Porter, T. M., Steeves, R., Emilson, E., et al. (2019a). Studying ecosystems with DNA metabarcoding: lessons from biomonitoring of aquatic macroinvertebrates. *Front. Ecol. Evol.* 7:434. doi: 10.1101/578591

Bush, A., Compson, Z., Monk, W., Porter, T. M., Steeves, R., Emilson, E., et al. (2019b). Studying ecosystems with DNA metabarcoding: lessons from aquatic biomonitoring. *bioRxiv* 578591. doi: 10.3389/fevo.2019.00434

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi: 10.1038/ismej.2017.119

Campos-Cerqueira, M., and Aide, T. M. (2016). Improving distribution data of threatened species by combining acoustic monitoring and occupancy modelling. *Methods Ecol. Evol.* 7, 1340–1348. doi: 10.1111/2041-210X.12599

Canhos, D. A. L., Sousa-Baena, M. S., de Souza, S., Maia, L. C., Stehmann, J. R., Canhos, V. P., et al. (2015). The importance of biodiversity e-infrastructures for megadiverse countries. *PLoS Biol.* 13:e1002204. doi: 10.1371/journal.pbio.1002204

Carvell, C., Roy, D. B., Smart, S. M., Pywell, R. F., Preston, C. D., and Goulson, D. (2006). Declines in forage availability for bumblebees at a national scale. *Biol. Conserv.* 132, 481–489. doi: 10.1016/j.biocon.2006.05.008

Cavallo, M., Borja, Á., Elliott, M., Quintino, V., and Touza, J. (2019). Impediments to achieving integrated marine management across borders: the case of the EU marine strategy framework directive. *Mar. Policy* 103, 68–73. doi: 10.1016/j.marpol.2019.02.033

Compson, Z. G., Monk, W. A., Curry, C. J., Gravel, D., Bush, A., Baker, C. J. O., et al. (2018). Linking DNA metabarcoding and text mining to create network-based biomonitoring tools: a case study on boreal wetland macroinvertebrate communities. *Adv. Ecol. Res.* 59, 33–74. doi: 10.1016/bs.aecr.2018.09.001

Compson, Z. G., Monk, W. A., Hayden, B., Bush, A., O'Malley, Z., Hajibabaei, M., et al. (2019). Network-based biomonitoring: exploring freshwater food webs with stable isotope analysis and DNA metabarcoding. *Front. Ecol. Evol.* 7:395. doi: 10.3389/fevo.2019.00395

Cordier, T., Forster, D., Dufresne, Y., Martins, C. I. M., Stoeck, T., and Pawlowski, J. (2018). Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Mol. Ecol. Resour.* 18, 1381–1391. doi: 10.1111/1755-0998.12926

Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., and Pawlowski, J. (2019). Embracing environmental genomics and machine learning for routine biomonitoring. *Trends Microbiol.* 27, 387–397. doi: 10.1016/j.tim.2018.10.012

Crampton-Platt, A., Yu, D. W., Zhou, X., and Vogler, A. P. (2016). Mitochondrial metagenomics: letting the genes out of the bottle. *GigaScience* 5:15. doi: 10.1186/s13742-016-0120-y

Cristescu, M. E., and Hebert, P. D. N. (2018). Uses and misuses of environmental DNA in biodiversity science and conservation. *Annu. Rev. Ecol. Evol. Syst.* 49, 209–230. doi: 10.1146/annurev-ecolsys-110617-062306

Culhane, F. E., Briers, R. A., Tett, P., and Fernandes, T. F. (2014). Structural and functional indices show similar performance in marine ecosystem quality assessment. *Ecol. Indic.* 43, 271–280. doi: 10.1016/j.ecolind.2014.03.009

Darling, J. A., and Mahon, A. R. (2011). From molecules to management: adopting DNA-based methods for monitoring biological invasions in aquatic environments. *Environ. Res.* 111, 978–988. doi: 10.1016/j.envres.2011.02.001

De Cáceres, M., and Legendre, P. (2009). Associations between species and groups of sites: indices and statistical inference. *Ecology* 90, 3566–3574. doi: 10.1890/08-1823.1

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 34, 2666–2669. doi: 10.1093/bioinformatics/bty149

Deagle, B., Thomas, A., McInnes, J., Clarke, L., Vesterinen, E., Clare, E., et al. (2019). Counting with DNA in metabarcoding studies: how should we convert sequence reads to dietary data? *Mol. Ecol.* 28, 391–406. doi: 10.1111/mec.14734

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., et al. (2017). Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol. Ecol.* 26, 5872–5895. doi: 10.1111/mec.14350

Derocles, S. A. P., Bohan, D. A., Dumbrell, A. J., Kitson, J. J. N., Massol, F., Pauvert, C., et al. (2018). Biomonitoring for the 21st century: integrating next-generation sequencing into ecological network analysis. *Adv. Ecol. Res.* 58, 1–62. doi: 10.1016/bs.aecr.2017.12.001

Desjardins-Proulx, P., Poisot, T., and Gravel, D. (2019). Artificial intelligence for ecological and evolutionary synthesis. *Front. Ecol. Evol.* 7:402. doi: 10.3389/fevo.2019.00402

Dickie, I. A., Boyer, S., Buckley, H. L., Duncan, R. P., Gardner, P. P., Hogg, I. D., et al. (2018). Towards robust and repeatable sampling methods in eDNA-based studies. *Mol. Ecol. Resour.* 18, 940–952. doi: 10.1111/1755-0998.12907

Doi, H., Fukaya, K., Oka, S.-I., Sato, K., Kondoh, M., and Miya, M. (2019). Evaluation of detection probabilities at the water-filtering and initial PCR steps in environmental DNA metabarcoding using a multispecies site occupancy model. *Sci. Rep.* 9:3581. doi: 10.1038/s41598-019-40233-1

Douglas, G. M., Beiko, R. G., and Langille, M. G. I. (2018). Predicting the functional potential of the microbiome from marker genes using PICRUSt. *Methods Mol. Biol.* 1849, 169–177. doi: 10.1007/978-1-4939-8728-3_11

Elbrecht, V., Vamos, E. E., Steinke, D., and Leese, F. (2018). Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ* 6:e4644. doi: 10.7717/peerj.4644

Enserink, M. (2018). European funders seek to end reign of paywalled journals. *Science* 361, 957–958. doi: 10.1126/science.361.6406.957

Evans, D. M., Kitson, J. J. N., Lunt, D. H., Straw, N. A., and Pocock, M. J. O. (2016). Merging DNA metabarcoding and ecological network analysis to understand and build resilient terrestrial ecosystems. *Funct. Ecol.* 30, 1904–1916. doi: 10.1111/1365-2435.12659

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., et al. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Mol. Ecol. Resour.* 15, 543–556. doi: 10.1111/1755-0998.12338

Ficetola, G. F., Taberlet, P., and Coissac, E. (2016). How to limit false positives in environmental DNA and metabarcoding? *Mol. Ecol. Resour.* 16, 604–607. doi: 10.1111/1755-0998.12508

Freilich, M., Wieters, E., Broitman, B., Marquet, P., and Navarrete, S. (2018). Species co-occurrence networks: can they reveal trophic and non-trophic interactions in ecological communities? *Ecology* 99, 690–699. doi: 10.1002/ecy.2142

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., et al. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods Ecol. Evol.* 7, 1299–1307. doi: 10.1111/2041-210X.12595

Gonzalez, G. H., Tahsin, T., Goodale, B. C., Greene, A. C., and Greene, C. S. (2016). Recent advances and emerging applications in text and data mining for biomedical discovery. *Brief. Bioinform.* 17, 33–42. doi: 10.1093/bib/bbv087

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., et al. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* 15, 475–476. doi: 10.1038/s41592-018-0046-7

Gutiérrez-Cánovas, C., Arribas, P., Naselli-Flores, L., Bennas, N., Finocchiaro, M., Millán, A., et al. (2019). Evaluating anthropogenic impacts on naturally stressed ecosystems: revisiting river classifications and biomonitoring metrics along salinity gradients. *Sci. Total Environ.* 658, 912–921. doi: 10.1016/j.scitotenv.2018.12.253

Hajibabaei, M., Porter, T. M., Wright, M., and Rudar, J. (2019). COI metabarcoding primer choice affects richness and recovery of indicator taxa in freshwater systems. *PLoS ONE.*14:e0220953. doi: 10.1371/journal.pone.0220953

Hallmann, C. A., Sorg, M., Jongejans, E., Siepel, H., Hofland, N., Schwan, H., et al. (2017). More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS ONE* 12:e0185809. doi: 10.1371/journal.pone.0185809

Hering, D., Borja, A., Jones, J. I., Pont, D., Boets, P., Bouchez, A., et al. (2018). Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. *Water Res.* 138, 192–205. doi: 10.1016/j.watres.2018.03.003

Hodgetts, T., Grenyer, R., Greenhough, B., McLeod, C., Dwyer, A., and Lorimer, J. (2018). The microbiome and its publics: a participatory approach for engaging publics with the microbiome and its implications for health and hygiene. *EMBO Rep.* 19:e45786. doi: 10.15252/embr.201845786

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., et al. (2016). A new view of the tree of life. *Nat. Microbiol.* 1:16048. doi: 10.1038/nmicrobiol.2016.48

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., et al. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nat. Ecol. Evol.* 3, 539–551. doi: 10.1038/s41559-019-0826-1

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., et al. (2008). Global trends in emerging infectious diseases. *Nature* 451, 990–993. doi: 10.1038/nature06536

Jonsson, T., Kaartinen, R., Jonsson, M., and Bommarco, R. (2018). Predictive power of food web models based on body size decreases with trophic complexity. *Ecol. Lett.* 21, 702–712. doi: 10.1111/ele.12938

Kamenova, S., Bartley, T. J., Bohan, D. A., Boutain, J. R., Colautti, R. I., Domaizon, I., et al. (2017). Invasions toolkit. *Adv. Ecol. Res.* 56, 85–182. doi: 10.1016/bs.aecr.2016.10.009

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

Kery, M., and Andrew Royle, J. (2015). *Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS: Volume 1:Prelude and Static Models*. London: Academic Press. doi: 10.1016/B978-0-12-801378-6.00001-1

Kissling, W. D., Walls, R., Bowser, A., Jones, M. O., Kattge, J., Agosti, D., et al. (2018). Towards global data products of essential biodiversity variables on species traits. *Nat. Ecol. Evol.* 2, 1531–1540. doi: 10.1038/s41559-018-0667-3

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., et al. (2018). Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* 16, 410–422. doi: 10.1038/s41579-018-0029-9

Larsson, A. J. M., Stanley, G., Sinha, R., Weissman, I. L., and Sandberg, R. (2018). Computational correction of index switching in multiplexed sequencing libraries. *Nat. Methods* 15, 305–307. doi: 10.1038/nmeth.4666

Lausch, A., Borg, E., Bumberger, J., Dietrich, P., Heurich, M., Huth, A., et al. (2018). Understanding forest health with remote sensing, part III: requirements for a scalable multi-source forest health monitoring network based on data science approaches. *Remote Sens.* 10:1120. doi: 10.3390/rs10071120

Lear, G., Dickie, I., Banks, J., Boyer, S., Buckley, H., Buckley, T., et al. (2018). Methods for the extraction, storage, amplification and sequencing of DNA from environmental samples. *N. Z. J. Ecol*. 42:10. doi: 10.20417/nzjecol.42.9

Leese, F., Bouchez, A., Abarenkov, K., Altermatt, F., Borja, Á., Bruce, K., et al. (2018). Why we need sustainable networks bridging countries, disciplines, cultures and generations for aquatic biomonitoring 2.0: a perspective derived from the DNAqua-net COST action. *Adv. Ecol. Res.* 58, 63–99. doi: 10.1016/bs.aecr.2018.01.001

Lefrançois, E., Apothéloz-Perret-Gentil, L., Blancher, P., Botreau, S., Chardon, C., Crepin, L., et al. (2018). Development and implementation of eco-genomic tools for aquatic ecosystem biomonitoring: the SYNAQUA French-Swiss program. *Environ. Sci. Pollut. Res. Int.* 25, 33858–33866. doi: 10.1007/s11356-018-2172-2

Leibold, M. A., and Chase, J. M. (2017). *Metacommunity Ecology*. Princeton, NJ: Princeton University Press. doi: 10.2307/j.ctt1wf4d24

Levy, S. E., and Myers, R. M. (2016). Advancements in next-generation sequencing. *Annu. Rev. Genomics Hum. Genet.* 17, 95–115. doi: 10.1146/annurev-genom-083115-022413

Li, L., Zheng, B., and Liu, L. (2010). Biomonitoring and bioindicators used for river ecosystems: definitions, approaches and trends. *Procedia environ. Sci.* 2, 1510–1524. doi: 10.1016/j.proenv.2010.10.164

Lugg, W. H., Griffiths, J., van Rooyen, A. R., Weeks, A. R., and Tingley, R. (2018). Optimal survey designs for environmental DNA sampling. *Methods Ecol. Evol.* 9, 1049–1059. doi: 10.1111/2041-210X.12951

Ma, A., Bohan, D. A., Canard, E., Derocles, S. A. P., Gray, C., Lu, X., et al. (2018a). A replicated network approach to "Big Data" in ecology. *Adv. Ecol. Res.* 59, 225–264. doi: 10.1016/bs.aecr.2018.04.001

Ma, A., Lu, X., Gray, C., Raybould, A., Tamaddoni-Nezhad, A., Woodward, G., et al. (2018b). Ecological networks reveal resilience of agro-ecosystems to changes in farming management. *Nat. Ecol. Evol.* 3, 260–264. doi: 10.1038/s41559-018-0757-2

Makiola, A., Dickie, I. A., Holdaway, R. J., Wood, J. R., Orwin, K. H., and Glare, T. R. (2019). Land use is a determinant of plant pathogen alpha-but not beta-diversity. *Mol. Ecol* 28, 3786–3789. doi: 10.1111/mec.15177

Makiola, A., Dickie, I. A., Holdaway, R. J., Wood, J. R., Orwin, K. H., Lee, C. K., et al. (2018). Biases in the metabarcoding of plant pathogens using rust fungi as a model system. *MicrobiologyOpen* 8:e00780. doi: 10.1002/mbo3.780

Mangul, S., Martin, L. S., Eskin, E., and Blekhman, R. (2019). Improving the usability and archival stability of bioinformatics software. *Genome Biol.* 20:47. doi: 10.1186/s13059-019-1649-8

Maruyama, A., Nakamura, K., Yamanaka, H., Kondoh, M., and Minamoto, T. (2014). The release rate of environmental dna from juvenile and adult fish. *PLoS ONE* 9:e114639. doi: 10.1371/journal.pone.0114639

McClenaghan, B., Compson, Z. G., and Hajibabaei, M. (2019). Validating metabarcoding-based biodiversity assessments with multi-species occupancy models: a case study using coastal marine eDNA. *bioRxiv* 797852. doi: 10.1101/797852

McGee, K. M., Robinson, C., and Hajibabaei, M. (2019). Gaps in DNA-based biomonitoring across the globe. *Front. Ecolo. Evol.* 7:337. doi: 10.3389/fevo.2019.00337

Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007

Morales-Castilla, I., Matias, M. G., Gravel, D., and Araújo, M. B. (2015). Inferring biotic interactions from proxies. *Trends Ecol. Evol.* 30, 347–356. doi: 10.1016/j.tree.2015.03.014

Muggleton, S. H., Schmid, U., Zeller, C., Tamaddoni-Nezhad, A., and Besold, T. (2018). Ultra-strong machine learning: comprehensibility of programs learned with ILP. *Mach. Learn.* 107, 1119–1140. doi: 10.1007/s10994-018-5707-3

Naderi, N., Kappler, T., Baker, C. J. O., and Witte, R. (2011). OrganismTagger: detection, normalization and grounding of organism entities in biomedical documents. *Bioinformatics* 27, 2721–2729. doi: 10.1093/bioinformatics/btr452

Olson, N. D., Treangen, T. J., Hill, C. M., Cepeda-Espinoza, V., Ghurye, J., Koren, S., et al. (2017). Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief. Bioinform.* 20, 1140–1150. doi: 10.1093/bib/bbx098

Ovaskainen, O., Abrego, N., Somervuo, P., Palorinne, I., Hardwick, B., Pitkänen, J.-M., et al. (2019). Monitoring fungal communities with the Global Spore Sampling Project. *Front. Ecol. Evol.* 7:511. doi: 10.3389/fevo.2019.00511

Papadopoulou, A., Taberlet, P., and Zinger, L. (2015). Metagenome skimming for phylogenetic community ecology: a new era in biodiversity research. *Mol. Ecol.* 24, 3515–3517. doi: 10.1111/mec.13263

Pauvert, C., Buée, M., Laval, V., Edel-Hermann, V., Fauchery, L., Gautier, A., et al. (2019). Bioinformatics matters: the accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. *Fungal Ecol.* 41, 23–33. doi: 10.1016/j.funeco.2019.03.005

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., et al. (2018). The future of biotic indices in the ecogenomic era: integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Sci. Total Environ.* 637–638, 1295–1310. doi: 10.1016/j.scitotenv.2018.05.002

Piñol, J., Senar, M. A., and Symondson, W. O. C. (2019). The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Mol. Ecol.* 28, 407–419. doi: 10.1111/mec.14776

Pocock, M. J. O., Evans, D. M., Fontaine, C., Harvey, M., Julliard, R., McLaughlin, Ó., et al. (2016). The visualisation of ecological networks, and their use as a tool for engagement, advocacy and management. *Adv. Ecol. Res.* 54, 41–85. doi: 10.1016/bs.aecr.2015.10.006

Poisot, T., Baiser, B., Dunne, J. A., Kéfi, S., Massol, F., Mouquet, N., et al. (2016). Mangal - making ecological network analysis simple. *Ecography* 39, 384–390. doi: 10.1111/ecog.00976

Poisot, T., Bruneau, A., Gonzalez, A., Gravel, D., and Peres-Neto, P. (2019). Ecological data should not be so hard to find and reuse. *Trends Ecol. Evol.* 34, 494–496. doi: 10.1016/j.tree.2019.04.005

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

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530, 228–232. doi: 10.1038/nature16996

Ransome, E., Geller, J. B., Timmers, M., Leray, M., Mahardini, A., Sembiring, A., et al. (2017). The importance of standardization for biodiversity comparisons: a case study using autonomous reef monitoring structures (ARMS) and metabarcoding to measure cryptic diversity on Mo'orea coral reefs, French Polynesia. *PLoS ONE* 12:e0175066. doi: 10.1371/journal.pone.0175066

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi: 10.1038/s42256-019-0048-x

Saito, V. S., Siqueira, T., and Fonseca-Gessner, A. A. (2015). Should phylogenetic and functional diversity metrics compose macroinvertebrate multimetric indices for stream biomonitoring? *Hydrobiologia* 745, 167–179. doi: 10.1007/s10750-014-2102-3

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., and Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nat. Ecol. Evol.* 3, 430–439. doi: 10.1038/s41559-018-0793-y

Schmidt-Traub, G., Obersteiner, M., and Mosnier, A. (2019). Fix the broken food system in three steps. *Nature* 569, 181–183. doi: 10.1038/d41586-019-0 1420-2

Schröter, M., Kraemer, R., Mantel, M., Kabisch, N., Hecker, S., Richter, A., et al. (2017). Citizen science for assessing ecosystem services: status, challenges and opportunities. *Ecosyst. Serv.* 28, 80–94. doi: 10.1016/j.ecoser.2017.09.017

Seegert, G. (2000). The development, use, and misuse of biocriteria with an emphasis on the index of biotic integrity. *Environ. Sci. Policy* 3, 51–58. doi: 10.1016/S1462-9011(00)00027-7

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.* 1962, 227–245. doi: 10.1007/978-1-4939-9173-0_14

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 15, 121–132. doi: 10.1038/nrg3642

Singer, G. A. C., Fahner, N. A., Barnes, J. G., McCarthy, A., and Hajibabaei, M. (2019). Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. *Sci. Rep.* 9:5991. doi: 10.1038/s41598-019-42455-9

Steenweg, R., Whittington, J., Hebblewhite, M., Forshner, A., Johnston, B., Petersen, D., et al. (2016). Camera-based occupancy monitoring at large scales: Power to detect trends in grizzly bears across the Canadian Rockies. *Biol. Conserv.* 201, 192–200. doi: 10.1016/j.biocon.2016.06.020

Stock, M., Poisot, T., Waegeman, W., and De Baets, B. (2017). Linear filtering reveals false negatives in species interaction data. *Sci. Rep.* 7:45908. doi: 10.1038/srep45908

Takahara, T., Minamoto, T., Yamanaka, H., Doi, H., and Kawabata, Z. (2012). Estimation of fish biomass using environmental DNA. *PLoS ONE* 7:e35868. doi: 10.1371/journal.pone.0035868

Tamaddoni-Nezhad A., Bohan D., Raybould A., and Muggleton S. (2015). "Towards machine learning of predictive models from ecological data," in *Inductive Logic Programming. Lecture Notes in Computer Science*, Vol. 9046, eds J. Davis and J. Ramon (Cham: Springer). doi: 10.1007/978-3-319-23708-4_11

Tamaddoni-Nezhad, A., Milani, G. A., Raybould, A., Muggleton, S., and Bohan, D. A. (2013). Construction and validation of food webs using logic-based machine learning and text mining. *Adv. Ecol. Res.* 49, 225–289. doi: 10.1016/B978-0-12-420002-9.00004-4

Tapolczai, K., Keck, F., Bouchez, A., Rimet, F., Kahlert, M., and Vasselon, V. et al. (2019). Diatom DNA metabarcoding for biomonitoring : strategies to avoid major taxonomical and bioinformatical biases limiting molecular indices capacities. *Front. Ecol. Evol.* 7:409. doi: 10.3389/fevo.2019.00409

Thomas, A. C., Deagle, B. E., Paige Eveson, J., Harsch, C. H., and Trites, A. W. (2016). Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material. *Mol. Ecol. Resour.* 16, 714–726. doi: 10.1111/1755-0998.12490

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., et al. (2017). A communal catalogue reveals earth's multiscale microbial diversity. *Nature* 551, 457–463. doi: 10.1038/nature24621

Vandewalle, M., De Bello, F., Berg, M. P., Bolger, T., Doledec, S., Dubs, F., et al. (2010). Functional traits as indicators of biodiversity response to land use changes across ecosystems and organisms. *Biodivers. Conserv.* 19, 2921–2947. doi: 10.1007/s10531-010-9798-9

Voulvoulis, N., Arpon, K. D., and Giakoumis, T. (2017). The EU water framework directive: from great expectations to problems with implementation. *Sci. Total Environ.* 575, 358–366. doi: 10.1016/j.scitotenv.2016.09.228

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

Wolpert, D. H., and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* 1, 67–82. doi: 10.1109/4235.585893

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., et al. (2019). DNA metabarcoding—need for robust experimental designs to draw sound ecological conclusions. *Mol. Ecol.* 28, 1857–1862. doi: 10.1111/mec.15060

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

BRIEF RESEARCH REPORT
published: 14 January 2020
doi: 10.3389/fevo.2019.00511

Monitoring Fungal Communities With the Global Spore Sampling Project

Otso Ovaskainen[1,2]*, *Nerea Abrego*[3], *Panu Somervuo*[1], *Isabella Palorinne*[3], *Bess Hardwick*[1,4], *Juha-Matti Pitkänen*[4,5], *Nigel R. Andrew*[6†], *Pascal A. Niklaus*[7], *Niels Martin Schmidt*[8,9], *Sebastian Seibold*[10,11], *Juliane Vogt*[10], *Evgeny V. Zakharov*[12], *Paul D. N. Hebert*[12], *Tomas Roslin*[3,4] and *Natalia V. Ivanova*[12]

[1] Organismal and Evolutionary Biology Research Programme, University of Helsinki, Helsinki, Finland, [2] Department of Biology, Centre for Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway, [3] Spatial Foodweb Ecology Group, Department of Agricultural Sciences, University of Helsinki, Helsinki, Finland, [4] Spatial Foodweb Ecology Group, Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden, [5] Forest Health and Biodiversity, Natural Resources Institute Finland (LUKE), Helsinki, Finland, [6] Zoology, University of New England, Armidale, NSW, Australia, [7] Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland, [8] Department of Bioscience, Aarhus University, Roskilde, Denmark, [9] Arctic Research Centre, Aarhus University, Aarhus, Denmark, [10] Terrestrial Ecology Research Group, Department of Ecology and Ecosystem Management, Technical University of Munich, Freising, Germany, [11] Field Station Fabrikschleichach, Department of Animal Ecology and Tropical Biology, Julius-Maximilians-University Würzburg, Rauhenebrach, Germany, [12] Centre for Biodiversity Genomics & Department of Integrative Biology, University of Guelph, Guelph, ON, Canada

OPEN ACCESS

Edited by:
Dominique Gravel,
Université de Sherbrooke, Canada

Reviewed by:
Primrose Boynton,
Max Planck Institute for Evolutionary Biology, Germany
Brian John Pickles,
University of Reading, United Kingdom

***Correspondence:**
Otso Ovaskainen
otso.ovaskainen@helsinki.fi

†ORCID:
Nigel R. Andrew
orcid.org/0000-0002-2850-2307

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Ecology and Evolution

Received: *10 July 2019*
Accepted: *17 December 2019*
Published: *14 January 2020*

Citation:
Ovaskainen O, Abrego N, Somervuo P, Palorinne I, Hardwick B, Pitkänen J-M, Andrew NR, Niklaus PA, Schmidt NM, Seibold S, Vogt J, Zakharov EV, Hebert PDN, Roslin T and Ivanova NV (2020) Monitoring Fungal Communities With the Global Spore Sampling Project. Front. Ecol. Evol. 7:511. doi: 10.3389/fevo.2019.00511

The kingdom Fungi is a megadiverse group represented in all ecosystem types. The global diversity and distribution of fungal taxa are poorly known, in part due to the limitations related to traditional fruit-body survey methods. These previous hurdles are now being overcome by rapidly developing DNA-based surveys. Past fungal DNA surveys have predominantly examined soil samples, which capture high species diversity but represent only the local soil community. Recent work has shown that DNA samples collected from the air with cyclone samplers provide information on fungal diversity at the scale of some tens of kilometers around the sampling location. To test the feasibility of air sampling for investigating global patterns of fungal diversity, we established a new initiative called the Global Spore Sampling Project (GSSP). The GSSP currently involves 50 sampling locations distributed on all continents, with each location collecting two 24-h samples per week. Here we describe the GSSP methodology, including the sampling, DNA extraction and sequencing protocols, and the bioinformatics pipeline. We further report results based on 75 pilot samples from five locations, of which three in Europe, one in Australia, and one in Greenland. The results show highly consistent patterns, suggesting that GSSP holds much promise for systematic global fungal monitoring. The GSSP provides highly standardized sampling across space and time, enabling much-improved estimation of total fungal diversity, the global distribution of different fungal groups, fungal fruiting phenology, and the extent of long-distance dispersal in fungi.

