Using environmental (e)DNA sequencing for aquatic biodiversity surveys: a beginner’s guide

Jennifer L. A. Shaw⁴, Laura Weyrich⁵ and Alan Cooper⁴

⁴Australian Centre for Ancient DNA (ACAD), University of Adelaide, Darling Building, North Terrace, Adelaide, SA 5000, Australia. ⁵Corresponding author. Present address: Commonwealth Scientific and Industrial Research Organisation, Land and Water Flagship, Prescott Building, Waite Road, Urrbrae, SA 5064, Australia. Email: jennifer.shaw@csiro.au

Abstract. Biological surveys are needed to monitor and assess the health of ecosystems and the species within them. However, morphology-based biodiversity surveys can be invasive, time consuming and financially expensive. Recently, environmental (e)DNA sequencing has been demonstrated as a potential alternative to morphological-based surveys because it enables the rapid and inexpensive detection of multiple taxa from DNA present in the environment. Numerous studies have shown that eDNA-based biodiversity surveys can provide considerable information about aquatic ecosystem function and health. Therefore, this molecular method has the potential to improve how current aquatic biological surveys are conducted. Currently, most eDNA literature is aimed at an audience with a moderate to advanced knowledge of DNA sequencing, creating a barrier for many ecologists who lack DNA sequencing expertise but wish to apply such methods to their research. The aim of this review is to provide guidance to non-geneticists regarding sequencing eDNA for aquatic biodiversity surveys and to highlight the requirements that need to be considered before the technique can be effectively incorporated into biomonitoring programs. Specifically, we provide details and recommendations on some of the major principles, from sample collection to bioinformatic analyses. For those areas where specific recommendations cannot be given, we have provided references to suitable literature.

Additional keywords: metabarcoding, metagenomics, methods, monitoring, NGS, summary.

Received 21 September 2015, accepted 10 November 2015, published online 1 March 2016

Introduction

Aquatic systems are under mounting pressure from climatic changes, pollution and overexploitation (Dudgeon et al. 2006; Crain et al. 2009). As a result, the health of both marine and freshwater ecosystems has declined globally, in terms of biodiversity loss (Butchart et al. 2010), species invasions (Strayer et al. 2006) and water quality (Foley et al. 2005). In order to prevent further declines, and perhaps improve the health and quality of these ecosystems in the future, regular monitoring and management is required. Currently, there are several commonly used biological monitoring methods for aquatic systems, each designed for a specific group of organisms. For example, monitoring small vertebrates or invertebrate species often involves invasive capture-based surveys, such as netting, whereas plant surveys are usually observational. Populations of larger vertebrates may be surveyed using a combination of both observational and capture methods (Karr and Chu 1999), whereas surveys of microscopic biota, such as bacteria, algae, fungi, archaea and viruses, may include a variety of laboratory protocols, such as microscopy, heterotrophic plate counts, flow cytometry (Hammes et al. 2008), denaturing gradient gel electrophoresis (Muyzer et al. 1993) or quantitative polymerase chain reaction (PCR) analysis (Arya et al. 2005). Although each of these survey methods is informative when applied to the appropriate taxa of interest, there are associated biases and limitations. For example, the invasive nature of capture-based sampling can damage ecosystems and increase predation risk to organisms (Resources Inventory Committee 1997), and difficulties associated with identifying cryptic species and juvenile life stages often confound traditional morphological identification-based methodologies. Further, a continuous decline in taxonomic expertise and the non-standardised skill levels of different taxonomists can limit surveys and introduce bias. In addition, because biological surveys of multiple trophic levels are often required for effective assessments of aquatic system health, multiple methods and ecologists specialising in a diverse range of organisms would be needed, further inflating financial costs. Because most monitoring programs are financially limited, the consequence is usually to scale back the extent of the survey, such as to reduce the number of locations surveyed or samples collected.

Molecular-based, environmental (e)DNA surveys are a potential alternative to traditional morphological-based biodiversity surveys. ‘eDNA’ simply refers to the DNA that is present
in an environmental sample, such as water, sediment, soil or faeces (Taberlet et al. 2012), and sequencing this eDNA enables identification of taxa from all life stages without morphological identification. To elaborate, larger organisms, such as mammals, birds, fish and reptiles, leave DNA traces within environments in the form of shedding skin, excretion and releasing gametes during reproduction, and DNA from smaller organisms can also be detected (e.g. bacteria, plankton etc). Using eDNA in this way provides a non-invasive and rapid way to identify many organisms from an environmental sample.

