Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis

Teresita M. Porter1,2 | Mehrdad Hajibabaei1

1Centre for Biodiversity Genomics, Biodiversity Institute of Ontario and Department of Integrative Biology, University of Guelph, Guelph, ON, Canada
2Natural Resources Canada, Great Lakes Forestry Centre, Sault Ste. Marie, ON, Canada

Correspondence
Mehrdad Hajibabaei, Centre for Biodiversity Genomics & Department of Integrative Biology, University of Guelph, Guelph, ON, Canada.
Email: mhajibab@uoguelph.ca

Abstract
The purpose of this review is to present the most common and emerging DNA-based methods used to generate data for biodiversity and biomonitoring studies. As environmental assessment and monitoring programmes may require biodiversity information at multiple levels, we pay particular attention to the DNA metabarcoding method and discuss a number of bioinformatic tools and considerations for producing DNA-based indicators using operational taxonomic units (OTUs), taxa at a variety of ranks and community composition. By developing the capacity to harness the advantages provided by the newest technologies, investigators can “scale up” by increasing the number of samples and replicates processed, the frequency of sampling over time and space, and even the depth of sampling such as by sequencing more reads per sample or more markers per sample. The ability to scale up is made possible by the reduced hands-on time and cost per sample provided by the newest kits, platforms and software tools. Results gleaned from broad-scale monitoring will provide opportunities to address key scientific questions linked to biodiversity and its dynamics across time and space as well as being more relevant for policymakers, enabling science-based decision-making, and provide a greater socio-economic impact. As genomic approaches are continually evolving, we provide this guide to methods used in biodiversity genomics.

KEYWORDS
biodiversity, bioinformatics, biomonitoring, DNA barcode, environment, high-throughput sequencing, metabarcoding, metagenomics

1 INTRODUCTION

Biodiversity encompasses the diversity of organisms, their relationships and their functions within ecosystems. Biodiversity assessment using traditional methods mainly involves identifying morphological characters whose states can be compared with taxonomic keys for species identification. It is not uncommon that characters needed for taxonomic assignment are not present or are difficult to discern even for highly experienced taxonomists. The process of identifying smaller taxa such as insects or microscopic organisms from environmental samples often continues well beyond the field collection season, and the results are taxonomic assignments with varying degrees of resolution that is dependent on the availability of taxonomic keys, expertise of the taxonomist and condition of the sample. The issues vary by organism: for insects, damaged or juvenile specimens may not contain the characters needed for identification (Sweeney, Battle, Jackson, & Dapkey, 2011); for fungi, bacteria and other microscopic organisms, it may be the difficulty in isolating and culturing individuals or the collection of samples in life stages that lack the characters needed for identification (Bridge & Spooner, 2001). These are just a few of the impediments to the morphology-based identification process that also affect downstream users of taxonomic data for biodiversity research (Ebach, Valdecasas, & Wheeler, 2011). In biomonitoring studies, sampling needs to be repeated across time and space. Essentially, the pace that samples are collected in large-scale biomonitoring programmes can quickly exceed the capacity for...
taxonomists to identify them in a timely manner and this is where DNA-based methods can help investigators identify more taxa using methods capable of processing large data sets. In this review, we specifically address how DNA-based methods with high-throughput potential are scalable, that is, methods that can be easily adapted to process larger numbers of samples collected across time and space as well as automatically identify larger numbers of taxa from large-scale studies.

Biodiversity genomics integrates different data types such as biological indicators from traditional specimen collection, biomass estimates and biological activity assessments; environmental indicators that describe site characteristics; as well as DNA-based indicators which are the focus of this review (Figure 1). The use of DNA-based methods to both detect organisms and assign putative functions without having to isolate or identify individuals from environmental samples has been revolutionary in many different fields (Aylagas, Borja, Irigoien, & Rodríguez-Ezpeleta, 2016; Elbrecht & Leese, 2017; Gibson et al., 2015; Hajibabaei, Shokralla, Zhou, Singer, & Baird, 2011; Horton & Bruns, 2001; Langille et al., 2013; Nguyen et al., 2016; O’Brien, Parrent, Jackson, Moncalvo, & Vilgalys, 2005; Taberlet, Coissac, Hajibabaei, & Rieseberg, 2012; Torsvik & Ovreas, 2002). In particular, high-throughput sequencing (HTS) of environmental DNA (eDNA) for biomonitoring applications has been referred to as BIOMONITORING 2.0 (Baird & Hajibabaei, 2012). We use the term eDNA loosely to include the free degraded DNAs in the environment; DNA emitted from living organisms through their DNA secretions, faeces and shed cells; DNA contained within dead or dormant cells such as seeds, spores and sclerotia; as well as the DNA from whole organisms that are also recovered during the extraction process (Bohmann et al., 2014; Pietramellara et al., 2009; Taberlet, Prud’homme, et al., 2012). Biodiversity genomic methods are analogous in many ways to microbiome studies in agricultural or biomedical studies. Methods that start with total DNA extraction from a bulk sample such as soil, water or sediments can detect what organisms are present or the genomic potential of a community (Bik et al., 2012). The natural persistence and turnover of DNA in the environment result in a snapshot of community diversity within a certain window of time that depends on specimen biomass and density, temperature and trophic status of the system (Dejean et al., 2011). Biological indicators such as presence, abundance, or relative abundance of OTUs, taxa or communities can be derived from eDNA using the methods described below (Table 1, Box 1).

2 | MICROARRAYS

As shown in Table 1, this is a fluorescence-based method that can be used to detect individuals or a community of taxa or genes. This method enables highly parallel monitoring by miniaturizing traditional tube- or plate-based assays. Microscopic spots of oligonucleotide probes are attached to a solid surface called a microarray "chip" (Sauer et al., 2005; Schena et al., 1998). A fluorescently labelled sample containing sequences complementary to the probes is added to the chip. When hybridization occurs, fluorescence is measured. The use of microarrays in biodiversity research transformed the field by allowing pan-community assays to occur in a single reaction
without the complication of preparing gels for visualization (Gardner, Jaing, McLoughlin, & Slezak, 2010; Metfies & Medlin, 2005; Schena et al., 1998). Microarrays also enable samples to be evaluated with replication, an important consideration in microbial ecology (DeSantis et al., 2007). DNA microarrays can be used to answer questions about the presence and relative abundance of a known set of taxa using rDNA sequences or genes from mixed-community samples. For example, various chips (e.g., Virochip, GreeneChipPm, PhyloChip, Lawrence Livermore Microbial Detection Array) have been designed to obtain a snapshot of viral and microbial (bacteria, protozoa, fungi) diversity without the cost associated with the older generation of HTS platforms and without the time-consuming bioinformatics needed to process raw sequence data (Chou et al., 2006; DeSantis et al., 2006, 2007; Gardner et al., 2010; Palacios et al., 2007). Another way to assess biodiversity is to view it through a phylogenetic lens, which can be performed with PhyloChip results using the

<table>
<thead>
<tr>
<th>Method</th>
<th>Uses</th>
<th>Pros</th>
<th>Cons</th>
<th>Scalability</th>
<th>Future outlook</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microarrays (F;I/C)</td>
<td>Detects the presence or relative abundance of suites of markers or expressed transcripts from an individual specimen or a community sample</td>
<td>Simplified data analysis compared with the bioinformatics processing required with HTS methods</td>
<td>Not an appropriate method for taxon discovery. Data produced are fluorescence signals not sequences</td>
<td>Requires one single-use chip per sample. Not easily scalable</td>
<td>Likely to be replaced by sequencing-based methods below as cost per sequence continues to decrease and bioinformatic pipelines become easier to use</td>
</tr>
<tr>
<td>Quantitative PCR (qPCR) (F;I/C)</td>
<td>Detects the presence, abundance or relative abundance of a single marker from an individual specimen or a community sample</td>
<td>More sensitive than PCR. Can be used to quantify starting template amounts</td>
<td>Can be complicated by the background mixture of organismal DNAs co-isolated from eDNA samples. Data produced are fluorescence signals not sequences</td>
<td>Highly scalable as many samples can be run in batches on plates</td>
<td>Likely to remain a useful tool where rapid monitoring is needed to detect target taxa until integrated microfluidic devices are further developed and become a more cost-effective option</td>
</tr>
<tr>
<td>Digital PCR (dPCR) (F;I/C)</td>
<td>See qPCR above</td>
<td>See qPCR above. Digital PCR is more sensitive than PCR or qPCR. As only a single-template molecule is amplified per reaction, this avoids the problem of PCR bias with eDNA samples. Can be used as a target enrichment method prior to HTS</td>
<td>Requires the purchase of specialized equipment. Data produced are fluorescence signals not sequences</td>
<td>Current platforms are scalable in terms of the number of parallel PCRs reactions that occur, but typically, only a single sample can be run at a time</td>
<td>This platform is ideal for analysing eDNA samples, but the need to purchase specialized equipment may be a barrier to widespread adoption among laboratories that already have established qPCR protocols</td>
</tr>
<tr>
<td>Metabarcoding (S;C)</td>
<td>Gene marker surveys are currently the most popular method for biodiversity and DNA-based biomonitoring from a community sample</td>
<td>Reference databases and taxonomic assignment tools are available for the most widely used DNA barcoding regions. Phylogenetic and statistical analysis of ESVs or OTUs can be conducted even without the identification of metabarcodes</td>
<td>Identification of metabarcodes from eDNA samples relies on comparisons to potentially incomplete or inaccurate reference sequence databases. A single DNA marker, a short sequence or a lack of similar reference sequences may hinder fine-level taxonomic assignments</td>
<td>Highly scalable when combined with HTS and bioinformatic analysis in a HPC environment, see Box 3</td>
<td>This method is likely to be superseded by whole-genome assembly from metagenomic sequencing as: (i) single-molecule sequencing accuracy is improved, and (ii) as whole-genome sequence reference databases become more representative of environmental diversity</td>
</tr>
</tbody>
</table>