Keywords: biomonitoring, cyclone sampler, environmental DNA, fungi, global diversity

INTRODUCTION

Species in the megadiverse fungal kingdom play fundamental roles in ecosystem functioning as mutualists, decomposers, and pathogens. Over 80% of land plants establish mutualistic associations with mycorrhizal fungi (Wang and Qiu, 2006), which facilitate mineral and water uptake (Smith and Read, 2008). Saprotrophic fungi are primary decomposers of the organic matter in terrestrial ecosystems (Baldrian and Valášková, 2008). Pathogenic fungi have important implications for human health and food production (Almeida et al., 2019). Despite their importance, our knowledge of global fungal diversity and biogeography is minimal. While ca. 100,000 species of fungi have been described, estimates of global species richness vary between 0.5 and 10 million (Hawksworth and Lücking, 2017). Furthermore, current knowledge on the diversity and ecology of fungi is biased toward those groups producing macroscopic structures (mostly those producing visible fruiting bodies), even if the diversity of microscopic fungi is vastly greater.

The recent proliferation of environmental DNA based studies have overcome many limitations of fruiting body-based surveys, advancing knowledge of large-scale patterns of fungal diversity (e.g., Sato et al., 2012; Tedersoo et al., 2014; Barberán et al., 2015; Davison et al., 2015). To date, most fungal biogeographical studies have focused on soil communities (but see Barberán et al., 2015) even if different substrates (e.g., wood, litter) support very different assemblages. Furthermore, large-scale fungal studies have mostly been based on samples acquired from distant sites, although it is known that fungal communities can exhibit high spatial variation at very small spatial scales (Hazard et al., 2012; Kubartová et al., 2012). Thus, there are major knowledge gaps regarding the large-scale distributions of fungi, in particular of fungi other than those inhabiting soil.

Since many fungi disperse by windborne spores, DNA surveys based on aerial samples provide an alternative for characterizing the regional fungal composition. As demonstrated by Abrego et al. (2018), aerial sampling can simultaneously sample fungi growing on diverse substrates, while providing a regional scale perspective of some tens of kilometers. To test the feasibility of air sampling for investigating the global patterns of fungal diversity, we established an initiative called the Global Spore Sampling Project (GSSP). We first made a preliminary survey to identify researchers who would be interested in joining the project as sampling teams, and then selected the sampling teams so that we achieved maximal global coverage. Additional partners were encouraged to join if they were able to purchase the sampling equipment themselves. The GSSP currently involves fifty sampling locations distributed across all continents, with each location collecting two 24-h samples per week.

The GSSP project is designed to address research questions of both fundamental and applied nature. To start with, by examining accumulation curves for operational taxonomical units (OTUs) within and among locations, it will improve estimates of fungal diversity. With the help of taxonomic placement of the unknown OTUs (Abarenkov et al., 2018), GSSP will further help to reveal those groups of fungi that are least represented in taxonomy and sequence reference databases. Most importantly, it will provide a much-improved view of global fungal biogeography, shedding light e.g. on how fungal diversity changes along latitude. Based on research on soil fungi, such patterns may deviate substantially from those in other organisms (Tedersoo et al., 2014). Furthermore, these global data may reveal the distributions and temporal dynamics of fungi affecting humans, such as pathogenic fungi causing diseases or infecting crop plants.

In this paper, we describe the GSSP methodology, including the sampling, DNA extraction, sequencing, and concentration estimation protocols, and the bioinformatics pipeline. We further report results based on 75 pilot samples from five locations, of which three in Europe, one in Australia, and one in Greenland.

METHODS

The Global Spore Sampling Project (GSSP) is a globally distributed network of sampling locations (**Figure 1**) equipped with a cyclone sampler (Burkard Cyclone Sampler for Field Operation, Burkard Manufacturing Co Ltd; http://burkard.co.uk/product/cyclone-sampler-for-field-operation; Emberlin and Baboonian, 1995). The current network includes fifty sampling locations that cover all continents, but is most dense in Europe (**Figure 1**). At each sampling location, two 24-h samples are taken each week. Although sampling was planned to start synchronously in October 2018, realized sampling was initiated earlier at a few localities and later in some other localities (**Figure 1**). The sampling locations represent varying latitudes and altitudes. Some samplers are located in urban environments, while others are positioned in natural environments (e.g., forests, tundra).

In this paper, we utilize 75 samples collected during the pilot phase of GSSP from five sites (**Figure 1**). More details on these samples, such as the description of the sampling locations as well as the timing of the sampling, are given in **Supplementary Information**.

Sampling Protocol

Air DNA samples were acquired using cyclone samplers placed at ground level to ensure free airflow through the sampler. The cyclone sampler (shown in **Figure 1**) orientates itself in the direction of the wind and collects all particles from the air with a single reverse flow cyclone. The sampler collects particles with size >1 µm from the air directly into a sampling tube, including spores, pollen, bacteria, and small insects. The sampler's average air throughput is 16.5 liters per minute for a total of 23,800 liters during each 24-h sampling period. Sterile 1.5 ml Eppendorf vials were used as sampling tubes. After sampling, the vial was removed from the cyclone sampler, the lid was closed, and the vials were labeled with the site code and week number. Likewise, the time and duration of the sampling, as well as notes about the presence of rainwater or larger objects (e.g., arthropods), were recorded. Every week, two 24-h samples (henceforth called Sample A and Sample B) were collected from each site. To avoid contamination, gloves were used while handling the samples or the device. Sampling teams were instructed to clean the cyclone part of the device monthly with water and soap and to rinse it

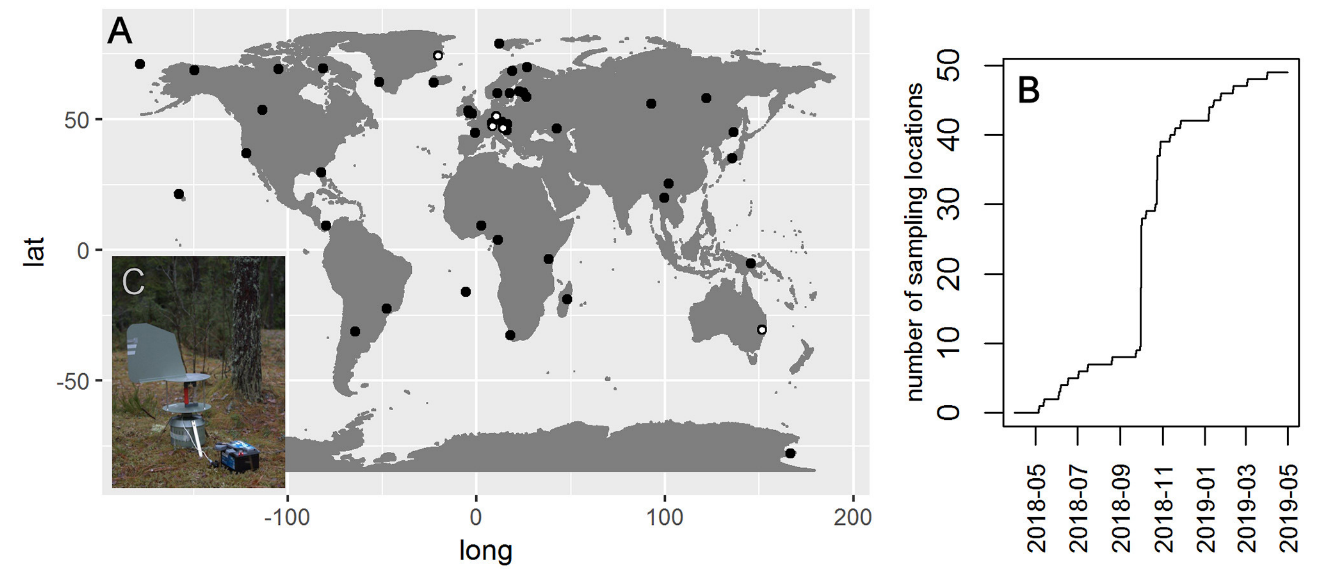


FIGURE 1 | GSSP sampling locations **(A)** and the increase in the number of sampling locations over time **(B)**. The five locations from which the pilot data analyzed in this paper originate are indicated by white dots. **(C)** Shows the cyclone sampler connected to a car battery.

with ethanol, or to sterilize it with dry-heat, chlorine or UV when such equipment was available.

The samples were stored at −20°C until they were shipped to the University of Helsinki, Finland. Shipping was at room temperature as the shipping time was relatively short and refrigerated transport would be costly. In Helsinki, visible arthropods were removed from the samples. To avoid losing fungal spores attached to the arthropod bodies, their surface was rinsed by adding sterile water into the sample tube and vortexing. After washing, the arthropods were removed with sterile tweezers. Samples with rainwater were dried in a freeze drier (24 h, −80°C, 0.57 mbar) covered with a porous Parafilm to avoid cross-contamination between samples. After drying, all samples were sent to the University of Guelph, Canada, for DNA extraction and sequencing.

DNA Extraction, Primers, and Sequencing

Upon receipt, each sample tube was accessioned and assigned a unique Process ID. DNA extraction followed Ivanova et al. (2008) with minor modifications. Two hundred seventy microlitre of ILB (700 mM GuSCN, 30 mM EDTA pH 8.0, 30 mM Tris–HCl pH 8.0, 0.5% Triton® X-100, 5% Tween-20) with 30 µl Proteinase K (20 mg/ml) was added to each collection tube before it was gently rotated to wash spores off the tube walls and lid, and the tube was then centrifuged at 11,000 g for 5 s. The resultant pellet was re-suspended by gentle pipetting, and the entire volume was transferred to a Lysing Matrix A tube (MP-BIO). Tissue was ground in a TissueLyser (Qiagen) for 2 min at 28 Hz. Samples were then incubated for 1 h at 56°C, followed by 1 h at 65°C. Lysates were transferred to a MN block containing 600 µl of 5M GuSCN Binding Buffer (5 M GuSCN, 16.66 mM EDTA pH 8.0, 8.33 mM Tris–HCl pH 6.4, 3.33% Triton® X-100), and the entire volume was transferred in two equal aliquots of 350 µl (each followed by centrifugation at 5,000 xg) onto AcroPrep 96 Filter Plates with 3.0 µm glass fiber media/0.2 µm Bio-Inert membrane, followed by two washes with WB buffer (60% ethanol, 50 mM NaCl, 10 mM Tris–HCl pH 7.4, 0.5 mM EDTA pH 8.0). DNA was eluted in 45 µl of 10 mM Tris-HCl pH 8.0.

Synthetic Positive Control

We applied a synthetic positive control approach (also called spiking approach), with the aim of translating the raw sequence counts into more quantitative estimates of DNA amount. The nine positive control plasmids were prepared from synthetic sequences that are generally consistent with fungal ITS sequences, yet different from all known natural sequences (Palmer et al., 2018). These contained ITS3 (GCATCGATGAAGAACGCAGC) and ITS4 (TCCTCCGCTTATTGATATGC) priming sites (White et al., 1990), and were synthetized as gBlocks at Integrated DNA Technologies (IDT). PCR products were amplified using Platinum Taq and cloned into TOPO4 vector using TOPO™ TA Cloning™ Kit for Sequencing, with One Shot™ TOP10 Chemically Competent *Escherichia coli* (Invitrogen) following manufacturer's instructions. Resulting clones were validated by Sanger sequencing, and each selected clone containing the desired insert was grown in 100 ml of liquid LB media with 150 µg/ml ampicillin. Plasmid DNA was extracted as described using standard protocols (Sambrook et al., 1989) with minor modifications. Plasmid DNA concentration was normalized using BR Qubit dsDNA kit and qPCR, resulting in a pool containing all nine plasmids at 0.01 ng/µl.

Estimating DNA Concentration With qPCR

As an alternative approach to the synthetic positive controls, we also estimated the DNA concentration on each sample using qPCR with a serial dilution of AMPtk plasmids. The target genetic marker ITS2 rDNA for fungi and plants was amplified using the Polymerase Chain Reaction (PCR) for 45 cycles on LightCycler96 (Roche). Each of the 12 μl reactions contained 6 μl of FastStart Essential DNA Green Master (Roche), 1.2 μl of each 10 μM primers (forward primer consisted of a cocktail of ITS_S2F and ITS3 mixed 1:1, ITS4 was used as single reverse primer), 1.6 μl of ddH$_2$O and 2 μl of genomic DNA. Standard curve DNA dilutions for 0.01 ng/μl, 0.001 ng/μl, 0.0001 ng/μ were generated in three replicates using AMPtk plasmids DNA. The thermocycling profile consisted of denaturation at 95°C for 10 min; 55 cycles of 95°C for 10 s, 51°C for 10 s, 72°C−40 s; melting: 95°C for 10 s, 65°C for 60 s, 97°C−1 s. Absolute quantification analysis was performed in LightCycler96 software v1.1.

Next Generation Sequencing Workflow

CCDB PCR Master Mix with Platinum Taq was prepared as described in Hebert et al. (2013). The total reaction volume was 12.5 μl and contained 10.5 μl of MMix and 2 μl of DNA template. The positive synthetic control (0.01 ng/ μl) containing nine plasmids was spiked into the PCR mastermix at a 1:100 ratio. The target genetic marker ITS2 rDNA was amplified using the Polymerase Chain Reaction (PCR) for 20 cycles with fusion primers ITS_S2F (Chen et al., 2010), ITS3 and ITS4 (White et al., 1990) tailed with Illumina adapters:

ITS_S2F-mis
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**ATG CGATACTTGGTGTGAAT**
ITS3-mis
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**GCA TCGATGAAGAACGCAGC**
ITS4-mis
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA G**TCCTCCGCTTATTGATATGC**

PCR products were diluted 2x, and a volume of 2 μl of the diluted product was used in second round PCR amplification with Index 1 and Index 2 fusion primers containing i5 and i7 Nextera indices for dual-indexed PCR: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-adapter-sequences-1000000002694-11.pdf.

The thermocycling for Platinum Taq PCR1 consisted of initial denaturation at 94°C for 2 min followed by 20 cycles of: denaturation at 94°C−1 min; annealing at 51°C– 1 min; extension at 72°C for 1 min; final extension at 72°C for 5 min. PCR2 with indexed Illumina primers consisted of 30 cycles of initial denaturation at 94°C for 2 min followed by 20 cycles of: denaturation at 94°C−1 min; annealing at 60°C– 1 min; extension at 72°C for 1 min; final extension at 72°C for 5 min. DNA template and indexed primer transfers were done on a Biomek FXP robot.

The library was pooled without normalization and purified using Ampure beads (Agencourt, Beckman Coulter) with 0.8:1 beads/PCR product ratio and sequenced on Illumina MiSeq following manufacturer's instructions with 25% spike of PhiX.

Bioinformatics Pipeline

Raw Illumina data were paired using Geneious Prime 2019.0.4, short sequences (<100 bp) were discarded and 5′-end and 3′-end were trimmed by quality (QV20) using BBDuk. The following bioinformatics workflow was used to process paired-end data: Cutadapt (v1.8.1) was used to trim primer sequences; Sickle (v1.33) was used for filtering (<200 bp) and Uclust (v1.2.22q) was used to cluster OTUs with 98% similarity threshold and minimum coverage 2.

We denote the total number of sequences obtained for sample i by n_i and henceforth refer to it as the sequencing depth. Only those samples for which the sequencing depth was at least 10,000 were included. Clusters that corresponded to the spikes were identified using Ublast with 95% similarity threshold. We denote by s_{ij} the number of sequences that were assigned to the spike j (with $j = 1, \ldots, 11$), by $s_i = \sum_j s_{ij}$ the total number of spike sequences in sample i, and by $m_i = n_i - s_i$ the number of sequences not assigned to the spikes. The amount of fungal DNA in each sample was estimated by $w_i = m_i/s_i$, with the caveat that some m_i sequences may not represent fungal DNA. With this definition, w_i measures the amount of fungal DNA in units relative to the spike (presuming no inhibition or competition for PCR reagents). As the total input of spike DNA was ca. 0.001 ng (1/100 volume of 0.01 ng/μl (10.4 μl) spiked into 1,040 μl PCR mix before adding DNA), e.g., $w_i = 24$ would theoretically correspond to a sample containing 0.024 ng of DNA in 2 μl of template DNA added to the reaction, indicating a DNA concentration of 0.012 ng/μl. During DNA extraction 78% of lysate was used for binding so total DNA yield can be calculated as 0.012 ng/μl *45 μl (elution volume), resulting in 0.54 ng in the 78% of the used lysate or 0.69 ng in the total initial sample. As the cyclone sampler processes 24 m^3 of air in 24 h, the value of $w_i = 24$ would theoretically correspond to 0.029 ng of fungal DNA per cubic meter of air. However, we note that this calculation assumes that all DNA amplifies equally across spikes and species, whereas many factors may cause variation in it (Polz and Cavanaugh, 1998; Sipos et al., 2007; Berry et al., 2011; Kennedy et al., 2014).

For the m_i sequences that were not assigned to the spikes and thus represented DNA, a probabilistic taxonomic placement was performed with PROTAX (Somervuo et al., 2017). The specific implementation of PROTAX to fungal identification (Abarenkov et al., 2018) is based on a taxonomy database that includes ca. 130,000 species, and a reference database with about 420,000 reference sequences that represent 22,300 species. We recorded for each query sequence the most likely taxonomic identity at the levels of phylum, class, order, family, genus, and species, and the uncertainty in these assignments as measured by probabilistic placement. We note that the uncertainty estimates of PROTAX account for the possibility that the species might be unknown to science (i.e., not included in the taxonomy database), or known to science but lacking reference sequences (Somervuo et al., 2017; Abarenkov et al., 2018). We followed Somervuo et al. (2017) in treating an identification as plausible if the probability of taxonomic placement was >50%, and reliable if the probability of taxonomic placement was >90%.

Statistical Analyses of the Pilot Data

The statistical analyses were aimed at testing the feasibility of the method in surveying local species communities. Given the lack of controlled mock-community data, the assessment on the reliability of the GSSP pipeline is based on the consistency of observed patterns, especially on the comparison of samples from the five sampling locations.

First, to examine how consistently different spikes were captured among the samples, we defined the relative spike proportion as $\hat{s}_{ij} = s_{ij}/s_i$. We performed an analysis of variance to examine how much of the variation among the relative spike proportions was explained by spike identity j.

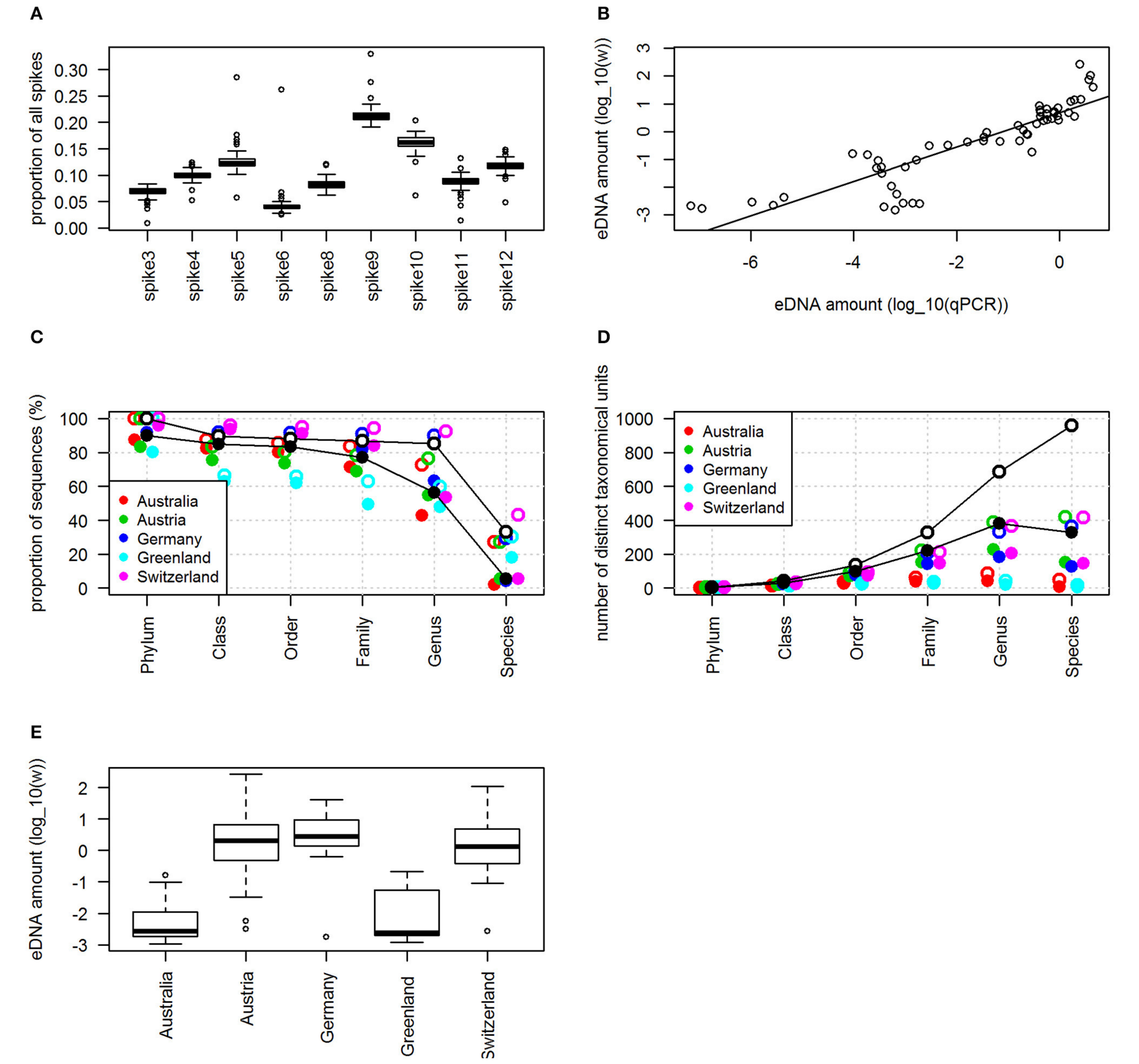


FIGURE 2 | Exploration of the pilot data. **(A)** Shows the relative spike proportions $\hat{s}_{ij}$. **(B)** Compares the amount of DNA as measured by qPCR (x-axis) to the sequencing-based measure w (y-axis), both $\log_{10}$-transformed (R^2 of linear regression 0.79). **(C)** Shows the proportion of samples and **(D)** the number of distinct taxa that could be identified at each taxonomic level. In **(C,D)** open circles refer to plausibly identified taxa (probability of correct taxonomic placement >50%), whereas closed dots refer to reliably identified taxa (probability of correct taxonomic placement >90%). In **(C,D)** black circles represent pooled samples, and other colors to locations-specific samples. **(E)** Shows the amount of DNA, measured by $\log_{10} w_i$, with samples sorted by location.

Second, to validate the accuracy of the sequence-based estimate of fungal DNA amount (w_i), we compared it to the qPCR-based estimate of fungal DNA amount (see above) with the help of linear regression. To identify variation in DNA amount over orders of magnitude, we $\log_{10}$ transformed both the sequence- and the qPCR-based estimates.

Third, we used generalized linear models to examine variation in the estimated amount of DNA (w_i) and OTU richness among the samples. For amount of DNA, we fitted a linear model where the response variable was $\log_{10} w_i$. For OTU richness, we fitted a negative binomial model. In both models, the explanatory variables were the sampling location, the presence of insects in the sample, the presence of water in the sample, and the $\log_{10}$ transformed number of sequences, i.e., the sequencing depth. While the samples were collected over several months and thus contain seasonal variation, we expected that most of the variation would be explained by the sampling location and that the three European sampling locations would yield similar amounts of DNA and OTU richness compared to the Greenland and Australian samples. We further expected that the treatment of the samples due to insects or water might influence the amount of fungal DNA and consequently OTU richness. As we normalized the samples by the spiking approach, we did not expect sequencing depth to have an influence on the estimated amount of DNA, but we expected that OTU richness may increase with sequencing depth.

Fourth, to examine how reliably the sequences could be identified, we computed the proportion of sequences that could be either plausibly or reliably identified for each taxonomic level and each sampling location. To examine how much diversity the samples contained, we computed for each taxonomic level and each sampling location the numbers of distinct OTUs that could be plausibly or reliably identified.

Fifth, to examine how community composition varied among the five locations, we employed a non-metric multidimensional scaling. For the community data we used g_{ij}, defined as the number of sequences in sample i that were plausibly (with probability $>$50%) identified to the genus j. To account for sequencing depth and the variation that covered many orders of magnitude, we transformed the data as $y_{ij} = \log\left((g_{ij}+1)/(n_i+1)\right)$. The non-metric multidimensional scaling was performed using Euclidian distance by the sammon function of the MASS R-package with the default settings as sammon(dist(y)). We further used the anosim function to test whether the five locations separated in the ordination space. We note that the NMDS analyses are aimed primarily for illustrating the data, and that we plan to apply more rigorous model-based analyses (see e.g., Gloor et al., 2017; Ovaskainen et al., 2017) after we have obtained data from all localities.

RESULTS

Among the 75 samples, 73 yielded $>$10,000 sequences with a mean of 52,000 sequences (standard deviation = 10,000; range= 25,000–67,000). The results below are based on the 3.8 million sequences represented by the 73 samples, whereas the two samples with $<$10,000 sequences are excluded.

Does the Use of Spikes Allow Accurate Estimation of DNA Amount?

Different spikes generated very different sequence proportions, ranging from ca. 0.05 to 0.20 (**Figure 2A**). These proportions were found to be stable since in the linear model spike identity explained $R^2 = 0.88$ ($p < 0.001$, $df = 8$, df-residual = 648, $F = 646$) of the variance among the proportions (reflected by the clear separation of the boxes in **Figure 2A**). The use of the spikes allowed very accurate estimation of the amount of fungal DNA, shown by the fact that the sequence-based estimate correlated very closely with the qPCR-based estimate (**Figure 2B**). The proportion of fungal DNA sequences (i.e., non-spike sequences) varied greatly among the samples: the median proportion was 24.9%, the minimum proportion 0.1%, and the maximum proportion 99.7%. Consequently, the estimated amount of DNA varied by five orders of magnitude among the samples, ranging from 0.001 to 264 units of spike DNA, corresponding to from 1.2e-6 to 0.3 ng of fungal DNA per cubic meter of air.

How Much Diversity Was Detected in the Samples?

The taxonomic placement was generally successful up to the family level, as 75% of the sequences were reliably, and 85% were plausibly, identified (**Figure 2C**). As expected, due to the incompleteness of the reference databases (Abarenkov et al.,

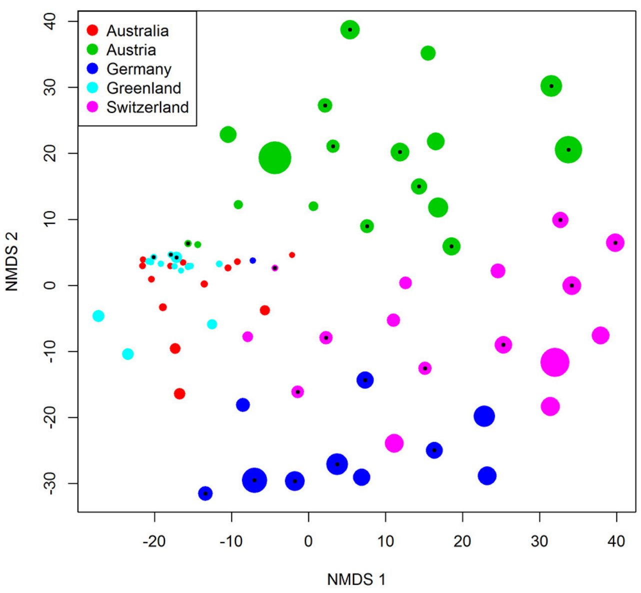


FIGURE 3 | NMDS ordination of the pilot data. Each dot corresponds to one sample. The color indicates sampling location, the size of the sample corresponds to the amount of DNA (measured by w_i), while the black dots mark samples that contained an insect. The final stress achieved in the NMDS was 0.061.