Using eDNA to detect organisms can refer to two different methods: (1) targeted sequencing, where a specific informative gene region is targeted and sequenced from within the DNA mixtures present in a sample; and (2) shotgun sequencing, where all DNA fragments within a sample are sequenced regardless of which gene they originate from. Shotgun sequencing may be useful for studies where there are multiple genes of interest, such as functional studies (Venter et al. 2004). However, because shotgun sequencing is non-discriminatory, a large majority of the sequence data may be uninformative. Targeted eDNA approaches can be further broken down into two main areas of research: (1) single species targeted eDNA detection, often used to detect a particular invasive or threatened species; and (2) multispecies targeted eDNA sequencing (also known as metabarcoding), which is used to obtain information about the biodiversity of an environment. Single species eDNA detection describes targeting a DNA sequence (also known as a genetic marker) that is unique to a particular organism from within an eDNA mixture. Single species targeted approaches have been experimentally validated for several important organisms and can provide semiquantitative abundance estimates (Laramie et al. 2015; Sigsgaard et al. 2015; Spear et al. 2015). However, because this approach is limited to detecting a single organism per reaction, it is not efficient when rapid identification of entire biological communities is required. In contrast, multispecies eDNA sequencing (metabarcoding) enables entire groups of organisms to be detected, such as vertebrates, fungi, plants or bacteria. It is achieved by targeting a less-specific genetic marker that comprises of two evolutionarily conserved DNA sequences (the target regions) interspaced by a highly variable species-specific sequence. This combination of conserved regions flanking a variable region makes multispecies identification possible in a single sequencing run, and therefore enables community biodiversity of an ecosystem to be characterised.

Multispecies eDNA sequencing was originally applied to bacterial communities (Chakravorty et al. 2007; Petrosino et al. 2009) but has since been extended to survey other taxonomic groupings, such as eukaryotic diversity (Chariton et al. 2010; Bik et al. 2012), vertebrates (Kelly et al. 2014; Thomsen et al. 2012a), invertebrates (Hajibabaei et al. 2011; Yu et al. 2012), fungi (Bellemain et al. 2010; Lim et al. 2010), plants (Burgess et al. 2011; Yoccoz et al. 2012) and viruses (Finkbeiner et al. 2008; Roosink 2012). At present, the method has been applied to a range of aquatic sample types, including marine ecosystems (Fonseca et al. 2010; Kimes et al. 2013), freshwater lakes and rivers (Sweeney et al. 2011; Drury et al. 2013), soils and sediments (Fierer et al. 2012) and drinking water distribution systems (Douterelo et al. 2013; Shaw et al. 2014). These studies have demonstrated that this method enables the detection of previously unprecedented biodiversity and can provide useful information about ecosystem functionality and health. Further, because a large number of samples can be processed in parallel on a single sequencing run and multiple species can be detected and identified from each sample, this approach is also comparably inexpensive and rapid compared with morphological-based methods (Baird and Hajibabaei 2012; Taberlet et al. 2012; Ji et al. 2013).

The benefits of metabarcoding from eDNA are clear, but many ecologists lack the expertise needed to optimise and validate the method for their field of research. Further, most eDNA literature assumes that readers have a firm understanding of genetic laboratory protocols, and many previous reviews have only discussed the possibilities and benefits of eDNA sequencing without explaining the methodological approaches directly. As a result, there is a lack of easily digestible published information detailing all the methodological steps for eDNA biodiversity surveys. This creates a barrier for many scientists who lack a strong genetic background but want to apply eDNA approaches. This review serves to provide an overview of eDNA survey procedures and aims to make this method more accessible to non-geneticists, allowing this approach to be more widely applied in aquatic ecosystems research. Specifically, we provide details and recommendations on some of the major principles, from sample collection to bioinformatic analyses, which require careful consideration before eDNA sequencing can be effectively used. Although this review focuses on aquatic environments, the underlying concepts also apply to other environments.

Overview of eDNA sequencing for biodiversity surveys

eDNA-based biodiversity surveys are comprised of five key steps (Fig. 1). Each step of the methodology should be optimised for the scientific question, the target taxa group and the inherent limitations associated with the sample type. The initial collection, processing and DNA extraction from the sample are all crucial steps for ensuring representative diversity is obtained. In the first part of the review we provide recommendations on the most efficient and effective sampling approaches, and then discuss the advantages and disadvantages of the most commonly used DNA concentration and extraction techniques.

After extraction, the target DNA in each sample is amplified by PCR (i.e. many identical DNA copies are made) using specially designed ‘tagged’ primers. The tagged primers bind to the target gene region (i.e. genetic marker) that is conserved across the taxon group of interest and the gene region is amplified across both the conserved and variable (i.e. species-specific) sections. During this process, a unique sample-specific tag is also incorporated into every amplified DNA sequence in that sample. This step enables multiple samples to be later combined and sequenced together, greatly reducing the throughput time and cost of sample processing. A brief description of PCR-amplification and information regarding the most commonly used genetic markers is provided later in the review.

Following PCR amplification, the samples are cleaned to remove all small non-target fragments left over from the PCR. Failure to remove the small non-target sequences can result in inefficient sequencing. After PCR, the tagged samples can be pooled together at equal molar concentrations to make a ‘DNA
library’. The DNA library is then sequenced using a next-generation sequencing (NGS) platform. The benefits and limitations of the (currently) most commonly used NGS platforms (454 Life Sciences, Roche, Branford, CT, USA; Ion Torrent PGM, Thermo Fisher, Carlsbad, CA, USA; and Illumina MiSeq, Illumina, San Diego, CA, USA) are reviewed herein, and we include information on how to design and construct tagged sequencing primers for generating DNA libraries on NGS platforms.

