VERTEBRATE METABARCODING **RESULTS**

Order number:	SO01111
Report number:	NM-RFU192
Company:	Scotland: The Big Picture
Contact:	James Nairne
Project:	Pilot water sampling Aberdeen
Sample type:	NatureMetrics eDNA disk filter
Date of report:	22-Nov-2022
Number of samples:	4

Thank you for sending your samples for analysis by NatureMetrics. Your samples have been metabarcoded following our eDNA survey - Vertebrate pipeline. **A taxon-by-sample table of your samples is attached to this report (NM-RFU192.SO01111.Vertebrate.xlsx)**. Each row in the table represents one taxon (OTU), shown with the lowest possible taxonomic assignment based on currently available reference data. Each column represents a sample, showing the proportion of sequence reads per detected OTU. Care should be taken in interpreting the numbers in terms of relative species abundance, but a high sequence proportion can be interpreted as lending greater confidence to a detection. This report contains biodiversity information that may be sensitive, particularly with respect to endangered or protected species. It is the responsibility of the client to ensure that due consideration is given to the data and that the information is shared in a responsible way.

Here we present an overview of the key results, followed by a more detailed report that starts with the taxonomic composition of the samples followed by a more detailed look at the steps taken to extract, amplify, sequence, and analyse your DNA. A glossary for terms in bold is provided at the end of the report to define key terms used within the report.

OVERVIEW OF YOUR RESULTS

- A total of 12 taxa were detected.
- Average taxon richness was 4.25 and ranged from 3 to 5.
- Most abundant sequences: three-spined stickleback (*Gasterosteus aculeatus*).
- Most commonly detected taxa: ring-necked pheasant (*Phasianus colchicus*) and three-spined stickleback (*Gasterosteus aculeatus*).
- Species of note: European eel (*Anguilla anguilla* - Critically Endangered).

FULL REPORT

Sample composition

A total of 12 taxa were detected (**Table 1**). 58.3% (7 taxa) were at least 99% similar to a species in the global reference databases, and species names are suggested. The remaining taxa were identified to the lowest possible taxonomic level: 25% to genus (3 taxa) and the remaining 16.7% to family (2 taxa). A total of 2 unique fish, 3 amphibians, 5 birds and 2 mammals were detected. The taxa belong to 9 orders, 12 families, and 10 genera.

Species of note include the: European eel (*Anguilla anguilla* - Critically Endangered).

The average taxon richness was 4.25 and ranged from 3 ('HSM1A') to 5 ('HSM1B' and 'HSM1D'). The relative proportion of the sequences found in each of the samples is shown in **Figure 1** and **Table 1** and the diversity is summarised in **Table 2** and **Table 3**.

Three-spined stickleback (*Gasterosteus aculeatus*), which accounted for 51.5% of the total sequence reads, was among the most abundant in terms of sequences. Among the most commonly detected species were ring-necked pheasant (*Phasianus colchicus*) and three-spined stickleback (*Gasterosteus aculeatus*), which were detected in 3 and 2 samples, respectively .

High-quality vertebrate sequence data were obtained for all 4 of the eDNA samples. All laboratory controls behaved as expected.

Table 1 (attached separately). Taxon-by-sample table.