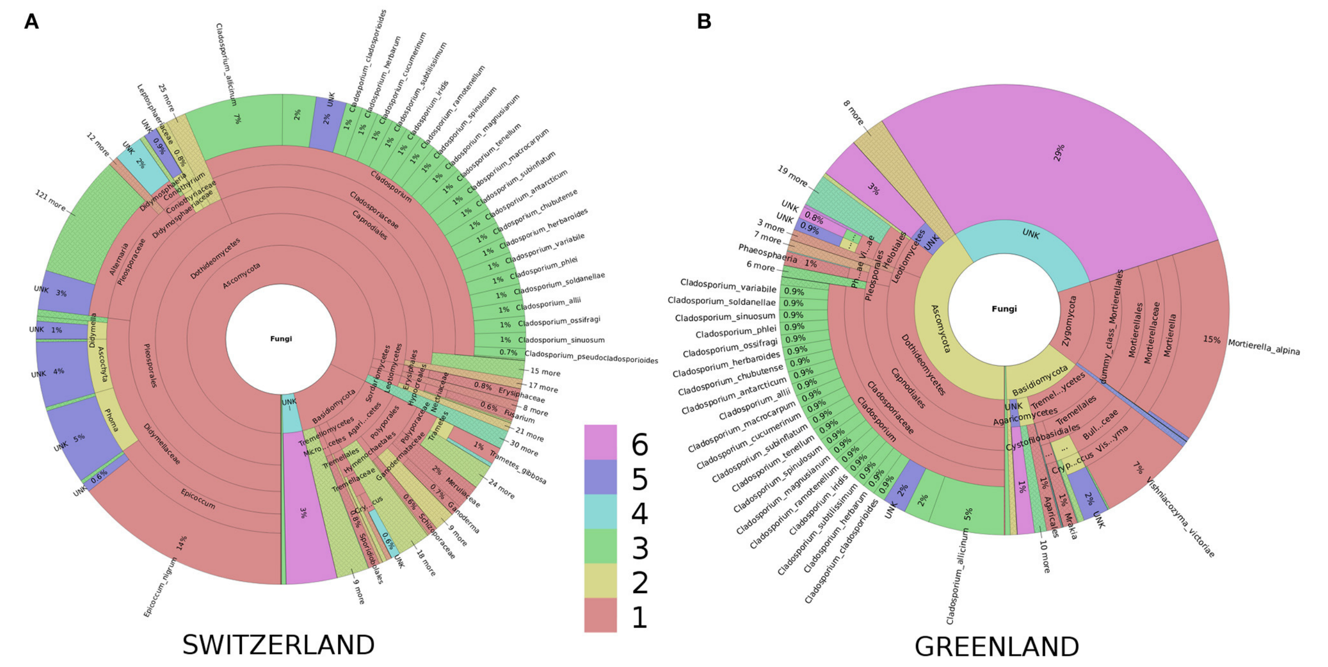


FIGURE 4 | Krona-wheels illustrating taxonomic distributions in the samples. Shown are the results for pooled samples from Switzerland **(A)** and pooled samples from Greenland **(B)**. Similar Krona-wheels for all five sampling locations (that can be interactively explored by the user e.g., to zoom to certain taxa) are given in **Supplementary Information**. The colors show the type and confidence level of each taxonomical placement. Colors 1–3 correspond to well-identified taxonomic units for which the proportion of reliable identifications is in the range [50%...100%] (Color 1), (0%...50%) (Color 2), or 0% (Color 3). Colors 4–6 correspond to unknown taxonomic units for which the proportion of reliable identifications is in the range [50%...100%] (Color 4), (0%...50%) (Color 5), or 0% (Color 6).

2018), a large proportion of the sequences remained unidentified at the genus and especially species levels. The samples yielded considerable taxonomic diversity, totalling about 1,000 species that could be plausibly identified (**Figure 2D**). The true diversity is likely much higher as many sequences remained unidentified at the species level and thus were not included in this estimate.

4088 OTU sequences representing 130,309 sequences (3.5% of all sequences) were assigned to unknown phyla in fungal classification. The best BLAST hits of these sequences against Genbank suggest that 66% of them belong to Viridiplantae, 18% to Fungi, 11% to Metazoa, 3% to Oomycetes, and 1% to Alveolata.

How Did Fungal Communities Vary Among the Sampling Locations?

The samples from the three European locations had consistently high levels of fungal DNA, whereas those from Australia and Greenland did not (**Figure 2E**). Variation in DNA amount was not explained by the presence of water ($p = 0.50$), insects ($p = 0.49$), or sequencing depth ($p = 0.19$), but location had a major effect ($p < 0.001$, $R^2 = 0.56$, $df = 4$, df-residual = 68, $F = 22$). Similarly, variation in OTU richness was not explained by the presence of water ($p = 0.08$), insects ($p = 0.69$), or sequencing depth ($p = 0.99$), but location had a major effect ($p < 0.001$, $df = 4$, df-residual = 68, null deviance = 161, residual deviance = 87). The NMDS suggested consistent variation in fungal community composition, as the five locations separated in the ordination space (anosim, $p < 0.001$), including the three European samples from neighboring countries (**Figure 3**). The Australian and Greenland samples were close in the NMDS plot because we chose to use a Euclidean distance metric and samples from both of these locations contained little DNA and low species diversity. The NMDS further suggested that the removal of insects did not influence the fungal communities e.g., due to introducing contamination, as the samples that contained insects did not deviate systematically from those that lacked them (**Figure 3**).

In terms of taxonomic composition, all samples were dominated by Ascomycota, but they also contained a substantial proportion of Basidiomycota, and the Greenland samples also Zygomygota (**Figure 4**). We note that samples from all five locations consistently contained a high proportion of the genus *Cladosporium*.

DISCUSSION

The results of the present pilot study are encouraging in three ways. First, DNA was obtained from nearly all samples, and sequence characterization recovered a substantial taxonomic diversity despite the small number of samples. Second, we could quantify DNA concentrations consistently by two methods, even if DNA amount was generally low and varied over five orders of magnitude. Third, the results showed a strong biological signal as samples were consistent at each site, suggesting low levels of contamination or other problems related to the workflow.

Community composition was well separated over space among the three neighboring countries of Europe. This result is consistent with an earlier study (Abrego et al., 2018), which showed that air samples represent the regional diversity of fungi procuding airborne spores at some tens of kilometers around the sampling site. However, the more precise spatial scale represented by the samples needs to be determined by further studies, such as sampling at different distances from known point sources.

The success of the pilot study suggests that the GSSP project has great promise for improving knowledge of fungal diversity and distribution at a global scale, with the caveat that GSSP is limited to those fungi only that can sampled from the air. Importantly, the results are quantitative in the sense that one can ask how many times more common (in terms of the number of ITS sequences) a species is in a sample from a particular location and time than in samples from another location and time. This will allow incorporating abundance information in many kinds of analyses. Since spores may disperse over very long distances, the presence of the species in a sample does not necessarily mean that it is part of the local fungal community. Thus, if using the data to construct "species lists," it will be important to separate local spore sources from long-distance dispersal. This can be done by accounting for sequence abundance, and in particular repeated occurrence of species at the same location. Conversely, examining the amount and origin of DNA that does not originate from the local community (e.g., samples acquired during winter in the arctic regions) will allow quantitative estimation of long-distance dispersal in fungi under current air circulation patterns.

The abundance of *Cladosporium* species at all sampling locations raises the possibility of contamination (Czurda et al., 2016). What speaks against this is that the total amount of DNA varied by five orders of magnitude among the samples, while the proportion of *Cladosporium* remained relatively stable. If its presence is due to contamination, one would expect the Greenland and Australian samples (with little DNA) to be dominated by it, whereas the other samples with much DNA would be less influenced. However, no such pattern was observed. *Cladosporium* spores are wind-dispersed and they are known to be very abundant in outdoor air (Harvey, 1967; Kurkela, 1997), and thus we consider their consistent presence in our samples biologically plausible.

To conclude, our pilot study suggests that the GSSP will contribute a fungal dimension to the increasing number of global trapping efforts combined with DNA methods, such as the Global Malaise Trap Program (https://biodiversitygenomics.net/projects/gmp/) or the African Soil Microbiology Project (Wild, 2016). Within the fungal realm, the program will add an important airborne perspective to ongoing perspectives focused on soil (Tedersoo et al., 2014; Bahram et al., 2018) and water (http://pk.emu.ee/en/structure/hydrobiologyandfishery/research/projects/international-projects/funaqua/). While we have focused here on fungal diversity, samples collected by the GSSP network can potentially also be utilized for other taxonomic groups such as plants where the same sampling method has been succesfully applied (Brennan et al., 2019). The GSSP network is open to new sampling teams, and thus we encourage new members to join (https://www.helsinki.fi/en/projects/lifeplan), especially from locations that are currently poorly covered.

DATA AVAILABILITY STATEMENT

Raw sequence data were deposited into ENA, accession PRJEB33255 (https://www.ebi.ac.uk/ena/browser/view/PRJEB33255).

AUTHOR CONTRIBUTIONS

OO, NA, PS, BH, and TR initiated the GSSP network. NRA, PN, NS, SS, and JV collected the data. IP, BH, and J-MP acted as project coordinators, and IP processed the samples in Helsinki. EZ and NI conducted DNA extraction, sequencing, and bioinformatics analyses. PS and OO conducted the statistical analyses. OO, NA, IP, and NI wrote the first draft of the manuscript. All authors contributed substantially to the revisions.

FUNDING

The research was funded by the Academy of Finland (Grant 309581 to OO, and postdoctoral Grant 308651 to NA), the Jane and Aatos Erkko Foundation, and the Research Council of Norway (SFF-III Grant 223257), the Australian Discovery Grant DP160101561 to NRA.

ACKNOWLEDGMENTS

We thank Victoria Eisenbach for acquiring the samples from Hainich, and Arusyak Abrahamyan for preparing plasmids for the synthetic positive control.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2019.00511/full#supplementary-material

REFERENCES

Abarenkov, K., Somervuo, P., Nilsson, H., Kirk, P., Huotari, T., Abrego, N., et al. (2018). PROTAX-fungi: a web-based tool for probabilistic taxonomic placement of fungal ITS sequences. *N. Phytol.* 220, 517–525. doi: 10.1111/nph.5301

Abrego, N., Norros, V., Halme, P., Somervuo, P., Ali-Kovero, A., and Ovaskainen, O. (2018). Give me a sample of air and I will tell which species are found from your region: molecular identification of fungi from airborne spore samples. *Mol. Ecol. Resour.* 18, 511–524. doi: 10.1111/1755-0998.12755

Almeida, F., Rodrigues, M. L., and Coelho, C. (2019). The still underestimated problem of fungal diseases worldwide. *Front. Microbiol.* 10:214. doi: 10.3389/fmicb.2019.00214

Bahram, M., Hildebrand, F., Forslund, S. K., Anderson, J. L., Soudzilovskaia, N. A., Bodegom, P. M., et al. (2018). Structure and function of the global topsoil microbiome. *Nature* 560, 233–237. doi: 10.1038/s41586-018-0386-6

Baldrian, P., and Valášková, V. (2008). Degradation of cellulose by basidiomycetous fungi. *FEMS Microbiol. Rev.* 32, 501–521. doi: 10.1111/j.1574-6976.2008.00106.x

Barberán, A., Ladau, J., Leff, J. W., Pollard, K. S., Menninger, H. L., Dunn, R. R., et al. (2015). Continental-scale distributions of dust-associated bacteria and fungi. *Proc. Natl. Acad. Sci. U.S.A.* 112, 5756–5761. doi: 10.1073/pnas.1420815112

Berry, D., Ben Mahfoudh, K., Wagner, M., and Loy, A. (2011). Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Appl. Environ. Microbiol.* 77, 7846–7849. doi: 10.1128/AEM.05220-11

Brennan, G. L., Potter, C., de Vere, N., Griffith, G. W., Skjøth, C. A., Osborne, N. J., et al. (2019). Temperate airborne grass pollen defined by spatio-temporal shifts in community composition. *Nat. Ecol. Evol.* 3, 750–754. doi: 10.1038/s41559-019-0849-7

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., et al. (2010). Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE* 5:e8613. doi: 10.1371/journal.pone.0008613

Czurda, S., Smelik, S., Preuner-Stix, S., Nogueira, F., and Lion, T. (2016). Occurrence of fungal DNA contamination in PCR reagents: approaches to control and decontamination. *J. Clin. Microbiol.* 54, 148–152. doi: 10.1128/JCM.02112-15

Davison, J., Moora, M., Öpik, M., Adholeya, A., Ainsaar, L., Bâ, A., et al. (2015). Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. *Science* 349, 970–973. doi: 10.1126/science.aab1161

Emberlin, J. C., and Baboonian, C. (1995). "The development of a new method of sampling airborne particles for immunological analysis," in: *16th European Congress of Allergology and Clinical Immunology* (Bologna: Monduzzi Editore S.p.A), 39–43.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 15:2224. doi: 10.3389/fmicb.2017.02224

Harvey, R. (1967). Air-spora studies at Cardiff: I. Cladosporium. *Transact. Br. Mycol. Soc.* 50, 479–495. doi: 10.1016/S0007-1536(67)80017-2

Hawksworth, D., and Lücking, R. (2017). "Fungal Diversity Revisited: 2.2 to 3.8 Million Species," in *The Fungal Kingdom*, eds J. Heitman, B. J. Howlett, P. W. Crous, E. H. Stukenbrock, T. Y. James, and N. A. R. Gow (Washington, DC: ASM Press), 79–95.

Hazard, C., Gosling, P., van der Gast, C. J., Mitchell, D. T., Doohan, F. M., and Bending, G. D. (2012). The role of local environment and geographical distance in determining community composition of arbuscular mycorrhizal fungi at the landscape scale. *ISME J.* 7, 498–508. doi: 10.1038/ismej.2012.127

Hebert, P. D. N., deWaard, J. R., Zakharov, E. V., Prosser, S. W. J., Sones, J. E., McKeown, J. T. A., et al. (2013). A DNA "barcode blitz": rapid digitization and sequencing of a natural history collection. *PLoS ONE* 8:e68535. doi: 10.1371/journal.pone.0068535

Ivanova, N. V., Fazekas, A. J., and Hebert, P. D. N. (2008). Semi-automated, membrane-based protocol for DNA isolation from plants. *Plant Mol. Biol. Report.* 26:186. doi: 10.1007/s11105-008-0029-4

Kennedy, K., Hall, M. W., Lynch, M. D. J., Moreno-Hagelsieb, G., and Neufeld, J. D. (2014). Evaluating bias of illumina-based bacterial 16S rRNA gene profiles. *Appl. Environ. Microbiol.* 80, 5717–5722. doi: 10.1128/AEM.01451-14

Kubartová, A., Ottosson, E., Dahlberg, A., and Stenlid, J. (2012). Patterns of fungal communities among and within decaying logs, revealed by 454 sequencing. *Mol. Ecol.* 21, 4514–4532. doi: 10.1111/j.1365-294X.2012.05723.x

Kurkela, T. (1997). The number of *Cladosporium* conidia in the air in different weather conditions. *Grana* 36, 54–61. doi: 10.1080/00173139709362591

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F. G., Duan, L., Dunson, D., et al. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* 20, 561–576. doi: 10.1111/ele.12757

Palmer, J. M., Jusino, M. A., Banik, M. T., and Lindner, D. L. (2018). Non-biological synthetic spike-in controls and the AMPtk software pipeline improve mycobiome data. *PeerJ.* 6:e4925. doi: 10.7717/peerj.4925

Polz, M. F., and Cavanaugh, C. M. (1998). Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.* 64, 3724–3730. doi: 10.1128/AEM.64.10.3724-3730.1998

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*. 2nd ed. New York, NY: Cold Spring Harbor Laboratory Press.

Sato, H., Tsujino, R., Kurita, K., Yokoyama, K., and Agata, K. (2012). Modelling the global distribution of fungal species: new insights into microbial cosmopolitanism. *Mol. Ecol.* 21, 5599–5612. doi: 10.1111/mec.12053

Sipos, R., Székely, A. J., Palatinszky, M., Révész, S., Márialigeti, K., and Nikolausz, M. (2007). Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol. Ecol.* 60, 341–350. doi: 10.1111/j.1574-6941.2007.00283.x

Smith, S. E., and Read, D. J. (2008). *Mycorrhizal Symbiosis*. New York, NY: Academic Press.

Somervuo, P., Yu, D., Xu, C., Ji, Y., Hultman, J., Wirta, H., et al. (2017). Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. *Methods Ecol. Evol.* 8, 398–407 doi: 10.1111/2041-210X.12721

Tedersoo, L., Bahram, M., Põlme, S., Kõljalg, U., Yorou, N. S., Wijesundera, R., et al. (2014). Global diversity and geography of soil fungi. *Science* 346:1256688. doi: 10.1126/science.1256688

Wang, B., and Qiu, Y. L. (2006). Phylogenetic distribution and evolution of mycorrhizas in land plants. *Mycorrhiza* 16, 299–363. doi: 10.1007/s00572-005-0033-6

White, T. J., Bruns, T. D., Lee, S., and Taylor, J. (1990). "Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics," in PCR Protocols: A Guide to Methods and Applications, eds M. A. Innis, D. H. Gelfaud, J. J. Sninsky, and T. J. White (London: Academic Press), 315–322. doi: 10.1016/B978-0-12-372180-8.50042-1

Wild (2016). Quest to map Africa's soil microbiome begins. *Nature* 10:152. doi: 10.1038/539152a

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 30 April 2020
doi: 10.3389/fevo.2020.00100

DNA Barcoding of Nematodes Using the MinION

Ineke E. Knot[1]*, George D. Zouganelis[2], Gareth D. Weedall[2], Serge A. Wich[1,2] and Robbie Rae[2]

[1] Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, Netherlands, [2] School of Biological and Environmental Sciences, Liverpool John Moores University, Liverpool, United Kingdom

Many nematode species are parasitic and threaten the health of plants and animals, including humans, on a global scale. Advances in DNA sequencing techniques have allowed for the rapid and accurate identification of many organisms including nematodes. However, the steps taken from sample collection in the field to molecular analysis and identification can take many days and depend on access to both immovable equipment and a specialized laboratory. Here, we present a protocol to genetically identify nematodes using 18S SSU rRNA sequencing using the MinION, a portable third generation sequencer, and proof that it is possible to perform all the molecular preparations on a fully portable molecular biology lab – the Bentolab. We show that both parasitic and free-living nematode species (*Anisakis simplex*, *Panagrellus redivivus*, *Turbatrix aceti*, and *Caenorhabditis elegans*) can be identified with a 96–100% accuracy compared to Sanger sequencing, requiring only 10–15 min of sequencing. This protocol is an essential first step toward genetically identifying nematodes in the field from complex natural environments (such as feces, soil, or marine sediments). This increased accessibility could in turn improve global information of nematode presence and distribution, aiding near-real-time global biomonitoring.

Keywords: MinION, DNA barcoding, biomonitoring, 18S (SSU) rRNA gene, *Anisakis simplex*, *Panagrellus redivivus*, *Turbatrix aceti*, *Caenorhabditis elegans*

OPEN ACCESS

Edited by:
*David Andrew Bohan,
INRA Centre Dijon Bourgogne
Franche-Comté, France*

Reviewed by:
*Virginia León-Règagnon,
Institute of Biology, National
Autonomous University of Mexico,
Mexico
Henrik Krehenwinkel,
University of Trier, Germany
Stefan Prost,
Centre for Translational Biodiversity
Genomics (LOEWE-TBG), Germany*

****Correspondence:***
*Ineke E. Knot
ineke@inekeknot.nl*

Specialty section:
*This article was submitted to
Environmental Informatics
and Remote Sensing,
a section of the journal
Frontiers in Ecology and Evolution*

Received: *07 June 2019*
Accepted: *27 March 2020*
Published: *30 April 2020*

Citation:
*Knot IE, Zouganelis GD,
Weedall GD, Wich SA and Rae R
(2020) DNA Barcoding of Nematodes
Using the MinION.
Front. Ecol. Evol. 8:100.
doi: 10.3389/fevo.2020.00100*

INTRODUCTION

Nematodes are one of the most abundant groups of metazoan organisms (Seesao et al., 2017). It is estimated that less than 4% of nematode species are currently known to science, with global species richness estimated between 10^6 and 10^8 (Lambshead, 2004). Many of these species are parasites that threaten the health of plants and animals, including humans. For example, the World Health Organization estimates that worldwide infections with soil-transmitted nematodes cause a human annual disease burden of 3.8 million years lost to disabilities (YLD), a disease burden in the same range as HIV/AIDS (4 million YLD) and twice as high as malaria (1.7 million YLD)[1].

Morphological identification is commonly used to identify nematode species, but also has significant drawbacks. For example, easily distinguishable morphological characters are scarce in nematodes, making identification difficult, time-consuming and often unsuccessful to genus or species level (Decraemer and Baujard, 1998; Lawton et al., 1998; Karanastasi et al., 2001; Lambshead, 2004; Hope and Aryuthaka, 2009). As a result, genetic identification is becoming

[1] Global Health Estimates 2016: Disease burden by Cause, Age, Sex, by Country and by Region, 2000–2016. Geneva, World Health Organization; 2018 [online] https://www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html.

increasingly important in nematology. There have been increased efforts in recent years to resolve the genetic taxonomy of nematodes and barcode nematode species using markers including the 18S small subunit ribosomal RNA gene (18S SSU rRNA), the 28S large subunit ribosomal RNA gene (28S LSU rRNA), the cytochrome oxidase I gene (COI) and the internal transcribed spacer (ITS) regions of the ribosomal RNA locus (Blaxter et al., 1998; Bhadury et al., 2006a,b; Hunt et al., 2016; O'Neil et al., 2017; Seesao et al., 2017; Pafčo et al., 2018).

The majority of these studies used Sanger sequencing but currently there are many sequencing technologies that have become more accessible and affordable for a wide array of applications (Kircher and Kelso, 2010; Van Dijk et al., 2014; Goodwin et al., 2016) such as high-throughput sequencing (HTS) and third generation sequencing (TGS). The latter is defined as single-molecule real-time sequencing (Van Dijk et al., 2014). Massive multiplexing of DNA barcode markers generates a great reduction of per sample sequencing costs and labor time compared to Sanger sequencing (Schuster, 2008; Shokralla et al., 2015). In this paper we explore TGS as an exciting opportunity for novel applications, such as near real-time biomonitoring of parasites, particularly nematodes.

A promising TGS platform is the MinION, introduced in 2014 by Oxford Nanopore Technologies (ONT). The MinION is a portable and compact USB-powered sequencer, generating long reads which can be base called in real-time (Jain et al., 2016). It utilizes a nanopore placed in a biological membrane through which DNA fragments are driven (Deamer et al., 2016), generating a difference in electrical current which can be measured and translated to different DNA bases. More in depth explanation of how the MinION works can be found in reviews by Plesivkova et al. (2019) and Krehenwinkel et al. (2019b). The MinION's portable nature makes it ideal for field research, proven by sequencing efforts in extreme conditions like the Arctic (Edwards et al., 2016; Goordial et al., 2017), Antarctic (Johnson et al., 2017) and the International Space Station (McIntyre et al., 2016). Shotgun genomic sequencing in a national park in Wales identified closely related plant species (Parker et al., 2017) and DNA barcoding (reviewed in Krehenwinkel et al., 2019b) has allowed for the identification of a variety of vertebrates in a rainforest in Ecuador (Pomerantz et al., 2018) and a rainforest in Tanzania (Menegon et al., 2017), all within hours of collection. Furthermore, the sequencer has successfully been used for real-time detection of Ebola virus during the 2014–2015 Ebola outbreak in West-Africa (Quick et al., 2016), Zika virus in Brazil (Faria et al., 2017; Quick et al., 2017), and the current outbreak of nCoV-2019[2], which is an important step toward actionable clinical diagnostics. The most popular application of the MinION sequencer so far is the identification of viral or bacterial populations through metagenomics of the 16S rRNA gene (e.g., Greninger et al., 2015; Quick et al., 2015; Benítez-Páez et al., 2016; Schmidt et al., 2017), but metazoan parasites such as nematodes have not yet been examined.

In this paper we highlight the first step toward sequencing nematodes *in situ*, by genetically identifying parasitic and free-living nematode species with the MinION and testing a portable molecular lab. Specifically, we had four objectives and we sought to: (1) optimize existing 18S SSU rRNA primer sets for MinION sequencing of nematodes; (2) genetically identify nematode species with the MinION; (3) compare the MinION sequencing data to Sanger sequencing data, to assess the quality of MinION data and; (4) test whether we could achieve these results using a portable molecular laboratory, the Bentolab.

MATERIALS AND METHODS

Barcode Testing With Known Species

We tested DNA barcoding on four different nematode species, *Anisakis simplex*, *Panagrellus redivivus*, *Turbatrix aceti*, and *Caenorhabditis elegans*. These species represent a subset of parasitic and free-living species with diverse lifestyles.

Anisakis simplex was dissected from fresh mackerel and stored in 70% ethanol, and one individual nematode was selected for DNA extraction. *A. simplex* is a marine parasite that uses crustaceans as intermediate hosts to infect teleosts and squids (Anderson, 2000). Although humans are accidental hosts for *Anisakis* spp., there has been a dramatic increase over the last decades in the reported prevalence of anisakiasis, a serious zoonotic disease (Chai et al., 2005).

Panagrellus redivivus was harvested from a fresh culture growing on oatmeal medium and used for DNA extraction. *P. redivivus* is a free-living nematode that has been used as a model system to study organ development, signal transduction, and toxicology and recently had its full genome and transcriptome sequenced (Srinivasan et al., 2013). The species is amongst others suggested as a comparative model for *Strongyloides*, as parasitic taxa are typically difficult to culture and analyze independently of their hosts (Blaxter et al., 1998).

Turbatrix aceti was harvested from a fresh culture in an apple cider vinegar medium and used for DNA extraction. The nematodes were washed in distilled water three times before DNA extraction, to mitigate an inhibiting effect of the vinegar medium on the subsequent Polymerase Chain Reaction (PCR). *T. aceti* is a free-living nematode that is mostly researched in relation to aging phenotypes, that are shared with other free-living nematodes such as *Caenorhabditis elegans* (Reiss and Rothstein, 1975). It is also used as live food in the larval stages of many fish species (Brüggemann, 2012). It lacks proper genetic studies, making it an interesting representative for the majority of nematode species that are mostly studied morphologically.

Caenorhabditis elegans strain N2 was grown on nematode growth medium (NGM) plates with *E. coli* OP50 for several days using standard procedures (Brenner, 1974) and subsequently harvested for DNA extraction.