The final step of an eDNA survey is data processing and visualisation of results. Currently, there are numerous software packages available, including some specifically developed for users with little programming knowledge and others that enable more experienced users to customise pipelines. Each step within the bioinformatic analyses requires user-specified parameters that can markedly influence results if not chosen appropriately; therefore, it is imperative to have a general understanding of bioinformatic processes to enable robust interpretation of the data. Some of the key bioinformatic steps are detailed in this review.

**Contamination considerations**

First and foremost, contamination should always be considered as a constant risk and minimised where possible when conducting any DNA sequencing study. Contamination can be introduced into eDNA datasets at multiple stages, the most critical being during sample collection, DNA extraction and PCR set-up, as these stages are performed when the DNA is in low concentration (Cooper and Poinar 2000). To minimise the amount of contaminant DNA, samples should be collected using sterile, DNA-free storage containers while wearing clean laboratory gloves and minimising handling of both container and sample (Daniel and van Oorschot 2011).

Sample preparation and DNA extractions should be performed in a clean environment using DNA-free equipment. A 70% ethanol solution is suitable for general cleaning of the laboratory (e.g. door handles, shelves, sinks); however, laboratory benches and equipment should be decontaminated more thoroughly using a 1–3% sodium hypochlorite solution (ideally left in contact with the surface for 15 min) and subjected to ultraviolet (UV) radiation.

---

**Fig. 1.** Basic workflow of a targeted environmental (e)DNA sequencing biodiversity survey. (1) Environmental samples are collected in sterile containers; (2) DNA is concentrated and extracted; (3) the target gene region (genetic marker) is amplified by polymerase chain reaction (PCR) using primers uniquely tagged for each sample; (4) multiple samples are pooled together for sequencing on a next-generation sequencing (NGS) platform; and (5) after sequencing, the sequences are split into separate sample files and the raw data is processed and analysed using various computer packages. Images 1, 2a, 3 and 4 are the authors’ own. Image 5 sourced from www.freimages.co.uk (reproduced with permission).
where possible (Gefrides et al. 2010; Fraise et al. 2013). It is imperative that extraction blank controls (where the sample is omitted or replaced with nuclease-free water) are extracted at the same time as the samples (Salter et al. 2014). The extraction blank controls should be carried through the entire process and sequenced alongside the samples to monitor any taxa introduced by laboratory contamination.

Further, specific pre- and post-PCR zones within the laboratory should be designated (ideally in separate rooms). This is important to reduce cross-contamination from high-copy, post-PCR products from previous projects (Kwok and Higuchi 1989). Ideally, movement of all equipment and personnel should be restricted between the two laboratory areas to a one-way direction, following the DNA concentration gradient. Full decontamination of all laboratory surfaces (and any equipment used) should be performed before and after each use.

Sample collection

Given the heterogeneous nature of ecosystems, it is difficult to predict levels of sampling sufficient to capture representative community diversity. Currently, there are no standards regarding the minimal representative sampling volume needed to capture the complete community diversity present in an aquatic ecosystem. Previous eDNA studies have found that sample volumes ranging from as small as 15 mL (Ficetola et al. 2008) up to 100 L (Gomez-Alvarez et al. 2012) can both yield diverse biological communities and a greater number of taxa compared with conventional methods. For example, a study by Thomsen et al. (2012a) successfully detected a larger number of marine fish using only thirty 50-mL samples of seawater pooled into a 1.5-L composite sample compared with conventional (non-molecular) survey techniques.

The sample volume required to identify representative biological communities is likely to be dependent on target taxon abundance, total biomass, community diversity and the properties of the sample media. Kelly et al. (2014) found that the abundance and biomass of the taxon of interest is highly correlated with the number of DNA sequences detected. For that reason, samples such as eutrophic waters, sewage effluent and wastewater, which may contain millions of algal, bacterial or viral cells per millilitre (Gerardi 2006; Burkholder 2001) may only require a small sample volume to characterise microbial community diversity. Further, a smaller volume may also be sufficient for sample types taken from extreme environments, such as highly acidic or saline materials, which may have low diversity due to the dominance of a few highly specialised taxa. Conversely, the DNA of comparatively less abundant organisms, such as vertebrates, may be more dilute and heterogeneous in a given environment, and larger sample volumes may be required as a result. As with traditional morphological-based surveys, there are ways to determine whether the level of sampling effort is sufficient. For example, by using permuted accumulation and rarefaction curves, where the number of individuals is substituted for the number of sequences associated with the target taxa (Fig. 2). The optimal sample volume will ultimately depend on the biological properties of the environment; it is recommended that pilot studies be performed unless a priori knowledge of the area is available.