(Continues)
<table>
<thead>
<tr>
<th>Method</th>
<th>Uses</th>
<th>Pros</th>
<th>Cons</th>
<th>Scalability</th>
<th>Future outlook</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organelle sequencing (S;I)</td>
<td>Sequencing of organelle DNA, for example, mtDNA or cpDNA from an individual specimen</td>
<td>Can provide additional sequence information for improved taxonomic assignment</td>
<td>All the markers are linked and may only reflect the evolution of the organelle as opposed to the organism as a whole</td>
<td>This method is not as scalable as methods that simply isolate eDNA without the need for prior isolation of individuals</td>
<td>As single-molecule sequencing technologies develop, whole genome sequencing will likely become more routinely applied in biodiversity studies. This technology shift may replace reduced representation methods such as organelle sequencing and genome skimming</td>
</tr>
<tr>
<td>Genome skimming (S;I)</td>
<td>Low-coverage genome sequencing of an individual specimen</td>
<td>Can provide high sequencing coverage of high copy number or highly repetitive regions such as mtDNA, rDNA or cpDNA to improve taxonomic assignments. Can also sample additional markers that may be useful for phylogenomics</td>
<td>The isolation of a sufficient amount of high-quality DNA from taxa that are not readily cultured using standard methods or for those with small body size may be difficult. Single-copy markers of interest may not be detected</td>
<td>See Organelle sequencing above</td>
<td>See Organelle sequencing above</td>
</tr>
<tr>
<td>Whole-genome sequencing (WGS) (S;I)</td>
<td>Genome sequencing of an individual specimen</td>
<td>Provides a reference sequence for the further development of molecular markers and to aid taxonomic assignment of sequences derived from a community sample</td>
<td>The isolation of a sufficient amount of high-quality DNA from taxa that are not readily cultured using standard methods or for those with small body size may be difficult</td>
<td>See Organelle sequencing above</td>
<td>See Organelle sequencing above</td>
</tr>
<tr>
<td>Metagenomics (S;C)</td>
<td>Shotgun DNA sequencing of a community sample</td>
<td>Provides an overview of the metabolic potential and taxonomic diversity of a community. For simple communities comprised of taxa with small genomes, it is possible to reconstruct WGSs</td>
<td>Pathway genes or markers of interest may only be recovered at low frequency in a metagenomic data set. Taxonomic assignments based on markers that are not commonly used may not be possible</td>
<td>Highly scalable when combined with HTS and bioinformatic analysis in a HPC environment</td>
<td>As HTS methods develop and reference databases grow, this method may become more useful in biodiversity studies and biomonitoring applications</td>
</tr>
<tr>
<td>Metatranscriptomics (S;C)</td>
<td>RNA-seq (cDNA sequencing) of a community sample</td>
<td>Provides an overview of the actual metabolic activity and taxonomic diversity of a community</td>
<td>It can be challenging to extract high-quality mRNA from environmental samples</td>
<td>See Metagenomics above</td>
<td>Likely to see more widespread use as WGS reference databases become more representative of environmental diversity</td>
</tr>
</tbody>
</table>

F = Fluorescence-based detection; S = Sequencing-based detection; I = Used on an individual specimen; C = Used on mixed-community samples.
BOX 1  Glossary

Amplicon: The short DNA sequence products of polymerase chain reaction (PCR) amplification using taxon- or gene-specific primers to target a particular region of the genome.

Biodiversity: The diversity of life, their relationships and their functions within ecosystems.

Biodiversity genomics: Biodiversity assessed using high-throughput DNA-based methods or data from whole genomes integrated with a broad array of metadata describing biological and environmental indicators.

Biomonitoring: Biodiversity analysis that is repeated across space and time that may focus on a target organism such as invasive or at-risk species, an assemblage such as the bioindicator groups (amphibians, birds, macroinvertebrates) as an indicator of ecosystem status.

cDNA: Messenger RNA reverse-transcribed into its complementary DNA sequence.

DNA Barcoding: A minimal standardized signature DNA sequence is used for species identification, for example, a 658-bp region of CO1 mtDNA is used for identification of animals. Other DNA barcode markers have been proposed for fungi, plants and protists. 16S rDNA has been used for the identification of bacteria.

eDNA: Environmental DNA comprised of free degraded DNAs in the environment as well as DNA co-extracted from whole organisms such as microscopic organisms, arthropods, nematodes; shed cells; faeces; as well as the DNA contained within dead or dormant cells such as seeds or spores.

ESV: Exact sequence variant. Also known as an amplicon sequence variant (ASV), zero-radius OTU (ZOTU) or simply an OTU defined by 100% sequence similarity.

Genome: The complete set of genetic data contained in an organism including organellar DNA.

Genomics: The sequencing and analysis of the genetic material of an organism.

HPC: High-performance computing, computer clusters can be used to run the same analysis for many samples in parallel, or splitting large jobs into many smaller ones for a quicker overall runtime. Available through private clusters or third-party cloud computing services.

HTS: High-throughput sequencing, sometimes referred to as next-generation sequencing or second-generation sequencing. Distinguished by the high number of sequencing reactions that occur in parallel.

mRNA: Messenger RNA that encodes for a gene product.

Marker: A gene or signature region of DNA with a known location in the genome and can be used to identify individuals or species.

Metadata: Supplementary data linked to DNA sequences that provide information in a standard and searchable way such as organismal or bulk environmental sample description.

Metagenomics: The study of genetic material isolated directly from environmental samples, such as water, soil or sediments, may also be referred to as environmental genomics, ecogenomics or community genomics.

Metatranscriptomics: The study of the expressed portion of genomes, mRNAs, isolated directly from an environmental sample that may be transcribed into cDNAs for high-throughput sequencing.

Mito-metagenomics: The assembly of whole mitochondrial DNA sequences from eDNA samples.

MiP: Molecular inversion probe used for target enrichment.

Multiplex sequencing: The addition of a unique DNA sequence tag to each sample, such as when multiple samples are pooled and sequenced at the same time, allows sequences from different samples to be distinguished from each other during data analysis.

Oligonucleotides: Relatively short nucleotide molecules used as primers for PCR, as probes on microarrays, or baits during target enrichment.

OTU: Operational taxonomic unit, a group of similar DNA sequences sometimes used as a proxy for “species” in diversity measures.

Primers: Short oligonucleotides that are complementary to a particular region of the genome and are a starting point for DNA replication by DNA polymerase during PCR.

rDNA: Ribosomal DNA that codes for the ribosomal RNA subunits that form ribosomes.


Taxon: An organism identified to any taxonomic rank (e.g., species to kingdom); plural taxa.

TE: Target enrichment.

WGS: Whole-genome sequencing involves determining the complete DNA sequence of an organism’s genome, also known as complete genome sequencing, full-genome sequencing, entire genome sequencing.
reaction (qPCR) on the chip and may be as specific as the species level using a microarray assay depends on the specificity of the oligonucleotides.

Microarrays are also useful for detecting short degraded sequences such as those extracted from eDNA and have even been successfully applied to the analysis of highly degraded ancient DNAs (Devault et al., 2014).

The most attractive feature of DNA microarrays for biodiversity genomics, the ability to compile pan-community profiles, has been surpassed by other DNA-based HTS methods described below. A microarray assay may cost hundreds of dollars per sample depending on the chip and provider. For comparison, the cost of an Illumina MiSeq run can be divided by 96 or more samples depending on the desired sequencing coverage bringing the cost per sample down to $100 or less if many samples are run together, notwithstanding the time and cost associated with HTS bioinformatics. The resolution of a microarray assay depends on the specificity of the oligonucleotides on the chip and may be as specific as the species level using “detection” probes or more general using “discovery” probes that only bind to highly conserved regions (Gardner et al., 2010). Unfortunately, microarrays are not suitable for detecting novel taxa or genes on their own (DeSantis et al., 2007). This is a major limitation as the extent of environmental biodiversity has yet to be fully described (Hawksworth, 1991; Torsvik & Ovreas, 2002). Another limitation of this method is that only hybridization patterns are recorded not the DNA sequences, which could be otherwise used for sequence-based inference methods for biodiversity analysis or the development of new probes. As microarrays are normally designed for a single use, the number of chips that are purchased as well as access to specialized equipment to analyse the chips may limit the number of samples and replicates that can be processed. Overall, the scalability of this method is poor compared with methods below that can be run in a multiwell format for parallel batch sample processing. As shown in Table 1, we think microarrays will eventually be phased out in favour of DNA sequence-based methods (below) as bioinformatics tools become easier to use.

### 3 QUANTITATIVE POLYMERASE CHAIN REACTION (qPCR)

As shown in Table 1, this is a fluorescence-based method that can be used to detect individual species or genes in biodiversity studies. This is a PCR-based method similar to standard endpoint PCR except that light is emitted and measured during every cycle as new DNA is synthesized (Arya et al., 2005). Also called real-time PCR, this method can be used in a quantitative manner with reference to a standard curve or a semi-quantitative manner among samples. This method can also be used with reverse-transcription qPCR to measure gene expression levels. Quantitative PCR to detect and quantify taxa and functional genes such as those involved in nutrient cycling or biodegradation from environmental samples is attractive for monitoring applications as well as more general biodiversity analysis (Smith & Osborn, 2009). This method is appealing for biodiversity studies because of the enhanced sensitivity of qPCR, compared with standard PCR, making this method more suitable for detecting rare species of interest for biosecurity (e.g., alien invasive species) or conservation efforts (e.g., endangered species) (Wilcox et al., 2013). The genes in pathways activated upon exposure to toxins or involved directly in nutrient cycling are obvious targets for qPCR. In a biomonitoring application, reverse-transcription qPCR was used to detect the expression of two sets of biomarker genes in response to heavy metal exposure in a marine mussel (Banni et al., 2007). In an ecotoxicological study, qPCR was adapted to target very long ampli-cons to assess DNA damage in different parts of the genome (Meyer, 2010). The premise behind this assay is that DNA damage caused by genotoxins may inhibit DNA polymerase progression along the template and results in reduced amplification. To address the question of how long DNA persists in the environment and the window of time captured in eDNA samples, a freshwater mesocosm experiment found that DNA became undetectable by qPCR 2 weeks after removal of animals (Thomsen et al., 2012). This method is more sensitive and more expensive than standard PCR and ideally suited for tracking single or small suites of target taxa. Taxonomic resolution as well as prevalence of false-positive and false-negative results depend on the specificity of the primers designed for qPCR but can be used to detect species even in the presence of congeneric spe-cies (Wilcox et al., 2013). However, caution must be exercised when closely related species might occur in the same sample at different concentrations as the abundant templates may be preferentially amplified (Wilcox et al., 2013). An advantage of this method for large-scale biodiversity studies is that the equipment needed to run qPCRs can accommodate 96-well or 384-well plate formats and this allows for easy scalability to parallelize the analysis of many samples at once. As a fluorescence-based detection method, however, it does not provide the sequence information that could be used for sequence-based biodiversity analyses or for the development of new molecular probes and primers.