DNA Extraction, PCR, and Sequencing

We extracted the DNA using the GeneJET Genomic DNA Purification Kit (ThermoFisher Scientific Ltd., Paisley,

[2] https://nanoporetech.com/about-us/news/novel-coronavirus-ncov-2019-information-and-updates; https://artic.network/ncov-2019

United Kingdom) according to manufacturer's instructions for mammalian tissue and rodent tail genomic DNA purification (protocol A), except that samples were lysed overnight (step 3) to ensure complete cuticle break down. DNA purity was measured on a NanoDrop 2000 spectrophotometer (software: NanoDrop2000, version 1.4.2; ThermoFisher Scientific Ltd., Paisley, United Kingdom).

We amplified an internal fragment of the 18S SSU rRNA gene from our DNA samples, using the primers and thermocycler protocol optimized by Floyd et al. (2005). This fragment is ∼900 bp in length and widely used for nematode species identification. According to ONT's instruction we adapted the primers from Floyd et al. (2005) to include an adapter tail at the 5′ end ("MinION tail," in lowercase), which is compatible with the MinION workflows. This resulted in the following forward primer: Nem_18S_F_MinION: 5′ tttctgttggtgctgatattgcCGCGAATRGCTCATTACAACAGC 3′ and reverse primer: Nem_18S_R_MinION: 5′ acttgcctgtcgctctatcttcGGGCGGTATCTGATCGCC 3′. A different primer pair, SSU18A and SSU26R (Floyd et al., 2002), was initially tested with the MinION tails, but resulted in no PCR amplification for these samples. Each 25-μl PCR mix contained 2 μl purified DNA extract, 0.5 μl each forward and reverse primers (10 μM; Sigma-Aldrich/Merck Ltd., Poole, United Kingdom), 9.5 μl nuclease free water (NFW; ThermoFisher Scientific Ltd., Paisley, United Kingdom), and 2X GoTaq Hot Start Colorless Master Mix (Promega, Southampton, United Kingdom). PCR was performed on a Bio-Rad T100 Thermal Cycler (Bio-Rad Laboratories Ltd., Watford, United Kingdom). The PCR protocol remained the same as Floyd et al. (2005): initial denaturation for 5 min at 94°C followed by 35 cycles of denaturation for 30 s at 94°C, annealing for 30 s at 54°C and extension for 1 min at 72°C, all followed by a final extension for 10 min at 72°C and cooling to 12°C.

Successful amplification was confirmed using a 2% agarose gel (Agarose I, Molecular Biology Grade; Thermo Fisher Scientific Ltd., Paisley, United Kingdom) made with 1x TBE buffer (Thermo Fisher Scientific Ltd., Paisley, United Kingdom), using 1 μl of NovelJuice nucleic acid stain (Sigma-Aldrich/Merck Ltd., Poole, United Kingdom) loaded with each sample and the size ladder. PCR products were purified using the GeneJET PCR Purification Kit (Thermo Fisher Scientific Ltd., Paisley, United Kingdom) following manufacturer's instruction and eluted in 50 μl of Elution Buffer. DNA purity was measured on a NanoDrop 2000 spectrophotometer (software: NanoDrop2000, version 1.4.2; Thermo Fisher Scientific Ltd., Paisley, United Kingdom) and DNA concentration on a Qubit 1.0 (Thermo Fisher Scientific Ltd., Paisley, United Kingdom), using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific Ltd., Paisley, United Kingdom). Both Nanodrop and Qubit measurements were measured twice per sample to confirm accuracy of the measurement.

We prepared the MinION library according to the 1D PCR barcoding amplicons/cDNA (SQK-LSK109) protocol from ONT (version PBAC12_9067_v109_revH_23MAY2018). This protocol incorporates a second PCR to attach ONT barcodes to our first-round PCR products as means of indexing, allowing multiple samples to be run on one flow cell and subsequent demultiplexing in the bioinformatics stage. Briefly, the PCR Barcoding Kit (EXP-PBC001; ONT Ltd., Oxford, United Kingdom) was used to prepare a 100-μl PCR mix containing 2 μl barcode (10 μM; ONT Ltd., Oxford, United Kingdom), 48 μl first-round PCR product, and 50 μl 2X LongAmp Taq Master Mix [New England BioLabs (NEB) Inc., Hitchin, United Kingdom].

We tried to prepare the first-round PCR products in equimolar concentrations for the barcoding PCR, but due to large variations in DNA concentrations between the samples we diluted the first-round PCR product of *A. simplex* and *P. redivivus* to between 100 and 150 fmol and used all the first-round PCR product for *T. aceti*. *A. simplex* received barcode number 05, *P. redivivus* barcode 06 and *T. aceti* barcode 07. PCR was performed on a Bio-Rad T100 Thermal Cycler (Bio-Rad Laboratories Ltd., Watford, United Kingdom). The PCR protocol for an amplicon length of ∼1,000 bp (including primers) was as follows: initial denaturation 3 min @ 95°C; denaturation 15 s at 95°C, annealing 15 s at 62°C, extension 50 s at 65°C (all 15 cycles); final extension 50 s at 65°C; hold at 4°C. The PCR products were cleaned up with 1X Agencourt AMPure XP beads (Beckman Coulter Inc., Indianapolis, IN, United States). Finally, 1 μl per purified second-round PCR product was quantified on the Qubit 1.0 (Thermo Fisher Scientific Ltd., Paisley, United Kingdom) using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific Ltd., Paisley, United Kingdom).

The concentration of *A. simplex* and *P. redivivus* DNA was too high for Qubit quantification, so we prepared and quantified a 1/5 dilution in NFW (Thermo Fisher Scientific Ltd., Paisley, United Kingdom) that was taken forward. The second-round PCR products were pooled in roughly equimolar concentrations in 47 μl NFW (Thermo Fisher Scientific Ltd., Paisley, United Kingdom).

Library preparation continued using the reagents from the Ligation Sequencing Kit (SQK-LSK109; ONT Ltd., Oxford, United Kingdom), according to manufacturer's instructions. Briefly, we prepared 325 ng pooled barcoded library in 47 μl NFW (ThermoFisher Scientific Ltd., Paisley, United Kingdom). Amplified product was end-repaired using NEBNext Ultra II End-Repair/dA-tailing Module (NEB Inc., Hitchin, United Kingdom) for 5 min at 20°C and 5 min at 65°C, after which it was cleaned up with 1X Agencourt AMPure XP beads (Beckman Coulter Inc., Indianapolis, IN, United States). Adapter ligation was performed using NEB Blunt/TA Ligation Master Mix (NEB Inc., Hitchin, United Kingdom) and reagents provided in the SQK-LSK109 kit. Ligation took place for 10 min at room temperature. DNA was eluted in 15 μl Elution Buffer after being purified with 0.4X AMPure XP beads and washed with the Short Fragment Buffer provided in the SQK-LSK109 kit. 1 μl of prepared library was quantified on the Qubit 1.0 (Thermo Fisher Scientific Ltd., Paisley, United Kingdom) using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific Ltd., Paisley, United Kingdom) and gave a measure of 6.36 ng/μl, which equates to a molarity of 102.9 fmol.

The protocol from ONT recommends loading 5-50 fmol of amplicon product onto the flow cell, so we diluted 5.44 μl of prepared library in 6.56 μl Elution Buffer to load 40 fmol of library onto the flow cell. The flow cell was primed for loading by flushing the flow cell with 1 ml priming mix (30 μl of Flush Tether in one tube of Flush Buffer), taking care to avoid the introduction of air bubbles. The library was prepared for loading by mixing 37.5 μl Sequencing Buffer, 25.5 μl Loading Beads and 12 μl diluted DNA library, after which the sample was added to a flow cell, type R9.5.1, through the SpotON sample port. Total library preparation time was estimated to be ∼3 h.

We performed the sequencing run using MinKNOW (version 3.4.5; ONT Ltd., Oxford, United Kingdom) on the MinIT (a small powerful computing unit that eliminates the need for a dedicated laptop; ONT Ltd., Oxford, United Kingdom), indicating the flow cell type and experimental kit used. As a metric of flow cell quality the MinKNOW software assesses flow cell active pore count, in the multiplexer (MUX) scan before each run. Higher active pore counts represent a high flow cell quality, with a maximum of 2,048 and a guaranteed level of 800. Our flow cell had 1,097 pores available for sequencing. The flow cell generated 116,620 reads in 10 min of sequencing, after which the run was stopped. The flow cell was subsequently washed using the Wash Kit (EXP-WSH002; ONT Ltd., Oxford, United Kingdom) with 150 μl Solution A, followed by 500 μl of Storage Solution, and stored in the fridge for re-use.

Portable DNA Extraction, PCR and Sequencing

In preparation for field work we tested whether the developed MinION procedure could also be performed on a fully portable system. We prepared the model organism *C. elegans* for MinION sequencing using a portable molecular lab, the Bentolab Pro[3] (Nature Biotechnology, 2016; Bento Bioworks Ltd., London, United Kingdom) and a multi tool (CMFTLi 10.8V Li-Ion Cordless Multifunction Tool, Clarke International Ltd., Epping, United Kingdom) as a low-cost handheld vortex. Most of the procedures are similar to above, but working on the Bentolab required some essential adaptations to protocols.

We extracted the DNA using the GeneJET Genomic DNA Purification Kit (ThermoFisher Scientific Ltd., Paisley, United Kingdom) according to manufacturer's instructions for mammalian tissue and rodent tail genomic DNA purification (protocol A). Adaptations to the procedure to make this protocol work on the Bentolab were as follows: In step 3, the sample was divided over two 0.2 ml PCR tubes and briefly spun down using the Bentolab's centrifuge (Bento Bioworks Ltd., London, United Kingdom). Subsequently, the two PCR tubes were incubated for 18 h at 56°C, using the Bentolab's thermocycler (Bento Bioworks Ltd., London, United Kingdom) as a heating block. The thermocycler protocol performed 18 cycles of 1 h at 56°C. In step 4, the lysate was then transferred to a 1.5 ml centrifuge, 20 μl RNase A was added and vortexed on the multi tool[4]. Vortexing on a multi tool can be achieved by attaching the "straight saw blade" to the multi tool. This blade provides enough space for up to four centrifuge tubes at the same time. As a safety measure the sharp end of the saw blade was covered with duck tape. Then the centrifuge tube was added to the blade with duck tape, ensuring a tight fit. The multi tool was turned on at the highest speed (21,000 strokes/minute), creating a similar effect as a lab vortex. The vortexing in step 5 and 6 was also performed using the multi tool. In step 7, the 2 ml collection tube of the GeneJet purification column was replaced by a 1.5 ml centrifuge tube with the cap cut off. The Bentolab's centrifuge can only handle 1.5 ml tubes; use of 2 ml collection/centrifuge tubes will lead to small plastic particles that can lead to reduced efficiency of the centrifuge lock system. Because of the reduced volume of the collection tube, the lysate was added to the prepared column at a maximum of 350 μl at a time, after which the sample was centrifuged for 1 min at 6,000 × g and the flowthrough discarded (with a total of three repeats necessary to complete step 7). In step 8, 250 μl Wash Buffer I was added at a time and centrifuged for 1 min at 8,000 × g and the flowthrough discarded (with a total of two repeats necessary to complete step 8). In step 9, 250 μl Wash Buffer II was added at a time and centrifuged for 4 min at 8,000 × g and the flowthrough discarded (with a total of two repeats necessary to complete step 9, and increased centrifuge time to compensate for the max 8,000 × g force of the Bentolab's centrifuge). An additional dry spin of 1 min at 8,000 × g was performed, after which the collection tube was discarded and replaced by a sterile 1.5 ml centrifuge tube. In step 10, 50 μl of Elution Buffer was added to the purification column.

PCR was prepared as described above, but this time performed using the Bentolab's thermocycler (Bento Bioworks Ltd., London, United Kingdom). Also, aluminum foil was used as a sterile work environment as an alternative for a PCR hood. Aluminum foil was taped to the bench space using masking tape. Bleach (1:10 ratio dilution in water) was sprayed on the surface, letting it sit for 3 min, and wiping the surface with clean paper tissue. This process was repeated twice to decontaminate, after which 70% ethanol was used to remove any residual bleach. The Nem_18S_F/R_MinION primers did not work for *C. elegans*, so the primers from Floyd et al. (2002) were used with MinION tails (in lowercase). The forward primer: SSU18A_MinION: 5′ tttctgttggtgctgatattgcAAAGATTAAGCCATGCATG 3′ and reverse primer: SSU26R_MinION: 5′ acttgcctgtcgctctatcttcCAT TCTTGGCAAATGCTTTCG 3′. Each 25-μl PCR mix contained 2 μl purified DNA extract, 0.5 μl each forward and reverse primers (10 μM; Sigma-Aldrich/Merck Ltd., Poole, United Kingdom), 9.5 μl nuclease free water (NFW; Thermo Fisher Scientific Ltd., Paisley, United Kingdom), and 2X GoTaq Hot Start Colorless Master Mix (Promega, Southampton, United Kingdom). PCR was performed on the Bentolab (Bento Bioworks Ltd., London, United Kingdom). The PCR protocol was adapted from Floyd et al. (2002): initial denaturation 5 min at 94°C; denaturation 1 min at 94°C, annealing 1.5 min at 60°C, extension 2 min at 72°C (all 35 cycles); final extension 10 min at 72°C; hold at 12°C.

Successful amplification was confirmed using a 2% agarose gel (Agarose I, Molecular Biology Grade; Thermo Fisher Scientific Ltd., Paisley, United Kingdom) made with 1x TBE buffer

[3] https://www.bento.bio

[4] Inspired by Holly Ganz' use of a multi tool for bead beating: https://youtu.be/Q7PM1xoMjiU.

(Thermo Fisher Scientific Ltd., Paisley, United Kingdom). For the Bentolab's small gel chamber (Bento Bioworks Ltd., London, United Kingdom) we used 27.5 ml of 1X TBE buffer with 0.5 g agarose. The need for a scale was eliminated by using an Eppendorf tube marked with the needed volume corresponding to 0.5 g agarose. Agarose was melted into the TBE buffer using a traditional coffee pot, which has a typical conical shape, on the hob. We have also found this method to work on a camping stove. The gel was then poured into the chamber and left to set for ∼15 min. The comb and shutters were removed and we added 45 ml 1x TBE buffer for the gel electrophoresis run, 60 min at 60V. We again used 1 μl NovelJuice (Sigma-Aldrich/Merck Ltd., Poole, United Kingdom) for the size ladder and per sample for DNA staining, as this DNA stain is safer to work with than traditional ethidium bromide and works both with UV transilluminators and with the blue LED transilluminator of the Bentolab (Bento Bioworks Ltd., London, United Kingdom).

The PCR product was cleaned up using GeneJET PCR Purification Kit (Thermo Fisher Scientific Ltd., Paisley, United Kingdom) following manufacturer's instruction for DNA purification using centrifuge (protocol A). Adaptations to the procedure to make this protocol work on the Bentolab were as follows: In step 3, the 2 ml collection tube of the GeneJet purification column was replaced by a 1.5 ml centrifuge tube with the cap cut off (see adaptations to DNA purification protocol for explanation). The solution of step 1 was added to the purification column, centrifuged for 1 min at 8,000 × g and the flowthrough discarded. In step 4, 350 μl Wash Buffer was added at a time and centrifuged for 1 min at 8,000 × g and the flowthrough discarded (with a total of two repeats necessary to complete step 4). In step 5, a dry spin of 1.5 min at 8,000 × g was performed. In step 6, the collection tube was discarded and replaced by a clean 1.5 ml centrifuge tube. 50 μl of Elution Buffer was added to the purification column and centrifuged for 1 min at 8,000 × g.

We prepared the MinION library according to the 1D PCR barcoding amplicons/cDNA (SQK-LSK109) protocol from ONT (version PBAC12_9067_v109_revH_23MAY2018). As mentioned above, this protocol incorporates a second PCR to attach ONT barcodes to our first-round PCR products as means of indexing. This not only allows multiple samples to be run on one flow cell, but also allows for demultiplexing in the bioinformatics stage when a flow cell is reused. Washing a flow cell after a run might leave some remnant DNA from previous runs. Therefore, the ONT barcodes help to identify the current sample in the bioinformatics stage. Briefly, the PCR Barcoding Kit (EXP-PBC001; ONT Ltd., Oxford, United Kingdom) was used to prepare a 100-μl PCR mix containing 2 μl barcode (10 μM; ONT Ltd., Oxford, United Kingdom), 2 μl first-round PCR product, 46 μl NFW (ThermoFisher Scientific Ltd., Paisley, United Kingdom) and 50 μl 2X LongAmp Taq Master Mix [New England BioLabs (NEB) Inc., Hitchin, United Kingdom]. *C. elegans* received barcode number 10. PCR was performed on the Bentolab thermocycler (Bento Bioworks Ltd., London, United Kingdom). The barcoding PCR protocol was slightly adjusted to accommodate the Bentolab's inability for setting cycles of 15 s and minimum thermocycler temperature of 10°C: initial denaturation 3 min @ 95°C; denaturation 20 s at 95°C, annealing 20 s at 62°C, extension 60 s at 65°C (all 12 cycles); final extension 50 s at 65°C; hold at 10°C. The PCR products were cleaned up with 1X Agencourt AMPure XP beads (Beckman Coulter Inc., Indianapolis, IN, United States) on a 3D-printed magnetic BOMB microtube rack[5] (Oberacker et al., 2019).

Library preparation continued using the reagents from the Ligation Sequencing Kit (SQK-LSK109; ONT Ltd., Oxford, United Kingdom), according to manufacturer's instructions. Since we wouldn't have an accurate way of quantifying DNA in the field we based the used volume of DNA on a previous MinION run (Knot, unpublished data), to prepare 33 μl barcoded library in 47 μl NFW (Thermo Fisher Scientific Ltd., Paisley, United Kingdom). Amplified product was end-repaired using NEBNext Ultra II End-Repair/dA-tailing Module (NEB Inc., Hitchin, United Kingdom) for 5 min at 20°C and 5 min at 65°C on the Bentolab thermocycler (Bento Bioworks Ltd., London, United Kingdom). The end-repaired library was cleaned up with 1X Agencourt AMPure XP beads (Beckman Coulter Inc., Indianapolis, IN, United States) on a 3D-printed magnetic BOMB microtube rack (Oberacker et al., 2019). Adapter ligation was performed using NEB Blunt/TA Ligation Master Mix (NEB Inc., Hitchin, United Kingdom) and reagents provided in the SQK-LSK109 kit. Ligation took place for 10 min at room temperature. DNA was eluted in 15 μl Elution Buffer after being purified with 0.4X AMPure XP beads and washed with the Short Fragment Buffer provided in the SQK-LSK109 kit.

The flow cell was primed for loading by flushing the flow cell with 1 ml priming mix (30 μl of Flush Tether in one tube of Flush Buffer), taking care to avoid the introduction of air bubbles. The library was prepared for loading by mixing 37.5 μl Sequencing Buffer, 25.5 μl Loading Beads and 12 μl DNA library, after which the sample was added to a flow cell, type R9.5.1, through the SpotON sample port. Total library preparation time was estimated to be ∼4.5 h.

We performed the sequencing run using MinKNOW (version 3.4.5; ONT Ltd., Oxford, United Kingdom) on the MinIT (a small powerful computing unit that eliminates the need for a dedicated laptop; ONT Ltd., Oxford, United Kingdom), indicating the flow cell type and experimental kit used. To test whether old flow cells can still be useful for sequencing small barcoded amplicon libraries, we used a flow cell that was used twice before, once in a 24 h run and once in a 2.5 h run. When reusing a flow cell the starting voltage has to be adjusted and we adjusted this to -225 V, equivalent to ONT's recommendation after ∼26 h previous run time. As mentioned above, higher active pore counts represent a high flow cell quality, with a maximum of 2,048 and a guaranteed level of 800 for new flow cells. The MUX scan indicated our flow cell had 43 pores available for sequencing. The flow cell generated 2,632 reads in 14 min of sequencing, after which the run was stopped.

Sanger Sequencing

Each of the samples used for the MinION sequencing was also sent for Sanger sequencing (GATC/Eurofins Genomics).

[5] https://bomb.bio/protocols/

Both forward and reverse strands were sequenced using the amplification primers as sequencing primers. Sanger sequence electropherograms were visually inspected and edited using 4Peaks version 1.8 (Nucleobytes B.V., Aalsmeer, the Netherlands). Edited forward strand and reverse complemented reverse strand sequences were aligned using Seaview version 4.7 (Gouy et al., 2010). Nucleotide mismatches were checked in the original electropherogram and resolved. A consensus sequence was derived for each sample and primer sequences trimmed from each end of it. The resulting sequences were 885 bp (*A. simplex*), 887 bp (*P. redivivus*), 832 bp (*T. aceti*), and 844 bp (*C. elegans*) long.

Bioinformatic Analyses

The raw fast5 MinION reads were basecalled and demultiplexed using Guppy version 3.2.4 + d9ed22f (ONT Ltd., Oxford, United Kingdom) to produce fastq files for each sample. Reads were classified as pass/fail based on a minimum quality score of 7. The fastq files were merged into one per sample and explored using Nanoplot (version 1.28.0[6]), creating plots displaying log transformed read length ("–loglength"). Barcode and primer trimming was performed using Porechop (version 0.2.4[7]). A second round of demultiplexing requiring barcodes at both ends of the reads ("–require_two_barcodes") was performed using Porechop. Subsequently, the MinION reads were processed using the default settings of the ONTrack pipeline (version 1.4.2[8]; Maestri et al., 2019). Briefly, Seqtk seq[9] was used to create fasta files complementary to the fastq files. Reads were clustered using VSEARCH (Rognes et al., 2016), after which the reads in the most abundant cluster were retained. Then 200 randomly sampled reads were used to produce a draft consensus sequence using Seqtk sample and aligned using MAFFT (Katoh et al., 2002). EMBOSS cons[10] was then used to retrieve a draft consensus sequence starting from the MAFFT alignment. Another 200 randomly sampled reads using Seqtk sample, different from the first iteration, were mapped to the draft consensus sequence using Minimap2 (Li, 2018) to polish the obtained consensus sequence. Samtools was used to filter and sort the alignment file and compress it to the bam format (Li et al., 2009). Nanopolish index and nanopolish variants – consensus modules from Nanopolish[11] were used to obtain a polished consensus sequence. The ONTrack pipeline was run with three iterations, the standard value of the pipeline. This resulted in three polished consensus sequences which were aligned with MAFFT to select the consensus sequence that was produced in the majority of times. All scripts of the pipeline were run within a virtual machine (as part of the ONTrack pipeline), emulating an Ubuntu v18.04.2 LTS operating system, on a Mac laptop without using any internet connection. All the code used for the bioinformatic analyses and additional files necessary to replicate the analyses can be found on https://github.com/ieknot/MinION-DNA-barcoding-of-nematodes. MinION fastq and Sanger fasta accession numbers are reported in the results.

To assess sequence accuracy, MinION raw reads and consensus reads were aligned to the corresponding Sanger-derived reference sequence using BLASTn (Altschul et al., 1990), with no sequence complexity masking ("-dust no-soft_masking false"). The consensus sequences were aligned to the corresponding Sanger sequence using the MUSCLE algorithm (Edgar, 2004) in Seaview version 4.7 (Gouy et al., 2010).

RESULTS

Sequencing Run Quality and Yield

The first and multiplexed flow cell had 1,097 pores available for sequencing. The flow cell generated 116,620 reads containing 6,033 Mb in 10 min of sequencing, after which the run was stopped. During basecalling 71.9% of these reads passed the minimum quality threshold. The basecalled reads were demultiplexed, producing 42,304 reads for analysis (**Table 1**). The mean read length was 1,015 bp for *A. simplex*, 1,011 bp for *P. redivivus* and 504 bp for *T. aceti*.

The second flow cell, used for the library prepared with a fully portable setup, had 43 pores available for sequencing. The flow cell generated 2,632 reads containing 1.94 Mb in 14 min of sequencing, after which the run was stopped. During basecalling 48.9% of these reads passed the minimum quality threshold. The basecalled reads were demultiplexed, producing 205 reads for analysis (**Table 1**). The mean read length was 833 bp for *C. elegans*.

To assess the usefulness of the sequence data for taxonomic identification of the samples, the raw reads for each sample were compared to the (Sanger) reference sequences using BLASTn. The distributions of percentage sequence identities for all pairwise read-reference comparisons are shown in **Figure 1**. The median percent identity was 88.5% for *A. simplex*, 87.7% for *P. redivivus*, 89.5% for *T. aceti* and 82.3% for *C. elegans*, indicating

TABLE 1 | Summary of number of reads per sample in subsequent bioinformatics steps.

MinION run	Species	Sample	Demultiplexed reads	Trimmed reads (%)	Reads used for consensus
1	*A. simplex*	BC05	8,059	3,294 (40.9%)	200
1	*P. redivivus*	BC06	13,802	5,515 (40.0%)	200
1	*T. aceti*	BC07	20,443	2,955 (14.5%)	200
2	*C. elegans*	BC10	483	205 (42.4%)	65

Trimmed reads are also represented as percentage of demultiplexed reads. For C. elegans the largest VSEARCH cluster contained only 65 reads, so the consensus sequence was generated with this amount of reads (see main text for explanation of the ONTrack bioinformatics pipeline).

[6] https://github.com/wdecoster/NanoPlot

[7] https://github.com/rrwick/Porechop

[8] https://github.com/MaestSi/ONTrack

[9] https://github.com/lh3/seqtk

[10] http://emboss.open-bio.org/rel/dev/apps/cons.html

[11] https://github.com/jts/nanopolish

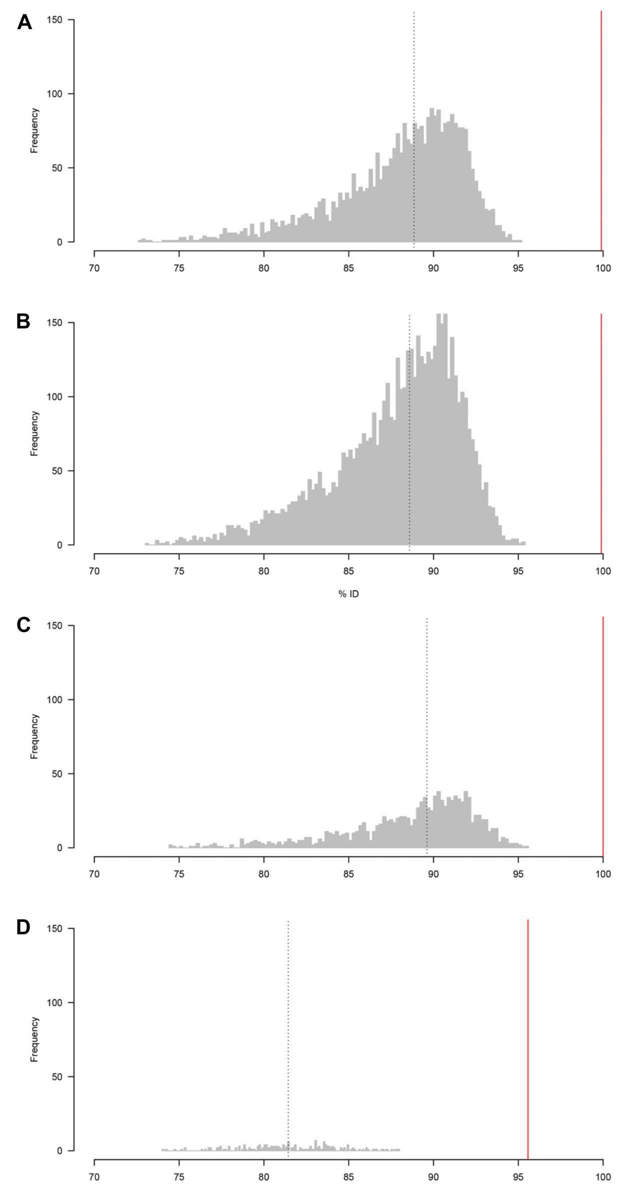


FIGURE 1 | Distribution of percentage sequence identity between raw reads and the (Sanger) reference sequence **(A)** *A. simplex*, **(B)** *P. redivivus*, **(C)** *T. aceti*, and **(D)** *C. elegans*. **(A–C)** were run on a new MinION flow cell, whereas **(D)** was run on a flow cell that had been used twice before (for a total of 26.5 h, see main text for more details). The median percent identity (indicated by a vertical dotted line) is 88.5% for *A. simplex*, 87.7% for *P. redivivus*, 89.5% for *T. aceti* and 82.3% for *C. elegans*. Bioinformatic analyses using the ONTrack pipeline generated a consensus sequence for every sample that had a 99.9% **(A,B)**, 100% **(C)**, and 95.6% **(D)** accuracy compared to their Sanger reads (indicated by a vertical red line).

a ∼11% error rate in sequencing for the first run and a ∼18% error rate in the second run (**Figure 1**).