Previous eDNA studies have suggested that the number of samples taken per site is more important than sample volume. Andersen et al. (2012) found that soil samples taken from within large zoo enclosures (40 × 40 m) accurately represented the megafaunal diversity present regardless of soil sample volume (6.5–32.5 g soil), and that large vertebrate taxa were infrequently detected within spatial replicates throughout a sample site, highlighting the importance of replication when working with heterogeneously distributed DNA. Similarly, a study that measured fish taxon biomass in a controlled marine mesocosm found that diversity was higher and more representative of the true diversity when three 1-L samples were assessed compared with one 15-L sample (Kelly et al. 2014). Therefore, we recommend that multiple spatial replicates of smaller volumes are collected in favour of fewer samples of larger volume. Replicates should ideally include experimental replicates (multiple samples taken per site) and technical replicates (multiple extractions completed on a single sample); however, the level of replication is often limited by time and financial constraints associated with the survey, and studies have shown that optimal levels of replication are highly dependent on the detection probability of the taxa (Ficetola et al. 2015).

A further consideration for eDNA biodiversity surveys is the persistence of DNA in the environment. A targeted species-specific eDNA study by Dejean et al. (2011) found that DNA of vertebrate taxa (frogs and fish) could be detected for 1 month after the organism was removed from an experimental freshwater pond. Similarly, Ficetola et al. (2008) and Thomsen et al. (2012b) detected the DNA of vertebrate taxa for up to 2 weeks after their removal from a controlled setting. This has vast implications for the optimal design of sampling events in the field, as the DNA of different taxa may persist for different periods of time.

Fig. 2. Rarefaction curves representing taxa (operational taxonomic unit, OTU) count increasing as number of total sequences or samples increases. Rarefaction curves represent the mean of repeated resampling of all sequences. A curve that plateaus (dashed line) indicates that most or all of the diversity has been characterised. A steep curve that does not plateau (solid line) indicates that more sampling or deeper sequencing is needed. In order to assign an even sequencing depth across all samples an arbitrary cut off value may be assigned, where the sequences from each sample are randomly subsampled up to a specific number of total sequences. However, this is only appropriate if all samples are reasonably well characterised (dashed lines).
implications for eDNA surveys of fast-flowing aquatic environments, because traces of an organism’s DNA could be sequenced further downstream in uninhabited areas, thus giving an inaccurate representation of the community. In addition to this, a recent study by Merkes et al. (2014) found that eDNA has the potential to be geographically transported by secondary vectors, such as the faeces of predators and vessels such as boats. Therefore, these factors should be taken into consideration when interpreting eDNA survey data. However, the persistence time of DNA is highly dependent on variables such as temperature, oxygen levels, pH, UV intensity and bacterial activity (Burger et al. 1999; Strickler et al. 2015). Therefore, in natural environments, the persistence time may be much less than in controlled settings, which may have unnaturally high concentrations of organism DNA.

Sample processing and DNA extraction
Aquatic eDNA surveys usually require sampling volumes of water that are too large for most extraction protocols. Therefore, samples may need to be concentrated before DNA extraction. Filtration and centrifugation are the most commonly used concentration methods, but both have associated limitations. For example, samples containing large amounts of particulate matter can obstruct filter pores during filtration, whereas centrifugation may not be able to reach the high speeds required to pellet small cells or extracellular DNA. Therefore, centrifugation may be more appropriate for samples with large amounts of particulate matter, and filtration may be more suitable for concentrating clearer water samples. Alternatively, a combination of both methods could be used. For example, following centrifugation, DNA can be extracted directly from the pellet and the supernatant retained and subjected to filtration to capture additional DNA. Centrifugation speed can be optimised according to target taxa, with high g centrifugation (10 000g for 15 min at 4°C) effectively concentrating large bacterial cells (Hoeﬂ et al. 2003), whereas alternative methods are required to pellet smaller viruses (Lewis and Metcalf 1988). Filtration has been more commonly used in aquatic system eDNA studies. When sequencing marine fish DNA from 1.5-L seawater samples, Thomsen et al. (2012a) filtered the seawater through a 0.47-μm membrane to concentrate the DNA before extracting the DNA directly from the filter membrane. Douterelo et al. (2013) and Hwang et al. (2012) both used a 0.22-μm membranes to filter 1 and 10 L of drinking water respectively for their studies on bacterial communities. To prevent particulate matter obstructing filter pores and optimise capture of eDNA, Turner et al. (2014) used multiple filter sizes to fractionate the suspended particulate matter and the eDNA from water collected from the Great Lakes. They found eDNA was most abundant when using 0.2-μm pore sizes, and suspended particulate matter was most abundant when a 100-μm pore size were used. Furthermore, this sequential ‘stacking’ of a range of filters has the added benefit of fractionating the sample into different phases, which may be beneficial for separating intracellular DNA within intact cells from degraded extracellular DNA.