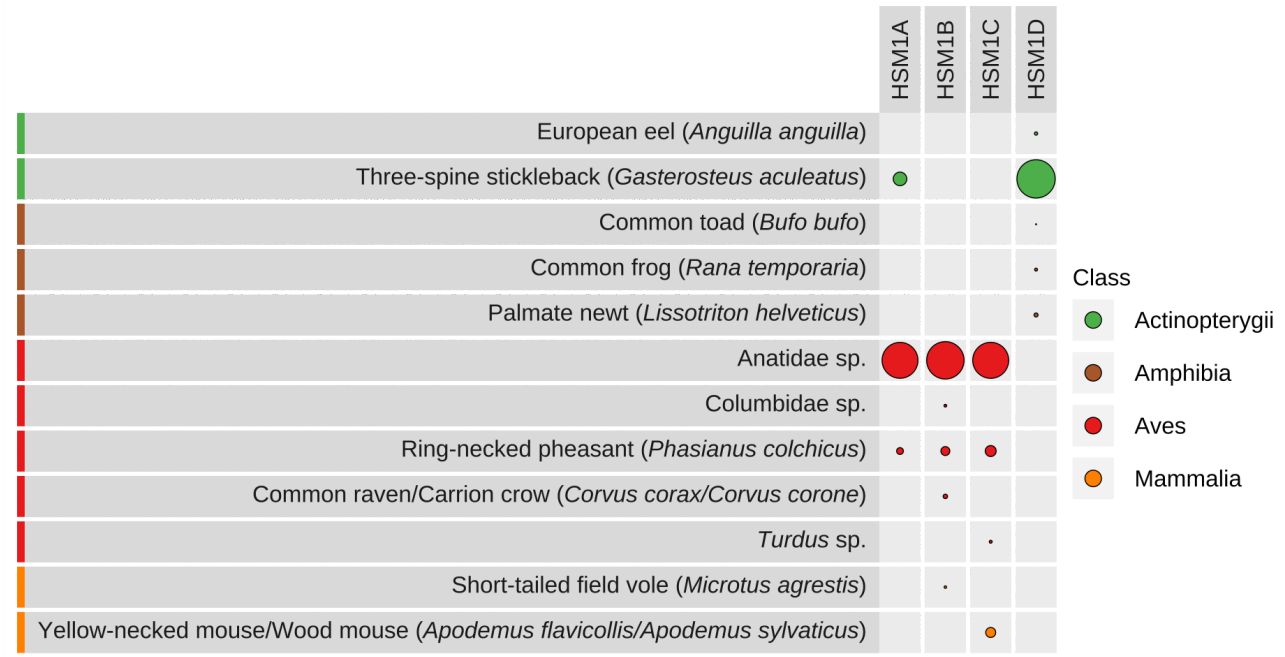

Figure 1. The proportion of the sequencing output allocated to the different taxa (rows) within each sample (columns). Each bubble per sample represents the proportion of DNA for each taxon for that sample. The size of the bubble is relative to the number of sequences from all taxa detected in that sample.

Table 2. Taxon richness among the samples.

Sample ID	Class	Order	Family	Genus	Taxa (Species)
HSM1A	2	3	3	2	3 (2)
HSM1B	2	5	5	3	5 (2)
HSM1C	2	4	4	3	4 (1)
HSM1D	2	4	5	5	5 (5)

Table 3 (attached separately). The frequency of occurrence of all detected families. Numbers correspond to the number of taxa belonging to those families in those samples.

METHODS

DNA from each filter was extracted using a commercial DNA extraction kit with a protocol modified to increase DNA yields. An extraction blank was also processed for the extraction batch. DNA was purified to remove PCR inhibitors using a commercial purification kit.

Comment: DNA yields were as expected.

Purified DNAs were amplified with PCR for a hypervariable region of the 12S rRNA gene to target vertebrates as part of the eDNA survey - Vertebrate pipeline. Our standard analysis includes 12 replicate PCRs per sample.

All PCRs were performed in the presence of both a negative control and a positive control sample. Amplification success was determined by gel electrophoresis.

Comment: PCR reactions were successful for all 4 samples. Electrophoresis bands were strong and of the expected size. Overall, 12 successful PCRs replicates were obtained for each of the 4 samples submitted for sequencing. No bands were observed on electrophoresis gels for the extraction blank or negative controls.

PCR replicates were pooled and purified, and sequencing adapters were added. Success was determined by gel electrophoresis.

Comment: All samples were successfully indexed, electrophoresis bands were strong and of the expected size. No repeat reactions were necessary.

Amplicons were purified and checked by gel electrophoresis, these were then quantified using a Qubit high sensitivity kit according to the manufacturer's protocol.

Comment: All amplicons were successfully purified.

All purified index PCRs were pooled into a final library with equal concentrations. The final library was sequenced using an Illumina MiSeq V3 kit at 10.5 pM with a 20% PhiX spike in.

Sequence data were processed using a custom bioinformatics pipeline for quality filtering, OTU clustering, and taxonomic assignment.

Comment: Both negative and positive controls were as expected. Very few sequences were discarded prior to dereplication, which is indicative of high-quality data with minimal PCR and sequencing errors. A total of 173,084 high-quality sequences were included in the final dataset.