As qPCR is a sensitive method, there are a number of considera-tions that need to be accounted for specifically when working with mixed-community samples. First, the limits of detection would need to be determined for different types of environmental samples. In fact, any mixed-template PCR-based method is susceptible to similar challenges as different types of eDNA samples will contain a different background of DNAs from a community of organisms in addition to PCR inhibitors such as polysaccharides, humic acids, tannins and heavy metals (Braid, Daniels, & Kitts, 2003; Schrader, Schielke, Eller-broek, & Johne, 2012). Additionally, this method on its own is not
suitable for discovering novel genes or species because the primers used for qPCR are specifically designed to target only known species or genes. The primary advantages of using qPCR are sensitivity, scalability, cost and speed for diagnostic screening of target taxa; however, issues caused by the complex eDNA background may be better circumvented using digital PCR (below). As noted in Table 1, qPCR is likely to remain the method of choice wherever rapid monitoring of target taxa is needed, but this method could be supplanted by integrated microfluidic devices if they become a more cost-effective option in future (see Section 12).

4 | DIGITAL PCR

As shown in Table 1, this is another fluorescence-based detection method that can be used to quantify individuals or be used as a target enrichment method prior to HTS. Digital PCR is an alternative to traditional qPCR where a sample is separated into thousands of parallel PCRs each with a single- or no-template molecule (Vogelstein & Kninger, 1999). This method can be used to determine starting copy number without the use of reference standards. Depending on the platform, reactions are carried out on a microfluidic device or in their own individual micelle droplets. Chamber digital PCR (cdPCR) is when a microfluidic device is used for digital PCR and can be used for real-time DNA quantification. Digital droplet PCR (ddPCR) is when digital PCR is carried out in micelle droplets, and this can be used for endpoint or real-time DNA quantification depending on the platform. In digital PCR, a single template is amplified on its own and can avoid problems from mixed-template PCR such as the generation of chimeric sequences, primer-template bias and template competition (Boers, Hays, & Jansen, 2015; Williams et al., 2006). ddPCR has been used to estimate eDNA concentration, fish abundance and biomass (Doi et al., 2015). ddPCR can also be used to target multiple markers in a single run for enrichment prior to HTS (see Section 11 below). In a variation of the above techniques, microfluidic multiplex digital PCR was used to co-amplify 16S rDNA and a metabolic gene from single bacterial cells (Ottesen, Hong, Quake, & Leadbetter, 2006). In this example, termite gut endosymbionts previously known from a metabolic gene survey were linked with their 16S rDNA sequence for the first time. An advantage of this method over traditional qPCR is that the complexity of background DNAs is reduced to a single-template strand per reaction. For example, in a study that directly compared ddPCR with qPCR, ddPCR was found to quantify eDNA, fish abundance and fish biomass more accurately than qPCR (Doi et al., 2015). For laboratories that already have qPCR protocols, conditions would need to be re-optimized for digital PCR. The sensitive nature of the method makes the problem of contaminants in laboratory products more pernicious, highlighting the importance of running negative controls (Salter et al., 2014). As shown in Table 1, this method is similar to traditional qPCR but is more sensitive and ideal for mixed templates derived from environmental samples. Due to the requirement for specialized equipment, this method may not be readily adopted by laboratories that already have qPCR equipment and protocols.

5 | DNA METABARCODING

The marker gene DNA sequencing technique used for the original prokaryote 16S ribosomal gene phylogenies and community surveys were quickly adapted to other markers to target fungi and then eukaryotes where the approach was rebranded as DNA metabarcoding with the defining goal of “species identification” from bulk environmental samples (Bik et al., 2012; O’Brien et al., 2005; Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012; Torsvik & Øvreas, 2002). DNA metabarcoding is rooted in efforts associated with DNA barcoding, where a standard species-specific marker gene such as mitochondrial cytochrome c oxidase 1 (CO1) is used for identifying single specimens of animals (Hebert, Cywinska, Ball, & deWaard, 2003). DNA metabarcoding involves PCR-coupled HTS of one or more DNA barcode markers (or other biodiversity markers) directly from mixed-community samples without the need to isolate individuals. The term “DNA metabarcoding” (Taberlet, Coissac, Pompanon, et al., 2012; Yu et al., 2012) has also been referred to as “DNA metagenetics” (Creer et al., 2010), “environmental barcoding” (Hajibabaei et al., 2011), “DNA metasystematics” (Hajibabaei, 2012), metagenomic amplicon sequencing or simply “marker gene surveys” (Bik et al., 2012). Essentially, these methods transformed the fields of microbial molecular ecology, biodiversity and biomonitoring by allowing whole communities of organisms to be targeted, at the same time, without the need to isolate individuals. Unlike the fluorescence or PCR-based methods discussed in the sections above, the DNA sequences produced by DNA metabarcoding provided greater resolution to distinguish among taxa and sparked discussions concerning the significance of the “rare biosphere” (Huse, Welch, Morrison, & Sogin, 2010; Reeder & Knight, 2009; Sogin et al., 2006). DNA metabarcodes are amenable to phylogenetic analysis and introduced a new way to analyse biodiversity using a phylogenetic diversity method that could be scaled up to keep pace with the newest HTS methods (Faith, Lozupone, Nipperess, & Knight, 2009; Hamady et al., 2010). Given the widespread application of DNA metabarcoding, we provide more details on key aspects of this approach as it is commonly used in biodiversity and biomonitoring studies.

5.1 | Mixed-template PCR

Target enrichment/amplification from mixed communities using PCR has been referred to as mixed-template or multitemplate PCR (Kalle, Kubista, & Rensing, 2014). PCR-coupled DNA metabarcoding is sensitive to the initial mixed-template PCR, including PCR cocktail composition, primers and cycling conditions. PCR bias caused by differential binding of PCR primers to template eDNA, the generation of artefacts (heteroduplexes, chimeric sequences, PCR duplicates), has been discussed at length in the literature (Bik et al., 2012; Shokralla, Spall, Gibson, & Hajibabaei, 2012; Tedersoo et al., 2015). Mixed-template PCR optimization often involves steps such as reducing the number of PCR cycles and increasing
extension time (Gohl et al., 2016; Haas et al., 2011; Ishii & Fukui, 2001; Kurata et al., 2004; O’Donnell, Kelly, Lowell, & Port, 2016; Suzuki & Giovannoni, 1996; Wang & Wang, 1997). The resulting amplicon sequences can be analysed as is or identified by comparison with a reference sequence database. PCR is not the only method that can be used for target enrichment prior to DNA metabarcoding, however, and these are described below, see Section 11.

5.2 Marker selection

Research communities focusing on a variety of taxonomic groups have identified their own signature DNA regions (Table 2) suitable for high-throughput taxonomic identification using a variety of methods (Table 3). The metabarcoding approach gained much momentum after the advancement of HTS technologies, and researchers focused on differing taxonomic groups have established their own markers and standard methods (Hajibabaei, 2012). Selection and standardization of marker genes used for DNA metabarcoding remain an active area of research as it can significantly influence results obtained in metabarcoding studies (Deagle, Jarman, Coissac, Pompanon, & Taberlet, 2014; Fahner, Shokralla, Baird, & Hajibabaei, 2016; Porter, Shokralla, Baird, Golding, & Hajibabaei, 2016). Marker choice is known to affect the subset of the community resolved due to differing levels of variability in sequences of different lengths, across taxonomic groups and primer bias (Bellemain et al., 2010; Claesson et al., 2010; Hollingsworth, Graham, & Little, 2011; Porter & Hajibabaei, 2017).

5.3 Metabarcoding versus traditional biomonitoring

In a study that directly compared traditional morphology-based and metabarcoding methods for surveying macroinvertebrates from river benthos, all species that comprised greater than 1% of the individuals in the sample mixture were detected (Hajibabaei et al., 2011). In fact, the metabarcoding approach has already become a key tool in some large-scale biomonitoring programmes looking to incorporate DNA-based methods into their existing regional or national programmes (Baird & Hajibabaei, 2012; Gilbert, Jansson, & Knight, 2014; GRDI-EcoBiomics, 2016). A key question for biodiversity analyses is how metabarcoding compares with traditional methods for community profiling. Despite differences in the exact taxa recovered using traditional methods and DNA metabarcoding (Hajibabaei et al., 2011), recent studies have found that metabarcoding of insects, birds, diatoms and zooplankton tends to recover more taxa than traditional methods, provide a finer level of resolution and can similarly be used as a DNA-based biological indicator (Ji et al., 2013; Pawlowski, Esling, Lejzerowicz, Cedhagen, & Wilding, 2014; Sweeney et al., 2011; Yang et al., 2017). Results from studies focusing on plants and animals have been reviewed in Deiner et al. (2017) and also found that DNA metabarcoding provided either complimentary

---

**TABLE 2** A list of the commonly used markers for DNA metabarcoding, databases, and tools for various taxonomic groups. This is not an exhaustive list, for generic tools we focus on those that seem to be most popular or are best suited for high-throughput preprocessing of amplicon reads