The often large number of samples in eDNA surveys projects creates a bottleneck at the extraction stage. Therefore, streamlined, standardised, high-throughput extraction kits are useful for speeding up processing times. Extraction kits, such as the Mo Bio (Mo Bio laboratories, Carlsbad, CA, USA) or Qiagen (Hilden, Germany) range, have been specifically designed to process difficult environmental samples, including those that contain high amounts of humic acids and other PCR inhibitors (i.e. sediments or wastewater samples; see manufacturer websites), in a reasonably rapid manner. However, despite the obvious benefits of using manufactured extraction kits, studies have shown that the DNA yield of commercial kits is generally lower than specially developed ‘in-house’ extraction techniques (Hurt et al. 2001; Martin-Laurent et al. 2001; Mahmoudi et al. 2011), although DNA quality is arguably more important than DNA quantity.

Different DNA extraction protocols can produce vastly different community profiles. For example, Deiner et al. (2015) found significantly different species detection rates ranging from 0 to 33% across different extraction protocols. These biases will likely be more significant in phenotypically diverse communities. For example, in microbial communities, fragile cells will lyse more easily than spore or cyst-forming taxa. If lysis is too gentle, then DNA from robust species will not be extracted; yet, if lysis is too vigorous, the DNA from more fragile species could become overly sheared. Both factors may skew the resulting community data. Therefore, we recommend that various extraction protocols are tested and optimised for specific taxa groups and sample types, and that a standardised extraction method is adopted across all samples.

Many eDNA extraction protocols for biodiversity studies feature a mechanical lysis step, such as bead beating, in the presence of a lysis buffer. Lysis buffer is a solution used for the purpose of lysing open the cells and minimises DNA damage once the DNA is released. Most lysis buffers contain salts (e.g. Tris-HCl or EDTA) to regulate the acidity and osmolality of the extraction solution, whereas detergents (e.g. Triton-X or sodium dodecyl sulphate (SDS)) are added to break up membrane structures. The choice of buffer and detergent depends on sample type and target organisms; for example, animal, bacterial and yeast cells all have differing requirements for optimal lysis because of the presence or absence of a cell wall. Other important considerations for optimal DNA extraction include temperature, duration of lysis and mechanical lysis choice.

There are many choices of bead for bead-based mechanical lysis of environmental cells. Garnet beads are non-uniform in size and shape and have sharp edges; this variation can help homogenise and break apart sediments while shearing open the cells within the sample. However, garnet beads are not suitable for prolonged periods of bead beating because they can easily break apart and lose the ability to lyse cells effectively. Ceramic, glass and metal beads are more robust and therefore suitable for longer bead beating periods, which may be needed to lyse more resilient cysts and spores. However, it must be noted that extended bead beating can result in heat damage to extracellular DNA, leading to lower yields or sheared DNA (von Atzigen and Kennedy 2011). Therefore, when the target group of interest contains species of varying fragility, it may be best to use two separate extraction protocols (more and less intensive) per sample and combine the extractions at a later stage in the process. Von Atzigen and Kennedy (2011) recommend a bead
beating speed of 4000 rpm for 45 s as a standard starting point when extracting DNA from complex environmental communities. This speed and the length of time should be increased when targeting taxa with more robust cell walls, such as spores or cysts, and reduced when dealing with more fragile cells. The speed and time required will also vary according to the bead type and size (von Atzigen and Kennedy 2011).

PCR amplification, primer design and genetic marker choice

eDNA-based biological surveys involve amplifying and then sequencing an appropriate gene region or ‘genetic marker’ to obtain informative DNA sequences from a target taxon group. An ideal genetic marker can be defined as a short species- or genera-specific genomic region flanked on either side by a genomic region that is conserved across multiple target taxa. Primer sets are designed to be complementary to the conserved region, causing the molecule to bind to the DNA fragment and enabling PCR amplification through the variable (species-specific) interior section. It is important that primer sets are equally complementary to all taxa within the target group and therefore amplify all DNA sequences with similar efficiency to avoid taxonomic bias (Polz and Cavanaugh 1998). Unique sample specific ‘tags’ (sometimes referred to as an index or barcoded tag) are also incorporated into the primer design to enable discrimination between sequences from different samples (Fig. 3). Tags are usually ~6–12 bp in length and must differ from other sample tags by at least two base pairs so that no single sequencing error can result in misidentification (Binladen et al. 2007; Shokralla et al. 2012).

There are already a range of commonly cited genetic markers for eDNA-based biodiversity surveys (Table 1), including the 16S genetic marker for bacteria, the 18S marker for eukaryotes, and the CO1 and 12S marker for vertebrates. By default the more commonly used genetic markers tend to have larger sequence databases, providing broad taxonomic reference collections to match sequences against. Genes that are highly conserved across taxonomic space tend to be hampered with a reduction in phylogenetic resolution for species identification. For example, 18S rDNA, which has been successfully used for eukaryote communities, and 16S rDNA, which has been similarly used for bacterial communities, are both conserved across a wide range of phyla but often lack the resolution to identify sequences beyond the family or genus level (Creer et al. 2010; Bik et al. 2012). Therefore, careful selection of an appropriate primer set is crucial for balancing the widest taxonomic coverage with the best phylogenetic resolution and minimal taxonomic bias.