Consensus taxonomic assignments were made for each OTU using sequence similarity searches against the NCBI nt (GenBank) reference database. Assignments were made to the lowest possible taxonomic level where there was consistency in the matches. Conflicts were flagged and resolved manually. Minimum similarity thresholds of 99%, 97%, and 95% were used for species-, genus- and higher-level assignments respectively. In cases where there were equally good matches to multiple species, public records from GBIF were used to assess which were most likely to be present in the United Kingdom. Higher-level taxonomic identifications or multiple potential identifications were reported in cases that could not be resolved in this way.

The OTU table was then filtered to remove low abundance OTUs from each sample (<0.025% or <10 reads, whichever is the greater threshold for the sample). Unidentified, non-target, and common contaminant sequences were then removed.

Note that unidentified or misidentified taxa can result from incomplete or incorrect reference databases, and taxa may be missed due to low quality DNA, environmental contaminants, or the dominance of other species in the sample.

Please note that the abundance of taxa cannot be directly inferred from the proportion of total sequence reads. While the proportion of sequence reads is a consequence of abundance, it is also impacted by biomass, activity, surface area, condition, distance from the physical sample, primer bias, and species-specific variation in the genome.

Table 4. Sample information table.

Kit ID	Sample ID	Volume filtered	Date received
ASD-01-01660	HSM1A	1800ml	26-Sep-2022
ASD-01-01659	HSM1B	1300ml	26-Sep-2022
ASD-01-01658	HSM1C	1900ml	26-Sep-2022
ASD-01-01661	HSM1D	1500ml	26-Sep-2022

END OF REPORT

Report issued by: Sophie Gooding

Contact: team@naturemetrics.co.uk

GLOSSARY

adapter	short, artificially synthesised nucleotide sequence which attaches to the ends of the target DNA or RNA sequences prior to sequencing. They are typically used to aid in attachment of the target sequence to other functional molecules/sequences.
amplicon	A DNA sequence which is the product of PCR amplification.
bioinformatics	An interface between genetics, computational biology, statistics, and programming in which DNA or other biological data is processed, analysed and integrated into research or communications.
bioinformatics pipeline	Refers to a data processing pipeline that takes the raw sequence data from high-throughput sequencing (often 20 million sequences or more) and transforms it into usable ecological data. Key steps for metabarcoding pipelines include quality filtering, trimming, merging paired ends, removal of sequencing errors such as chimeras, clustering of similar sequences into molecular Operational Taxonomic Units, and matching one sequence from each cluster against a reference database. The output is a OTU-by-sample table showing how many sequences from each sample were assigned to each OTU.
BMWP	Short for biological monitoring working party, an index that can be used to measure water quality by scoring the presence of aquatic invertebrate indicator taxa. The index is reliant on taxa that are less tolerant of polluted water bodies (e.g. Ephemeroptera, Plecoptera, Trichoptera).
BOLD	Barcode Of Life Database; a specialised database of eukaryote COI reference sequences.
contaminant sequences	The sensitivity of high-throughput sequencing of eDNA means that contamination is always a concern that needs to be minimised. The sources of contamination are threefold: Natural - Examples of natural contaminants include: frequent visitors to site, faecal discharge from predators, livestock, wastewater, and fishing bait. This type of contamination is typically unavoidable and very difficult to quantify. Sequences of this type are typically flagged and conservatively removed from the sequencing output. Typical contaminant species include cow, pig, dog, cat, sheep, etc. Sampling - Human contamination of sampling equipment can reduce the efficiency of the sequencing. This type of

	contamination can be minimised by stringent contamination protocols, such as PPE. Laboratory - Residual DNA can contaminate other samples processed at the same time in other labs. At NatureMetrics this is mitigated by a designated eDNA laboratory, strict decontamination procedures, negative controls, and good laboratory practices.
dereplication	The identification of unique sequences so that only one copy of each sequence is reported.
eBioAtlas	A global partnership between IUCN and NatureMetrics to map the world's biodiversity using DNA from water samples as a foundation for the Global Biodiversity Framework and to enable IUCN Red List Assessments.
eDNA	Short for 'environmental DNA'. Refers to DNA deposited in the environment through excretion, shedding, mucous secretions, saliva etc. This can be collected in environmental samples (e.g. water, sediment) and used to identify the organisms that it originated from. eDNA in water is broken down by environmental processes over a period of days to weeks. It can travel some distance from the point at which it was released from the organism, particularly in running water. eDNA is sampled in low concentrations and can be degraded (i.e. broken into short fragments), which limits the analysis options.
extraction blank	A DNA extraction with no sample added to assess potential contamination during the DNA extraction process.
gel electrophoresis	The process in which DNA is separated according to size and electrical charge via an electric current, while in a gel. The process is used to confirm the successful amplification of a specifically sized fragment of DNA.
high-throughput sequencing	Technology developed in the 2000s that produces millions of sequences in parallel. Enables thousands of different organisms from a mixture of species to be sequenced at once, so community DNA can be sequenced. Various different technologies exist to do this, but the most commonly used platform is Illumina's MiSeq. Also known as Next-Generation Sequencing (NGS) or parallel sequencing.
inhibitors/inhibition	Naturally-occurring chemicals/compounds that cause DNA amplification to fail, potentially resulting in false negative results. Common inhibitors include tannins, humic acids and other organic compounds. Inhibitors can be overcome by either diluting the DNA (and the inhibitors) or by additional cleaning of the DNA, but

dilution carries the risk of reducing the DNA concentration below the limits of detection. At NatureMetrics, inhibition is removed using a commercial purification kit.

invasive

Invasive species are defined using GRIIS (Global Register of Introduced and Invasive Species) which is a checklist of Introduced and Invasive species for each country. The IUCN describes an Introduced species as a species outside of its natural range and dispersal potential, and an Invasive species as an introduced species which becomes established in a habitat, is an agent of change or threatens native biological diversity.

IUCN Red List

The IUCN (International Union for the Conservation of Nature) is a global union of government and civil organisations that disseminates information to assist conservation. The IUCN Red List of Threatened Species is an inventory of the conservation status of over 100,000 species worldwide. The Red List evaluates data such as population trends, geographic range and the number of mature individuals in order to categorise species based on their extinction risk:

Extinct (EX) - No individual of this species remains alive.

Extinct in the Wild (EW) - Surviving individuals are only found in captivity.

Critically Endangered (CE) - species faces an extremely high risk of extinction in the wild. e.g. Population size estimated at fewer than 50 mature individuals.

Endangered (EN) - species faces a very high risk of extinction in the wild. e.g. Population size estimated at fewer than 250 mature individuals.

Vulnerable (VU) - species faces a high risk of extinction in the wild. e.g. Population size estimated at fewer than 10,000 mature individuals and declining.

Near Threatened (NT) - species is below the threshold for any of the threatened categories (CE, E, V) but is close to this threshold or is expected to pass it in the near future.

Least Concern (LC) - species is not currently close to qualifying for any of the other categories. This includes widespread and abundant species.

Data Deficient (DD) - There is currently insufficient data available to make an assessment of extinction risk. This is not a threat category - when more data becomes available the species may be recategorised as threatened.

Jaccard similarity index

This index is a calculation that compares two samples to see which taxa are shared and which are distinct. The higher the percentage,

	the more similar two samples are in their community composition.
metabarcoding	Refers to identification of species assemblages from community DNA using barcode genes. PCR is carried out with non-specific primers, followed by high-throughput sequencing and bioinformatics processing. Can identify hundreds of species in each sample, and 100+ different samples can be processed in parallel to reduce sequencing cost.
NCBI nt	National Centre for Biotechnology Information nucleotide database; a general reference database.
negative control	Used to determine whether PCR reactions are contaminated.
NMDS	Non-metric multidimensional scaling (NMDS) is a method that allows visualisation of the similarity of each sample to one another. The dissimilarity between each sample is calculated, taking into account shared taxa (Jaccard similarity index), and then configured into a 2D ordinal space that allows the similarity-based relationship between each sample to be plotted. Samples which are closer together are more similar to one another in terms of community composition, while samples which are further apart are less similar. This type of clustering analysis allows you to see if certain types of samples, for example, those from a particular habitat type, are more clustered together and therefore more similar to one another compared to other groups.
nucleotide	An individual unit of genetic material which, when strung together constitutes a DNA (or RNA) strand/sequence.
OTU	Operational Taxonomic Unit; similar sequences are clustered into OTUs at a defined similarity threshold. OTUs are approximately equivalent to species and are treated as such in our analyses. Species-level taxonomic assignments may or may not be possible, depending on the availability of reference sequences and the similarity between closely related species in the amplified marker. It may be possible to refine the taxonomic assignment for an OTU later as more sequences are added to reference databases.
PCR	Polymerase Chain Reaction; a process by which millions of copies of a particular DNA segment are produced through a series of heating and cooling steps. Known as an 'amplification' process. One of the most common processes in molecular biology and a precursor to most sequencing-based analyses.

positive control	Used to determine whether the PCR is working correctly.
primers	Short sections of synthesised DNA that bind to either end of the DNA segment to be amplified by PCR. Can be designed to be totally specific to a particular species (so that only that species' DNA will be amplified from a community DNA sample), or to be very general so that a wide range of species' DNA will be amplified. Good design of primers is one of the critical factors in DNA-based monitoring.
rarefaction curve	A plot showing the number of taxa as a function of the sequencing depth (number of reads). Rarefaction curves grow rapidly at first as common species are found then reach a plateau as only the rarest species remain to be detected. Rarefaction curves can provide an indication as to whether the species being studied have been comprehensively sampled.
rarefy	A normalisation technique which transforms the data to remove biases associated with uneven sampling depth (number of reads) across samples. The sampling depth of each sample is standardised to a specified number of reads (usually that of the sample with the lowest depth) by random resampling.
reference databases	Over time, the DNA sequences of many species have been compiled into publicly accessible databases by scientists from around the world. These databases serve as a reference against which unknown sequences can be queried to obtain a species identification. The most commonly accessed database is NCBI, which is maintained by the US National Institute of Health. Anyone can search for DNA sequences at https://www.ncbi.nlm.nih.gov.
richness	The total number of taxa within a sample.
rRNA	Ribosomal RNA.
SAC species	Typically the presence of these species potentially elevates the conservation status of a site to a Special Area of Conservation (SAC). Special Areas of Conservation (SACs) are strictly protected sites designated under the EC Habitats Directive.
sequence(s)	A DNA sequence is made up of four nucleotide bases represented by the letters A, T, C & G. The precise order of these letters is used to compare genetic similarity among individuals or species and to identify species using reference databases. In high-throughput sequencing analyses (e.g. metabarcoding), many identical copies of the same sequence are obtained for each species in the sample. The number of copies obtained per species is known as the number of sequence reads, and this is often -

	although not always - related to the relative abundance of the species.
SILVA	SILVA is a database of small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA sequences for all three domains of life (Bacteria, Archaea and Eukarya).
taxon (s.) / taxa (pl.)	Strictly, a taxonomic group. Here we use the term to describe groups of DNA sequences (OTUs) that are equivalent to species. We do not use the term species because we are unable to assign complete identifications to all of the groups at this time due to gaps in the available reference databases.
taxonomy	The branch of science concerned with classification of organisms. species (s./pl.) - A group of genetically similar organisms that show a high degree of overall similarity in many independent characteristics. Related species are grouped together into progressively larger taxonomic units, from genus to kingdom. Homo sapiens (human) is an example of a species. genus (s.) / genera (pl.) - A group of closely related species. Each genus can include one or more species. Homo is an example of a genus. family (s.) / families (pl.) - A group of closely related genera. Homo sapiens is in the Family Hominidae (great apes). order (s.) / orders (pl.) - A group of closely related families. Homo sapiens is in the Order Primates. class (s.) / classes (pl.) - A group of closely related orders. Homo sapiens is in the Class Mammalia. phylum (s.) / phyla (pl.) - A group of closely related classes. Homo sapiens is in the Phylum Chordata.
UKBAP species	UK Biodiversity Action Plan species have been identified as being the most threatened and requiring conservation action under the UK Biodiversity Action Plan.
UNITE	A ribosomal RNA database for identification of fungi.