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Marker</th>
<th>Reference databases</th>
<th>Software tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prokaryotes</td>
<td>16S rDNA</td>
<td>GreenGenes (DeSantis et al., 2006)</td>
<td><strong>PICRUSt</strong> (Langille et al., 2013)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ribosomal Database Project (RDP) (Cole et al., 2014)</td>
<td>RDP classifier (Wang, Garrity, Tiedje, &amp; Cole, 2007)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SILVA (Pruesse et al., 2007)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>UNITE (Abarenkov et al., 2010)</td>
<td>ITx (Bengtsson-Palme et al., 2013)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RDP classifier</td>
</tr>
<tr>
<td>Animals</td>
<td>CO1 mtDNA</td>
<td>BOLD (Ratnasingham &amp; Hebert, 2007)</td>
<td><strong>BOLD</strong> (Ratnasingham &amp; Hebert, 2007)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CO1 Arthropod classifier (Porter &amp; Hajibabaei, 2017)</td>
<td>RDP classifier</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CO1 Insect classifier (Porter et al., 2014)</td>
<td></td>
</tr>
<tr>
<td>Plants</td>
<td>rbCl +</td>
<td>See <strong>Generic databases below</strong></td>
<td>See <strong>Generic pipelines below</strong></td>
</tr>
<tr>
<td></td>
<td>matK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other Eukaryotes</td>
<td>18S rDNA</td>
<td>GreenGenes</td>
<td>See <strong>Generic pipelines below</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SILVA</td>
<td></td>
</tr>
<tr>
<td>Any</td>
<td>Any</td>
<td><strong>Generic databases</strong>:</td>
<td><strong>Generic pipelines</strong>:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>International Nucleotide Sequence Database (INSD) Collaboration*</td>
<td>DADA2 (Callahan et al., 2016)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Galaxy (Goecks, Nekrutenko, &amp; Taylor, 2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>MOTHUR</strong> (Schloss et al., 2009)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>QIIME</strong> (Caporaso et al., 2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RDP pipeline (Cole et al., 2014)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>USEARCH</strong> package (Edgar, 2013, 2016)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>VSEARCH</strong> (Rognes, Flouri, Nichols, Quince, &amp; Mahé, 2016)</td>
</tr>
</tbody>
</table>

*The INSD is an international initiative between the National Centre for Biotechnology Information (NCBI), the DNA Data Bank of Japan (DDBJ) and the European Nucleotide Archive (ENA).
TABLE 3 Commonly omitted methods for taxonomic assignment of signature DNA sequences from DNA metabarcoding studies. In this table, we have specifically omitted species delineation methods that should not be conflated with taxonomic assignment methods. Additionally, some of these methods were originally developed for the taxonomic assignment of metagenomic reads but can be applied to amplicon sequences.

<table>
<thead>
<tr>
<th>Taxonomic assignment method</th>
<th>Description</th>
<th>Programs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity-based</td>
<td>Includes methods that use a score calculated from pairwise sequence alignments or a comparison between a sequence and a profile hidden Markov model (HMM) (generated from a multiple sequence alignment)</td>
<td>BLAST (Altschul, Gish, Miller, Myers, &amp; Lipman, 1990)</td>
</tr>
<tr>
<td>Phylogeny-based</td>
<td>Based on the concept of orthology and phylogenetic theory, this method requires a multiple sequence alignment and a model of sequence evolution. The results are a phylogenetic hypothesis of evolutionary relatedness and a measure of statistical support for branch points</td>
<td>MEGA (Matsen, Kodner, &amp; Armbrust, 2010)</td>
</tr>
<tr>
<td>Composition-based</td>
<td>Query and reference sequences are broken down into libraries of shorter words of size “k.” Taxonomic assignment is based on k-mer frequencies in the query and reference library sequences</td>
<td>KNN Classify.seqs method (Schloss et al., 2009)</td>
</tr>
<tr>
<td>Hybrid methods</td>
<td>Combines different approaches from above, sometimes with other methods where indicated, to make taxonomic assignments</td>
<td>Fuzzycld2 (Shi et al., in press)</td>
</tr>
<tr>
<td>Bayesian tree-less methods</td>
<td>Based on the coalescent theory of speciation, uses Bayesian tree-less methods that use either the coalescent or a proxy for the coalescent, the number of segregating sites</td>
<td>Coalescent Assigner (Abdo &amp; Golding, 2007)</td>
</tr>
<tr>
<td>Machine learning</td>
<td>Uses theories from machine learning such as support vector machines and artificial neural networks</td>
<td>BPS (Zhang, Sikes, Muster, &amp; Li, 2008)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SVM classifier (Seo, 2010)</td>
</tr>
</tbody>
</table>

or increased richness compared with traditional methods. Although many studies have successfully used metabarcoding to find differences in sites that are known to be distinct, a recent analysis used metabarcoding to assess similar sites subject to natural variation and low-intensity management (Emilson et al., 2017). This study showed that freshwater invertebrate biodiversity obtained from metabarcoding the CO1 BR5 region was positively correlated with stream condition gradients (Emilson et al., 2017). Additionally, DNA metabarcoding compares favourably to traditional methods for zooplankton and diatom community profiling by alleviating the taxonomic impediment—the bottleneck implicated in traditional morphology-based specimen identification (Ji et al., 2013; Yang et al., 2017; Zimmermann, Glückner, Jahn, Enke, & Gemeinholzer, 2015). Although fieldwork is still a costly and time-consuming endeavour, once samples are subsampled into tubes or plates, this method becomes highly scalable and amenable to parallelization and automation (Box 2).

5.4 | Quantitative or not?

Because PCR is commonly used for amplification of target genes in DNA metabarcoding, read abundance does not necessarily reflect organismal abundance, biomass, or activity (Amend, Seifert, & Bruns, 2010; Elbrecht & Leese, 2015; Klappenbach, Saxman, Cole, & Schildt, 2001; Polz & Cavanaugh, 1998; Tedersoo et al., 2010; Větrovský & Baldrian, 2013). There have been reports, however, that read abundance does scale with estimated biomass for fungi and zooplankton under certain conditions (Amend et al., 2010; Yang et al., 2017). Generally speaking, because of issues with primer bias and mixed-template PCR bias, natural variation in copy number, as well as variation in biomass and density among organisms, it has been suggested that a conservative approach is to treat DNA metabarcoding data as presence–absence data only (Elbrecht & Leese, 2015; Hajibabaei, Spall, Shokralla, & van Konyenenburg, 2012; Hajibabaei et al., 2011). In a study of marine benthic fauna, ecological indices using abundance or presence-only data performed similarly (Ranasinghe, Stein, Miller, & Weisberg, 2012).

5.5 | Bioinformatics

With advances from new HTS platforms and a growing need for more efficient data analysis and interpretation, the bioinformatics considerations are varied and constantly evolving (Box 3). Bioinformatic challenges related to plant and animal metabarcoding are different than those faced by DNA barcoding methodologies (Coissac, Riaz, & Puillandre, 2012). The taxonomic resolution gained by this
method depends on the resolution of the signature DNA sequence targeted (e.g., species level for DNA barcodes) and availability of sequences in reference databases. There are a variety of platforms available for biodiversity analysis, some particularly well-suited for beginners in the field because they can provide smooth wrappers around commonly used command-line tools as well as well-documented usage examples in online forums (Table 4).

Often, results from DNA metabarcoding are used to inform researchers as to which samples are worth a deeper look using metagenomic, metatranscriptomic or target-enriched sequencing. The data generated using DNA metabarcoding are more tractable in amount and computational resources required for analysis than those generated from whole metagenome or metatranscriptome sequencing, and the most popular platform is currently the Illumina MiSeq (Box 5). For current biomonitoring applications, DNA metabarcoding leads the way over other methods of biodiversity analysis in terms of scalability, cost and tractability. As summarized in Table 1, this method is likely to remain the preferred choice for routine biodiversity and biomonitoring applications until the cost of sequencing whole genomes is decreased to the point that it can become routine so that genome databases become more representative of the biodiversity found in nature.

**BOX 2 Scaling up**

DNA metabarcoding is a highly scalable method that can be used to increase both the breadth and depth of sampling in biodiversity and biomonitoring studies as shown in Box Figure 1. The barcoding marker is enriched from eDNA samples in the metabarcoding step. Using kits for substrate-specific DNA extraction can increase efficiency and reproducibility, as well as reduce laboratory-to-laboratory variability across large projects. Using automation, liquid-handling robots, can reduce pipetting errors. Working with samples in multi-well plates can improve throughput by conducting many reactions in parallel. When samples are prepared for high-throughput sequencing (HTS), sequencing breadth and depth can be adjusted by varying the number of pooled (multiplexed) samples loaded on an Illumina MiSeq lane. For instance, samples collected from different field sites or multiple markers from the same sample can be tagged, pooled, and run together. To adjust the depth of sequencing, the total number of samples run in a lane can be varied up or down. The large amount of data generated by HTS often necessitates analysis via high-performance computing (HPC). Cloud computing, in particular, can be used to store and analyze data in an environment with more computing resources than typically available to most individual researchers. Bioinformatic pipelines can be run on batches of samples across multiple runs in parallel to reduce computational time. As data analysis methods develop, data flows can be developed, documented, and versioned to ensure analysis reproducibility (Falster, FitzJohn, Pennell, & Cornwell, 2017).

**BOX FIGURE 1** DNA metabarcoding is a scalable DNA-based biodiversity analysis method. At each step, methods can be easily scaled upwards to accommodate more samples if needed. HPC, high-performance computing; HTS, high-throughput sequencing.
BOX 3  Bioinformatics considerations for using DNA metabarcoding for biodiversity analysis

Two key resources for making DNA metabarcoding a suitable tool for biodiversity analysis are as follows: (i) comprehensive metadata and (ii) comprehensive reference databases.

METADATA
The deposition of rich metadata ensures the long-term utility of published amplicon sequences and encourages comparative studies (Bik et al., 2012). As a result, sequence annotation standards can facilitate the comparison of sequence data across studies (Yilmaz et al., 2011).