Despite the existence of these commonly used genetic markers, it is important that we do not become complacent with genetic marker choice. Discovery and design of new, more-effective genetic markers could lead to improved eDNA surveys in future. To facilitate the discovery of optimal gene regions for the target taxa, several bioinformatic tools have been developed,
Table 1. List of commonly used targeted multispecies environmental (e)DNA genetic markers for various taxon groups and references citing their use

<table>
<thead>
<tr>
<th>Genetic marker</th>
<th>Target taxon group</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>16S rRNA</td>
<td>Bacteria</td>
<td>Claesson et al. (2010), Chakravorty et al. (2007)</td>
</tr>
<tr>
<td>18S rRNA</td>
<td>Eukaryotes</td>
<td>Charnot et al. (2010), Medinger et al. (2010), Bik et al. (2012)</td>
</tr>
<tr>
<td>12S rRNA</td>
<td>Vertebrates</td>
<td>Kelly et al. (2014), Hardy et al. (2011), Riaz et al. (2011)</td>
</tr>
<tr>
<td>Mitochondrial CO1</td>
<td>Vertebrates and invertebrates</td>
<td>Hebert et al. (2003), Hajibabaei et al. (2011), Yu et al. (2012)</td>
</tr>
<tr>
<td>ITS</td>
<td>Fungi and algae</td>
<td>Nilsson et al. (2008), Schoch et al. (2012)</td>
</tr>
<tr>
<td>Chloroplast rbcL</td>
<td>Plants</td>
<td>Hollingsworth et al. (2009)</td>
</tr>
<tr>
<td>Chloroplast trnL</td>
<td>Plants</td>
<td>Hollingsworth et al. (2009), Yoccoz et al. (2012)</td>
</tr>
</tbody>
</table>

such as PRIMER 3 (Rozen and Skaletsky 1999), ecoPrimer (Riaz et al. 2011) and GENEIOUS (Kearse et al. 2012). These tools also enable the in silico simulated amplification of target taxa against potential primer sets, enabling the amplification efficiency and genetic resolution of the primers to be studied before purchase. However, it is important to note that in silico amplification may not always accurately reflect in vivo applications (Clarke et al. 2014), and primers should therefore be experimentally validated before use. Currently, the adoption of new multispecies genetic markers in eDNA studies is hindered by the limited reference sequences available for some genomic regions within genetic repositories, and this should be considered during experimental design.

NGS: primer construct and NGS platform choice

One of the major benefits of eDNA sequencing surveys is the ability to sequence many samples in parallel, greatly reducing the throughput time and cost of biomonitoring. The incorporation of sample-specific tags within the primer set construct enables many samples to be physically combined and analysed in parallel, and subsequently separated during data processing. In addition to the sample-specific tag, a sequencing platform adaptor sequence must also be incorporated into the design (Fig. 3). The platform adaptor sequence enables the NGS platform to recognise and sequence the DNA fragments. The adaptor sequences can be obtained from the platform manufacturer websites, along with other useful instructions regarding DNA library preparation. Once the amplified DNA sequences within a sample contain the sample-specific tag and a platform adaptor sequence, multiple samples can be pooled in equimolar concentrations to form the ‘DNA library’. The DNA library is then quantified, diluted to a concentration appropriate for the NGS platform and sequenced.

There are two main methods for constructing eDNA libraries. The simplest approach involves a single PCR amplification using so-called fusion primers, which are composed of the platform adaptor sequence, the unique sample-specific tag and the sequence complementary to the target gene region (Fig. 3a). The second approach involves two PCR amplification steps (Fig. 3b). The first PCR step simply amplifies the target gene and adds a short artificial linker sequence; the second PCR step then uses primers that bind to the linker sequence to add the full-length platform-specific adaptor and sample-specific tag to the target sequence (de Cárcer et al. 2011). The second PCR step should be constrained to just a few cycles (approximately five to seven cycles) because unnecessary amplification can skew the proportion of sequences within the amplification reaction further (PCR bias; Polz and Cavanaugh 1998). Each approach has benefits and limitations. The simplicity of the single PCR approach minimises PCR bias and contamination risk, but there are considerable financial costs for generating a large number of fusion primers each with different sample-specific tags. In the two-step approach, a sample-specific tag could be incorporated into both the first and second primer sets, creating a larger number of sample–tag combinations, which would reduce the costs associated with obtaining many uniquely tagged primers. Furthermore, the two-step approach allows versatility in the choice of NGS platform because the platform-specific adaptor sequence is incorporated during the second PCR step; therefore, the same primers can be used on multiple platforms.

Several NGS platforms have been marketed since the first widely available 454 Life Sciences (Roche) system (Mardis 2008; Shokralla et al. 2012; Shendure and Ji 2008). Currently, the dominant Illumina (MiSeq, NextSeq) and Life Technologies (Ion Torrent PGM) systems are restricted to sequencing regions between 100 and 600 bp in length, albeit at very economic prices and enormous capacity. The estimated data output of each platform constrains the number of samples that can be run in parallel, because a greater number of samples reduces the number of sequences per sample. For example, at the time of writing, the Miseq platform could generate 25 million sequences (on average 15 Gb of data) per run; therefore, a sequencing run containing 100 samples would yield ~250 000 sequences per sample. However, NGS platforms undergo regular hardware and software upgrades; therefore, it is recommended that up-to-date information is obtained directly from the manufacturer’s website before making a decision regarding platform choice. Other NGS platforms are associated with more specialist requirements, such as the long read lengths available using 454 (Life Sciences), or single molecule sequencing using Pacific Bioscience’s Sequel system (Pacific Biosciences, Menlo Park, CA, USA).