Bioinformatics Analyses

The second demultiplexing round in Porechop, which included the trimming step to remove primers and ONT barcodes, produced the following number of reads per sample to be taken forward in the ONTrack bioinformatics pipeline (Maestri et al., 2019): 3,294 reads for *A. simplex*, 5,515 reads for *P. redivivus*, 2,955 reads for *T. aceti* and 205 for *C. elegans* (**Table 1**). These reads were used in the VSEARCH clustering step of the ONTrack pipeline, where contaminating sequences were removed by only taking the largest cluster of reads forward. *T. aceti* raw data showed an unexpected short mean read length of 504 bp. The clustering step in the ONTrack pipeline removed contaminating sequences, after which the mean read length of *T. aceti* improved to 773 bp. From these clusters 200 reads were subsampled per sample for consensus sequence generation. However, for *C. elegans* the largest VSEARCH cluster contained only 65 reads, so the consensus sequence was generated with this amount of reads (**Table 1**).

The default setting of the ONTrack pipeline is to run three iterations of the pipeline, generating three consensus sequences per sample. Subsequently, it aligns the consensus sequences generated during each round and selects the final consensus sequence based on the majority rule. Two species, *A. simplex* and *P. redivivus*, generated three consensus sequences which were all different. Since they all had the same statistical probability of being correct, the first consensus sequence was randomly selected. *T. aceti* had a consensus sequence supported by two iterations, and the *C. elegans* consensus sequence was supported by all three iterations of the pipeline.

Median percent identity between consensus sequences and the reference sequence was significantly improved in all cases (**Figure 1**). For *A. simplex* and *P. redivivus* the accuracy improved to 99.9%, for *T. aceti* to 100% and for *C. elegans* to 95.6% (**Figure 2**). Compared to the raw MinION sequences this is an improvement of 11.4, 12.2, 10.5, and 13.3%, respectively. The MinION datasets generated for this study can be found in the European Nucleotide Archive (ENA) under the project ID PRJEB37489 (samples ERS4397495, ERS4397496, ERS4397497, and ERS4397498). MinION consensus sequences are available in the **Supplementary Material.**

Sanger sequence read lengths were 885 bp for *A. simplex*, 887 bp for *P. redivivus*, 832 bp for *T. aceti* and 844 bp for *C. elegans*. The Sanger reads of all samples matched 100% to a sequence of the correct species on NCBI (**Figure 2**). Sanger consensus sequences are available at GenBank under the accession numbers MT246663, MT246664, MT246665, and MT246666.

DISCUSSION

We successfully genetically identified four nematode species using 18S SSU rRNA barcoding on the MinION. We also proved that this can be accomplished using a fully portable molecular lab. This was possible by successfully adapting 18S SSU rRNA primers (Floyd et al., 2002, Floyd et al., 2005) with MinION tails. The read lengths of both samples fall within the expected range. Our first run yielded three successful species

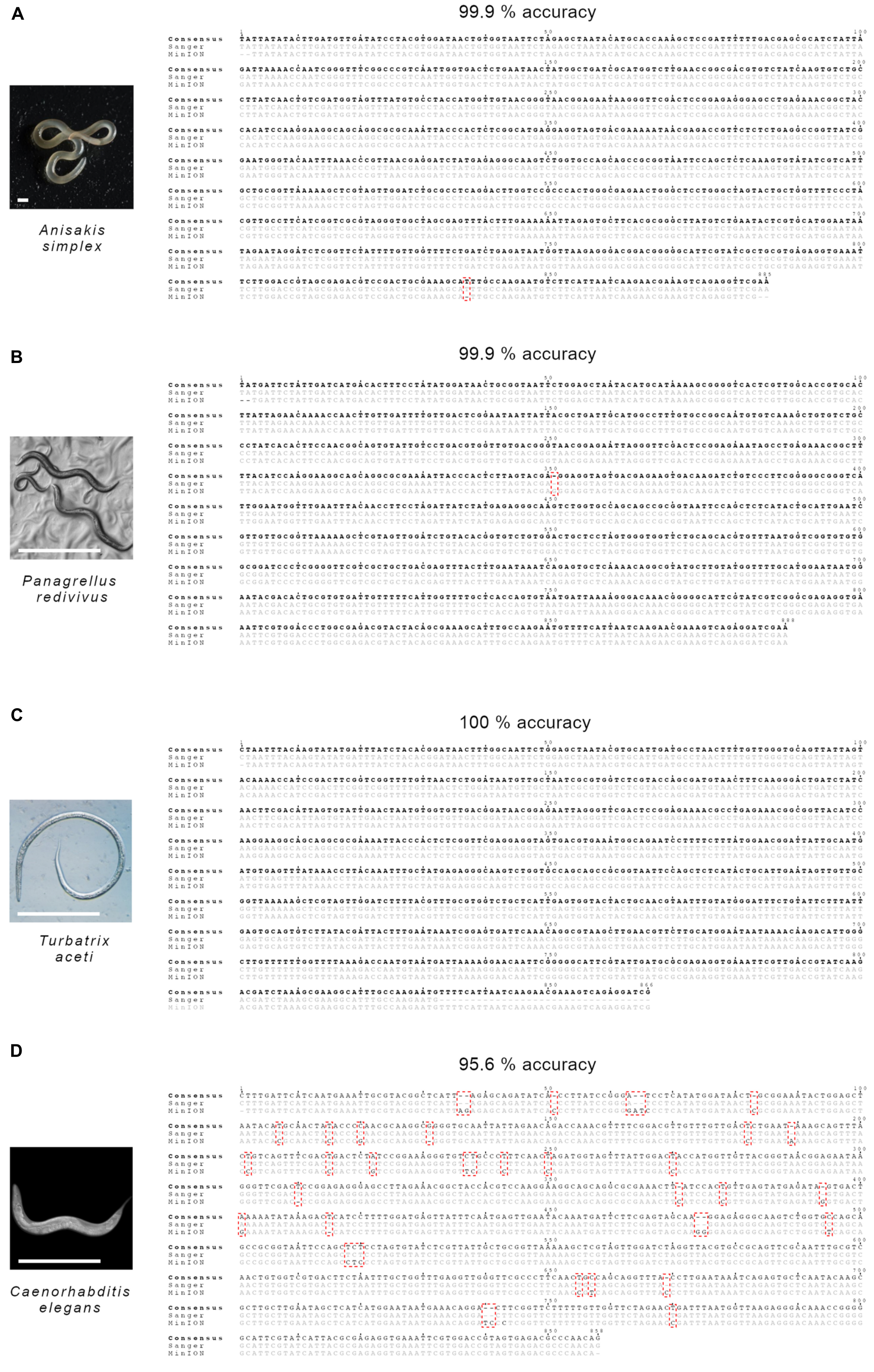


FIGURE 2 | Species investigated and nucleotide alignments of MinION and Sanger sequences comparing consensus accuracy for **(A)** *A. simplex*, **(B)** *P. redivivus*, **(C)** *T. aceti*, and **(D)** *C. elegans*. Sanger sequences have a 100% accuracy. Accuracy shown is the accuracy of the MinION consensus reads. Comparison against accession numbers MF072711.1 (*A. simplex*), AF083007.1 (*P. redivivus*), AF202165.2 (*T. aceti*), and MN519140.1 (*C. elegans*). The scale bar in the photographs of **(A–C)** represents 1 mm and in **(D)** represents 0.5 mm.

identifications from MinION reads that have an accuracy of 99.9–100%, when compared to their respective Sanger reads. Our second run yielded a successful species identification from the MinION consensus read that has an accuracy of 95.6%, when compared to Sanger sequencing. The Sanger reads of all samples matched 100% to a sequence of the correct species on NCBI.

The *C. elegans* data gives a considerably lower accuracy of its MinION consensus sequence than the three other species. We are confident that this is unrelated to the preparation of the sample on a portable setup, as we have applied this setup in a field situation and got MinION consensus sequences that matched closer to the correct species on NCBI (Knot et al., unpublished data). The *C. elegans* run was set up to test the portable setup before bringing all the equipment out to the field. As such, the primary objective was to prove that everything worked, with data quality being a lower priority. Hence, when the library was loaded and the MinION flow cell MUX scan indicated only 43 working pores, we continued the experiment nonetheless. The MinION generated data for 15 min, after which no active pores remained and the run was stopped. The flow cell that was used twice before, once in a 24 h run and once in a 2.5 h run, multiplexing nine samples in two runs, generating 1,950,657 reads containing 2.11 Gb in total. We therefore feel that the limited accuracy in the *C. elegans* run has more to do with the limited lifespan of a flow cell, than with the sample preparation or the portable sample preparation. We used the EXP-WSH002 wash kit from ONT, which has now been succeeded by the EXP-WSH003 kit. This latest kit incorporates a nuclease to digest and remove nucleic acid that has been loaded onto a flow cell previously and has proven to be much more efficient in maintaining flow cell quality after a wash than the previous wash kit[12]. We therefore do not expect future MinION flow cells to deteriorate as much after using the wash kit as was the case for our *C. elegans* run.

Tackling a large phylum like Nematoda presents challenges that other phyla might be less affected by Kumar et al. (2012). For example, we started exploring the primers developed by Floyd et al. (2005), because these primers are optimized for a wide phylogenetic range of nematodes. However, addition of the MinION tails seems to alter the efficiency of these primers. The tailed primers amplified *A. simplex*, *P. redivivus* and *T. aceti* successfully, but the addition of MinION tails prohibited the primers to amplify *C. elegans* successfully. We then switched to a primer optimized specifically for soil nematodes (Floyd et al., 2002), and found that this primer with MinION tails amplified *C. elegans* without problems. Future work will benefit from testing a wider array of nematode primers.

The potential throughput of a MinION R9 chemistry flow cell is ∼20 Gb (Krehenwinkel et al., 2019b), and has been shown to be sufficient to generate new draft genomes of nematodes through shotgun sequencing (Eccles et al., 2018; Fauver et al., 2019). Future barcoding work could focus on maximizing the utility of each flow cell by multiplexing samples in one run or harvesting the long-read potential unique to TGS platforms like PacBio and ONT. For example, Srivathsan et al. (2019) have developed an improved low-cost MinION pipeline where they multiplexed 3,500 samples per flow cell. However, preparing so many samples for sequencing requires significant labor time (Krehenwinkel et al., 2019b; Piper et al., 2019; Srivathsan et al., 2019). Heeger et al. (2018) used PacBio circular consensus sequencing to show the feasibility of long-read metabarcoding of environmental samples using a ∼4,500 bp ribosomal DNA marker that included most of the eukaryote SSU and LSU rRNA genes and the complete ITS region. Krehenwinkel et al. (2019a,b) showed that long-read barcoding using the MinION of a similar ribosomal DNA region, spanning ∼4,000 bp, has great potential for *in situ* species identification too, although degraded DNA can be a limiting factor in generating long-read barcodes. Small scale projects that do not require such high throughput could alternatively focus on using the newly released Flongle flow cell instead of a traditional MinION flow cell, at a cost of $90 instead of $475-$900 (depending on number of flow cells purchased), respectively. The membrane in this flow cell contains less nanopores to generate a throughput of 1–2 Gb, to accommodate projects with lower throughput demands (Krehenwinkel et al., 2019b). There is a trade-off between flexibility, where a project can sequence samples whenever they want, and cost-effectiveness, where a project can sequence as many samples as possible to get the lowest possible costs per sample. The latter is highly unlikely to be necessary and achievable in very remote regions, given the previously mentioned time restrictions this places on projects.

CONCLUSION

The use of the MinION opens up exciting possibilities for next-generation biomonitoring. The high efficiency of the MinION consensi compared to the Sanger sequences shows that the MinION can be used to identify diverse nematode species. Extrapolating our results to potential application in a field setting, our results suggest that barcoding with the MinION can generate enough reads for reliable identification within 15 min, assuming good DNA quality and depending on the number of samples that are multiplexed. Our study shows the potential for barcoding eukaryotes and can aid biomonitoring of invertebrate species. Optimizing portable sequencing methods for nematode identification is the first step to sequencing nematode species in the field. One of the challenges ahead for TGS of nematode species lies in the identification of nematodes species from mixed samples from complex natural environments like soil, marine sediments or feces. This challenge could be overcome in several ways. Improvements in the underlying MinION technologies is crucial and will improve accuracy and decrease error rates, just as previous improvements have already shown (Eisenstein, 2019). Further optimization of the bioinformatics analyses is also of high importance. Improved algorithms will lead to higher accuracy in species identification. These improvements will open up possibilities like near real-time genetic identification of nematodes from e.g., soil or feces, which would allow for analyses

[12] https://store.nanoporetech.com/flow-cell-wash-kit-r9.html

of soil nematodes as indicator of soil environment disturbance or rapid parasite identification.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are publicly available. The MinION dataset can be found in the European Nucleotide Archive under the project ID PRJEB37489 (samples ERS4397495, ERS4397496, ERS4397497, ERS4397498; and runs ERR4030416, ERR4030415, ERR4030417, ERR4030418, respectively). The Sanger sequences are available at GenBank under the accession numbers MT246663, MT246664, MT246665, and MT246666. All the code used for the bioinformatic analyses and additional files necessary to replicate the analyses, including a detailed explanation of dataset content, can be found on https://github.com/ieknot/MinION-DNA-barcoding-of-nematodes.

AUTHOR CONTRIBUTIONS

SW, RR, and IK designed the study. IK performed DNA extraction, prepped the Sanger sequencing samples, and performed the nanopore experiments. GZ, RR, and IK optimized the primers. GZ and IK performed PCRs. GW and IK optimized the bioinformatics pipeline. IK, GW, and RR wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was financially supported by the World Wildlife Fund-Netherlands and the University of Amsterdam.

ACKNOWLEDGMENTS

We thank Anne-Sophie Crunchant and Noémie Bonnin for discussions about and feedback on this manuscript, and Yannis Ancele for his help with some of the bioinformatics. We thank Anna de Luis Otero, Prof. Arthur Edison and Prof. Ian Chin-Sang for permission to use the pictures of *A. simplex, P. redivivus*, and *C. elegans*, respectively.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.00100/full#supplementary-material

The MinION consensus FASTA files correspond to the species as follows: **Data Sheet 1** is *A. simplex*, **Data Sheet 2** is *P. redivivus*, **Data Sheet 3** is *T. aceti* and **Data Sheet 4** is *C. elegans*.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Dj Lipman. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Anderson, R. C. (2000). "Nematode parasites of vertebrates," in *Their Development and Transmission*, ed. R. C. Anderson (Wallingford: CABI Publishing).

Benítez-Páez, A., Portune, K. J., and Sanz, Y. (2016). species-level resolution of 16S RRNA gene amplicons sequenced through the MinIONTM portable nanopore sequencer. *GigaScience* 5, 1–9. doi: 10.1186/s13742-016-0111-z

Bhadury, P., Mc Austen, D. T., Bilton, P. J. D., Lambshead, A. D., Rogers, A. D Smerdon, G. R., et al. (2006a). Development and evaluation of a DNA-barcoding approach for the rapid identification of nematodes. *Mar. Ecol. Prog. Ser.* 320, 1–9. doi: 10.3354/meps320001

Bhadury, P., Mc Austen, Bilton, D. T., Lambshead, P. J. D., and Rogers, A. D. (2006b). Molecular detection of marine nematodes from environmental samples. overcoming eukaryotic interference. *Aqu. Microb. Ecol.* 44, 97–103. doi: 10.3354/ame044097

Blaxter, M. L., De Ley, J. R., Garey, L. X., Liu, P., Scheldeman, A., Vierstraete, J. R., et al. (1998). A molecular evolutionary framework for the phylum nematoda. *Nature* 392, 71–75.

Brenner, S. (1974). The genetics of caenorhabditis elegans. *Genetics* 7, 71–94. doi: 10.1002/cbic.200300625

Brüggemann, J. (2012). Nematodes as live food in larviculture – a review. *J. World Aquacul. Soc.* 43, 739–763. doi: 10.1111/j.1749-7345.2012.00608.x

Chai, J. Y., Murrel, K. D., and Lymbery A. J. (2005). Fish-borne parasitic zoonoses. Status and issues. *Int. J. Parasitol.* 35, 1233–1254. doi: 10.1016/j.ijpara.2005.07.013

Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. *Nat. Biotechnol.* 34, 518–524. doi: 10.1038/nbt.3423

Decraemer, W., and Baujard, P. (1998). A polytomous key for the identification of species of the family trichodoridae thorne, 1935 (Nematoda. Triplonchida). *Fundam. Appl. Nematol.* 21, 37–62.

Eccles, D., Chandler, J., Camberis, M., Henrissat, B., Koren, S., Le Gros, G., et al. (2018). De novo assembly of the complex genome of nippostrongylus brasiliensis using MinION long reads. *BMC Biol.* 16:6. doi: 10.1186/s12915-017-0473-4

Edgar, R. C. (2004). MUSCLE, multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Edwards, A., Ar Debbonaire, S. M., Nicholls, S. M. E., Rassner, B., Sattler, J. M., Cook, T., et al. (2016). In-field metagenome and 16S RRNA gene amplicon nanopore sequencing robustly characterize glacier microbiota. *BioRxiv* 1–40. [Preprint]. doi: 10.1101/073965

Eisenstein, M. (2019). Playing a long game. *Nat. Methods* 16, 683–686. doi: 10.1038/s41592-019-0507-7

Faria, R. N., Quick, J., Morales, I., Thézé, J., Jesus, J. G., Giovanetti, M., et al. (2017). Establishment and cryptic transmission of zika virus in Brazil and the Americas. *Nature* 546, 406–410. doi: 10.1038/nature22401

Fauver, J. R., Martin, J., Weil, G. J., Mitreva, M., and Fischer, P. U. (2019). De novo assembly of the brugia malayi genome using long reads from a single MinION flowcell. *Sci. Rep.* 9, 1–10. doi: 10.1038/s41598-019-55908-y

Floyd, R., Abebe, E., Papert, A., and Blaxter, M. (2002). Molecular barcodes for soil nematode identification. *Mol. Ecol.* 11, 839–850.

Floyd, R. M., Ad Rogers, P. J. D., Lambshead, and Cr Smith. (2005). Nematode-specific PCR primers for the 18S small subunit RRNA gene. *Mol. Ecol. Notes* 5, 611–612. doi: 10.1111/j.1471-8286.2005.01009.x

Goodwin, S., McPherson, JD, and McCombie, WR. (2016). Coming of age, ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi: 10.1038/nrg.2016.49

Goordial, J. I, Altshuler, K., Hindson, K., Chan-Yam, E., Marcolefas, and Whyte, LG. (2017). In situ field sequencing and life detection in remote (79°26'N) canadian high arctic permafrost ice wedge microbial communities. *Front. Microbiol.* 8:2594. doi: 10.3389/fmicb.2017.02594

Gouy, M., Guindon, S., and Gascuel, O. (2010). Sea view version 4, A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* 27, 221–224. doi: 10.1093/molbev/msp259

Greninger, A. L., Sn Naccache, S., Federman, G., Yu, P., Mbala, V., Bres, D., et al. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med.* 7, 1–13. doi: 10.1186/s13073-015-0220-9

Heeger, F., Yurkov, A., Mazzoni, C. J., Bourne, E. C., Spröer, C., Overmann, J., et al. (2018). Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. *Mol. Ecol. Resour.* 18, 1500–1514. doi: 10.1111/1755-0998.12937

Hope, W. D., and Aryuthaka, C. (2009). A partial revision of the marine nematode genus elzalia (Monhysterida, Xyalidae) with new characters and descriptions of two new species from khung kraben bay, east thailand. *J. Nematol.* 41, 64–83.

Hunt, V. L., Ij Tsai, A., Coghlan, A. J., Reid, N., Holroyd, B. J., Foth, A., et al. (2016). The Genomic Basis of Parasitism in the Strongyloides Clade of Nematodes. *Nat. Genet.* 48, 299–307. doi: 10.1038/ng.3495

Jain, M., Olsen, H. E., Paten, B., and Akeson, M. (2016). The Oxford nanopore MinION, delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17, 1–11. doi: 10.1186/s13059-016-1103-0l

Johnson, S. S., Zaikova, E., Goerlitz, D. S., Bai, Y., and Tighe, S. W. (2017). Real-Time DNA sequencing in the antarctic dry valleys using the Oxford nanopore sequencer. *J. Biomol. Tech.* 28, 2–7. doi: 10.7171/jbt.17-2801-009

Karanastasi, E., Decraemer, W., Zheng, J., Martins De Almeida, M. T., and Brown, D. (2001). Interspecific differences in the fine structure of the body cuticle of trichodoridae thorne, 1935 (Nematoda, Diphtherophorina) and review of anchoring structures of the epidermis. *Nematology* 3, 525–533. doi: 10.1163/156854101753389130

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT, a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Kircher, M., and Kelso, J. (2010). High-throughput DNA sequencing – concepts and limitations. *BioEssays* 32, 524–536. doi: 10.1002/bies.200900181

Krehenwinkel, H., Pomerantz, A., Henderson, J. B., Kennedy, S. R., Lim, J. Y., Swamy, V., et al. (2019a). Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. *GigaScience* 8, 1–16. doi: 10.1093/gigascience/giz006

Krehenwinkel, H., Pomerantz, A., and Prost, S. (2019b). Genetic biomonitoring and biodiversity assessment using portable sequencing technologies. *Curr. Future Dir.. Genes* 10, 1–16. doi: 10.3390/genes10110858

Kumar, S., Koutsovoulos, G., Kaur, G., and Blaxter, M. (2012). Toward 959 nematode genomes. *Worm* 1, 42–50. doi: 10.4161/worm.19046

Lambshead, P. J. D. (2004). "Marine nematode biodiversity," in *Nematology, Advances and Perspectives*, eds Z. X. Chen, S. Y. Chen, and D. W. Dickson (Oxford University Press), 1.