REFERENCE DATABASES
High quality reference databases for taxonomic assignment are essential for successful DNA metabarcoding. There are several issues that affect current reference database quality:

- Insufficiently identified records. This is a problem that affects all metabarcoding markers and public records can have differing levels of taxonomic annotation from strain or BIN or species to a variety of more inclusive taxonomic ranks. It has been shown for fungal ITS sequences that with the increasing use of high-throughput DNA metabarcoding from eDNA, the number of unnamed anonymous DNA sequences has been accumulating in GenBank and vastly exceeds the number of known taxonomically identified reference sequences (Hibbett et al., 2011; Nilsson, Kristiansson, Ryberg, & Larsson, 2005).
- Incorrectly annotated records. This is an issue that affects all metabarcoding markers. Unexpected annotation errors can arise due to contamination or misidentification by collectors. For fungal ITS sequences, the reliability of taxonomic records in public databases have been questioned and there has even been a call for third-party annotation of GenBank records (Bidartondo, 2008; Nilsson et al., 2006).
- Biased record collection. For any metabarcoding marker, there can be variability in taxon representation across different geographic locations, habitats, and taxonomic groups. For instance, CO1 sequences in the BOLD database are dominated by Diptera (Insecta) and Lepidoptera (Insecta) sequences from Canada (Porter et al., 2014). This is particularly problematic for metabarcoding studies in areas where endemic diversity is expected to be high, such as in the tropics or for microbes—particularly in forest soils (Basset et al., 2012; Hawksworth, 1991; Tedersoo et al., 2014).
- Incomplete databases. It is known that sequence-based identification of samples can be hindered by incomplete reference databases (Porter & Golding, 2011, 2012; Porter et al., 2014; Sundquist et al., 2007; Taberlet, Coissac, Pompanon, et al., 2012). Most studies assume that their taxa will be identified through metabarcoding, however, the extent of database incompleteness is not known for every group. For example, despite Insecta being one of the largest groups of CO1 sequences in GenBank, it has been estimated that only about 12% of extant insect genera are represented by a CO1 sequence (Porter et al., 2014).

The significance of these issues lies in the level of undetected false-positive taxonomic assignments in metabarcoding studies. Type II error, incorrectly rejecting a true null hypothesis, is also referred to as a false-positive assignment and has been discussed in the literature (Virgilio, Backeljau, Nevado, & De Meyer, 2010; Porter et al., 2014). See Taxonomic assignment below.

SEQUENCE READ ERRORS
PCR and sequencing errors can produce artefactual sequences. The removal of these sequences is essential to avoid inflating richness counts and including sequence artefacts in community analyses (Huse et al., 2010; Reeder & Knight, 2009). There are several methods for handling sequence errors in metabarcoding studies, especially those that use PCR for enrichment.

The use of paired-end sequencing, especially for short amplicons, can provide sequence overlap between the forward and reverse reads and provide redundancy in the regions where sequence error rates start to increase (i.e., at the ends of each read). Chimeric sequences can be generated during PCR. A chimera removal step is included in many of the widely used bioinformatic pipelines (Table 4). Clustering reads into OTUs can absorb highly similar reads with sequence errors. Clustering reads into OTUs defined by some distance threshold (or sequence similarity cutoff) is often the default approach in most metabarcoding bioinformatic pipelines (Table 4). Additionally, a generic method to remove sequence artefacts is to simply remove low frequency OTUs such as singletons and doubletons (Huse et al., 2010; Tedersoo et al., 2010). A step including the removal of low frequency OTUs are integrated into many of the most widely used bioinformatic pipelines (Table 4).

The term denoising was first introduced as a sequencing platform-specific method to remove sequences with distinctive errors produced during the sequencing step (Quince, Lanzen, Davenport, & Turnbaugh, 2011). Since then, it has become a more general term to describe the process of removing reads with predicted errors from any source. The denoising method can be as simple as removing low frequency OTUs. For example, ‘rare’ OTUs such as singletons and doubletons or OTUs containing a particularly low
frequency of total reads. This step is implemented in most bioinformatic pipelines (Table 4). The USEARCH UNOISE algorithm attempts to predict and remove low read frequency amplicons as containing possible sequence errors if there is a high similarity high frequency amplicon present (Edgar, 2016). Part of the denoising process can also include the removal of contaminant sequences. The USEARCH unoise algorithm will remove PhiX contaminant sequences automatically, however, the removal human, host, or common lab contaminants requires extra steps outside of this pipeline (Edgar, 2016). Finally, the removal of nonspecific amplification products (e.g., nuclear pseudogenes of mitochondrial genes; NUMTs) can be challenging to implement in high-throughput pipelines, but for protein coding genes, such as CO1, these can include removing reads that contain frameshifts or indels that disrupt the open reading frame (Song, Buhay, Whiting, & Crandall, 2008).

DETERMINING THE BASIC UNIT FOR DNA-BASED BIODIVERSITY ANALYSES

There are three main types of DNA-based indicators generated by DNA metabarcoding: (i) taxa (at one or several taxonomic ranks), (ii) operational taxonomic units (OTUs), or (iii) exact sequence variants (ESVs). For simple biological inventories, taxonomy-based lists are probably the most relevant output (see Table 3). For a simple calculation of richness, one could count the number of unique taxa, ESVs, or OTUs. Moving from taxon lists to community analyses, however, involves the creation of data matrices. These matrices may be based simply on taxonomy (at any rank or variable ranks), or they can be based on OTUs or ESVs. OTUs (or ESVs) in turn can be global OTUs (clustered from reads from all samples at once) or OTUs created for a single sample at a time. The advantage of using global OTUs is that these are directly comparable across all samples. The advantage of OTU-based analyses is that all the data can be analysed together whether or not they can be taxonomically assigned with confidence. The advantage of taxonomy-based analyses is that known taxonomy can be used to help narrow down a complex data set into groups of indicator taxa that are known be ecologically relevant such as members of the insect orders Ephemeroptera, Plecoptera and Trichoptera (EPT) that have been shown to be sensitive to water pollution as they require clean water with a high level of dissolved oxygen or members of the Chironomidae that can be an indicator of poor water quality as they are rapid colonizers and have been previously shown to be tolerant of high water pollution (Buss, Baptista, Silveira, Nessimian, & Dorvillé, 2002; Emilson et al., 2017).

CHOOSING A SEQUENCE CLUSTERING METHOD

Metabarcoding clustering can be performed using a variety of methods: phylogenetic clustering (Box 4); single-linkage, average-linkage, or furthest-linkage clustering, or other hybrid methods. Reads can be clustered against a reference in open- or closed-reference methods. Closed-reference clustering is when reads are clustered against a reference database directly. Open-reference clustering is much like closed-reference clustering except that reads that do not cluster with the references are then clustered de novo into their own new clusters using a distance threshold as a cut-off. For example, the Galaxy, MOTHUR, QIIME, and USEARCH pipelines all provide methods to perform reference-based clustering using the 165 rDNA Green Genes or SILVA databases, or the ITS reference databases. Reference-based clustering is popular with markers where extensive reference databases exist, but a recent study has shown that OTU richness and beta-diversity are often greatly exaggerated with these methods (Edgar, 2017). De novo clustering, on the other hand, may outperform closed- and open-reference clustering with 16S sequences by better representing actual distances between sequences (Westcott & Schloss, 2015). De novo clustering is when all reads are clustered among themselves using any of a variety of algorithms, then subsequently taxonomically assigned (Table 3). A consideration with this approach is the cut-off that is used to delineate the OTUs. A range of 1–3% sequence dissimilarity is popular in the literature and was once used to approximate species units for measuring or estimating diversity, but it is now recognized to be somewhat arbitrary as sequence variation within and among species varies across taxa. To avoid the ‘lumping’ of similar reads from different species into a single OTU, the use of exact sequence variants (ESVs) has been proposed (Callahan, McMurdie, & Holmes, 2017; Edgar, 2017b). The DADA2 and USEARCH unoise method both specifically produce ESVs as output (Callahan et al., 2016; Edgar, 2016). Another hybrid clustering method is SWARM and it can be run on its own or as a part of the QIIME pipeline (Mahé, Rognes, Quince, de Vargas, & Dunthorn, 2014). A comparison of OTU clustering methods with 16S rDNA data showed that the most stable and accurate OTUs could be produced using different algorithms and may vary according to data set (Westcott & Schloss, 2015).

TAXONOMIC ASSIGNMENT

One of the simplest taxonomic assignment methods is incorporated into the reference-based clustering approaches described above. Alternatively, there are a variety of methods for taxonomic assignment that use different parameters for optimization such as similarity, phylogeny, or composition among others (Table 3). The most popular method for metabarcoding taxonomic assignment is the similarity-based top BLAST hit approach. Though the method is relatively easy to implement and run in parallel, studies that have shown that type II error (false-positive rates) are high with this method and that the top hit is not always the closest phylogenetic
neighbour (Koski & Golding, 2001; Virgilio et al. 2010). One way to reduce false positive rates is to use a method that produces a measure of statistical confidence to filter out uncertain taxonomic assignments (Munch, Boomsma, Willerslev, & Nielsen, 2008; Munch, Boomsma, Willerslev, et al., 2008; Porter & Hajibabaei, 2017; Wang et al., 2007). Additionally, the top BLAST hit method is slow compared to other widely used methods. In large-scale biodiversity and biomonitoring studies, taxonomic assignment methods need to be fast and provide a statistical measure of confidence for taxonomic assignments at all ranks. A popular method is the Ribosomal Database Project naïve Bayesian classifier which is implemented in most popular bioinformatics pipelines (Table 4) or can be run on its own. This method can be trained for any metabarcode marker and reference sets already exist for the prokaryote 16S, fungal ITS and LSU rDNA regions, as well as for the CO1 animal barcode marker (Cole et al., 2009; Liu, Porras-Alfaro, Kuske, Eichorst, & Xie, 2012; Porter et al., 2014; Porter & Hajibabaei, 2017; Wang et al., 2007).