Bioinformatic analysis

NGS of eDNA produces extremely large datasets that need to be carefully processed to avoid misleading conclusions. Several processing steps are required before the data can be interpreted and analysed (Fig. 4), and each step can affect the results (Kunin et al. 2010), making it important to fully understand, justify and report all parameters used. The first step generally involves...
sequences together as a single taxonomic cluster before speed, many researchers opt to cluster all highly similar sequences from forming centroids. Other de novo algorithms have adopted useful strategies to avoid both fixed clustering thresholds, and sequence input-order dependency (e.g. Swarm clustering method; Mahé et al. 2014).

Alternative, clustering methods include closed and open reference picking. Closed reference clustering algorithms use taxonomically verified sequences from genetic databases as the OTU centroids. Sequences within the data pool that are within the similarity threshold to species within the genetic database are clustered together. Although this method offers a more stringent approach to OTU picking, it is limited to the sequences available within the genetic database, and biases could be introduced by database incompleteness (Bik et al. 2012). To overcome this, open reference OTU picking operates by carrying out closed reference OTU clustering, followed by de novo

Fig. 4. Workflow of bioinformatic processing steps used to analyse metagenomic data. (OTU, operational taxonomic unit.)
clustering on any remaining sequences outside the specified threshold.

Finally, the filtered and clustered OTU sequences are compared with a genetic reference database to identify the taxa present. Either the most representative (abundant) sequence in an OTU or a consensus of all the sequences within an OTU is selected to search and match to taxonomically verified sequences within a genetic database. GenBank is a comprehensive online genetic database (http://www.ncbi.nlm.nih.gov/genbank, accessed January 2015); however, it is not as extensively curated as other smaller databases and, as such, sequence matches may be inaccurate or uninformative. More heavily curated databases, such as Greengenes (bacterial 16S rRNA; http://greengenes.secondgenome.com, accessed January 2015), UNITE (fungal internal transcribed spacer (ITS); http://www2.dpes.gu.se/project/ unite/UNITE_intro.htm, accessed January 2015), SILVA (eukaryotic 18S/16S rRNA; http://www.arb-silva.de/, accessed January 2015) and BOLD (eukaryotic CO1; http://www.barcodinglife.com/, accessed January 2015) enable accurate and reliable taxonomic identification for eDNA data. However, for studies involving a specific group of taxa (e.g. fish from a specific region), it may be appropriate to select a subset of GenBank sequences and create a personally curated local database to increase computational efficiency at this step. Once taxonomic identities have been assigned to each OTU cluster, a species abundance table (or OTU table) can be generated that describes the number of sequences attributed to each OTU or a consensus of all the sequences within an OTU is present. Either the most representative (abundant) sequence in each sample to compare species richness and evenness across samples. Box plots, phylogenetic trees, heat maps, principal variate analysis software tools such as Primer-E and R, at this stage eDNA data should be easily transferrable to researchers from non-genetic backgrounds.

Bioinformatic tools

Custom building a set of command line scripts to process eDNA data is a flexible option that encourages an advanced level of analytical comprehension. Custom bioinformatics pipelines can be built using tools such as GALAXY (Blankenberg et al. 2010), BioPerl (http://perl doc.perl.org/, accessed January 2015), Bioperl (http://doc.bioperl.org/, accessed January 2015) and R (R Foundation for Statistical Computing, Vienna, Austria; see http://www.R-project.org, accessed January 2015), followed by programs such as MEGAN (Meta Genome Analyzer; Huson et al. 2007), EXPLICET (Robertson et al. 2013) and R (see http://www.R-project.org) to visualise the data. However, development of customised pipelines requires a moderate to advanced knowledge of computer scripting and, as a result, many researchers may opt for online data processing programs, such as the Ribosome Database Project (RDP; Cole et al. 2005) and MG-RAST (Meyer et al. 2008). These programs are specifically developed to allow standardised eDNA data processing and analysis for scientists with few bioinformatics skills. However, although these programs are useful, a lack of understanding of the processes and key assumptions made during the processing steps can result in incorrect data interpretation. The publicly available QIIME software (Caporaso et al. 2010) is a popular choice for scientists who may not be confident enough to create their own customised pipelines but who want to maintain control and flexibility when analysing eDNA data. QIIME provides a user-friendly, but customisable, workflow that involves selecting from a set of ready-made scripts and parameters in the pipeline. It encourages researchers to process data themselves by providing guidance for each step of the process. Within QIIME, the UniFrac software provides useful diversity estimates based on phylogenetic divergence between taxa (Lozupone and Knight 2005; Lozupone et al. 2006) and can be used to compare samples and reveal patterns (Lozupone and Knight 2008). Regardless of the data processing approaches used, it is crucial that each of the bioinformatics steps is justified and selected according to the research question applied. It is recommended that where possible scientists analyse their own data to minimise misinterpretation of data, using bioinformatic tools that suit their individual skill level.