Lawton, J. H., De Bignell, B., Bolton, G. F., Bloemers, P., Eggleton, P. M., Hammond, M., et al. (1998). Biodiversity inventories, indicator taxa and effects of habitat modification in tropical forest. *Nature* 391, 72–76. doi: 10.4103/1658-354x.84113

Li, H. (2018). Minimap2, Pairwise Alignment For Nucleotide Sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Maestri, S., Cosentino, E., Paterno, M., Freitag, H., Garces, J. M., Marcolungo, L., et al. (2019). A rapid and accurate minion-based workflow for tracking species biodiversity in the field. *Genes* 10:468. doi: 10.3390/genes10060468

McIntyre, A. B. R., Rizzardi, L., Yu, A. M., Alexander, N., Rosen, G. L., Botkin, D. J., et al. (2016). Nanopore sequencing in microgravity. *NPJ Microgravity* 2, 1–9. doi: 10.1038/npjmgrav.2016.35

Menegon, M., Cantaloni, C., Rodriguez-Prieto, A., Centomo, C., Abdelfattah, A., Rossato, M., et al. (2017). On site DNA barcoding by nanopore sequencing. *PLoS One* 12:e0184741. doi: 10.1371/journal.pone.0184741

Nature Biotechnology (2016). *Bento Lab*. doi: 10.1038/nbt0516-455

Oberacker, P., Stepper, P., Bond, D. M., Höhn, S., Focken, J., Meyer, V., et al. (2019). Bio-on-magnetic-beads (BOMB), Open platform for high-throughput nucleic acid extraction and manipulation. *PLoS Biol.* 17: e3000107. doi: 10.1371/journal.pbio.3000107

O'Neil, N. J., Olsen, H. E., Snutch, T. P., Tyson, J. R., Hieter, P., and Jain, M. (2017). MinION-based long-read sequencing and assembly extends the caenorhabditis elegans reference genome. *Genome Res.* 28, 266–274. doi: 10.1101/gr.221184.117

Pafčo, B., Èížková, D., Kreisinger, J., Hasegawa, H., Vallo, P., Shutt, K., et al. (2018). Metabarcoding analysis of strongylid nematode diversity in two sympatric primate species. *Sci. Rep.* 8, 1–11. doi: 10.1038/s41598-018-24126-3

Parker, J., Helmstetter, A. J., Devey, D., Wilkinson, T., and Papadopulos, AST. (2017). Field-based species identification of closely-related plants using real-time nanopore sequencing. *Sci. Rep.* 7, 1–8. doi: 10.1038/s41598-017-08461-5

Piper, A. M., Batovska, J., Cogan, N. O. I., Weiss, J., Cunningham, J. P., Rodoni, B. C., et al. (2019). Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. *GigaScience* 8, 1–22. doi: 10.1093/gigascience/giz092

Plesivkova, D., Richards, R., and Harbison, S. (2019). A Review Of The Potential Of the MinIONTM single-molecule sequencing system for forensic applications. *Wiley Interdiscip. Rev. Forensic Sci.* 1:e1323. doi: 10.1002/wfs2.1323

Pomerantz, A., Peñafiel, N., Arteaga, A., Bustamante, L., Pichardo, F., Coloma, L. A., et al. (2018). Real-time DNA barcoding in a rainforest using nanopore sequencing, opportunities for rapid biodiversity assessments and local capacity building. *GigaScience* 7, 1–14. doi: 10.1093/gigascience/giy033

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., et al. (2015). Rapid Draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol.* 16, 1–14. doi: 10.1186/s13059-015-0677-2

Quick, J., Nd Grubaugh, S. T., Pullan, I. M., Claro, A. D., Smith, K., Gangavarapu, G., et al. (2017). Multiplex PCR method for MinION and illumina sequencing of zika and other virus genomes directly from clinical samples. *Nat. Protoc.* 12, 1261–1266. doi: 10.1038/nprot.2017.066

Quick, J., Nj Loman, S., Duraffour, J. T., Simpson, E., Severi, L., Cowley, J., et al. (2016). Real-Time, portable genome sequencing for ebola surveillance. *Nature* 530, 228–232. doi: 10.1038/nature16996

Reiss, U., and Rothstein, M. (1975). Age related changes in isocitrate lyase from the free living nematode, turbatrix aceti. *J. Biol. Chem.* 250, 826–830.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH, a versatile open source tool for metagenomics. *PeerJ* 4:e2584. doi: 10.7717/peerj.2584

Schmidt, K., Mwaigwisya, S., Crossman, L. C., Doumith, M., Munroe, D., Pires, C., et al. (2017). Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J. f Antimicrob. Chemother.* 72, 104–114. doi: 10.1093/jac/dkw397

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nat. Methods* 5, 16–18. doi: 10.1038/nmeth1156

Seesao, Y., Gay, M., Merlin, S., Viscogliosi, E., Aliouat-Denis, C. M., and Audebert, C. (2017). A review of methods for nematode identification. *J. Microbiol. Methods* 138, 37–49. doi: 10.1016/j.mimet.2016.05.030

Shokralla, S., Porter, T. M., Gibson, J. F., Dobosz, R., Janzen, D. H., Hallwachs, W., et al. (2015). Massively parallel multiplex dna sequencing for specimen identification using an illumina MiSeq platform. *Sci. Rep.* 5, 1–7. doi: 10.1038/srep09687

Srinivasan, J., Ar Dillman, M. G., Macchietto, L., Heikkinen, M., Lakso, K. M., Fracchia, I., et al. (2013). The draft genome and transcriptome of panagrellus redivivus are shaped by the harsh demands of a free-living lifestyle. *Genetics* 193, 1279–1295. doi: 10.1534/genetics.112.148809

Srivathsan, A., Hartop, E., Puniamoorthy, J., Lee, W. T., Kutty, S. N., Kurina, O., et al. (2019). Rapid, large-scale species discovery in hyperdiverse taxa using 1D MinION sequencing. *BMC Biol.* 17:96. doi: 10.1186/s12915-019-0706-9

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten Years of next-generation sequencing technology. *Trends Genet.* 30, 418–426. doi: 10.1016/j.tig.2014.07.001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

SYSTEMATIC REVIEW
published: 19 May 2020
doi: 10.3389/fevo.2020.00135

A Systematic Review of Sources of Variability and Uncertainty in eDNA Data for Environmental Monitoring

Chloé Mathieu[1†], *Syrie M. Hermans*[2], *Gavin Lear*[2], *Thomas R. Buckley*[2,3], *Kevin C. Lee*[1] *and Hannah L. Buckley*[1*]

[1] *School of Science, Auckland University of Technology, Auckland, New Zealand,* [2] *School of Biological Sciences, University of Auckland, Auckland, New Zealand,* [3] *Manaaki Whenua—Landcare Research, Auckland, New Zealand*

Environmental DNA (eDNA) is becoming a standard tool in environmental monitoring that aims to quantify spatiotemporal variation for the measurement and prediction of ecosystem change. eDNA surveys have complex workflows encompassing multiple decision-making steps in which uncertainties can accumulate due to field sampling design, molecular biology lab work, and bioinformatics analyses. We conducted a quantitative review of studies published prior to December 2017 ($n = 431$) that had sampled eDNA from a variety of ecosystems and that had explicitly accounted for variability and uncertainty associated with eDNA workflows, either in their study design (e.g., replication) or data analysis (e.g., statistically modeling the spatiotemporal variation). We recorded differences among research studies in their spatial and temporal study design, the detected scales of natural variation in the study taxa, and how researchers measured and addressed the multiple sources of variability and uncertainty associated with the eDNA workflow. We show that relatively few studies used eDNA to understand temporal variation in biodiversity compared to spatial variation, and fewer described how uncertainties were addressed. We recommend increasing the number of temporal studies and to account for both natural variation and sources of uncertainty, such as imperfect detection, when undertaking eDNA surveys. Of studies that quantified spatiotemporal variation, this review identified gaps in the scales over which researchers have observed these patterns. Increasing the number of long-term and broad-scale eDNA studies will improve understanding of how useful eDNA is at scales relevant for monitoring the effects of environmental changes such as climatic shifts or land use change. Even where sources of spatiotemporal variation and uncertainty were accounted for, the effort in quantifying this variation differed among the different steps in the eDNA process, from field, to laboratory and bioinformatics procedures, depending on the type of community studied (micro- vs. macro-organism communities). We recommend more consistent experimental and modeling methods, accounting for spatiotemporal variation, and uncertainty in eDNA collection, and analysis, and incorporation of prior knowledge of sources of variability via Bayesian modeling approaches to account for uncertainties such as imperfect detection, to generate robust diversity estimates and increase the comparability of eDNA datasets for environmental monitoring across space and time.

Keywords: bioinformatics, community, eDNA, experimental design, single taxon, spatiotemporal scale, uncertainty, variability

Edited by:
David Andrew Bohan,
INRA Centre Dijon Bourgogne
Franche-Comté, France

Reviewed by:
Zacchaeus Greg Compson,
University of New Brunswick
Fredericton, Canada
Agnès Bouchez,
INRA Centre Alpin de Recherche sur
les Réseaux Trophiques des
Ecosystèmes Limniques, France

***Correspondence:**
Hannah L. Buckley
hannah.buckley@aut.ac.nz

†Present address:
Chloé Mathieu,
Observatoire des Galliformes
de Montagne, Sévrier, France

Specialty section:
This article was submitted to
Environmental Informatics and
Remote Sensing,
a section of the journal
Frontiers in Ecology and Evolution

Received: *12 June 2019*
Accepted: *23 April 2020*
Published: *19 May 2020*

Citation:
Mathieu C, Hermans SM, Lear G,
Buckley TR, Lee KC and Buckley HL
(2020) A Systematic Review of
Sources of Variability and Uncertainty
in eDNA Data for Environmental
Monitoring. Front. Ecol. Evol. 8:135.
doi: 10.3389/fevo.2020.00135

INTRODUCTION

Understanding spatial and temporal community patterns and processes is fundamental for disentangling the processes underpinning current and past biodiversity (Levin, 1992; Eme et al., 2015), and for predicting future biodiversity patterns (Sandel and Smith, 2009); it is thus fundamental for conservation and environmental management (Levin, 1992). However, because organisms respond differently to processes operating at a variety of spatial and temporal scales, quantifying spatiotemporal patterns and making generalizations about the causal processes remains challenging (Levin, 1992). Our ability to fully and accurately estimate sources of uncertainty in biodiversity measurements is key to advancing our understanding of these spatiotemporal scales of variation. Sources of variation in spatiotemporal data may be due to natural variation in biodiversity through space and time, but may also be due to errors introduced during field sampling and other steps in the biodiversity estimation process (Chen et al., 2013). The advent of new molecular tools, accessible at an increasingly reasonable price, has facilitated the use of environmental DNA (eDNA) to quantify biodiversity patterns (Chave, 2013). This has been especially useful for improving our understanding of spatiotemporal variation in environmental microbial communities (Shade et al., 2018) since prior methods to classify taxa based on traits such as morphology, frequently group taxa that are unrelated by descent (Kysela et al., 2016). For molecular biodiversity assessments, taxa are instead normally identified by the analysis of short, but taxonomically informative, DNA regions (Taberlet et al., 2018). In comparison to traditional approaches in community ecology, DNA surveys can allow detection of multiple taxa, including cryptic species, simultaneously. Environmental DNA (or eDNA) that is excreted or shed from live and dead organisms can be extracted from environmental samples (such as soil, water, air, and feces) without the need to isolate or even sight a specific taxon (Taberlet et al., 2012) with the potential to improve detection of nocturnal, rare and transient species, as well as species dwelling in less visible habitats, such as underground. Consequently, eDNA methods can be less environmentally damaging (Rees et al., 2014) and more accurate than field sampling (Janosik and Johnston, 2015). Particularly following the development of next-generation sequencing, eDNA methods can also be cheaper than some traditional methods and are thus they are increasingly considered as viable for the monitoring of both micro- and macro-organisms (Holdaway et al., 2017; Taberlet et al., 2018). While eDNA methods have been applied to study microbial communities for over two decades (Ranjard et al., 2000) leading to well established methods and a wide range of studies, molecular research to investigate macro-organisms is still in its infancy, with ongoing method development and optimisation.

Despite growing enthusiasm among the scientific community for biodiversity analysis using eDNA methods, species detection from eDNA studies is imperfect (Schmidt et al., 2013). Environmental DNA surveys have complex workflows encompassing multiple decision-making steps over which errors can accumulate (**Table 1**). Detection errors can occur due to: (1) the incorrect detection of target taxa when they are absent (false positives) and (2) failing to detect the target taxa when they are present (false negatives; Darling and Mahon, 2011). These detection problems can be attributed to either method or process errors. DNA-based method errors include all the errors resulting from the multiple steps used in performing the eDNA survey protocol (Zinger et al., 2019). This encompasses field sampling, laboratory sample processing and bioinformatics steps. DNA-based process errors comprise all the errors due to natural variability in species' DNA concentration in space and time (Darling and Mahon, 2011). For instance, a site with a higher concentration of DNA for the species α at time t should have a higher probability of detection than another site, or than the same site at time $t + 1$ if the DNA concentration decreases over time. Consequently, the design and use of appropriate eDNA surveys are highly context-dependent and so the development of a standard protocol that can account for these uncertainties (i.e., that measures the error due to the study design) for all taxa and all conditions is challenging (Taberlet et al., 2018). Nevertheless, some recommendations for reducing errors and increasing research reproducibility have been made (Dickie et al., 2018; Zinger et al., 2019). Improvement in our understanding of how these sources of error vary over time and space is necessary to avoid poor estimates of diversity (Carini et al., 2017) and to correctly interpret ecosystem functioning. If not properly accounted for, these sources of uncertainty can bias our understanding of biodiversity patterns and potentially misinform critical management and conservation decisions (Darling and Mahon, 2011; Chen et al., 2013; Furlan et al., 2016). For example, in invasive species surveillance and monitoring, the risk of false positives is a significant concern for managers and stakeholders who are concerned with minimizing both expenditure and any inconvenience caused by having to implement unnecessary pest control actions (Darling and Mahon, 2011).

In this systematic review, we conduct a gap analysis of the literature to quantify scales of natural spatiotemporal variation detected by eDNA studies and identify the uncertainties introduced by field sampling design, laboratory choices and bioinformatics procedures that may impact the accuracy and reproducibility of present-day diversity assessments conducted using eDNA. We reviewed empirical research that used eDNA sampled from natural environments to assess (1) the disparity among these studies due to differences in their spatial and temporal scales of observation and (2) if, and how, researchers have measured and addressed the multiple sources of uncertainty associated with eDNA sampling, laboratory processing and bioinformatics. Our approach was to follow a quantitative review methodology (Pickering and Byrne, 2014) to detect trends and gaps in how uncertainty and variability are detected and dealt with in eDNA studies. Our review methods were designed to capture the variety of research methods that have been employed in eDNA studies (i) on macro-organisms (length body $\geq 500\,\mu$m) and micro-organisms (length body $< 500\,\mu$m) (Martiny et al., 2006), (ii) across levels of organization, from a single taxon to multi-taxon communities, and (iii) across ecosystems, from above- and below-ground terrestrial systems, to both freshwater and marine aquatic ecosystems. Our aim

TABLE 1 | Steps and decisions required as part of a typical eDNA workflow, and the potential biases or uncertainties introduced at each of these steps.

Decisions required at each step of the eDNA workflow	Uncertainty or biases introduced	Relevant references
Sample collection		
Number of samples to collect Volume of samples to collect Spatial distribution of samples Temporal distribution of samples	False negative detection of taxa due to: • Insufficient number or volume of samples to capture the true diversity • Spatial or temporal design insufficient to capture true diversity False positive detection of taxa due to: • Cross contamination between samples during collection	Cantera et al., 2019 Davis A. J. et al., 2018 Dickie et al., 2018 Goldberg et al., 2016 Zinger et al., 2019
Molecular laboratory processes		
Storage of samples after collection Sample pre-processing eDNA extraction method PCR protocol to use (primer selection, reagents, cycling conditions) DNA sequencing methods qPCR methods	False negative detection of taxa due to: • Improper storage of samples leading to DNA degradation • DNA extraction biases • PCR inhibitors present • Unsuitable primers or PCR protocol False positive detection of taxa due to: • Contamination, especially cross-contamination between samples • Relic DNA present • Index/barcode jumping Misleading abundance values due to: • DNA extraction biases • Preferential amplification of DNA from some organisms over others	Clarke et al., 2014 Davis N. M. et al., 2018 Dopheide et al., 2019 Goldberg et al., 2016 Hermans et al., 2018 Schnell et al., 2015 Taberlet et al., 2018 Zinger et al., 2019
Bioinformatic processes		
Quality control thresholds Algorithms for chimera removal and sequence clustering Databases and thresholds to use for taxonomic assignments	False negatives/lower biodiversity detection due to: • Excessive quality filtering leads to too few sequences remaining • Sequence clustering threshold inappropriate/too high • Inadequate coverage of target taxa in databases False positives/inflated biodiversity detection due to: • Insufficient quality filtering steps passing too many reads with sequencing errors • Insufficient chimera removal • Misclassification of reads, possible misclassification in database	Brown et al., 2015 Coissac et al., 2012 Nearing et al., 2018 Zinger et al., 2019

for identifying trends and existing knowledge gaps in the understanding of spatiotemporal variability and uncertainty in eDNA research was to provide clear recommendations as to how researchers may adapt future study designs to best account for spatiotemporal variation and uncertainty arising from the collection and analysis of eDNA data.

MATERIALS AND METHODS

Systematic, quantitative reviews (i) identify the research question, (ii) identify and test appropriate keywords by searching databases, (iii) review and consistently record data from papers identified in the searches, and (iv) summarize and record patterns emerging from the resultant data (Pickering and Byrne, 2014). Following this approach, our research question, "How are variability and uncertainty measured and accounted for in eDNA studies?", was answered by conducting four different topic searches with ISI Web of Science Core Collection in November 2017 to extract four types of studies from the literature (**Table 2**): (topic 1) ecological studies using eDNA, (topic 2) metagenomics studies applied in ecology, (topic 3) studies of spatiotemporal variation and (topic 4) studies on the quantification of uncertainty in the eDNA process. We also included articles cited by six papers that have reviewed eDNA methodologies (Jansson and Tas, 2014; Boetius et al., 2015; Cavicchioli, 2015; Zeglin, 2015; Battin et al., 2016; Fierer, 2017). We obtained in total an initial list of 2,589 articles, from which we excluded all studies without either any spatial replication or temporal replication, and those without measurements of variability detailed either in the main text or any supplementary material, i.e., those studies that did not take replicate measurements and therefore, spatiotemporal variation and uncertainties cannot be calculated (point samples). As this review focused on the use of eDNA in environmental monitoring, we also excluded studies that were reviews, meta-analyses, laboratory-based experiments, within-organism (microbiome) studies, relic DNA studies focusing on the composition of historic communities, not on eDNA, and all studies where samples were collected from living or artificial substrates from the initial list. A number of papers ($n = 31$) conducted independent analyses to address questions at multiple spatiotemporal scales [e.g., (Chen et al., 2014) used differing approaches to assess spatial variation occurring across small (i.e., cm) vs. large (i.e., km) spatial scales] and were therefore entered as multiple studies in the final results database. This resulted a final list of 399 papers, which provided data for 431 studies. While this list contains the studies that

TABLE 2 | Search strategies used for the selection of the reviewed studies.

Topic	Targeted studies	Keywords searched in web of science	Additional articles from recent review papers
1	Ecological studies using eDNA	(eDNA OR "environmental DNA" OR metabarcoding OR barcoding) AND (variation OR scale OR gradient OR change OR evolution OR dynamic) AND (structure OR distribution OR pattern OR temporal OR spatial OR biogeography OR macroecology OR geographic) NOT (gut)	None
2	Metagenomic studies in applied ecology	(metagenomic) AND (ecology) AND (variation OR scale OR gradient OR change OR evolution OR dynamic) AND (structure OR distribution OR pattern OR temporal OR spatial OR biogeography OR macroecology OR geographic) NOT (gut)	None
3	Studies of spatial and temporal variation in applied ecology	("microbial ecology" OR "environmental ecology") AND (spatial OR "temporal variation" OR "temporal scale" OR scale OR temporal OR dynamic) AND (pattern OR distribution OR temporal OR spatial OR biogeography OR macroecology OR geographic OR composition) NOT (gut) NOT (virus)	Fierer, 2017, Jansson and Tas, 2014 Boetius et al., 2015, Battin et al., 2016, Cavicchioli, 2015; Zeglin, 2015
4	Studies on improving eDNA methods	(eDNA OR "environmental DNA" OR metabarcoding OR barcoding) AND: ("Error detection" OR "uncertainty source" OR "eDNA relic" OR "uncertainty level" OR "imperfect sensitivity" OR "probability of detection" OR detection OR uncertainty OR "source of variation" OR "sources of variation" OR "experimental variability")	None

met our search criteria (**Table 2**) and conditions outlined above, it is inevitable that some eDNA research was not captured, for example due to them not including any of the keywords we used in our search. Every article was read to a level sufficient to extract all required data by the primary author (CM) with consultation with other co-authors on specific methods or terminology, where required. Data gathered from each article related to (1) the spatial and temporal design of the study and (2) how sources of uncertainty were accounted for. All articles included in the quantitative review are contained in **Table S1** and the raw data extracted are included in **Table S2**.

We recorded the "type of variation" (spatial, temporal, both, or none) and size of the organisms studied (micro-, macroscopic, or both). Definition of the "type of variation" was based on the results presented. For example, if a site was sampled more than once over time, but the results presented only spatial variation in the data, i.e., temporal variation was averaged for each spatial replicate, or was otherwise ignored in the analysis, it was scored as a spatial study. All "space-for-time" a.k.a. chronosequence studies, i.e., studies using spatial samples to infer temporal variation, were considered to be spatial. For instance, three successional vegetation stages (grassland, mosaic, and forest) of green alder (*Alnus viridis*) encroachment were sampled by Schwob et al. (2017) at a single time to study the temporal dynamics of microbial communities in subalpine soils. Other attributes extracted were the type of ecosystem (terrestrial or aquatic) and the taxonomic level studied, namely "community" (studying more than 10 taxa), "group of taxa" (studying fewer or 10 taxa) or a single taxon (studying one unique taxon or unique species). The category "group of taxa" represents a selection of taxa based on author defined similarities in their morphology, life history traits, or conservation status.

To study spatiotemporal design more specifically, studies without either temporal or spatial replication, i.e., studies simulating method uncertainties ($n = 17$), were excluded, resulting in a subset of 414 studies out of the original 431. From those articles remaining, the type of eDNA outputs used were extracted, i.e., if researchers conducted taxon based analysis [e.g., using a single qPCR assay as in (Erickson et al., 2016)], community-based analysis [e.g., using "metabarcoding" as in Dulias et al. (2017)] or assessed the diversity of a broader array of genes [e.g., the analysis of functional gene or shotgun metagenomics data as in Dopheide et al. (2015) and Jeffries et al. (2016), respectively]. We also extracted both the spatial and temporal extents of studies. The spatial extent was defined as an area polygon (in km^2) encompassing all samples collected in the horizontal plane. For example, for a publication studying fish along a depth gradient, the study extent was defined as the maximal surface delimited by all the sampling sites (horizontal surface); we did not take the depth (vertical plane) gradient into account. If the sampling sites were not clearly identified, we used the area or region given in the paper. If the area was not named, but a map or satellite image provided, we used appropriate tools (e.g., Google Earth Pro (https://www.google.com/earth) and QGIS QGIS Development Team (2018) to identify and measure the spatial extent. The temporal extent corresponded to the duration of the study, i.e., the amount of time that elapsed between the first and last temporal replicates.

The type of diversity used to describe a community was also extracted. Three categories were defined based on the number of taxa (taxonomic diversity), the evolutionary history (phylogenetic diversity), or the functional traits in a community (functional diversity). We measured the scales of spatial and temporal variation in the community, i.e., the distance in space and/or time of a community when a shift was observed. The temporal scale of community variations is the time elapsed (in days) between t_0 (initial time) and t_1 (time when a community shift was observed). The spatial variation scale corresponds to the average distance between places (in km), where a community shift was observed. For instance, if the authors reported a spatial change at the site scale, the spatial scale variation here is the average distance between sites. We used different tools including the R software environment (R Core Team, 2019), Google Earth Pro and QGIS to measure this spatial variation when it was not clearly specified in the text. We also noted when the authors did not observe any significant spatial or temporal variation in the community metric ("no change"), or where the nature of the spatial or temporal variation was not clearly significant ("not clear") or where there were "mixed effects", e.g., dependent on the taxon considered. In some other cases, authors noted the presence of significant spatial and/ or temporal variation in communities without providing the required information to quantify it ("unquantifiable changes").

To study if and how researchers measured and explicitly addressed the multiple sources of uncertainty associated with eDNA sampling in their data analysis and interpretation, we considered only those studies accounting for sources of uncertainty (164 studies out of the original 431), i.e., (a) the studies measuring uncertainty by experimentation, modeling or statistical analysis (quantification) and (b) the studies comparing eDNA survey results with previous knowledge or traditional surveys (comparison). We extracted (1) the type of errors studied (false negative, false positive, or both), (2) their sources (error process or method), and (3) the workflow step at which the errors were studied (sample collection, molecular laboratory work, or bioinformatics processing). For the studies quantifying sources of uncertainty, we categorized the type of uncertainty measured following the workflow steps outlined in **Table 1**. More details about each category are given in **Table S3**. If several sources of uncertainties were quantified in the same study, the study was counted more than once. For instance, if a study measured sources of uncertainty due to DNA temporal variability and storage conditions, then this study was counted twice. We also extracted the suggestions given by authors where possible (if there were none, we recorded "no clear suggestion") and categorized them by their degree of generalization, i.e., "weakly" (case-specific suggestions), "moderately" (suggestions only applicable to a large group of taxa) and "strongly" generalizable (suggestions that are potentially applicable to all studies). In the count of the studies providing suggestions, we counted any study providing at least one suggestion in any of these categories. In addition, based on the highly generalizable suggestions made by articles quantifying eDNA uncertainty, for all the studies (431 articles), we extracted four parameters for communities to see how these main suggestions were applied to decrease uncertainty; namely (1) the number of studies using an occupancy model (yes, no, or other), (2) the number of genes and primer sets used, (3) the sampling intensity in relation to the study extent (temporal and spatial), and (4) the number of replicates used during the lab work process (PCR and DNA extraction replicates). Non-parametric tests were used to test for significant differences among taxonomic levels and communities studied (macro- and micro-organisms) in terms of the number of genes, primer sets, and number of replicates used (Wilcoxon test and Kruskal-Wallis test with a Bonferroni correction). All analyses were conducted and figures generated within the R environment for statistical computing (R Core Team, 2019) implementing the tidyverse, dplyr, ggplot, ggalt, and gridExtra packages. Since molecular methods have changed radically over the years, substantially increasing the number of temporal and spatial samples that can be feasibly collected and analyzed, we not only undertook our analyses on the full dataset, but also on a datasets restricted only to next-generation sequencing and qPCR studies to ensure that that any observations and recommendations made for historical data are still accurate for interpretations of modern day methods; these additional analyses made little difference to the pattern of results and so are presented in **Figures S1–S3**.

RESULTS

The 431 research studies that accounted for spatiotemporal variation and/ or uncertainty in their eDNA workflow in some way spanned 21 years. The number of publications using eDNA increased over time (**Figure 1**), but remain dominated by microbial research (66%), compared to studies focusing on larger organisms (34%). Aquatic ecosystems (marine and freshwater) were the most frequently studied (76%), compared to terrestrial ecosystems (23%). Only 1% of studies sampled both ecosystem types. After 2010, the selected literature consisted mostly of studies at the community level (**Figure 1B**); most community analysis used high throughput DNA sequencing, whereas most single taxon studies used qPCR or targeted multiple genes to provide information the abundance of multiple taxa. Taxonomic-based investigations were used most frequently (92%); only a few studies used a metagenomic approach (3%). Where diversity patterns were assessed, these community level studies most often explored taxonomic diversity (92%), with functional and/or phylogenetic diversity less often reported (**Figure 2**). Additionally, most studies reported only one (69%) or two (27%) types of diversity measures.

Scales of Spatial and Temporal Variation

Most studies investigated spatial variation (67.8%); papers focusing only on temporal changes represented 4.2% of the studies, whereas 27.7% considered both spatial and temporal variation (**Figure 3**). Of those studying temporal variation, 68.7% were short-term studies (a year or less), while medium- (2–5 years) and long-term (>5 years) studies represented 25.4 and 6.0%, respectively. In contrast, spatial variation was studied more consistently across a range of scales.

A total of 272 articles working at the community level quantified the scale of variation across space and/or time. Among these, the majority assessed communities of microscopic organisms (∼90%). Significant spatial variation was observed

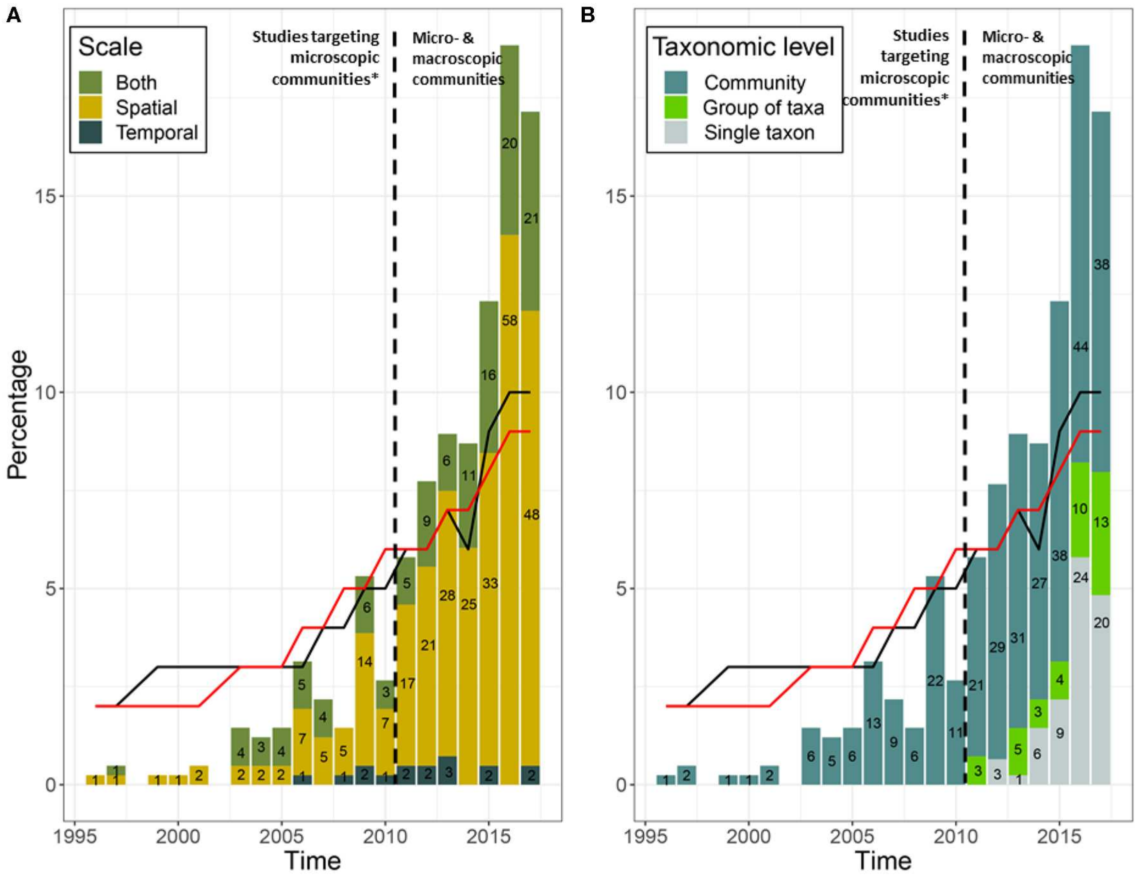


FIGURE 1 | (A) Spatiotemporal scales and **(B)** taxonomic levels studied (community, group of taxa, and single taxon) over time using eDNA. "Group of taxa" is defined as a group of fewer than 10 taxa clustered based on similarities in their morphology, life history traits, or conservation status. Numbers on the histogram bars represent the number of studies during that year. Black dashed lines show the beginning of eDNA studies on macro-organisms, with the exception of one study targeting macro-organisms by Bhadury et al. (2006). Black and red lines show the percentage of published papers in all scientific journals from 1996 to 2017 about micro- ($n = 26{,}535$) and macro-organisms ($n = 152{,}288$), respectively; data extracted from Web of Sciences using the key words "Microbiology" and "Ecology NOT Microbiology" respectively.

in both micro- and macroscopic communities (**Figure 4**, inner rings). In microscopic communities, significant spatial variation was observed at small (within 10 km), medium (between 10 and 1,000 km) and large (from 1,000 to 100,000 km) spatial scales (**Figure 4A**, outer ring). Macroscopic communities mostly showed spatial variation at small and medium scales (**Figure 4B**, outer ring). Few studies reported no change, unclear, or unquantifiable changes, or mixed effects depending on the specific taxa, land-uses, treatments, regions, sites or genes studied, or mixed effects depending on the laboratory, or bioinformatic methods, or the type of diversity measured (**Figure 4**, intermediate rings).