**DATA NORMALIZATION**

It has been shown that uneven sequencing effort can skew community comparisons because false negatives (undersampling) can emphasize differences between communities (Ghiring, Green, & Schadt, 2012). There are methods to normalize or accommodate uneven sequence numbers among samples to avoid library size bias. Most recently, it has been pointed out that metabarcoding data are inherently compositional. As such, the most appropriate normalization technique would be a log ratio transformation (Gloor, Macklaim, Pawlowsky-Glahn, & Egozcue, 2017). Traditionally, however, randomly subsampling all libraries down to the smallest library size prior to OTU creation has been a simple solution to avoid diversity estimator sample size bias; however, this involves throwing away sequence information. Alternatively, rarefaction estimates of samples can be compared at some common level of sequencing depth. Both these methods also work when comparing obviously different sample types using presence-absence data. For finding species showing different abundances among samples, however, it has also been shown that rarefaction, as well as the common approach where sequence numbers are converted to simple proportions for each library, results in a high rate of false positives (McMurdie & Holmes, 2014). Instead, the use of alternative methods such as ANCOM (a compositional approach) or DESEQ2 may be able to more accurately identify differential taxon abundances among samples (Mandal et al., 2015; McMurdie & Holmes, 2014). A comparison of these methods using 16S rDNA has shown that rarefaction is an effective method for normalizing data and that for comparing abundances the best method to use may depend on data characteristics such as the total number of samples and how uneven the library sizes are, as well as composition of microbial communities among samples (Weiss et al., 2017).

### 6 | ORGANELLE SEQUENCING

As shown in Table 1, this is a sequencing-based method suitable for characterizing individual organelle genomes. In this method, organelle DNA is either isolated from an individual or the organelle genome can be bioinformatically assembled from an individual’s shotgun genomic sequences with or without prior enrichment (Cronn et al., 2008; McPherson et al., 2013; Nock et al., 2011; Parks, Cronn, & Liston, 2009). Individual or combinations of organelle genes from mitochondria and plastids have a long history of use in biosystematics as a basis for classification (Hollingsworth et al., 2011). The term “super-barcoding” has been used when whole-organelle genomes are used specifically for taxonomic assignment (Li et al., 2015). For plants in particular, this has been shown to circumvent the lack of variation seen in some groups when single barcode markers are used for taxonomic assignment (Li et al., 2015). Recent advances in sequencing technologies now enable whole chloroplast or mitochondrial genomes to be sequenced and used for a wide range of population genetic to phylogenetic studies not only from individuals but also from eDNA samples. Specifically, researchers have proposed a mito-metagenomics approach for eDNA (Tang et al., 2014). This involves computationally assembling all mitochondrial genes sequenced from eDNA into whole mitochondria. With respect to the amount of DNA sequence data generated, organelle sequencing sits between DNA metabarcoding and whole-genome sequencing. It provides a way to obtain more genetic information at a much lower cost compared with whole-genome analysis. Unfortunately, even if a whole mitochondrial genome is sequenced and used for comparative analysis, it essentially behaves as a single marker because all the mitochondrial genes are linked. The utility of organellar genomes for addressing challenging biosystematics questions is limited because the bulk of genetic information comes from unlinked nuclear markers whose evolution can follow a different evolutionary trajectory (Hollingsworth, Li, van der Bank, & Twyford, 2016). Also, in an eDNA framework, mito-metagenomics has not been applied to large sample sizes and varied sample types presumably because the sequencing depth needed for whole mitochondrial genome reconstruction is high and reference databases for mitochondrial genomes are very small compared with reference databases for the commonly used DNA metabarcoding markers (Table 2). As summarized in Table 1, this method will likely be phased out in favour of whole-genome sequencing as single-molecule methods develop, accuracy improves and costs decrease.

### 7 | GENOME SKIMMING

As shown in Table 1, this is a DNA sequencing-based detection method applied to individual specimens. This method differs from
Phylogenetic methods can be used to analyse biodiversity data in several ways: (i) for OTU delimitation, (ii) for single-marker taxonomic assignments, (iii) for phylogenomic taxonomic assignments and (iv) for comparing beta-diversity across communities. These methods are included here because phylogenetics can complement the use of traditional biodiversity metrics by taking into consideration the evolutionary history of the sampled lineages across sites (Faith, 1996, 2013; Hamady et al., 2010; Parks, Porter, et al., 2009).

**OTU DELIMITATION**

There are instances when defining an OTU based on sequence similarity alone can be problematic. For example, choosing a sequence similarity cut-off to define an OTU is arbitrary, often based on community consensus, or chosen to approximate "species" even though it is known that Linnaean taxa do not necessarily coincide with DNA-based OTUs across lineages. It is possible instead to use a phylogeny-based method to define OTUs, like a single gene application of the phylogenetic species concept where the members of a terminal clade may comprise one OTU or taxon, although this is usually a manual process, difficult to apply in large-scale studies as this method can be computationally time consuming as the number of sequences analysed grows. This approach is best used to refine OTU membership for target taxa of interest as opposed to clustering whole communities of organisms. Once OTUs are delimited, they can be analysed as is or they can be taxonomically assigned (Table 3).

**TAXONOMIC ASSIGNMENT**

Phylogenetic methods can be used to make taxonomic assignments. For example, the sequence variability in the fungal barcode marker, the ITS rDNA region, is an advantage for making fine-level taxonomic assignments but is problematic for phylogeny-based taxonomic assignments across diverse groups of taxa. It is common that only congeneric taxa can be analysed in the same multiple sequence alignment and thus to analyse a community of diverse taxa would require many independent alignments and phylogenetic trees to identify their closest relatives for taxonomic assignment (Porter & Golding, 2011). To overcome this difficulty, phylogeny-based programs are available that can automate this process for any marker (Table 3). These methods can help investigators apply consistent cut-offs based on statistical support values to identify taxonomic assignments that they can be confident in. A major drawback of these methods is that they can be relatively slow and computationally intensive compared with alternative methods.

**PHYLOGENOMICS**

When markers are sampled by genome mining from whole-genome sequences, genome skimming or transcriptome sequencing, combined to make phylogenies, this is referred to as phylogenomics (Eisen, 1998). With respect to biodiversity studies, phylogenomics may be most useful for refining species identifications of target taxa as well as to understand their evolutionary histories at multiple taxonomic levels (Steele & Pires, 2011). For example, highly resolved phylogenies of yeast species have been produced by concatenating up to 106 genes (Rokas, Williams, King, & Carroll, 2003). Although most studies aim to identify taxa using single signature DNA markers, some groups can be problematic and may require multiple markers for correct species assignments. For example, to correctly identify plant taxa, multiple regions can be used (Straub et al., 2012). Comparative analyses can identify key genes and pathways that can be used as markers in future work (Riley et al., 2014). Comparisons among taxa in this way can help predict clades representing feature diversity worthy of special consideration for conservation efforts and can be used to guide topological restrictions on trees based on less data (see Phylogenetic diversity below).

**PHYLOGENETIC DIVERSITY**

Perhaps the most practical use for phylogenetic methods in biodiversity analysis is to use this as a way to compare phylogenetic diversity across samples providing another window on beta-diversity. The concept of phylogenetic diversity is based on the assumption that branch lengths among taxa from different samples represent the underlying feature diversity of these taxa (Faith, 2013). Unlike the traditional richness measure where each species is given an equal weight of 1, using a phylogenetic diversity metric, communities can be compared according to the amount of unique branch lengths they represent (Tucker & Cadotte, 2013). In such a scenario, it is possible that a site with high richness may reflect low phylogenetic diversity if the species are closely related. This may have an impact in conservation studies where decisions on which sites to protect are driven by how diversity is measured. To facilitate the processing of large data sets, software tools need to be chosen carefully. For example, heuristics exist to construct very large trees such as with FastTree (Hamady et al., 2010; Price, Dehal, & Arkin, 2009). Branch lengths can then be used to calculate phylogenetic beta-diversity in large data sets using Fast UniFrac (Faith et al., 2009). A drawback of this method is that it may not be possible to produce a good phylogeny based on a single marker. Topological restrictions based on previous multi-marker work, however, can be used to structure single-gene trees. Fortunately, the calculation of phylogenetic diversity using the UniFrac method has been shown to be robust to sequencing effort and phylogenetic method (Lozupone, Hamady, Kelley, & Knight, 2007).
organelle sequencing because there is no need for prior isolation or targeting of organelle DNA prior to sequencing individuals. Genome skimming involves shallow- or low-coverage sequencing of an organism's genome (including organelle genomes) to obtain sequence data that can be used to address biodiversity-related questions. Experimental work has shown that even shallow sequencing (e.g., 1–2 Gb) can provide surprisingly deep coverage of high copy number organelle DNA (plastids, mitochondria) and other repetitive sequences such as the full ribosomal cistron (Straub et al., 2012) often used in biosystematics (Hollingsworth et al., 2016; Li et al., 2015). This method has primarily been employed in plants due to the difficulty in obtaining species-level resolution using DNA barcodes (Hollingsworth et al., 2016). Genome skimming can provide sequence data suitable for biodiversity analysis without the need to pick and choose marker genes or optimizing PCR protocols for individual genes. With museum or eDNA samples with degraded DNAs, focusing on high copy regions by genome skimming may be more successful than targeting low copy regions of the genome (Dodsworth, 2015). In plants where the variation in plastid DNA can be especially valuable for taxonomic assignment, the natural variation of cpDNA per cell in different life stages of a leaf can result in more or less coverage of cpDNA versus other repetitive regions such as rDNA (Dodsworth, 2015). Additionally, the genome skimming approach may not be easily applicable to smaller organisms (e.g., small insects) and difficult to cultivate organisms where it could be difficult to extract enough genomic material. These issues, combined with the need to isolate individuals before sequencing, could make scalability a problem for large-scale ecological investigations. As summarized in Table 1, as single-molecule sequencing methods advance further, accuracy improves and cost reduces, this method is likely to be phased out in favour of WGS.

### Whole-genome sequencing projects

Numerous projects are contributing to the population of whole-genome sequences in databases such as the i5K initiative that aims to...
TABLE 4 Commonly used generic bioinformatics pipelines and packages to analyse signature DNA regions from metabarcoding studies

<table>
<thead>
<tr>
<th>Pipeline</th>
<th>Description</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>DADA2</td>
<td>Runs in the R environment. Processes data from fastq files, removes errors and chimeras, and produces sample abundances and taxonomic assignments</td>
<td>Produces exact sequence variants (ESVs) instead of OTUs for greater resolution than OTU-based methods</td>
</tr>
<tr>
<td>Galaxy</td>
<td>Provides an environment where scripts can be assembled into pipelines to assist with raw data processing</td>
<td>Graphical user interface</td>
</tr>
<tr>
<td>MOTHUR</td>
<td>Command-line driven. Semi-automated pipeline allows raw sequence data to be processed through to community analysis using OTU- or taxonomy-based methods</td>
<td>Initially offered a way to create OTUs, remove putatively chimeric sequences using a variety of methods, calculate ecological indices and create Venn diagrams. Now also offers a variety of pipelines to process raw reads and make taxonomic assignments</td>
</tr>
<tr>
<td>QIIME</td>
<td>Command-line driven. Semi-automated pipeline allows raw sequence data to be processed through to community analysis</td>
<td>Wrapper for many commonly used programs for analysing DNA metabarcoding reads, particularly 16S sequences. Pipeline automatically formats the input and output files to work with a variety of programs to allow easy comparison of results using the most popular methods. QIIME2 also comes with an easy to use graphical user interface and an application programmer interface for data scientists</td>
</tr>
<tr>
<td>RDP pipeline</td>
<td>Provides a graphical user interface to process amplicon sequences from raw reads, performs 16S and 28S rDNA taxonomic assignments, as well as provides 16S and 28S secondary structure, and diversity analysis tools</td>
<td>Provides access to the RDP classifier for classifying SSU rDNA for bacteria and archaea, as well as ITS and LSU rDNA for fungi</td>
</tr>
<tr>
<td>USEARCH</td>
<td>Command-line driven, normally used for sequence clustering into operational taxonomic units, but can also be used for sequence similarity searches, denoising and newer versions can handle raw sequence data</td>
<td>Initially offered a way to cluster reads into operational taxonomic units (OTUs), a method to search for similar sequences and identify putatively chimeric sequences. This package now offers pipelines to process raw reads, denoise reads, and cluster reads while automatically removing chimeric sequences, sequence errors and PhiX reads. 32-bit version is available for all users free of charge, but is limited to 4-Gb memory at most. 64-bit version available and allows users to use all the memory available on a 64-bit computer</td>
</tr>
<tr>
<td>VSEARCH</td>
<td>Performs many of the functions available in USEARCH except denoising</td>
<td>Open-source software available free of charge and allows users to use all the memory available on a 64-bit computer</td>
</tr>
</tbody>
</table>

sequence 5000 arthropod genomes (i5K Consortium, 2013), the 1000 fungal genome project (approximately 800 fungal genomes are currently available through the U.S. Department of Energy's Joint Genome Institute MycoCosm portal) (Grigoriev et al., 2014), the GIGA project targeting 7000 noninsect and non-nematode invertebrates (mostly marine taxa) for sequencing (GIGA Community of Scientists, 2014), the Genome 10K project that aims to sequence one individual from every vertebrate genus (Koepfli, Paten, & O'Brien, 2015), the Genomic Encyclopedia of Bacteria and Archaea (GEBA) initiative that sequenced and released 1000 bacterial and archaeal genomes (Mukherjee et al., 2017), as well as other projects targeting plant and crop genomes (Li, Wang, & Zeigler, 2014). All of these data are essential resources for the further development of molecular primers, probes and, in some cases, identification of eDNA sequences generated by other genomic methods discussed in this review.

In a biodiversity or biomonitoring context, WGS data from single organisms are useful for both taxonomic or functional assignments as well as marker and primer development for qPCR, digital PCR or phylogenetics (Box 4). As summarized in Table 1, as single-molecule sequencing methods become more available and accurate, this method may become as routine as single gene sequencing is today in biodiversity studies and biomonitoring applications. At present, WGS of organisms from eDNA is only feasible for microbes with the smallest genomes and simplest organization (a single or few circular chromosomes) (see Sections 6 and 9).

9 | METAGENOMICS

As shown in Table 1, this is a DNA sequencing-based method that can used to profile mixed communities. Metagenomics involves sequencing all the genomic material from many different taxa whose bulk DNA was extracted directly from environmental samples such as soil, biofilms, water, sediments, benthos and air (Venter et al., 2004). This approach is also known as "shotgun sequencing," "environmental genomics," "ecogenomics" or "community genomics." Although this approach was first introduced based on Sanger sequencing technology, the advances of HTS approaches have generated much momentum in metagenomics research and applications. Typically, this technique samples genes from across the genome, not just DNA barcode regions, so the functional potential of a sample can be explored. Comparative metagenomics can be used to compare the metabolic potential of organisms from different environmental samples and even determine taxonomic and functional
profiles from the thousands of gene markers from communities of organisms (Tringe et al., 2005). This method has been proposed as a way to detect uncultured organisms that are difficult to identify by traditional means (Handelsman, 2004). Genes and gene families can be identified from metagenomic sequences. Identifying the taxa that these genes belong to, however, can be challenging. As signature DNA regions suitable for taxonomic analysis will also be present in the sample, these can be used to identify individual taxa (Liu, Gibbons, Ghodsi, & Pop, 2010; Manichanh et al., 2008). Reconstruction of individual genomes is also possible depending on the sequencing depth, taxonomic complexity and size of organisinal genomes in the sample. In a recent study, nearly 8,000 metagenome-derived prokaryote genomes were assembled from 1,500 public metagenomes (Parks et al., 2017). This type of achievement is not yet possible for eukaryotes due to the size and complexity of their genomes, but it may be in future as sequencing technologies and bioinformatics methods progress. Metagenomics has found application in ancient DNA studies looking at the evolution of antibiotic resistance, studies of the microbes involved in honey bee colony collapse disorder (Cox-Foster et al., 2007; D’Costa et al., 2011). Metagenomics is a widely used technique to explore microbiomes on a small scale and can be scaled upwards for broad-scale ecological surveys (The Human Microbiome Project Consortium, 2012; Venter et al., 2004). An advantage of this method is that amplification-free metagenomic sample preparation avoids the PCR bias that other methods may otherwise be subject to. A challenge with this method is that with sequencing effort spread over all genomic regions, not just the signature DNA regions suitable for taxonomic assignment, there may be a reduced set of taxa that can be identified with confidence. Unfortunately, taxonomic assignment of nonsignature DNA regions may be biased towards organisms whose whole genomes are present in databases (false positives). The sequencing depth required to capture a community would be much higher than the sequencing depth required to saturate taxon sampling using DNA metabarcoding. As summarized in Table 1, as HTS and single-molecule sequencing technologies advance, output grows, and costs decrease (Box 5), this method is likely to be even more widely used as amplification-free methods are very appealing to many investigators hoping to circumvent the many known issues with mixed template PCR in PCR-coupled DNA metabarcoding. As with many other methods, as the number of annotated genomes in public databases grows, the ability to annotate metagenomic samples should continue to improve.

10 | METATRANSCRIPTOMICS

As shown in Table 1, this is a sequencing-based detection tool suitable for identifying genes from individuals in a community. Whereas metagenomics can provide information on taxonomic composition and metabolic potential, metatranscriptomics can be used to provide a snapshot of the metabolic activity in a community. Metatranscriptomics involves HTS of reverse-transcribed complementary DNA (cDNA) from messenger RNA (mRNA) isolated directly from environmental samples (Carvalhais, Dennis, Tyson, & Schenk, 2012; Mason et al., 2012). This method has already been used to look at functional diversity of microbes and eukaryotes in soil (Bailly et al., 2007; Urich et al., 2008). It has been shown that while metagenomics can show metabolic potential (e.g., of deep-sea microbial communities), the results from metatranscriptomics may yield very different insights as to which genes are actually being expressed (Mason et al., 2012). Whereas reverse transcriptase PCRs can only detect the expression of a single gene at a time, metatranscriptomics is a high-throughput method that can survey thousands of genes at a time. Unfortunately, the proportion of an RNA extraction that contains mRNA is very low (2–3%) and may need to be amplified to obtain enough material for sequencing. If a PCR-based amplification step is used, then the diversity in downstream steps may not reflect initial relative abundances. Obtaining high-quality samples with intact mRNA may be challenging for many types of environmental samples. As the number of organisms with sequenced genomes increases, so too should the ability of investigators to annotate their metatranscriptomes. As this method provides a snapshot of the genes and pathways that are expressed in an environmental sample, this is a very attractive method of generating a very large set of functional gene information from across a community of organisms playing a variety of functional roles while using rather generic methods. When coupled with HTS, this method is highly scalable. As summarized in Table 1, as WGS reference databases become more representative of environmental diversity, this method is likely to become a more reliable source of functional community profiling.

11 | TARGET ENRICHMENT

The terms “target” or “targeted” enrichment refers to a general technique that can be used in combination with many of the above-mentioned methods. Target enrichment resides between single gene metabarcoding and whole-genome sequencing approaches because it allows a suite of markers to be targeted and enriched prior to sequencing. Commonly used enrichment methods include the following: (i) hybrid capture, (ii) selective circularization and (iii) PCR amplification. The use of target enrichment is to increase the efficiency of HTS for biodiversity analyses (Mamanova et al., 2010; Mertes et al., 2011). Generally, these methods are used to enrich for taxa/genes present at low abundance in a sample (e.g., parasites/pathogens), or to reduce the detection of taxa genes present at high abundance in a sample (e.g., the host). Because this method relies on designing oligonucleotides to capture target sequences, this method may limit the detection of new taxa not currently represented in public databases.