Visualisation and interpretation of eDNA data

An OTU abundance table is the common starting point for data visualisation and interpretation. From this matrix, the species detected, community diversity and patterns in community structure can be examined. However, there are a few important steps to consider. First, it is important to normalise the number of sequences across samples, because the number of sequences in each sample can vary considerably because of pipetting error or unavoidable sequencing biases. Normalisation can be achieved by converting the total sequence counts per sample into proportions or percentages. Alternatively, the number of sequences per sample could be rarefied to a specified value. Rarefaction randomly selects a specified number of sequences from the sequence pool in a permutational manner so that each sample contains an even number of sequences. Rarefaction curves, plotting the number of OTUs as a function of the number of sequences, can be viewed to determine whether a cut-off value (limiting the number of sequences used per sample) is appropriate for the dataset (Fig 2); a steep slope indicates that a large fraction of the species diversity remains to be discovered, a flatter slope indicates that a reasonable number of individual sequences have been considered in the rarefaction. Therefore, specifying a rarefaction cut-off value should be avoided if the curve slope for any one of the samples is disproportionately steep.

In addition to normalising the data, it may be necessary to apply transformations to the data before carrying out statistical analyses to down weight highly abundant taxa or vice versa, as multispecies eDNA data rarely follow a normal distribution. There are several software programs that enable informative visualisation of multispecies eDNA data (e.g. QIIME, EXPLICET, MEGAN, R). Biodiversity measures, such as the Shannon–Wiener and Chao diversity indices, may be applied to each sample to compare species richness and evenness across samples. Box plots, phylogenetic trees, heat maps, principal coordinate analysis (PCoA) plots and non-metric multidimensional scaling (nMDS) plots can be used to visualise patterns in the community structure. These visualisation tools enable data to be easily interpreted by ecologists that may not have a strong genetic background. OTU tables can often be exported into statistical software programs that were not initially designed for interpretation of eDNA data (e.g. Primer-E, Clarke 1993). Because most biomonitoring organisations already use multivariate analysis software tools such as Primer-E and R, at this stage eDNA data should be easily transferrable to researchers from non-genetic backgrounds.
Sequence read abundance and species richness

Strictly quantitative organism abundance measures based on the sequence read abundance is not recommended for multispecies eDNA data. This is due to several reasons. For example, most organisms contain multiple copies of the target gene in their cells, and these copy numbers can vary by an order of magnitude across species (Rooney and Ward 2005; Lee et al. 2009). Further, several steps in the eDNA protocol can skew the relative abundance of a sequence read even further, such as DNA extraction (DeSantis et al. 2005; Feinstein et al. 2009), primer choice (Jumpponen 2007; Engelbrektson et al. 2010), PCR bias (Polz and Cavanaugh 1998) and sequencing biases and errors (Kumin et al. 2010). However, several studies have found that read counts are semiquantitative, where differences in the proportional abundance of sequences for a given species across samples reflects the actual proportional abundance of that species in the environment (Porazinska et al. 2009; Kelly et al. 2014; Elbrecht and Leese 2015). Despite this, the ability to infer species abundance from multispecies eDNA datasets is still regarded as highly contentious (Deagle et al. 2013), especially for larger eukaryotic organisms, which may have considerably variable biomasses and gene copy numbers. In particular, a study by Amend et al. (2010) examined the relationship between read abundance and biological abundance for known quantities and identities of fungi along a dilution gradient and found that read abundance is approximately quantitative within species and across samples, but between-species comparisons are more tenuous. Further, if validation of proportional species abundances is required, multispecies data could be verified using single species genetic markers, because several studies have identified strong positive relationships between read abundance and population densities using species-specific targeted eDNA markers (Takahara et al. 2012; Thomsen et al. 2012b; Pilliod et al. 2013).

Concluding remarks

NGS technologies and bioinformatics programs have provided revolutionary new ways to assess biodiversity in complex ecosystems (Binladen et al. 2007). To date, eDNA has been used to successfully assess community structure and biodiversity in several aquatic habitats, including marine, lakes, river systems and drinking water distribution systems. However, in order for this technique to become a regular biomonitoring tool, important concepts and considerations need to be better understood by ecologists by making DNA-based protocols more transparent and understandable to non-geneticists. Recent improvements in sequencing platforms and bioinformatic tools, as well a reduction in the sequencing time and financial costs of high-throughput genomics (Logue et al. 2008; Pennisi 2011), offer strong potential for eDNA approaches to be used as a biodiversity monitoring tool. This review provides a beginners guide to the practical application of multispecies eDNA sequencing to aquatic systems and highlights several important considerations for integrating this approach into routine biomonitoring programs.

References


A guide to eDNA-based biodiversity surveys