From the 29% of articles studying temporal changes within microbial communities, most quantified community changes occurring over short time scales (within 1 year). A few studies observed mixed effects depending on taxa, biome, site, or the gene studied (**Figure 4A**, intermediate ring). Temporal shifts in the composition of communities of macro-organisms were reported only across short timescales (1–6 months) in seven articles (**Figure 4B**, intermediate and outer rings).

Identifying Uncertainties Introduced During the eDNA Workflow

Our research identified 164 studies that explicitly measured and/ or modeled sources of uncertainty in their eDNA data. Of these, 50% measured uncertainty by quantification (using experimentation, modeling or statistical analysis), 42% by comparison (eDNA survey results vs. previous knowledge or traditional survey results; such data are easier to obtain for some groups of macro-organisms) and 8% using both approaches. The proportions of articles accounting for uncertainty in each ecosystem type were 44 and 17% for aquatic and terrestrial ecosystems, and 80% for both. Most studies measured uncertainty either at the community (41%) or single taxon (38%) level compared to 21% of studies that focused on a group of taxa (e.g., bony fish; Clusa et al., 2017). In 45% of the articles, the type of error measured was not clearly specified. However, 39% of articles reported examining both false positive and false negative detections compared to fewer articles which reported examining false positive (5%) and negative (10%) detections separately.

Within the sampling phase, both the spatiotemporal variability of the eDNA and uncertainty introduced by the experimental design were considered (**Figure 5**). Publications studying microbial communities paid more attention to experimental sampling design, sources of uncertainty due to size of the samples (volume/area sampled) and the number of field replicates. In contrast, papers studying communities of macro-organisms were focused on the detection of spatiotemporal variability related to sample collection. Sources of uncertainty due to the effect of pooling samples and the presence of positive controls were tested only for communities of macro-organisms (**Figure 5**). Within the laboratory work phase, the sources of uncertainty due to the storage conditions, selection of molecular parameters, extraction, and amplification protocols, as well as the sequencing were quantified; errors produced by sample pre-processing were only studied for macro-organisms [e.g., differences in eDNA detection comparing the collection of sample DNA that had been either filtered or centrifuged, as in Vörös et al. (2017)]. Within the bioinformatics phase, the quality control, clustering of operational taxonomic units (OTUs) and taxonomy assignment were investigated for both microbial and macro-organism communities; uncertainties due to the chimera detection were only investigated for microbial community data. Only 15 articles out of the 431 studies statistically modeled uncertainty; all of these were macroorganism studies. Of these, 14 used occupancy models, and one article used a "simulation and resampling" model (Deiner et al., 2016).

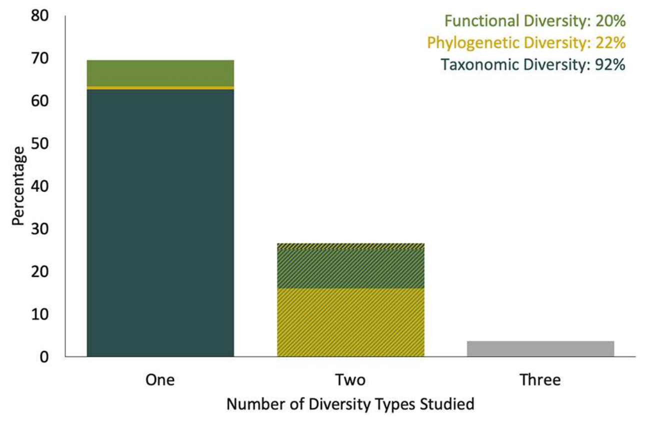


FIGURE 2 | Percent of reviewed community eDNA spatiotemporal studies that used different numbers and types of diversity measure (total number of studies = 293). Single diversity measures recorded were (bottom layer of stacked bar) taxonomic diversity, (center) phylogenetic and (top) functional diversity; combinations of two diversity measures studies included (bottom) taxonomic and phylogenetic diversity, (center) taxonomic and functional diversity and (top) functional and phylogenetic diversity. A similar plot, excluding studies that included no high-throughput DNA sequencing or qPCR data is available in **Figure S1**.

Suggestions of Ways to Account for Uncertainty in the eDNA Workflow

From the 164 articles quantifying sources of uncertainty, 60% did not provide any clear suggestions, 10, 17, and 14% gave weakly-, moderately- and strongly-generalisable suggestions, respectively (**Table 3** presents the strongly-generalisable suggestions; moderately- and weakly- generalisable suggestions are summarized in **Table S4**). The weakly-generalisable

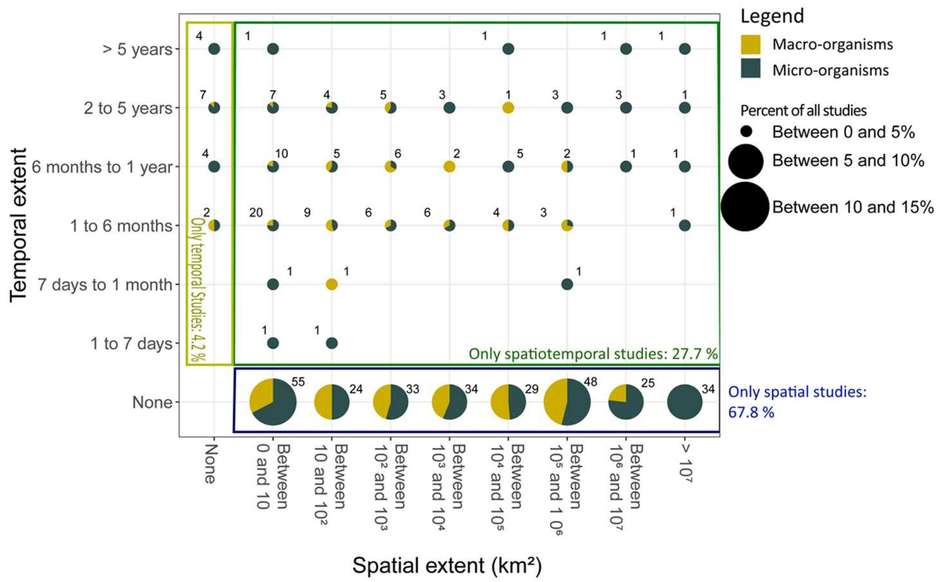


FIGURE 3 | The percent of studies using the eDNA of microscopic (dark green) and macroscopic (yellow) organisms that fell into different categories of spatial and temporal extent. The numbers within the plot correspond to the number studies reviewed in each category. Studies investigated spatial variation only (blue rectangle), temporal variation only (light green rectangle), or both spatial, and temporal variation (dark green rectangle). A breakdown of the spatial and temporal extent reported by the subset of studies which used "modern" molecular methods (next-generation sequencing or qPCR) is shown in **Figure S2**.

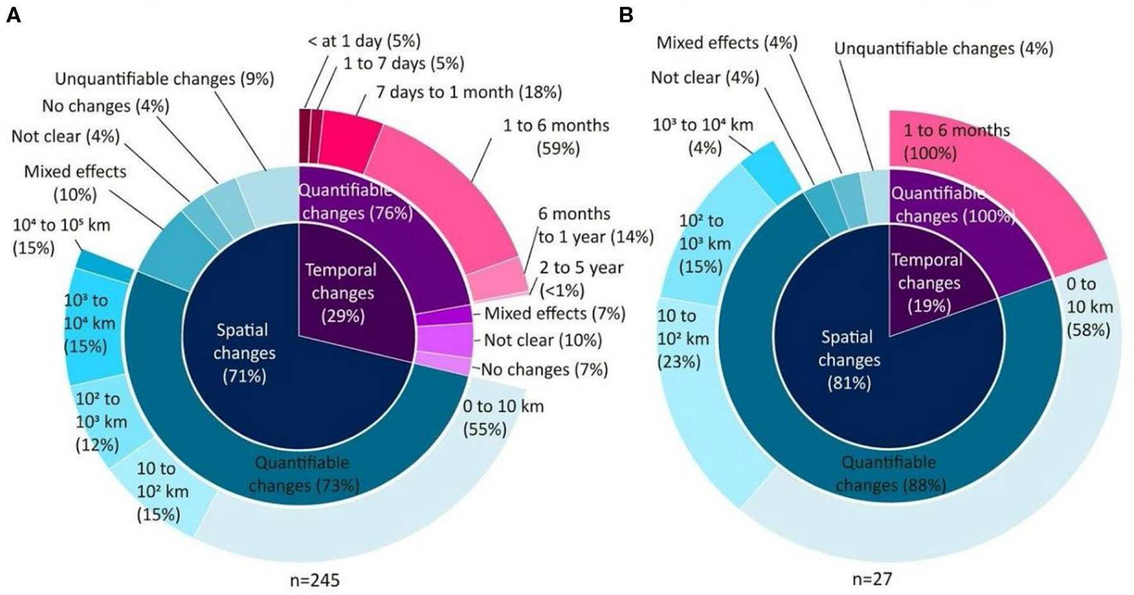


FIGURE 4 | Spatial and temporal scales of variation recorded from 272 articles working at the community level researching **(A)** microscopic and **(B)** macroscopic organisms. The inner circle indicates whether the variation was spatial (blue) or temporal (pink), the intermediate ring indicates whether or not the variation was quantified, and if quantified, the significant scales of the variation is indicated in the outer ring. The scale at which spatial and temporal variation was reported by the subset of studies which used "modern" molecular methods (next-generation sequencing or qPCR) is shown in **Figure S3**.

suggestions were often specific to a single taxon while moderately-generalisable suggestions were often applicable to either micro- or macroscopic communities (**Table S4**).

Highly-generalisable suggestions encouraged the use of pilot studies to obtain initial estimates of variability (Machler et al., 2016) and to facilitate adaptation of the study design for each specific study context (Deiner et al., 2015; Minamoto et al., 2016). Moreover, several studies suggested using hierarchical occupancy models, to evaluate if the level of replication is adequate to minimize detection errors (Ficetola et al., 2015; Lahoz-Monfort et al., 2016; Guillera-Arroita, 2017) and to use at least two independent sources of data (such as field observations in addition to eDNA) to account for false positive detections (Guillera-Arroita, 2017). Multiple marker genes were recommended regarding sequencing library preparation for community studies (Guardiola et al., 2016; Evans et al., 2017) or multiple primer sets targeting the same gene (Jeon et al., 2008). Schloss (2010) advised variable region selection should be based on the availability of conserved PCR primers and the presence of databases with adequate data for taxonomic identification from those regions (Schloss, 2010). Ideas and recommendations regarding accounting for, or reducing uncertainty in the bioinformatics phase were primarily focused on the quality control steps, where authors remove low-quality reads by investigating technical and analytical aspects (Huse et al., 2007; Schloss, 2010). Recommendations regarding OTU clustering were to develop a group-specific clustering threshold (Brown et al., 2015) and the use of more-complete taxonomic and sequence reference databases (Brown et al., 2015; Somervuo et al., 2017).

DISCUSSION

By analyzing the patterns in spatiotemporal variation observed in eDNA research to date, we have revealed important research gaps, and therefore highlighted where our future research efforts might best be dedicated. Overall, our results corroborate the findings of previous work (Strayer et al., 2006; Fierer, 2017), that the study of temporal variation has been neglected and the proportion of temporal studies did not increase over the reviewed period. In addition, relatively few studies have measured temporal variation over scales longer than 1 year. Second, there were relatively few studies that explicitly quantified and accounted for spatiotemporal variation and/ or uncertainties introduced by the eDNA sampling and analysis workflow, e.g., by statistical modeling or comparison of different protocols. Even where such sources of spatiotemporal variation were accounted for, the effort in quantifying this variation varied widely between the different steps in the eDNA process, from sampling, to laboratory, to bioinformatics procedures, depending on the type of organisms studied (micro- vs. macro-organisms). Finally, in contrast to taxonomic diversity, there were no studies showing how sources of spatial or temporal variation from eDNA studies affect functional and phylogenetic diversity estimates; such studies can be conducted using readily-available functional trait (e.g., the plant trait database TRY, https://www.try-db.org/TryWeb/Home.php; PICRUSt, Langille et al., 2013) and phylogenetic (e.g., the fish tree of life, https://fishtreeoflife.org/) data and would greatly inform the design and usefulness of eDNA studies for environmental monitoring. The above findings hold true when considering only newer methods, i.e., high-throughput

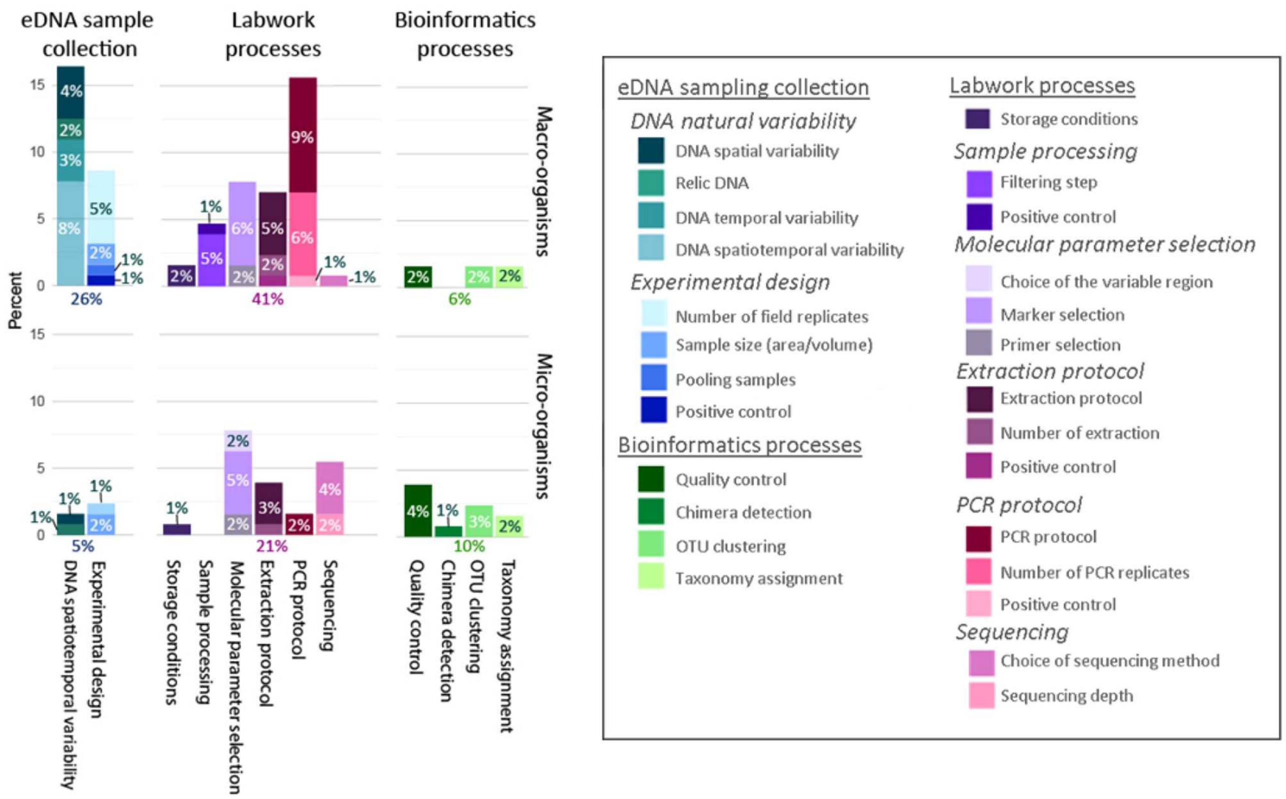


FIGURE 5 | How uncertainty was quantified differently across all reviewed studies on microscopic ($n = 41$ studies) and macroscopic ($n = 123$ studies) communities. The percent of studies that quantified uncertainty at each of the different steps in the eDNA workflow: eDNA field sample collection (blue colors), laboratory processing (pink colors), and bioinformatics (green colors).

DNA sequencing. Below, we discuss the gaps we identified in our systematic reviews, and importantly, make recommendations on ways to mitigate, and account for, spatiotemporal variation and uncertainties for future eDNA research that aims to inform environmental monitoring.

When designing appropriate eDNA studies for environmental monitoring, understanding the natural spatiotemporal scales of variation among organisms (i.e., where are targeted organisms distributed and how do these patterns vary over time?) and their DNA (i.e., where is eDNA distributed and what is its longevity in different environments?) sampled by eDNA surveys will greatly assist in ensuring sampling and replication are targeting the most informative sources of spatiotemporal variation (e.g., Lear et al., 2014; Ellis et al., 2015; Barata et al., 2017). While our review shows that spatiotemporal research in both micro- and macro-organisms has been conducted at various scales, improving our understanding requires an increase in the number of spatiotemporal studies, particularly over longer temporal scales. To be able to detect and quantify environmental changes for a given system, it is vital that we understand the ratios of spatial to temporal variation. For example, when using eDNA to monitor environmental change due to land use change or climatic change, it is crucial to understand how much spatial variation is expected between replicates compared to that expected over time. If spatial variation is relatively high, it may be difficult to detect temporal change, even over longer time periods.

Additional eDNA-specific effects need to be considered when monitoring environmental change. For example, part of the natural scale of temporal variation in eDNA data is the ability of DNA to persist in the soil or other substrates after an organism has left or died (sometimes referred to as relic DNA). Measures of natural spatiotemporal variation can be included in statistical models estimating taxon occupancy and diversity. However, temporal variation was accounted for, or measured in, only three studies in this review (Pilliod et al., 2014; Balasingham et al., 2017; Carini et al., 2017). For example, in their study, Carini et al. (2017) showed that relic DNA can increase the richness estimation of prokaryotic and fungal soil communities up to 55% and can also bias relative abundance estimates of taxa. This can affect our understanding of present-day biodiversity patterns. The analysis of environmental RNA rather than DNA provides additional opportunities for the analysis of present day diversity owing to the more transient nature of single-stranded RNA molecules, as suggested by Pochon et al. (2017) and Zaiko et al. (2018); to date however, the analysis of environmental RNA has been poorly investigated as a tool for biodiversity monitoring.

For eDNA surveys, uncertainty in the detection of taxa depends not only on natural DNA variability within the

TABLE 3 | Table of highly generalizable suggestions from the reviewed literature for measuring variability and reducing uncertainties in eDNA data at each point in the eDNA workflow.

Type of variability or uncertainty quantified	Suggestions	References from database (See Table S1)
Sample collection		
DNA natural variability: temporal	Remove relic DNA.	Carini et al., 2017
DNA natural variability: spatial	Effective eDNA sampling methods should be informed by species' distributions.	Eichmiller et al., 2014
Experimental design: Number of replicates	(1) Run occupancy models, (2) evaluate the rate of false positives, and (3) evaluate if the level of replication level is appropriate to control for false negatives. Use at least two sources of data to validate results, i.e., from two different types of survey.	Ficetola et al., 2015; Guillera-Arroita et al., 2017
Experimental design: Size of the samples	Conduct a pilot study to measure the variation so that an adequate number of samples can be determined.	Machler et al., 2016
Experimental design: Controls	Use negative controls.	Furlan and Gleeson, 2017
Experimental design: Pooled samples	Avoid pooling when estimating richness, except when comparing among sites.	Sato et al., 2017
Molecular laboratory processes		
Storage conditions	Use consistent treatment of samples, either freezing (at similar temperature) or unfrozen and process the samples quickly.	Docherty et al., , 2015; Takahara et al., 2015; Weltz et al., 2017
eDNA extraction: Protocol	Consider the biases caused by the extraction protocols and adjust according to the research question or context.	Deiner et al., 2015; Minamoto et al., 2016
eDNA extraction: Controls	Use negative controls when extracting DNA from samples of water, air etc.	Furlan and Gleeson, 2017; Spens et al., 2017
eDNA amplification: Marker selection	Use a marker that has a well-developed reference database of sequences. Use multiple markers, where appropriate, e.g., where many different taxa are being targeted.	Clarke et al., 2017; Evans et al., 2017
eDNA amplification: Choice of the variable region	The region selection should be based on the availability of conserved PCR primers and on the availability of database sequences for that region.	Schloss, 2010
eDNA amplification: Primer selection	Use multiple PCR primer sets to increase sequence coverage.	Jeon et al., 2008
eDNA amplification: PCR protocol	When using qPCR: (1) a primer set targeting plant chloroplast that evaluates the presence of amplifiable DNA from field samples to increase confidence in a negative result, (2) an animal group primer set to increase confidence in the assay result, and (3) a species-specific primer set to assess presence of DNA from the target species.	Veldhoen et al., 2016
eDNA amplification: Number of PCR replicates	Run occupancy models to estimate detection probabilities and rate of false presences. This can be used to evaluate whether the level of replication is adequate to control for false negatives. If necessary, "uncertain presences," not confirmed by multiple PCRs, can be removed. Occupancy models can incorporate prior information regarding the presence of organisms from an independent survey method that is not prone to false-positive errors.	Ficetola et al., 2015; Lahoz-Monfort et al., 2016; Guillera-Arroita et al., 2017
eDNA amplification: Controls	Use a secondary, generic primer designed to co-amplify endogenous DNA sampled during species-specific eDNA surveys.	Furlan and Gleeson, 2017
Sequencing	Use next-generation sequencing methods.	Terrat et al., 2015
Bioinformatic processes		
Quality control	Investigate the effects of bioinformatics protocols on the ability to accurately generate high-quality sequences and classify them. Remove all reads containing one or more single ambiguous base and ones whose lengths are outside the main distribution. Consider the effect of fragment length.	Huse et al., 2007; Schloss, 2010
Chimera detection	Use and compare several different algorithms.	Quince et al., 2011
OTU clustering and taxonomic assignment	Develop and use well-populated and regulated sequence databases that allow individual reads to be used directly for taxonomic assignment, without the need for OTU clustering. Consider developing a group-specific clustering threshold for clustering OTUs.	Brown et al., 2015, 2016; Somervuo et al., 2017

spatiotemporal scales studied, but also the eDNA survey method itself (Furlan et al., 2016). An understanding of the levels of variability in the field and lab is critical for determining the appropriate number of replicates required to reduce the variability in diversity estimates or probability of occupancy. Despite this, less than 40% of the studies in our review accounted for the uncertainty related to technical aspects of eDNA research. The most reliable and reproducible way to do this is to conduct pilot studies to quantify the spatiotemporal variation and its effects on diversity or occupancy measurements in advance of designing an eDNA survey. Our review shows that, although there are knowledge gaps particularly at the larger scales, significant information exists in the literature regarding spatiotemporal variation in eDNA that could be used to justify spatiotemporal study design choices and/ or incorporated into the subsequent analysis of eDNA data, thus avoiding additional costs by drawing on this existing knowledge. Similarly, methodological studies of technical variation introduced during the lab work or bioinformatics phases of the workflow can generate similarly useful data (e.g., Edgar, 2017, 2018; Davis A. J. et al., 2018; Hermans et al., 2018; Nearing et al., 2018; Dopheide et al., 2019). These data can be used to design a pilot study for a system that has not been previously worked on or used as prior information in statistical modeling of variation and uncertainty, e.g., to inform Bayesian priors and other forms of statistical probability inference.

Hierarchical Bayesian occupancy models can use occurrence or abundance data obtained from eDNA surveys to quantify natural spatiotemporal variation while statistically accounting for uncertainty at multiple spatial and temporal scales and can be applied in studies at the single-taxon, groups or community levels (e.g., Tyre et al., 2003; Kéry and Royle, 2009; Kery et al., 2009; Yamaura et al., 2012; Guillera-Arroita, 2017; Doi et al., 2019; Wineland et al., 2019). The development of these methods is an active area of research (e.g., Hui, 2016; Ovaskainen et al., 2017; Tobler et al., 2019), and as computing power continues to increase, such modeling has increasing potential to make powerful contributions to our understanding because they allow the incorporation of differences in organism detection, which, when ignored, can bias estimates of diversity (Iknayan et al., 2014). Few studies in our review had used this type of model (14 out of 331 articles). Similarly, Kellner and Swihart (2014) have shown that the majority (77%) of ecological studies using traditional survey methods do not account for imperfect detection. Within the relatively small number of studies in this review that accounted for uncertainty, only a few (24%) were on communities of micro-organisms. About half of the articles accounting for uncertainty compared eDNA survey results with previous knowledge or traditional survey results. This way of accounting for uncertainty is primarily used by researchers working with macroorganisms, since taxonomic analyses of microbial community composition are almost exclusively undertaken using molecular methods. One requirement of hierarchical occupancy models is that they have the appropriate data for spatiotemporal variation and uncertainties to quantify detection probabilities for taxa under the study conditions (Guillera-Arroita, 2017). However, obtaining detection probability data is not as difficult as may be expected. Indeed, as we have shown, some knowledge of the spatiotemporal scales of variation from eDNA studies exists and such data can readily be incorporated into explanatory and predictive modeling of eDNA data as Bayesian priors.