Hybrid capture uses long oligonucleotides, either bound to a microarray or to beads in solution, to capture target sequences (Mamanova et al., 2010). This is sometimes referred to as “PCR-free” enrichment. For biodiversity analyses where the objective is to detect as many different taxa as possible, hybridization capture has been shown to recover a greater diversity of arthropod and insect
orders compared with traditional morphological taxonomic assignment methods and PCR-coupled metabarcoding (Shokralla et al., 2016). Hybrid capture is a reproducible method, produces relatively uniform coverage of target sequences and has good capture rates (Tewhey et al., 2009). This method may be able to generate sufficient template for library preparation that the initial mixed-template PCR step in DNA metabarcoding can be avoided (Shokralla et al., 2016). Additionally, hybrid capture tends to select for short fragments with higher specificity than longer fragments. This is because longer fragments will have a higher proportion of off-target sequence compared with the probe and because of possible cross-hybridization within longer fragments (Mamanova et al., 2010). This bias towards short fragments makes hybrid capture suitable for processing potentially degraded samples from eDNA and producing the short sequence libraries typically prepared for Illumina MiSeq DNA metabarcoding (Box 5). Bead-based hybridization is highly scalable across many samples and can be used to detect thousands of targets at a time. Baits targeting highly repetitive elements are known to work especially well compared with baits targeting low copy regions, but this may not be a problem for studies targeting rDNA or mtDNA across many taxa. This method is likely to see increased use in biomonitoring and ecological studies, particularly those studies targeting relatively low abundance taxa such as arthropods from soil samples where bacteria and fungi tend to dominate DNA metabarcoding libraries. Hybridization-based capture is easy to implement and is a low-cost approach to improve the efficiency of high-throughput DNA metabarcoding. It has been suggested that the integration of hybridization enrichment in biodiversity analyses of signature DNA regions could mean a shift to a more meaningful interpretation of read numbers for CO1 metabarcoding studies, that is, reflecting biomass, but this needs further study as data on mitochondrial number variation and body size variation can be quite different even across a single taxonomic group such as the Insecta (Shokralla et al., 2016). Bead-based hybridization in solution can be conducted in 96-well plates and is more scalable than on-array enrichment, which also requires special equipment (Mamanova et al., 2010).

Selective circularization using molecular inversion probes (MIPs) works much like hybridization capture except that a universal sequence is flanked by target-specific sequences, such as restriction sites, and these constructs hybridize to sheared or digested DNA-forming loops. Once the MIPs have hybridized to their targets, nucleotides are added to fill the gap and ligation closes the circles. This method is highly specific. It has been shown that a large portion of sequences, however, map to the universal sequence and target-specific sequences (Tewhey et al., 2009). This method has the potential to detect fewer taxa for biodiversity analyses compared with hybridization capture so is less likely to be adopted by the molecular ecology community for biodiversity studies.

Target enrichment using PCR is often the first step in PCR-coupled DNA metabarcoding. Digital PCR (discussed above) can also be used for target enrichment prior to HTS (Tewhey et al., 2009). The main advantage of using PCR is its low cost, ease of implementation and the production of large volumes of template for library prior to HTS. However, there are many issues regarding amplification bias and subsequent changes from the original template ratios in mixed-template reactions and have already been discussed (Metabarcoding, above). Additionally, even digital PCR has its own biases and requires careful optimization. Because of the biases associated with mixed-template PCR, any method that avoids this is an attractive option for investigators who want to see a less-biased view of biodiversity in their samples.

## 12 | Future Outlook and Challenges

The purpose of this review was to provide a guide to commonly used as well as newer and lesser-known methods for genomics analysis of biodiversity data. Along with this, we also presented databases, tools and methods used with the widely popular and highly scalable DNA metabarcoding method for conducting biodiversity and biomonitoring studies. Despite widespread use of many of the techniques we review here, there remain challenges to DNA-based biodiversity analyses that need to be addressed before the field can move from descriptive works to a form that can be used to inform policy and management decisions or be utilized in long-term large-scale studies: (i) continued development of highly scalable laboratory methods, (ii) improving bioinformatic algorithms and their accessibility through robust software tools, (iii) large-scale integration of different data types and (iv) growth of reference databases.

### 12.1 | Scalable Laboratory Methods

The most popular data generation methods for high-throughput biodiversity and biomonitoring studies are scalable; that is, they can accommodate increases in number of samples to be processed because they are amenable to automation and parallelization. Kits are currently available to process samples from DNA extraction through to sequencing in plate-formatted batches. Microfluidics, however, can further miniaturize a reaction's footprint to microscopic lengths and to microlitre or picolitre volumes. Microfluidics, or lab-on-a-chip solutions, could play a role in biodiversity studies by reducing sample sizes, decreasing reaction times, increasing automation and eventually reducing cost (Dutse & Yusof, 2011; Liu & Zhu, 2005; Wu, Kodzius, Cao, & Wen, 2014). An integrated microfluidic solution that allows for DNA extraction, PCR and DNA fragment size detection on a single chip already exists (Easley et al., 2006). In the future, we could see how an integrated microfluidic solution that manages nucleic acid extraction through to sequencing could become the "sample-in-answer-out" holy grail for truly high-throughput biomonitoring that is rapid, reproducible, and eventually portable and easy to use by nonspecialists.

### 12.2 | Bioinformatics

We use the term bioinformatics to include not just raw sequence processing, but the implementation of algorithms for the analysis of
large-scale data sets. Current bioinformatic methods are a moving target, continually striving to keep up with the increasingly large data sets being generated by HTS platforms. Current challenges include improving the existing taxonomic and functional assignment tools, generally moving away from similarity- and phylogeny-based assignments in large-scale studies and moving towards composition-based, machine learning, and other hybrid methods that are faster and produce meaningful confidence values for assignments. We predict the next generation of algorithms will not only classify sequences, but attempt to predict which ones represent new species (Lan, Wang, Cole, & Rosen, 2012). The newest trends are random forest classifiers that can be used, for example, to predict sample origin based on community composition, that is, classification of whole communities as opposed to single taxa. Additionally, Bayesian classifiers can be used not only for taxonomic assignment but also for determining source/sink environmental interactions. For example, the Earth Microbiome Project analysed 2.2 billion 16S rDNA sequence reads from more than 23,000 samples, and they used a portion of this extensive microbial catalog to train a random forest sample classifier to predict the origin of the remaining samples (Thompson et al., 2017). They also used a leave-one-out cross-validated model with all source environments to determine which other environments were most similar. Another bioinformatic bottleneck is the production of reports and visualizations in an intuitive manner without the need for extensive programming skills. For example, a drag-and-drop type platform that allows users to explore different data visualizations, such as from microbiome studies, is already being developed (Bik, 2014). The ability to reduce large amounts of data into usable results, a process that can take just as long or even longer than the sampling process, will go a long way towards understanding complicated systems, and informing management decisions in a more timely manner.

12.3 | Integration of different data types

Biodiversity studies greatly benefit from databases containing DNA sequences (National Center for Biotechnology Information (NCBI), 1988; Ratnasingham & Hebert, 2007; Cole et al., 2014). Sequence data are not particularly meaningful on its own, however, without their metadata. A future challenge will involve strengthening linkages among the usual biodiversity metadata such as taxonomy, geographic information, local biotic and abiotic measurements, as well as incorporate earth observation data such as numerical weather data as well as photograph, radar and sonar imagery. For instance, addressing management impacts on a large scale to inform science-based decision-making will require marrying environmental data from Earth observation with biodiversity information for comprehensive modelling (Bush et al., 2017).

12.4 | Growth of reference databases

In the future, the ability to concurrently sample large numbers of unlinked markers from individuals as well as from eDNA samples in large-scale biodiversity studies will likely come from PCR-free techniques such as target enrichment, metatranscriptome, and metagenome sequencing (Hollingsworth et al., 2016). Each of these methods allows multiple regions of the genome to be captured, increasing the DNA sequence information per taxon and increasing the chances of detecting the greatest number of taxa. This information can only be fully leveraged when comprehensive reference sequence databases are richly annotated as well as designed to allow for efficient data mining and report generation.

Focusing on individual specimens and alpha taxonomy has been the tradition in biodiversity surveys of macroscopic organisms. Although specimens are necessary for assembling vouchers and reference sequence libraries, biomonitoring projects have gained momentum by including genomic analysis of environmental samples. It has already been shown that techniques such as DNA barcoding and metabarcoding can make significant contributions to biodiversity and biomonitoring studies (Janzen et al., 2005; Meier, Wong, Srivathsan, & Foo, 2016; Shokralla, Hellberg, Handy, King, & Hajibabaei, 2015). For better or worse, DNA-based methods are supplementing and, in some cases, even supplanting individual specimen-based collection for large-scale biomonitoring (Baird & Hajibabaei, 2012). Although multi-omics are often considered the future of community studies, in the microbial world, the thinking has come full circle. There has been a call for more work on isolating and cultivating specimens together with ecological observations to improve their understanding of microbial communities (Vilanova & Porcar, 2016). To provide some perspective, we borrow the analogy used by E.O. Wilson (Wilson, 2017), that DNA-based biodiversity and biomonitoring studies are like an aerial-survey; what we need are more “boots-on-the-ground”. Ultimately, the continued growth of high-quality reference sequences will only be possible in collaboration with taxonomists who have the expertise to find, collect, culture, and identify new specimens for DNA barcoding and WGS. If every “metabarcoder” reached out to include such experts in their projects, this could help to build a stronger foundation for the community as a whole. We hope this review provides some insight on how scalable DNA-based methods are currently becoming the leading source for acquiring biodiversity information.

AUTHOR CONTRIBUTIONS

T.M.P. and M.H. wrote the manuscript.

ORCID

Teresita M. Porter http://orcid.org/0000-0002-0227-6874
Mehrdad Hajibabaei http://orcid.org/0000-0002-8859-7977

REFERENCES

microbes in honey bee colony collapse disorder. Science, 318, 283-287. https://doi.org/10.1126/science.1146498


---

**How to cite this article:** Porter TM, Hajibabaei M. Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Mol Ecol*. 2018;00:1–26. https://doi.org/10.1111/mec.14478