Further work needs to be undertaken on how to incorporate uncertainty from occupancy modeling into modeling diversity measurements in a way that is useful for environmental monitoring of communities and ecosystems (Denes et al., 2015; Dorazio et al., 2015). More than half of the studies on communities in this review (52%) analyzed only one type of diversity measure and, in most cases, it was taxonomic diversity. However, a focus on other measurements of diversity (functional and phylogenetic diversity) is required to better understand, not only spatiotemporal variation in community composition, but also the functional role and evolutionary history of these communities and ultimately the complex interactions among composition, function, and the evolutionary processes that shape their assembly over time and space (Pavoine and Bonsall, 2011; Fierer, 2017). No studies in this review investigated the relationship between uncertainty in eDNA methods and temporal variation in functional and phylogenetic diversity measures; indeed our search terms identified only nine studies using metagenomics methods to explore the functional diversity of communities via eDNA analysis. This is a key knowledge gap in eDNA study design. For example, if functional redundancy among taxa is present, fewer field, and laboratory replicates might be required to detect functional shifts in the studied ecosystem. Quantifying the sources of uncertainty across spatiotemporal scales in taxonomic, functional, and phylogenetic measurements of diversity will help to improve study designs and, therefore, make better recommendations for environmental management.

Research using eDNA spans a wide range of questions and requires an interdisciplinary methodological approach encompassing many methods, meaning it can be difficult to make widely applicable recommendations (Zinger et al., 2019). Nonetheless, we have collated suggestions that can be applied to improve accuracy and reproducibility at all steps of the eDNA workflow (**Table 3**); the largest portion of recommendations focusses on the molecular process, which is often inconsistent between research labs. Some researchers, such as Lear et al. (2018), have proposed standardized sets of methods to overcome bias impeding data comparison. However, it is clear that we need to continue putting more effort into improving eDNA methods, including the expansion of reference databases, improving targeted gene regions, and accounting for the errors and biases associated with sequencing technologies, many of which are quickly developing. In particular, researchers working on communities of macro-organisms need to do more to quantify sources of uncertainty due to bioinformatics. Crucially, the natural scales of variability and sources of uncertainty need to continue to be monitored throughout the eDNA process so we can better understand their implications.

CONCLUSIONS

Improving our understanding of both temporal and spatial variability of communities and their DNA will help us to answer both crucial methodological (e.g., what are the best sampling scales to detect environmental shifts? Where and when to sample?), and theoretical questions related to environmental monitoring (e.g., what are the factors driving temporal and spatial patterns? How can these patterns help us to measure ecosystem "health" or restoration success?). Moreover, regardless of the question that the researcher wishes to address, imperfect detection should be considered when working with ecological data using eDNA surveys, all the way from study design, through data collection, lab-work and bioinformatic processes. Multiple sources of uncertainty are present in all eDNA surveys, but robust replication in the field and laboratory can help quantify and minimize the detection errors. The combined use of prior knowledge of sources of variability from the literature or pilot studies within flexible statistical models that can incorporate these sources of information, will lead to more robust predictions of diversity and occupancy. Such experimental and modeling frameworks will also allow us to further explore the sensitivity of other biodiversity measures, such as functional and phylogenetic diversity, to detect errors and spatiotemporal variability at multiple scales. Environmental DNA is a promising method for environmental monitoring, but more research needs to be done to understand and quantity both natural spatiotemporal variation and technical variation introduced by study design and methods.

DATA AVAILABILITY STATEMENT

The raw data extracted from the 431 studies are available in **Table S2**.

AUTHOR CONTRIBUTIONS

HB and CM: conception and design of the study and interpretation. CM: data acquisition and analysis. CM, HB, GL, TB, SH, and KL: contributed to the writing of the manuscript.

FUNDING

This work was supported by funding from the New Zealand Ministry of Business, Innovation and Employment (New Zealand's Biological Heritage NSC, C09X1501).

ACKNOWLEDGMENTS

We thank Dave Bohan, Corinne Vacher, and five reviewers for many constructive comments that improved the paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.00135/full#supplementary-material

REFERENCES

Balasingham, K. D., Walter, R. P., and Heath, D. D. (2017). Residual eDNA detection sensitivity assessed by quantitative real-time PCR in a river ecosystem. *Mol. Ecol. Resour.* 17, 523–532. doi: 10.1111/1755-0998.12598

Barata, I. M., Griffiths, R. A., and Ridout, M. S. (2017). The power of monitoring: optimizing survey designs to detect occupancy changes in a rare amphibian population. *Sci. Rep.* 7:16491. doi: 10.1038/s41598-017-16534-8

Battin, T. J., Besemer, K., Bengtsson, M. M., Romani, A. M., and Packmann, A. I. (2016). The ecology and biogeochemistry of stream biofilms. *Nat. Rev. Microbiol.* 14, 251–263. doi: 10.1038/nrmicro.2016.15

Bhadury, P., Austen, M. C., Bilton, D. T., Lambshead, P. J. D., Rogers, A. D., and Smerdon, G. R. (2006). Molecular detection of marine nematodes from environmental samples: overcoming eukaryotic interference. *Aquat. Microbial. Ecol.* 44, 97–103. doi: 10.3354/ame044097

Boetius, A., Anesio, A. M., Deming, J. W., Mikucki, J. A., and Rapp, J. Z. (2015). Microbial ecology of the cryosphere: sea ice and glacial habitats. *Nat. Rev. Microbiol.* 13, 677–690. doi: 10.1038/nrmicro3522

Brown, E. A., Chain, F. J., Zhan, A., MacIsaac, H. J., and Cristescu, M. E. (2016). Early detection of aquatic invaders using metabarcoding reveals a high number of non-indigenous species in Canadian ports. *Div. Distribut.* 22, 1045–1059. doi: 10.1111/ddi.12465

Brown, E. A., Chain, F. J. J., Crease, T. J., MacIsaac, H. J., and Cristescu, M. E. (2015). Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? *Ecol. Evol.* 5, 2234–2251. doi: 10.1002/ece3.1485

Cantera, I., Cilleros, K., Valentini, A., Cerdan, A., Dejean, T., Iribar, A., et al. (2019). Optimizing environmental DNA sampling effort for fish inventories in tropical streams and rivers. *Sci. Rep.* 9:3085. doi: 10.1038/s41598-019-39399-5

Carini, P., Marsden, P. J., Leff, J., Morgan, E. E., Strickland, M. S., and Fierer, N. (2017). Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat. Microbiol.* 2:16242. doi: 10.1038/nmicrobiol.2016.242

Cavicchioli, R. (2015). Microbial ecology of Antarctic aquatic systems. *Nat. Rev. Microbiol.* 13, 691–706. doi: 10.1038/nrmicro3549

Chave, J. (2013). The problem of pattern and scale in ecology: what have we learned in 20 years? *Ecol. Lett.* 16, 4–16. doi: 10.1111/ele.12048

Chen, G., Kéry, M., Plattner, M., Ma, K., and Gardner, B. (2013). Imperfect detection is the rule rather than the exception in plant distribution studies. *J. Ecol.* 101, 183–191. doi: 10.1111/1365-2745.12021

Chen, Y., Zhen, Y., He, H., Lu, X., Mi, T., and Yu, Z. (2014). Diversity, abundance, and spatial distribution of ammonia-oxidizing β-proteobacteria in sediments from Changjiang estuary and its adjacent area in East China Sea. *Microb. Ecol.* 67, 788–803. doi: 10.1007/s00248-013-0341-x

Clarke, L. J., Beard, J. M., Swadling, K. M., and Deagle, B. E. (2017). Effect of marker choice and thermal cycling protocol on zooplankton DNA metabarcoding studies. *Ecol. Evol.* 7, 873–883. doi: 10.1002/ece3.2667

Clarke, L. J., Soubrier, J., Weyrich, L. S., and Cooper, A. (2014). Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Mol. Ecol. Resour.* 14, 1160–1170. doi: 10.1111/1755-0998.12265

Clusa, L., Ardura, A., Fernandez, S., Roca, A. A., and Garcia-Vazquez, E. (2017). An extremely sensitive nested PCR-RFLP mitochondrial marker for detection and identification of salmonids in eDNA from water samples. *PeerJ* 5:e3045. doi: 10.7717/peerj.3045

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

Darling, J. A., and Mahon, A. R. (2011). From molecules to management: adopting DNA-based methods for monitoring biological invasions in aquatic environments. *Environ. Res.* 111, 978–988. doi: 10.1016/j.envres.2011.02.001

Davis, A. J., Williams, K. E., Snow, N. P., Pepin, K. M., and Piaggio, A. J. (2018). Accounting for observation processes across multiple levels of uncertainty improves inference of species distributions and guides adaptive sampling of environmental DNA. *Ecol. Evol.* 8, 10879–10892. doi: 10.1002/ece3.4552

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., and Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 6:226. doi: 10.1186/s40168-018-0605-2

Deiner, K., Fronhofer, E. A., Machler, E., Walser, J. C., and Altermatt, F. (2016). Environmental DNA reveals that rivers are conveyer belts of biodiversity information. *Nat. Commun.* 7:12544. doi: 10.1038/ncomms12544

Deiner, K., Walser, J-C., Maechler, E., and Altermatt, F. (2015). Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biol. Conserv.* 183, 53–63. doi: 10.1016/j.biocon.2014.11.018

Denes, F. V., Silveira, L. F., and Beissinger, S. R. (2015). Estimating abundance of unmarked animal populations: accounting for imperfect detection and other sources of zero inflation. *Methods Ecol. Evol.* 6, 543–556. doi: 10.1111/2041-210X.12333

Dickie, I. A., Boyer, S., Buckley, H., Duncan, R. P., Gardner, P., and Hogg, I. D. (2018). Towards robust and repeatable sampling methods in eDNA based studies. *Mol. Ecol. Resour.* 18, 940–952. doi: 10.1111/1755-0998.12907

Docherty, K. M., Borton, H. M., Espinosa, N., Gebhardt, M., Gil-Loaiza, J., Gutknecht, J. L., et al. (2015). Key edaphic properties largely explain temporal and geographic variation in soil microbial communities across four biomes. *PLoS ONE* 10:e0135352. doi: 10.1371/journal.pone.0135352

Doi, H., Fukaya, K., Oka, S-I., Sato, K., Kondoh, M., and Miya, M. (2019). Evaluation of detection probabilities at the water-filtering and initial PCR steps in environmental DNA metabarcoding using a multispecies site occupancy model. *Sci. Rep.* 9:3581. doi: 10.1038/s41598-019-40233-1

Dopheide, A., Lear, G., He, Z., Zhou, J., and Lewis, G. D. (2015). Functional gene composition, diversity and redundancy in microbial stream biofilm communities. *PLoS ONE* 10:e123179. doi: 10.1371/journal.pone.0123179

Dopheide, A., Xie, D., Buckley, T. R., Drummond, A. J., and Newcomb, R. D. (2019). Impacts of DNA extraction and PCR on DNA metabarcoding estimates of soil biodiversity. *Methods Ecol Evol.* 10,120–133. doi: 10.1111/2041-210X.13086

Dorazio, R. M., Connor, E. F., and Askins, R. A. (2015). Estimating the effects of habitat and biological interactions in an avian community. *PLoS ONE* 10:e0135987. doi: 10.1371/journal.pone.0135987

Dulias, K., Stoof-Leichsenring, K. R., Pestryakova, L. A., and Herzschuh, U. (2017). Sedimentary DNA versus morphology in the analysis of diatom-environment relationships. *J. Paleolimnol.* 57, 51–66. doi: 10.1007/s10933-016-9926-y

Edgar, R. C. (2017). UNBIAS: an attempt to correct abundance bias in 16S sequencing, with limited success. *BioRxiv* 124149. doi: 10.1101/124149

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* 34, 2371–2375. doi: 10.1093/bioinformatics/bty113

Eichmiller, J. J., Bajer, P. G., and Sorensen, P. W. (2014). The relationship between the distribution of common carp and their environmental DNA in a small lake. *PLoS ONE* 9:e112611. doi: 10.1371/journal.pone.0112611

Ellis, M. M., Ivan, J. S., Tucker, J. M., and Schwartz, M. K. (2015). rSPACE: spatially based power analysis for conservation and ecology. *Methods Ecol. Evol.* 6, 621–625. doi: 10.1111/2041-210X.12369

Eme, D., Zagmajster, M., Fišer, C., Galassi, D., Marmonier, P., Stoch, F., et al. (2015). Multi-causality and spatial non-stationarity in the determinants of groundwater crustacean diversity in Europe. *Ecography* 38, 531–540. doi: 10.1111/ecog.01092

Erickson, R. A., Rees, C. B., Coulter, A. A., Merkes, C. M., McCalla, S. G., Touzinsky, K. F., et al. (2016). Detecting the movement and spawning activity of bigheaded carps with environmental DNA. *Mol. Ecol. Res.* 16, 957–965. doi: 10.1111/1755-0998.12533

Evans, N. T., Li, Y., Renshaw, M. A., Olds, B. P., Deiner, K., Turner, C. R., et al. (2017). Fish community assessment with eDNA metabarcoding: effects of sampling design and bioinformatic filtering. *Can. J. Fish. Aquat. Sci.* 74, 1362–1374. doi: 10.1139/cjfas-2016-0306

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., et al. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Mol. Ecol. Resour.* 15, 543–556. doi: 10.1111/1755-0998.12338

Fierer, N. (2017). Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat. Rev. Microbiol.* 15:579. doi: 10.1038/nrmicro.2017.87

Furlan, E. M., and Gleeson, D. (2017). Improving reliability in environmental DNA detection surveys through enhanced quality control. *Marine Freshw. Res.* 68, 388–395. doi: 10.1071/MF15349

Furlan, E. M., Gleeson, D., Hardy, C. M., and Duncan, R. P. (2016). A framework for estimating the sensitivity of eDNA surveys. *Mol. Ecol. Resour.* 16, 641–654. doi: 10.1111/1755-0998.12483

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., et al. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods Ecol. Evol.* 7, 1299–1307. doi: 10.1111/2041-210X.12595

Guardiola, M., Wangensteen, O. S., Taberlet, P., Coissac, E., Jesus Uriz, M., and Turon, X. (2016). Spatio-temporal monitoring of deep-sea communities using metabarcoding of sediment DNA and RNA. *PeerJ* 4:e2807. doi: 10.7717/peerj.2807

Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography* 40, 281–295. doi: 10.1111/ecog.02445

Guillera-Arroita, G., Lahoz-Monfort, J. J., van Rooyen, A. R., Weeks, A. R., and Tingley, R. (2017). Dealing with false-positive and false?negative errors about species occurrence at multiple levels. *Methods Ecol. Evol.* 8, 1081–1091. doi: 10.1111/2041-210X.12743

Hermans, S. M., Buckley, H. L., and Lear, G. (2018). Optimal extraction methods for the simultaneous analysis of DNA from diverse organisms and sample types. *Mol. Ecol. Resour.* 18, 557–569. doi: 10.1111/1755-0998.12762

Holdaway, R. J., Wood, J. R., Dickie, I. A., Orwin, K. H., Bellingham, P. J., Richardson, S. J., et al. (2017). Using DNA metabarcoding to assess New Zealand's terrestrial biodiversity. *New Zeal. J. Ecol.* 41, 251–262. doi: 10.20417/nzjecol.41.28

Hui, F. K. (2016). boral–Bayesian ordination and regression analysis of multivariate abundance data in R. *Methods Ecol. Evol.* 7, 744–750. doi: 10.1111/2041-210X.12514

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., and Mark Welch, D. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 8:R143. doi: 10.1186/gb-2007-8-7-r143

Iknayan, K. J., Tingley, M. W., Furnas, B. J., and Beissinger, S. R. (2014). Detecting diversity: emerging methods to estimate species diversity. *Trends Ecol. Evol.* 29, 97–106. doi: 10.1016/j.tree.2013.10.012

Janosik, A. M., and Johnston, C. E. (2015). Environmental DNA as an effective tool for detection of imperiled fishes. *Environ. Biol. Fish.* 98, 1889–1893. doi: 10.1007/s10641-015-0405-5

Jansson, J. K., and Tas, N. (2014). The microbial ecology of permafrost. *Nat. Rev. Microbiol.* 12, 414–425. doi: 10.1038/nrmicro3262

Jeffries, T. C., Schmitz Fontes, M. L., Harrison, D. P., Van-Dongen-Vogels, V., Eyre, B. D., Ralph, P. J., et al. (2016). Bacterioplankton dynamics within a large anthropogenically impacted urban estuary. *Front. Microbiol.* 6:1438. doi: 10.3389/fmicb.2015.01438

Jeon, S., Bunge, J., Leslin, C., Stoeck, T., Hong, S., and Epstein, S. S. (2008). Environmental rRNA inventories miss over half of protistan diversity. *BMC Microbiol.* 8:222. doi: 10.1186/1471-2180-8-222

Kellner, K. F., and Swihart, R. K. (2014). Accounting for imperfect detection in ecology: a quantitative review. *PLoS ONE* 9:e111436. doi: 10.1371/journal.pone.0111436

Kery, M., Dorazio, R. M., Soldaat, L., Van Strien, A., Zuiderwijk, A., and Royle, J. A. (2009). Trend estimation in populations with imperfect detection. *J. Appl. Ecol.* 46, 1163–1172. doi: 10.1111/j.1365-2664.2009.01724.x

Kéry, M., and Royle, J. A. (2009). "Inference about species richness and community structure using species-specific occupancy models in the national Swiss breeding bird survey MHB," in *Modeling Demographic Processes in Marked Populations*, eds D. L. Thomson, E. G. Cooch, M. J. Conroy (Boston, MA: Springer), 639–656. doi: 10.1007/978-0-387-78151-8_28

Kysela, D. T., Randich, A. M., Caccamo, P. D., and Brun, Y. V. (2016). Diversity takes shape: understanding the mechanistic and adaptive basis of bacterial morphology. *PLoS Biol.* 14:e1002565. doi: 10.1371/journal.pbio.1002565

Lahoz-Monfort, J. J., Guillera-Arroita, G., and Tingley, R. (2016). Statistical approaches to account for false-positive errors in environmental DNA samples. *Mol. Ecol. Resour.* 16, 673–685. doi: 10.1111/1755-0998.12486

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

Lear, G., Bellamy, J., Case, B. S., Lee, J. E., and Buckley, H. L. (2014). Fine-scale spatial patterns in bacterial community composition and function within freshwater ponds. *ISME J.* 8, 1715–1726. doi: 10.1038/ismej.2014.21

Lear, G., Dickie, I., Banks, J., Boyer, S., Buckley, H. L., Buckley, T. R., et al. (2018). Methods for the extraction, storage, amplification and sequencing of DNA from environmental samples. *New Zeal. Ecol. Soc.* 42, 10–50A. doi: 10.20417/nzjecol.42.9

Levin, S. A. (1992). The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology* 73, 1943–1967. doi: 10.2307/1941447

Machler, E., Deiner, K., Spahn, F., and Altermatt, F. (2016). Fishing in the water: effect of sampled water volume on environmental DNA-based detection of macroinvertebrates. *Environ. Sci. Technol.* 50, 305–312. doi: 10.1021/acs.est.5b04188

Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., et al. (2006). Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* 4, 102–112. doi: 10.1038/nrmicro1341

Minamoto, T., Naka, T., Moji, K., and Maruyama, A. (2016). Techniques for the practical collection of environmental DNA: filter selection, preservation, and extraction. *Limnology* 17, 23–32. doi: 10.1007/s10201-015-0457-4

Nearing, J. T., Douglas, G. M., Comeau, A. M., and Langille, M. G. I. (2018). Denoising the denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ* 6:e5364. doi: 10.7717/peerj.5364

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., et al. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* 20, 561–576. doi: 10.1111/ele.12757

Pavoine, S., and Bonsall, M. B. (2011). Measuring biodiversity to explain community assembly: a unified approach. *Biol. Rev.* 86, 792–812. doi: 10.1111/j.1469-185X.2010.00171.x

Pickering, C., and Byrne, J. (2014). The benefits of publishing systematic quantitative literature reviews for PhD candidates and other early-career researchers. *High. Educ. Res. Dev.* 33, 534–548. doi: 10.1080/07294360.2013.841651

Pilliod, D. S., Goldberg, C. S., Arkle, R. S., and Waits, L. P. (2014). Factors influencing detection of eDNA from a stream-dwelling amphibian. *Mol. Ecol. Resour.* 14, 109–116. doi: 10.1111/1755-0998.12159

Pochon, X., Zaiko, A., Fletcher, L. M., Laroche, O., and Wood, S. (2017). Wanted dead or alive? Using metabarcoding of environmental DNA and RNA to distinguish living assemblages for biosecurity applications. *PLoS ONE* 12:e0187636. doi: 10.1371/journal.pone.0187636

QGIS Development Team (2018). QGIS Geographic Information System. Open Source Geospatial Foundation Project. Available online at: http://qgis.osgeo.org.

Quince, C., Lanzen, A., Davenport, R. J., and Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics.* 12:38. doi: 10.1186/1471-2105-12-38

R Core Team. (2019). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/ (accessed May, 2020).

Ranjard, L., Poly, F., and Nazaret, S. (2000). Monitoring complex bacterial communities using culture-independent molecular techniques: application to soil environment. *Res. Microbiol.* 151, 167–177. doi: 10.1016/S0923-2508(00)00136-4

Rees, H. C., Bishop, K., Middleditch, D. J., Patmore, J. R. M., Maddison, B. C., and Gough, K. C. (2014). The application of eDNA for monitoring of the great crested newt in the UK. *Ecol. Evol.* 4, 4023–4032. doi: 10.1002/ece3.1272

Sandel, B., and Smith, A. B. (2009). Scale as a lurking factor: incorporating scale-dependence in experimental ecology. *Oikos* 118, 1284–1291. doi: 10.1111/j.1600-0706.2009.17421.x

Sato, H., Sogo, Y., Doi, H., and Yamanaka, H. (2017). Usefulness and limitations of sample pooling for environmental DNA metabarcoding of freshwater fish communities. *Sci. Rep.* 7, 1–12. doi: 10.1038/s41598-017-14978-6

Schloss, P. D. (2010). The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLOS Comput. Biol.* 6:e1000844. doi: 10.1371/journal.pcbi.1000844

Schmidt, B. R., Kery, M., Ursenbacher, S., Hyman, O. J., and Collins, J. P. (2013). Site occupancy models in the analysis of environmental DNA presence/absence surveys: a case study of an emerging amphibian pathogen. *Methods Ecol. Evol.* 4, 646–653. doi: 10.1111/2041-210X.12052

Schnell, I. B., Bohmann, K., and Gilbert, M. T. P. (2015). Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. *Mol. Ecol. Resour.* 15, 1289–1303. doi: 10.1111/1755-0998.12402

Schwob, G., Roy, M., Manzi, S., Pommier, T., and Fernandez, M. P. (2017). Green alder (*Alnus viridis*) encroachment shapes microbial communities in subalpine soils and impacts its bacterial or fungal symbionts differently. *Environ. Microbiol.* 19, 3235–3250. doi: 10.1111/1462-2920.13818

Shade, A., Dunn, R. R., Blowes, S. A., Keil, P., Bohannan, B. J. M., and Herrmann, M. (2018). Macroecology to unite all life, large and small. *Trends Ecol. Evol.* 33, 731–744. doi: 10.1016/j.tree.2018.08.005

Somervuo, P., Yu, D. W., Xu, C. C. Y., Ji, Y., Hultman, J., Wirta, H., et al. (2017). Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. *Methods Ecol. Evol.* 8, 398–407. doi: 10.1111/2041-210X.12721

Spens, J., Evans, A. R., Halfmaerten, D., Knudsen, S. W., Sengupta, M. E., Mak, S. S., et al. (2017). Comparison of capture and storage methods for aqueous macrobial eDNA using an optimized extraction protocol: advantage of enclosed filter. *Methods Ecol. Evol.* 8, 635–645. doi: 10.1111/2041-210X.12683

Strayer, D. L., Eviner, V. T., Jeschke, J. M., and Pace, M. L. (2006). Understanding the long-term effects of species invasions. *Trends Ecol. Evol.* 21, 645–651. doi: 10.1016/j.tree.2006.07.007

Taberlet, P., Coissac, E., Hajibabaei, M., and Rieseberg, L. H. (2012). Environmental DNA. *Mol. Ecol.* 21, 1789–1793. doi: 10.1111/j.1365-294X.2012.05542.x

Taberlet, P., Bonin, A., Zinger, L., and Coissac, E. (2018). *Environmental DNA: For Biodiversity Research and Monitoring*. Oxford: Oxford University Press. doi: 10.1093/oso/9780198767220.001.0001

Takahara, T., Minamoto, T., and Doi, H. (2015). Effects of sample processing on the detection rate of environmental DNA from the Common Carp (Cyprinus carpio). *Biol. Conserv.* 183, 64–69. doi: 10.1016/j.biocon.2014.11.014

Terrat, S., Plassart, P., Bourgeois, E., Ferreira, S., Dequiedt, S., Adele-Dit-De-Renseville, N., et al. (2015), Meta-barcoded evaluation of the ISO-11063 standard. *Microb. Biotechnol.* 8: 131–142. doi: 10.1111/1751-7915.12162

Tobler, M. W., Kéry, M., Hui, F. K., Guillera-Arroita, G., Knaus, P., and Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology* 100:e02754. doi: 10.1002/ecy.2754

Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K., and Possingham, H. P. (2003). Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecol. Appl.* 13, 1790–1801. doi: 10.1890/02-5078

Veldhoen, N., Hobbs, J., Ikonomou, G., Hii, M., Lesperance, M., and Helbing, C. C. (2016). Implementation of novel design features for qPCR-based eDNA assessment. *PLoS ONE* 11:e0164907. doi: 10.1371/journal.pone.0164907

Vörös, J., Márton, O., Schmidt, B. R., Gál, J. T., and Jelić, D. (2017). Surveying Europe's only cave-dwelling chordate species (proteus anguinus) using environmental DNA. *PLoS ONE* 12:e0170945. doi: 10.1371/journal.pone.0170945

Weltz, K., Lyle, J. M., Ovenden, J., Morgan, J. A., Moreno, D. A., and Semmens, J. M. (2017). Application of environmental DNA to detect an endangered marine skate species in the wild. *PLoS ONE* 12:e178124. doi: 10.1371/journal.pone.0178124

Wineland, S. M., Welch, S. M., Pauley, T. K., Apodaca, J. J., Olszack, M., Mosher, J. J., et al. (2019). Using environmental DNA and occupancy modelling to identify drivers of easrtern hellbender (*Cryptobranchus alleganiensis alleganiensis*) extirpation. *Freshwater Biol.* 64, 208–221. doi: 10.1111/fwb.13210

Yamaura, Y., Royle, J. A., Shimada, N., Asanuma, S., Sato, T., Taki, H., et al. (2012). Biodiversity of man-made open habitats in an underused country: a class of multispecies abundance models for count data. *Biodiv. Conserv.* 21, 1365–1380. doi: 10.1007/s10531-012-0244-z

Zaiko, A., Pochon, X., Garcia-Vazquex, E., Olenin, S., and Wood, S. A. (2018). Advantages and limitations of environmental DNA/RNA tools for marine biosecurity: management and surveillance of non-indigenous species. *Front. Mar. Sci.* 5:322. doi: 10.3389/fmars.2018.00322

Zeglin, L. H. (2015). Stream microbial diversity in response to environmental changes: review and synthesis of existing research. *Front. Microbiol.* 6:454. doi: 10.3389/fmicb.2015.00454

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., et al. (2019). DNA metabarcoding—need for robust experimental designs to draw sound ecological conclusions. *Mol. Ecol.* 28, 1857–1862. doi: 10.1111/mec.15060

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership