Natural England Commissioned Report NECR286

Developing DNA applications within Natural England for freshwater mussels, saline lagoon species and terrestrial invertebrates in 2017/18

First published 21 May 2020

www.gov.uk/natural-england

Foreword

Natural England commission a range of reports from external contractors to provide evidence and advice to assist us in delivering our duties. The views in this report are those of the authors and do not necessarily represent those of Natural England.

Background

DNA based applications have the potential to significantly change how we monitor biodiversity and which species and taxa we monitor. These techniques may provide cheaper alternatives to existing species monitoring, an ability to detect species that we do not currently monitor effectively and the potential to develop new measures of habitat and ecosystem quality.

Natural England has been supporting the development of DNA techniques for a number of years. The use of environmental DNA (eDNA) to determine the presence or absence of great crested newts in ponds is now a standard tool for developers and consultants.

There are still significant limitations to the use of this technology in other areas and in 2017/18 Natural England worked with NatureMetrics to prove the concept of using DNA for monitoring terrestrial invertebrates, saline lagoon species and freshwater mussels. This report presents the results from those proof of concept studies, which were promising, and suggests areas where further development could be targeted.

This report should be cited as:

TANG, C.Q., CRAMPTON-PLATT, A., & SMITH, T. 2020. Developing DNA applications within Natural England for freshwater mussels, saline lagoon species and terrestrial invertebrates in 2017/18. Natural England Commissioned Reports, Number 286

Natural England Project Managers– Gavin Measures, Maija Marsh, Jon Webb and Andy Nisbet

Contractor - NatureMetrics

Keywords – DNA, eDNA, terrestrial invertebrates, metabarcoding, saline lagoon invertebrates, freshwater mussels,

Further information
This report can be downloaded from the Natural England Access to Evidence Catalogue: **http://publications.naturalengland.org.uk/** . For information on Natural England publications contact the Natural England Enquiry Service on 0300 060 3900 or e-mail **enquiries@naturalengland.org.uk.**

ISBN 978-1-78354-580-3

Executive Summary

In 2017/18, Natural England commissioned NatureMetrics to conduct proof of concept studies using DNA for:

1. Developing eDNA sampling methods to monitor freshwater mussel populations in England

The UK is home to 6 species of freshwater Unionoida mussels, each with a different distribution. Of these 6 species, the depressed river mussel (*Pseudanodonta complanata*) and the pearl mussel (*Margaritifera margaritifera*), have experienced significant declines during the 20th century, and are now protected species. Traditional survey methods however are expensive, time consuming, and have a significant risk of missing their presence at low densities. Environmental DNA (eDNA) methods may offer a viable alternative solution that is quick and sensitive.

A set of primers that specifically amplify British Unionid mussels were developed in this study which can be used detect the 6 native freshwater Unionid mussels in the UK. When the sample is likely to contain only a single species, the resulting amplicons can be Sanger sequenced to provide a quick and reliable species determination, however the assay requires further validation.

2. Lagoon invertebrate eDNA

The starlet sea anemone, *Nematostella vectensis*, and the lagoon sand shrimp, *Gammarus insensibilis*, are saline lagoon specialist protected species. However, like other lagoon specialist species, monitoring these species using traditional survey techniques can be notoriously difficult, due to their small size and cryptic nature.

In this study a set of primers were developed that specifically amplified *Nematostella vectensis*. These primers appear specific to *N. vectensis* but require more validation. Specific primers to use for a *Gammarus insensibilis* eDNA assay were failed to be developed, due to the paucity of genetic reference data available for this species. Obtaining the first reference sequences relies on the liberal use of universal primers and a great deal of sequencing effort. More research is required to start the development of this assay.

3. Metabarcoding of invertebrate samples from vane traps and from riparian pitfall traps

This study investigated the use of metabarcoding to identify invertebrate samples from vane traps and riparian pitfall traps, and compared the results with traditional morphological identification. Metabarcoding was found to currently be less successful than traditional methods, however the study recognises that the metabarcoding method shows promise and further development is required.

Developing eDNA sampling methods to monitor freshwater mussel populations in England

Introduction

The UK is home to 6 species of freshwater Unionoida mussels (Table 1). Each of these species has a different distribution (Figure 1). Of these 6 species, the depressed river mussel (*Pseudanodonta complanata*) and the pearl mussel (*Margaritifera margaritifera*), have experienced significant declines during the 20th century and both are now included on a list of the most threatened UK species drawn up as part of the Government's Biodiversity Action Plan. The pearl mussel is also protected under Schedule 5 of the Wildlife and Countryside Act (1981) and the Wildlife (Northern Ireland) Order 1985, and is listed on Annexes II and V of the EU Habitats and Species Directive and Appendix III of the Bern Convention, and the IUCN Invertebrate Red List, where its status is described as endangered (IUCN Red List of Threatened Species 2018). Classified as a priority species by the UK Biodiversity Steering Group, a national Species Action Plan has been prepared to encourage measures for its survival.

Table 1. Diversity of UK freshwater mussels and their IUCN status.

Species	Common	Status	Native
Anodonta anatina	Duck mussel	LC	Yes
Anodonta cygnea	Swan mussel	LC	Yes
Margaritifera	Pearl mussel	EN	Yes
Pseudanodonta complanata	Depressed river mussel	VU	Yes
Unio pictorum	Painter's mussel	LC	Yes
Unio tumidus	Swollen river mussel	LC	Yes

Of the most critically important is the freshwater pearl mussel (*Margaritifera margaritifera*). In order to better conserve this species is it imperative that surveys are conducted as part of any proposed development work that may impact their habitat. Traditional survey methods however are expensive, time consuming, and have a significant risk of missing their presence at low densities. Environmental DNA (eDNA) methods may offer a viable alternative solution that is quick and sensitive. Here, we design new 16S rDNA primers to amplify British Unionoida species from eDNA samples.

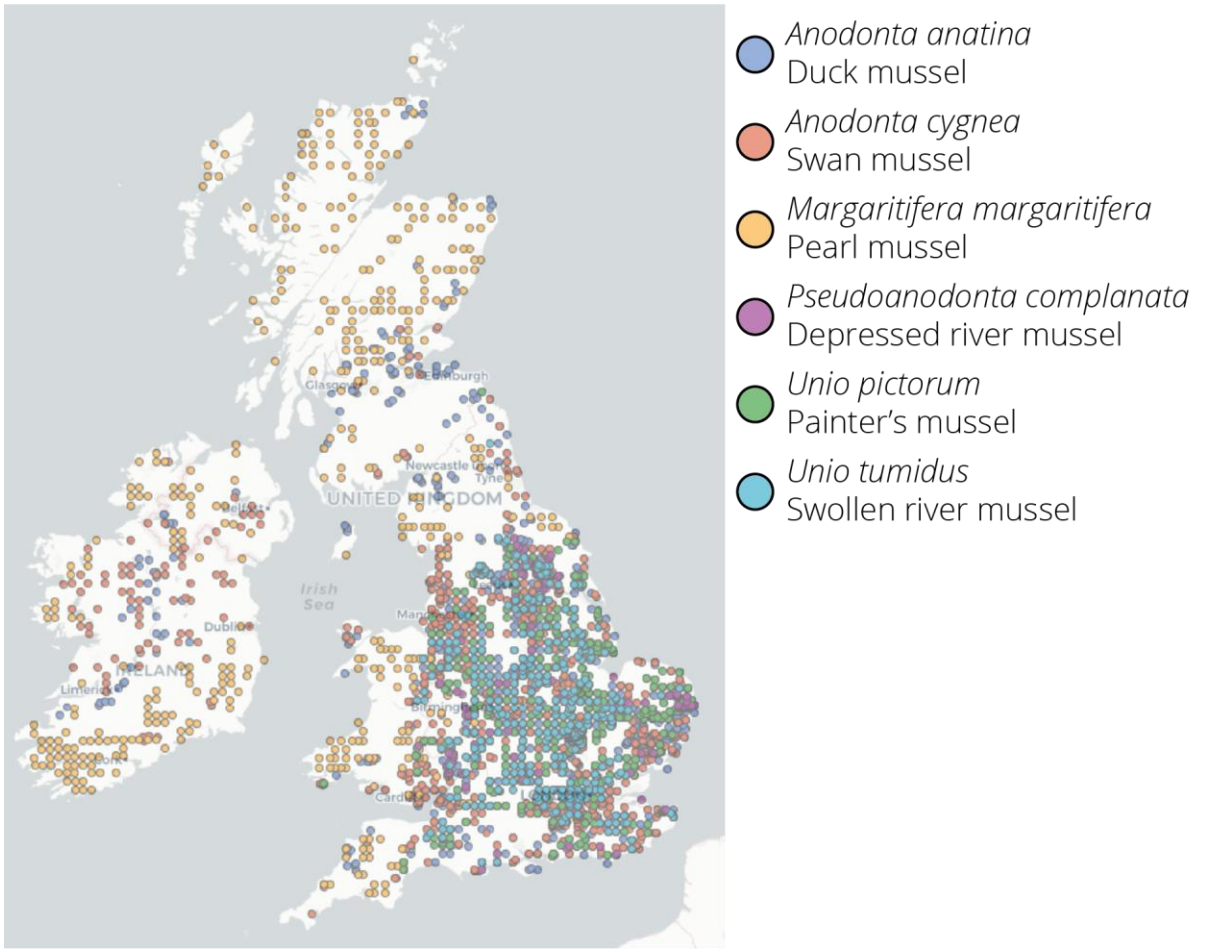


Figure 1. Freshwater mussel UK distribution. Map generated from NBN Atlas (nbnatlas.org).

Methods

Primer design

Primers were designed to target the six native Unionoida freshwater mussel species in the UK. The initial target genomic regions to design primers were all mitochondrial (owing to the likely higher copy number of genes present in eDNA): COI mtDNA, cytB mtDNA, and 16S rDNA.

PrimerMiner was used to search GenBank for mitochondrial genomes and COI mtDNA, cytB mtDNA, and 16S rDNA sequences belonging to British Unionoida species. A total of 352 COI, 13 cytB, and 244 16S sequences were downloaded. These sequences were aligned and used to identify conserved regions suitable for primer design. Only 16S yielded regions that were adequately conserved for primer binding and variable enough for species delimitation. The specificity of these primers were checked using primerBLAST.

In vitro tests

Sampling

The new Unionoida primers were used specifically to detect the presence of freshwater pearl mussels (*Margaritifera margaritifera*) in rivers across England. A total of 51 eDNA samples were collected by a Natural England team led by Gavin Measures between the 4th and 20th of September 2017 using the NatureMetrics Sterlitech eDNA filters (see Annex 2 for sampling protocol). These samples were taken at varying distances from known mussel beds in Cumbria and Shropshire (Table 2).

Molecular work

DNA from each filter was extracted using a commercial DNA extraction kit with a protocol modified to increase DNA yields. Samples were extracted in three batches (Table 2): 1) River Ehen 2284 - 2298 (received on the 18th of September), 2) Rivers Irt, Bleng, Clun, and Teme 2299 - 2322 (received on the 21st and 22nd of September), and 3) Rivers Kent and Gowan 2340 - 2351 (received on the 27th of September). DNA was purified to remove PCR inhibitors using a commercial purification kit.

Purified DNAs were amplified with 12 replicate PCRs for a ~478 bp hypervariable region of the 16S rDNA gene using newly designed primers. The PCR mixture comprised 1x concentration of Phusion Green Mastermix, 0.3 µM of Flexor-Mussel, 0.3 µM of Rectus-Mussel, 1.5 mM of $MgCl_2$, 0.6 mg/ml of BSA, 3% volume of DMSO, 1 µl of uninhibited DNA, and ddH_2O up to a total volume of 15 µl. PCR conditions comprised an initial denaturation phase of 95°C for 3 minutes, followed by 10 cycles of 98°C for 20 seconds, 69°C for 15 seconds, and 72°C for 15 seconds, a 10 cycle touchdown phase with a - 0.5°C annealing temperature, and finally 25 cycles of 98°C for 20 seconds, 65°C for 15 seconds, and 72°C for 15 seconds. All PCRs were performed in the presence of a negative template control. Amplification success was determined by gel electrophoresis.

Table 2. Sample locations collected from A) Cumbria and B) Shropshire

River	County	No. Samples
River Ehen	Cumbria	15
River Irt	Cumbria	9
River Bleng	Cumbria	6
River Kent	Cumbria	5
River Gowan	Cumbria	4
River Brathay	Cumbria	3
River Clun	Shropshire	5
River Teme	Shropshire	4

Sanger sequencing

The successful PCR products were purified and Sanger sequenced. Raw sequences were trimmed and cleaned of low quality base calls to increase the confidence of identification. After cleaning, sequences were identified by comparing the sequences to the NCBI reference database.

High throughput sequencing

While Sanger sequencing is quicker than high throughput sequencing, it is only able to provide a single sequence, which is problematic if there is more than one Unionoida species present. High throughput sequencing would be required when more than one Unionoida species is expected to present as the signal from multiple species would result in unclear and messy signals.

The first 11 samples were prepared for high throughput sequencing. PCR replicates were pooled and purified, and sequencing adapters were added. Success was determined by gel electrophoresis. Amplicons were purified and checked by gel electrophoresis, these were then quantified using a Qubit high sensitivity kit (Table 3) according to the manufacturer's protocol. All purified index PCRs were pooled into a final library with equal concentrations. The final library was sequenced using an Illumina MiSeq V2 kit at 15 pM with a 10% PhiX spike in.

Sequence data was processed using a custom bioinformatics pipeline for quality filtering, dereplication, and taxonomic assignment. After filtering, taxa were identified by comparing those sequences to the NCBI reference database. The presented species-level identification is the top hit on the databases based on species identity.

Table 3. Volume of water filtered and the resultant concentration of purified DNAs and index PCRs. Concentrations for the index PCRs are presented for the first 11 samples, which were prepared for high throughput sequencing.

NM ID	Sample ID	Volume filtered	DNA (ng/µl)	Index (ng/µl)	Batch
2284	Upper River Ehen 1	1000 ml	1.3	14	1
2285	Upper River Ehen 2	1000 ml	1.16	13.4	1
2286	Upper River Ehen 3	NA	2.31	12.8	1
2287	Ennerdale Bridge 1	1000 ml	0.564	12.1	1
2288	Ennerdale Bridge 2	1000 ml	0.864	10.8	1
2289	Ennerdale Bridge 3	NA	1.38	11.8	1
2290	Ennerdale Bridge 4	1000 ml	1.1	11.3	1
2291	Ennerdale Bridge 5	1000 ml	1.71	13.5	1
2292	Ennerdale Bridge 6	1000 ml	0.86	14.1	1
2293	Upper River Ehen 4	1000 ml	2.53	15.2	1
2294	Upper River Ehen 5	1000 ml	1.99	13.5	1

NM ID	Sample ID	Volume filtered	DNA (ng/µl)	Index (ng/µl)	Batch
2295	Upper River Ehen 6	1000 ml	1.58	NA	1
2296	Lower River Ehen 2	1000 ml	0.412	NA	1
2297	Lower River Ehen 1	1000 ml	0.748	NA	1
2298	Lower River Ehen 3	1000 ml	0.612	NA	1
2299	River Irt (Lower 1)	1250 ml	0.106	NA	2
2300	River Irt (Lower 2)	1250 ml	0.125	NA	2
2301	River Irt (Lower 3)	1250 ml	0.116	NA	2
2302	River Irt (Upper 1)	1250 ml	0.103	NA	2
2303	River Irt (Upper 2)	1250 ml	0.105	NA	2
2304	River Irt (Upper 3)	1250 ml	0.116	NA	2
2305	River Irt (Santon Bridge 1)	1250 ml	0.136	NA	2
2306	River Irt (Santon Bridge 2)	1250 ml	0.104	NA	2
2307	River Irt (Santon Bridge 3)	1250 ml	0.0916	NA	2
2308	River Bleng 1	1250 ml	0.0848	NA	2
2309	River Bleng 2	1250 ml	0.0872	NA	2
2310	River Bleng 3	1250 ml	0.096	NA	2
2311	River Bleng (Blengdale 1)	1250 ml	0.0936	NA	2
2312	River Bleng (Blengdale 2)	1250 ml	0.084	NA	2
2313	River Bleng (Blengdale 3)	1250 ml	0.0936	NA	2
2314	River Clun (Leintwardine 1)	1440 ml	0.134	NA	2
2315	River Teme 1a	1600 ml	0.197	NA	2
2316	River Teme 2a	1620 ml	0.147	NA	2
2317	River Clun (Jays Bridge 1)	1600 ml	0.238	NA	2
2318	River Clun (Leintwardine 3)	1500 ml	0.0956	NA	2
2319	River Teme 1b	150 ml	0.212	NA	2
2320	River Clun (Jays Bridge 2)	1600 ml	0.492	NA	2
2321	River Teme 2b	1620 ml	0.243	NA	2
2322	River Clun (Leintwardine 2)	1440 ml	0.162	NA	2
2340	River Kent (Staveley 1)	1010 ml	0.704	NA	3
2341	River Kent (Staveley 2)	1000 ml	0.327	NA	3
2342	River Kent (Gowan)	1000 ml	0.444	NA	3
2343	River Kent (Ulthwaite Bridge)	1000 ml	0.552	NA	3
2344	Upper River Kent 2	1000 ml	0.226	NA	3
2345	Upper River Kent 2	1000 ml	1.58	NA	3
2346	River Kent headwaters 1	1000 ml	0.205	NA	3
2347	River Kent headwaters 2	1000 ml	0.15	NA	3
2348	River Kent headwaters 3	1000 ml	0.152	NA	3
2349	River Brathay (Clappersgate)	1000 ml	0.342	NA	3
2350	River Brathay 1	1000 ml	0.284	NA	3
2351	River Brathay 2	1000 ml	0.312	NA	3

Results

Primer design

COI and cytB were far too variable to design universal primers. 16S on the other hand had stretches of conserved DNA suitable for priming. The balance between variability of conservedness in 16S rDNA made it the prime target region for primer design. A set of primers were designed that had zero mismatches to any of the 244 16S sequences (Flexor-Mussel: 5' - TTA GCG TGA GCG TGC TAA GGT AG - 3' and Rectus-Mussel: 5' - CTT AAG CCA ACA TCG AGG TCG - 3').

In silico analysis of the flexor and rectus mussel primers bind perfectly to the 6 native Unionoida species with zero mismatches at those priming sites. The region is sufficiently different to differentiate between the 6 species (Figure 2). The primers are a perfect match to 193 different freshwater mussel species (Annex 1), but none of these are native to the UK. The primers do not amplify for any non-target British species for which 16S has been sequenced and made publically available on NCBI.

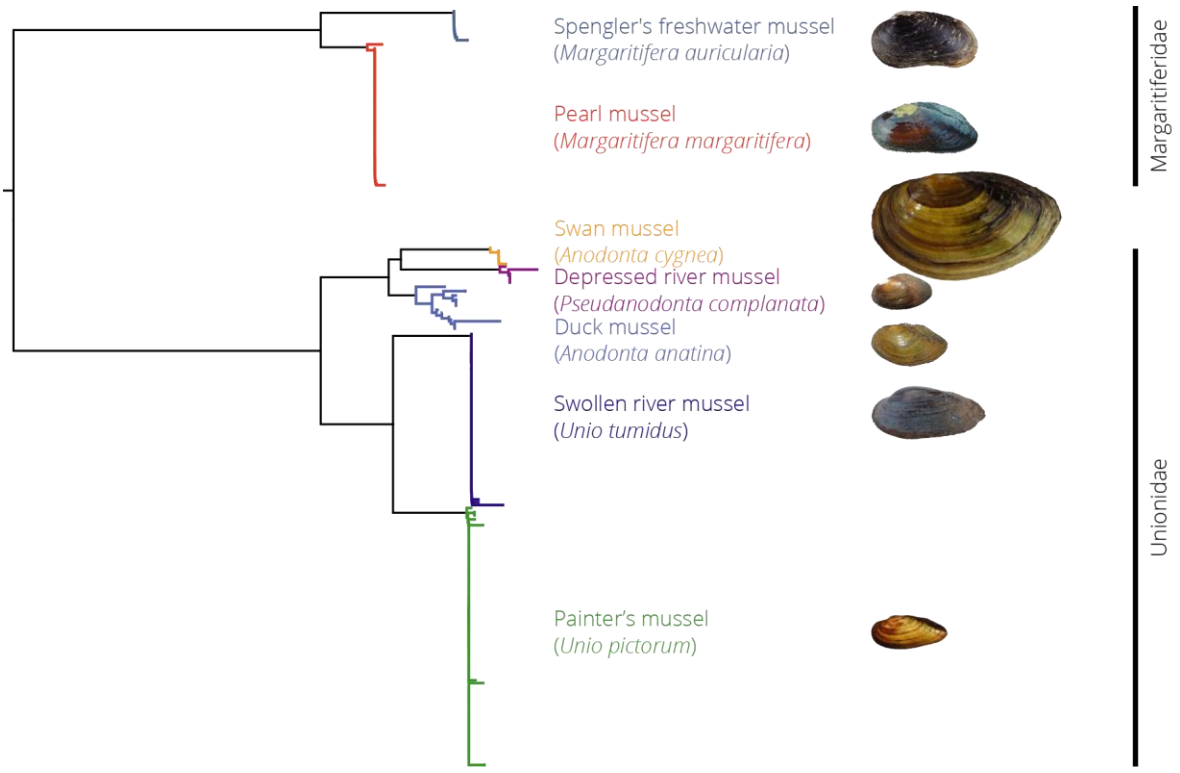


Figure 2. 16S gene tree of the UK freshwater mussel species (Duck mussel *Anodonta anatina* [16 sequences], Swan mussel *Anodonta cygnea* [7], Depressed river mussel *Pseudanodonta complanata* [7], Painter's mussel *Unio pictorum* [91], Swollen river mussel *Unio tumidus* [61], Pearl mussel *Margaritifera* [50], and Spengler's freshwater mussel *Margaritifera auricularia* [12]).

In vitro tests

DNA extraction was successful for the first batch (average DNA concentration = 1.27 ng/µl), which was 9x higher than batch 2 (0.14 ng/µl) and 3x higher than batch 3 (0.44 ng/µl). Unfortunately all downstream processes for batches 2 and 3 were limited by the lack of DNA extracted from the filters, we hypothesise this was a result of either defective reagents in the extraction process, or an incorrect extraction method was performed. PCR reactions were consistently successful for 15 of the 51 samples. Electrophoresis bands were strong and of the expected size and controls performed as expected (Figure 3). The failure to amplify DNA for the remaining 36 samples was consistent with the idea that the second and third batches of DNA extractions failed. Several attempts to repeat these amplifications with different reaction volumes, concentrations, and timings, resulted in real but inconsistent amplification. Due to this inconsistency, we deem the results from these 36 samples as inconclusive.

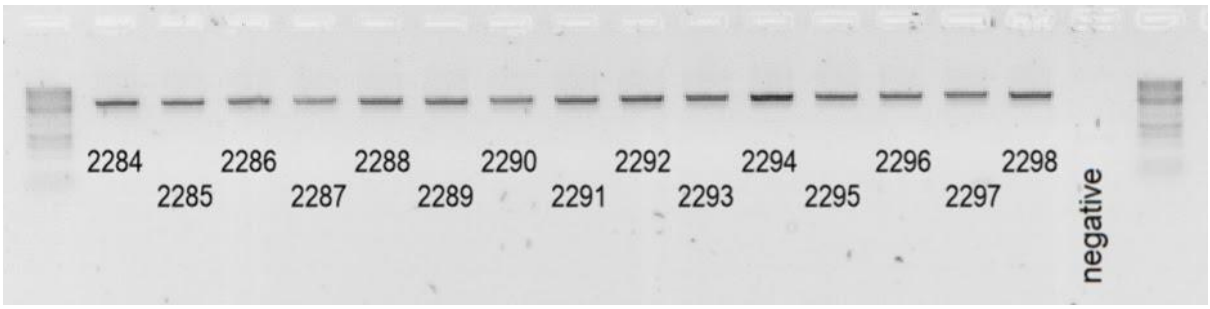


Figure 3. Gel electrophoresis image of the purified PCR products. Numbers correspond to NM IDs provided in Table 3.

Sanger sequencing

The positive PCR products were successfully Sanger sequenced; the quality of the sequence was very high and clear enough to identify the sample (>90% HQ; Figure 4). The sequence length obtained was sufficient for conclusive identification (>400 bp). The presented species-level identification is the top hit on the NCBI database based on species identity (Table 4). All of the sequences were identical and were each a perfect match (100%) to freshwater pearl mussel (*Margaritifera margaritifera*).

Figure 4. Raw chromatograms from the 15 successfully amplified samples.

Table 4. The identification results are based on the percentage similarity (% ID) of the generated sequences to the reference database.

ID	Sample ID	Binomial	Vernacular	% ID
2284	Oxbow River Ehen by Sonde 10m u/s	*M. margaritifera*	Pearl mussel	100
2285	Oxbow River Ehen	*M. margaritifera*	Pearl mussel	100
2286	Oxbow River Ehen by Sonde	*M. margaritifera*	Pearl mussel	100
2287	Ennerdale bridge - No.2 - 30m from bridge	*M. margaritifera*	Pearl mussel	100
2288	Ennerdale Bridge - 3 - 5m u/s crossable	*M. margaritifera*	Pearl mussel	100
2289	Ennerdale Bridge - Ehen	*M. margaritifera*	Pearl mussel	100
2290	Ennerdale Church 1	*M. margaritifera*	Pearl mussel	100
2291	Ennerdale Church 2	*M. margaritifera*	Pearl mussel	100
2292	Ennerdale Church - Ehen	*M. margaritifera*	Pearl mussel	100
2293	Horseshoe - Ehen	*M. margaritifera*	Pearl mussel	100
2294	Horseshoe - Ehen No.2	*M. margaritifera*	Pearl mussel	100
2295	Horseshoe - Ehen No.3	*M. margaritifera*	Pearl mussel	100
2296	River Ehen at Longlands Lake No.2	*M. margaritifera*	Pearl mussel	100
2297	River Ehen at Longlands Lake	*M. margaritifera*	Pearl mussel	100
2298	River Ehen at Longlands Lake No.3	*M. margaritifera*	Pearl mussel	100

High throughput sequencing

The 11 samples intended for high throughput sequencing were a successfully indexed with strong electrophoresis bands of the expected size. All amplicons were successfully purified and were of high yield (Table 3).

The MiSeq paired-end sequencing of the 11 samples yielded 1,000,854 reads, of which 99.3% passed our internal quality filter. Very few sequences were discarded prior to dereplication, which is indicative of high quality data with minimal PCR and sequencing errors.

Only freshwater pearl mussel was detected in the 11 samples. No other taxa were detected in the samples. Plotting the number of reads against the distance from the mussel bed seems to have a quantitative pattern. This was not explained by the volume of water filtered (which was assumed to be 1,000 ml for each sample; Table 3). The trend exhibits a weak logarithmic pattern ($R^2 = 0.235$) with an outlier at 40 m, which has fewer than expected number of reads given the short distance (Table 5). The fit of the trendline improves to $R^2 = 0.587$ when the outlier is removed (Figure 5).

Table 5. The number of pearl mussel sequence reads and the distance from the mussel beds the samples were taken.

NM ID	Sample ID	Distance from mussel bed (m)	Number of Pearl mussel reads
2284	Oxbow River Ehen by Sonde 10m upstream	30	85847
2285	Oxbow River Ehen	5	95346
2286	Oxbow River Ehen by Sonde	40	92509
2287	Ennerdale bridge - No.2 - 30m from bridge	30	92277
2288	Ennerdale Bridge - 3 - 5m u/s crossable	40	63184
2289	Ennerdale Bridge - Ehen	10	101638
2290	Ennerdale Church 1	100	92727
2291	Ennerdale Church 2	150	81321
2292	Ennerdale Church - Ehen	120	87882
2293	Horseshoe - Ehen	30	91158
2294	Horseshoe - Ehen No.2	5	98944

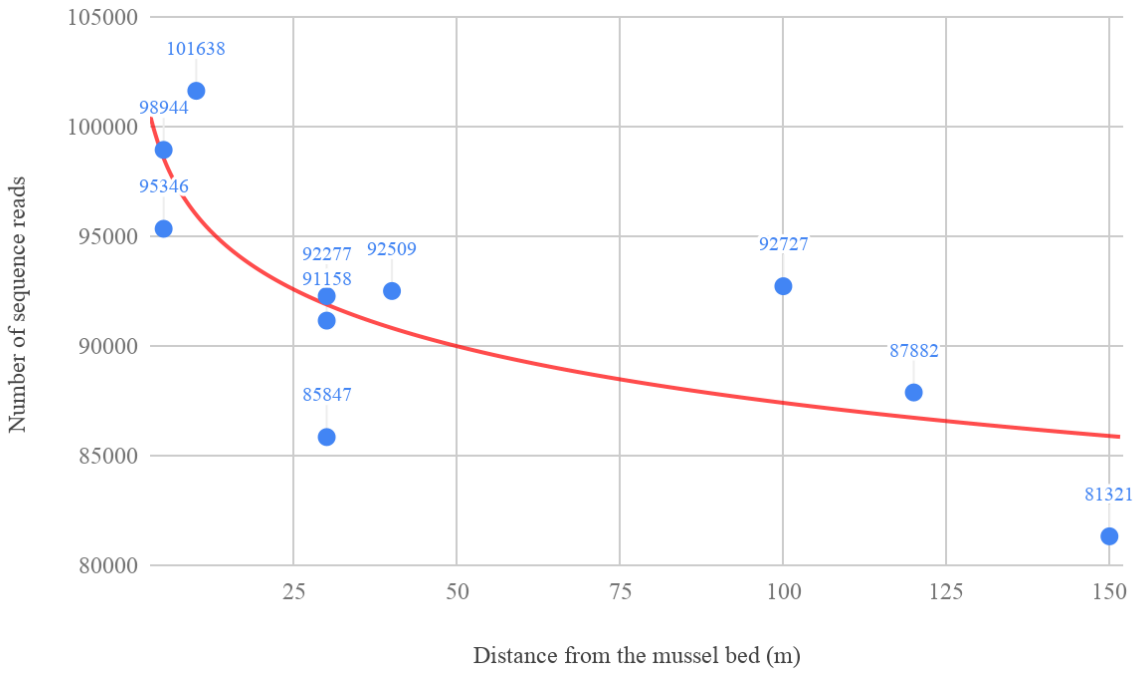


Figure 5. Relationship between the number of pearl mussel sequence reads and the distance from the mussel beds the samples were taken. The fit of the trend has an R^2 of 0.587, but this excludes 2288 as an outlier.

Discussion

Here we have developed and started to validate a set of primers that specifically amplify British Unionid mussels. These primers can be used detect the 6 native freshwater Unionid mussels in the UK (Duck mussel *Anodonta anatina*, Swan mussel *Anodonta cygnea*, Depressed river mussel *Pseudanodonta complanata*, Painter's mussel *Unio pictorum*, Swollen river mussel *Unio tumidus*, Pearl mussel *Margaritifera margaritifera*), but also bind perfectly to another 193 species that have sequences available online (Annex 1).
We have shown there that when the sample is likely to contain only a single species, the resulting amplicons can be Sanger sequenced to provide a quick and reliable species determination. An added benefit of these primers is that we have designed the primers to be group specific rather than species-specific, which means that multiple species present in the same sample can be differentiated in a single reaction. For example, we recently used these primers to detect 5 species of Unionidae mussels (Duck mussel *Anodonta anatina*, Swan mussel *Anodonta cygnea*, Depressed river mussel *Pseudanodonta complanata*, Painter's mussel *Unio pictorum*, and Swollen river mussel *Unio tumidus*) from some Swedish lakes and rivers (Figure 6).

The positive results we show here match with what is known for the River Ehen sampling sites. We were able to show that both Sanger and high throughput sequencing methods are useful for the detection of pearl mussels. The River Ehen supports the largest freshwater pearl mussel *Margaritifera* population in England. Exceptionally high densities (greater than 100 m^2) are found at some locations, with population estimates for the entire river exceeding 500,000. Unfortunately the DNA extraction for samples taken from the Rivers Irt, Bleng, Clun, Teme, Kent and Gowan failed. These samples included known positives and negatives, and were taken from much smaller populations of freshwater pearl mussels. We hypothesise that either defective reagents were used in the extraction process, or an incorrect extraction method was performed. These failures resulted in a complete overhaul of our extraction methodology and traceability, and no such problem has happened since. While we have begun evaluating this particular assay, it still requires a greater deal of validation. In particular the limits of detection and the true specificity of the primers (particularly with bacterial DNA sources that are not on the reference databases) have not been fully explored. Additional validation tests could include:

1. A comprehensive *in vitro* specificity testing other Unionoida tissue samples and Dressenid samples would be advised to complement the *in silico* analyses that were done.
2. More extensive field testing at sites of known presence and absence of freshwater pearl mussels as well as other mussel species.
3. A series of artificial or laboratory dilution assays to determine proper limits of detection.

As it stands, a positive sample that has been sequenced can be interpreted as a positive detection, but more work is required to confirm whether negative assays equate to truly negative samples.

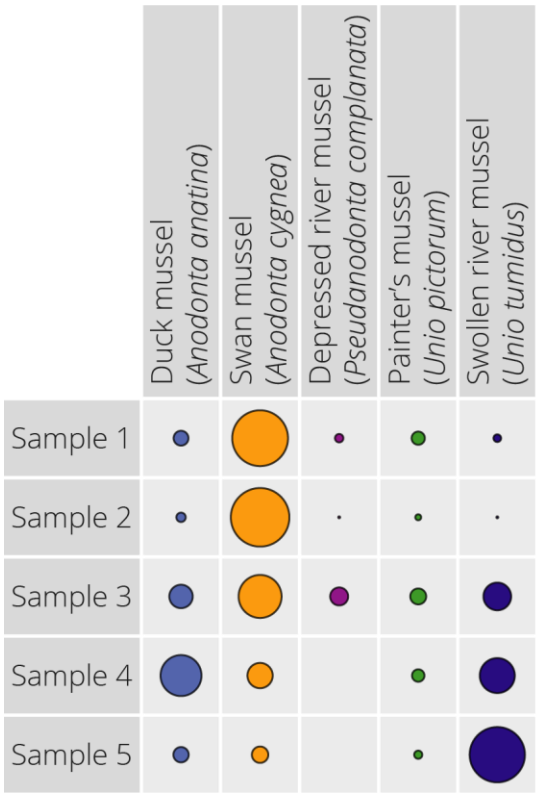


Figure 6. Proportion of the sequencing output allocated to five different Unionoida species in five different lake and river samples. The size of the bubbles correspond to the proportion of the sequencing output for that sample.

Annex 1

List of species with perfectly complementary primer-binding sites to the 16S primers designed here. This list includes only those species that have been sequenced for this specific region of 16S, it will not include those species that have not been sequenced.

Actinonaias ligamentina, *Actinonaias pectorosa*,
Aculamprotula tientsinensis,
Amblema elliottii, *Amblema perplicata*, *Amblema plicata*,
Anemina sp.,
Anodonta anatina, Anodonta arcaeformis, Anodonta couperiana, Anodonta cygnea, Anodonta euscaphys, Anodonta exulcerata, Anodonta lucida, Anodonta sp., Anodonta woodiana,
Anodontoides radiatus,
Arconaia lanceolata,
Contradens contradens, Contradens sp.,
Cristaria plicata,
Cuneopsis pisciculus,
Cyprogenia aberti,
Cyrtonaias tampicoensis,
Dahurinaia dahurica,
Dromus dromas,
Ellipsaria lineolata,
Elliptio complanata, Elliptio dariensis, Elliptio dilatata, Elliptio sp.,
Elliptoideus sloatianus,
Ensidens sagittarius, Ensidens sp.,
Epioblasma brevidens, Epioblasma capsaeformis, Epioblasma torulosa, Epioblasma triquetra,
Fusconaia barnesiana, Fusconaia cerina, Fusconaia cor, Fusconaia flava, Fusconaia subrotunda, Fusconaia succissa,
Glebula rotundata,
Hemistena lata,
Hyriopsis cumingii, Hyriopsis schlegelii, Hyriopsis sp.,
Indonaia andersoniana,
Lamellidens brandti, Lamellidens exolescens, Lamellidens generosus, Lamellidens indawgyiensis, Lamellidens marginalis, Lamellidens sp.,
Lamprotula caveata, Lamprotula coreana, Lamprotula fibrosa, Lamprotula leai, Lamprotula leaii, Lamprotula scripta, Lamprotula tortuosa,
Lampsilis altilis, Lampsilis australis, Lampsilis cardium, Lampsilis ornata, Lampsilis ovata, Lampsilis perovalis, Lampsilis radiata, Lampsilis siliquoidea, Lampsilis subangulata, Lampsilis teres,
Lanceolaria grayana,
Lasmigona complanata, Lasmigona compressa, Lasmigona costata,
Leguminaia sp., Leguminaia wheatleyi,
Lemiox rimosus,
Leoparreysia canefrii, Leoparreysia tavoyensis,
Lepidodesma languilati,
Leptodea fragilis, Leptodea leptodon, Leptodea ochracea,
Lexingtonia dolabelloides,
Ligumia nasuta, Ligumia recta,
Margaritifera auricularia, Margaritifera dahurica, Margaritifera falcata, Margaritifera hembeli,

Margaritifera homsensis, Margaritifera laevis, Margaritifera laosensis, Margaritifera margaritifera, Margaritifera marocana, Margaritifera marrianae, Margaritifera middendorffi, Margaritifera togakushiensis,
Medionidus acutissimus, Medionidus conradicus,
Megalonaias nervosa,
Mutela dubia, Mutela hargeri,
Nodularia douglasiae,
Obliquaria reflexa,
Obovaria olivaria, Obovaria subrotunda, Obovaria unicolor,
Oxynaia pugio, Oxynaia sp.,
Parreysia olivacea, Parreysia sp., Parreysia tavoyensis,
Physunio sp.,
Pilsbryoconcha exilis, Pilsbryoconcha sp.,
Plectomerus dombeyanus,
Pleurobema chattanoogaense, Pleurobema clava, Pleurobema collina, Pleurobema decisum, Pleurobema georgianum, Pleurobema hanleyianum, Pleurobema oviforme, Pleurobema pyriforme, Pleurobema rubellum, Pleurobema strodeanum, Pleurobema taitianum, Pleurobema troschelianum,
Popenaias popeii,
Potamilis alatus, Potamilus alatus, Potamilus purpuratus,
Potomida littoralis,
Pronodularia japanensis,
Pseudanodonta complanata,
Pseudodon bogani, Pseudodon cf., Pseudodon cumingii, Pseudodon manueli, Pseudodon mouhotii, Pseudodon sp., Pseudodon vondembuschianus,
Ptychobranchus fasciolaris, Ptychobranchus greenii, Ptychobranchus jonesi, Ptychobranchus occidentalis, Ptychobranchus subtentum,
Ptychorhynchus ptisteri,
Pyganodon grandis,
Quadrula apiculata, Quadrula metanerva, Quadrula,
Quincuncina burkei, Quincuncina infucata,
Radiatula bonneaudii, Radiatula mouhoti, Radiatula sp.,
Reginaia ebena, Reginaia rotulata,
Scabies crispata, Scabies sp.,
Sinanodonta sp., Sinanodonta woodiana,
Solenaia carinata, Solenaia oleivora, Solenaia sp., Solenaia triangularis,
Strophitus undulatus,
Toxolasma texasiensis,
Truncilla truncata,
Unio bonellii, Unio cf., Unio crassus, Unio delphinus, Unio douglasiae, Unio durieui, Unio elongatulus, Unio foucauldianus, Unio gibbus, Unio japanensis, Unio mancus, Unio pictorum, Unio ravoisieri, Unio tigridis, Unio tumidiformis, Unio tumidus,
Uniomerus declivus, Uniomerus obesus,
Utterbackia imbecilis, Utterbackia imbecillis,
Venustaconcha ellipsiformis,
Villosa delumbis, Villosa iris, Villosa vanuxemensis, Villosa villosa.

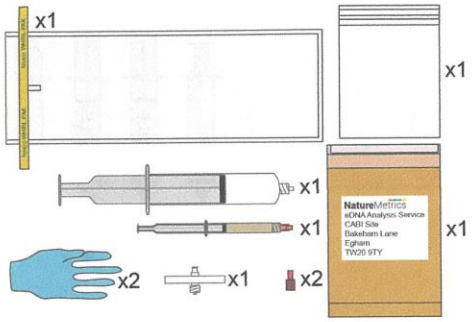
Annex 2

NatureMetrics
DNA-Based Monitoring

eDNA from Water: Sampling Protocol

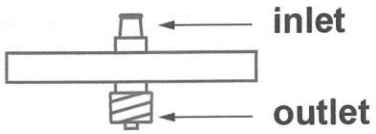
Kit Contents

1x Whirlpak bag
1x 50 ml luer lock plastic syringe
1x filter
2x red luer lock caps
2x nitrile gloves
1x small syringe with preservative
1x resealable zip lock bag
1x NatureMetrics addressed jiffy bag



Sample Collection

1. Put on the gloves. *The purpose of the gloves is to avoid contact with the water or sampling equipment with your skin during the sampling process, as this can result in large amounts of human DNA being introduced to the sample, which could decrease the chances of detecting rare wildlife species.*
2. Tear off the top of the Whirlpak bag. Fill the bag by holding it open at the water's surface and pointing against the direction of the flow.
3. Roll down the top of the bag and shake to mix the water.
4. Draw up 50 ml of water from the bag into the large syringe.
5. Attach the syringe to the filter inlet and press the plunger to push the water through the filter. *The syringe should easily twist onto the inlet side of the filter.*



6. Repeat steps 4 – 5 until the filter is clogged or the bag is empty. Make a note on the sample sheet of the total volume processed. *By the end it may take some force to push the water through the filter, particularly if there's a lot of sediment.*
7. Detach the syringe from the filter and pull back the plunger to fill the syringe with air. Reattach the filter and push the air through to expel any water trapped inside the filter. *It helps to gently shake the filter as you do this.*

Lagoon invertebrate eDNA

Introduction

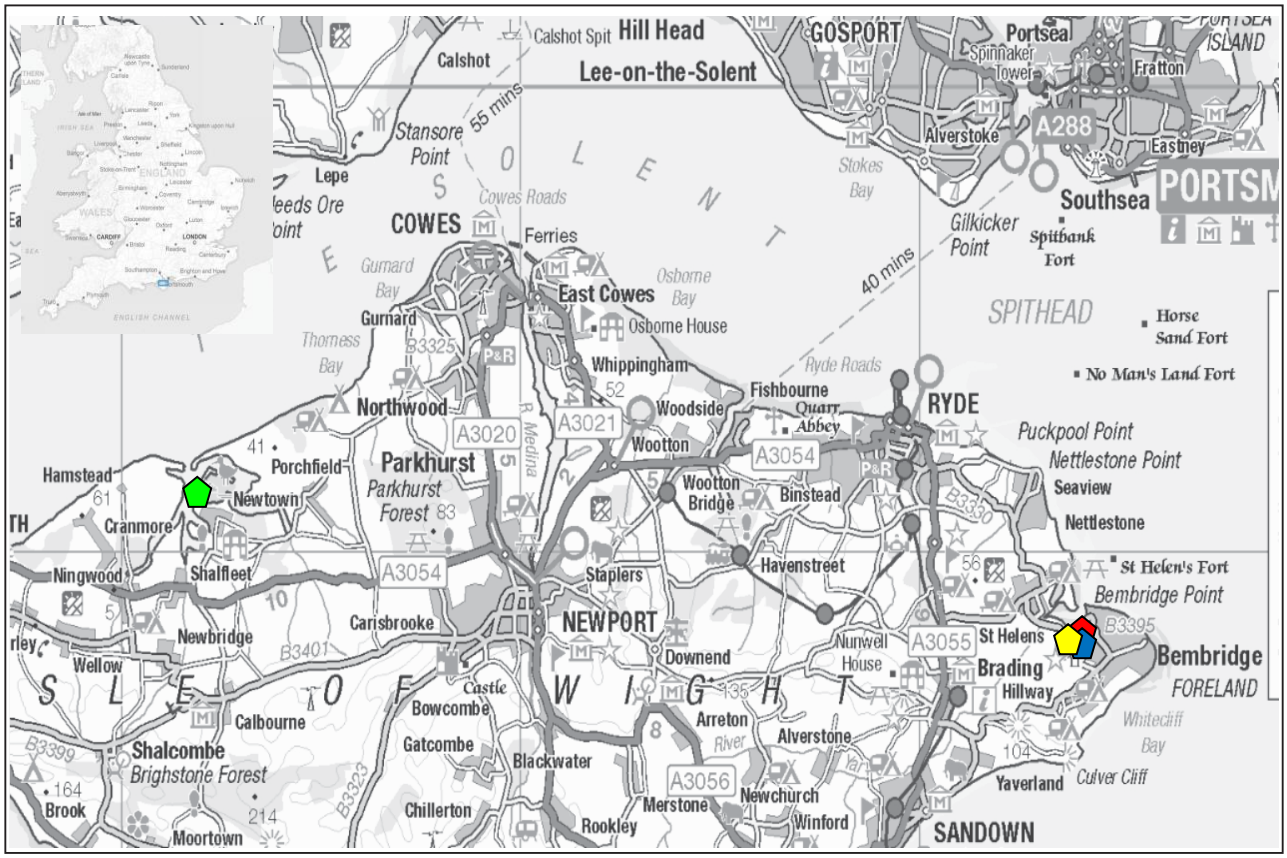
The starlet sea anemone, *Nematostella vectensis*, and the lagoon sand shrimp, *Gammarus insensibilis*, are lagoon specialist species protected under Schedule 5 of the Wildlife and Countryside Act 1981 Sections 9 and 9(4)(a), respectively. Both species are also Marine Conservation Zone Species of Conservation Interest (MCZ SOCIs), and hence afford protection under the Marine and Coastal Access Act (2009) in sites where they are designated features. It is within Natural England's remit to consider these (and other lagoon protected species) in their advice, to safeguard their habitats and populations.

However, like other lagoon specialist species, monitoring *N. vectensis* and *G. insensibilis* using traditional survey techniques can be notoriously difficult, due to their small size and cryptic nature. For example, *N. vectensis* is small (a maximum length of 15 mm), translucent, and buries itself under the substrate, making it difficult to find using conventional sampling techniques. As a result, the ecology, population size and dynamics, and distribution of these species are often poorly understood, and consequently these species are likely to be under-recorded in the UK.

Historically, Natural England's lagoon protected species monitoring has involved the use of traditional survey methods, such as morphological identification of infaunal and epifaunal invertebrates from cores, sediment grabs, buckets, nets and, in the case of *G. insensibilis*, individuals collected from the water column together with their associated species of algae (*Chaetomorpha linum*). However, the success of these techniques in capturing some of the rare and conspicuous lagoon specialist species such as *N. vectensis* and *G. insensibilis* has been variable and expensive. Molecular (environmental DNA, or eDNA) approaches may provide a more cost-effective and reliable alternative to sampling lagoon specialist invertebrates. This will have benefits both in terms of ecological data collection and also when addressing statutory casework, in areas where protected lagoon species are present or believed to be present.

Previous work using an eDNA metabarcoding approach with commonly used universal barcoding primers (targeting variable regions of the COI mtDNA and 18S rDNA genes) failed to detect the presence of *N. vectensis* or any *Gammarus* species, likely because eDNA metabarcoding using these primers is hindered by the sometimes overwhelming amount of sequences associated with more broad-spectrum universal assays (e.g. COI eDNA metabarcoding is often inundated by bacterial amplification). Here we trial species-specific polymerase chain reaction (PCR)-based assays to circumvent these issues.
This study investigates the use of species-specific PCR-based assays in detecting *N. vectensis* and *G. insensibilis* on the Isle of Wight lagoons. Isle of Wight has six lagoons in total, including those at Bembridge Lagoons and Newtown Harbour (considered in this study, see Figure 1), and Yar Lagoon (not considered in this study). All of the Isle of Wight lagoons are Sites of Species Scientific Interest (SSSIs), the lagoons at Bembridge Harbour are also

part of the Solent and Isle of Wight Lagoons Special Area of Conservation (SAC), and the lagoons at Newtown Harbour are part of the Solent Maritime SAC.



NATURAL ENGLAND

Map Produced from WebMap2 on 18/02/20 Map Projection: British National Grid
Map Scale at A3: 1:78,882

- Bembridge Harbour Lagoon
- Harbour Farm 1
- Harbour Farm 2
- Newtown Quay

Figure 1: Sampling sites on the Isle of Wight

Nematostella vectensis eDNA assay

Methods

Primer design

Three different sets of primers were designed to target different sections of the variable mitochondrial 16S rRNA gene region specific to *N*. *vectensis*. Target regions were chosen using a primer mining software (primerBLAST) aimed at the complete mitochondrial genome (GenBank accession: DQ643835.1). Three primer pairs were chosen *in silico* based on that:

1. They bound to *N. vectensis* but no other sequences on the NCBI reference database.
2. They yielded PCR products between 150 - 300 bp in length.
3. The forward and reverse primers had a similar melting temperature to facilitate efficient amplification.

In vitro tests

Each primer pair was tested *in vitro* using DNA extracted from *N. vectensis* tissue samples collected from Bembridge Harbour Lagoon on the 28th of September 2016. Various tests using different annealing temperatures and PCR additives were conducted to optimise the PCR protocol for these primer pairs. The final optimised PCR reaction comprised 1 x DreamTaq Mastermix, 0.5 µM of each primer, 0.8 mg/mL of BSA, and 2 µL of purified DNA to a total volume of 20 µL. The PCR cycling conditions were an initial denaturation at 95 ºC for 3 minutes, followed by 50 cycles of 95 ºC for 30 seconds, 60 ºC for 30 seconds, 72 ºC for 30 seconds, and a final elongation step of 72 ºC for 5 minutes.

In vivo tests: Bembridge Harbour, Newtown Harbour and Harbour Farm 1 & 2 Lagoons

Fieldwork methodology

Figure 2 shows the saline lagoons chosen for this study. The number of sampling stations within each lagoon is shown in Table 1. Within each sampling station, 20 subsamples of lagoon water were taken from the water column near the sediment and poured into a Whirlpak bag to make up a total of 1 litre of water. After ensuring the water was adequately mixed in the bag, a 50 mL syringe was used to plunge the water through a filter into a sterile outlet. This was repeated until the filter was clogged with sediment or the sample bag was empty. Longmire's DNA preservation solution was added and the filter outlet was capped and sealed in a plastic bag and posted to the eDNA laboratory for analysis.

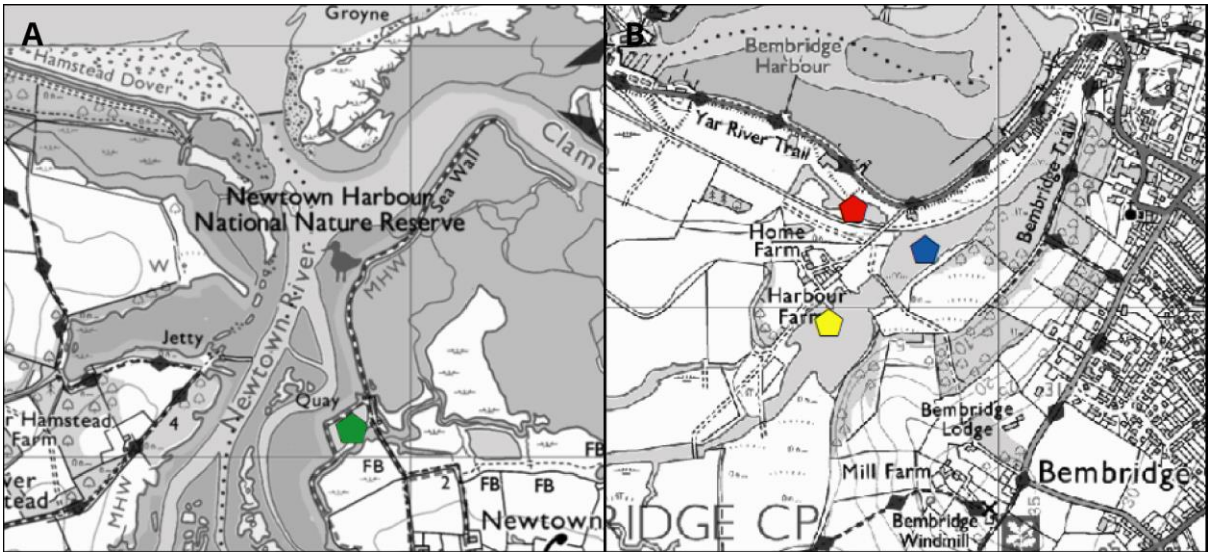


Figure 2. Saline lagoons sampled for eDNA, including a) Newtown Quay lagoon in green, and b) Bembridge lagoons, with Harbour Farm 1 in blue; Harbour Farm 2 in yellow; and Bembridge Harbour lagoon in red.

Table 1. Number of sampling stations within the saline lagoons chosen for this study.

Lagoon	Number of sampling stations
Bembridge Harbour lagoon	10 (+1 from 2016)
Harbour Farm 1	10
Harbour Farm 2	5
Newtown Quay	5

Thirty one eDNA filter samples were provided to NatureMetrics from Bembridge Harbour, Harbour Farm 1, Harbour Farm 2, and Newtown Harbour (Table 2). DNA from each filter was extracted using a commercial DNA extraction kit with a protocol modified to increase DNA yields. DNA was purified to remove PCR inhibitors using a commercial purification kit. Extracted eDNA samples were amplified with 8 replicate PCRs for a ~150 bp region of the 16S rRNA gene to target *N. vectensis*. All PCRs were performed in the presence of both a negative control and a positive control sample (DNA from *N. vectensis* tissue sample, diluted 1:1000 times in water). Amplification success was determined by gel electrophoresis.
A single replicate from each positive sample was purified and Sanger sequenced. Raw sequences were trimmed and cleaned of low quality base calls to increase the confidence of identification. After cleaning, sequences were identified via BLAST searches to the NCBI nucleotide database. The presented species-level identification is the top hit on the NCBI database based on species identity.

Table 2. Volume of water filtered and the resultant concentration of DNA for each of the 31 eDNA samples.

ID	Lagoon	Volume filtered	DNA (ng/µl)
1981	Bembridge Harbour	Unknown	0.904
2761	Bembridge Harbour	300 ml	14.8
2766	Bembridge Harbour	360 ml	14.9
2767	Bembridge Harbour	290 ml	0.704
2768	Bembridge Harbour	240 ml	16
2776	Bembridge Harbour	240 ml	19.3
2781	Bembridge Harbour	240 ml	20.4
2782	Bembridge Harbour	230 ml	0.7
2784	Bembridge Harbour	270 ml	>20
2785	Bembridge Harbour	180 ml	22
2787	Bembridge Harbour	330 ml	20
2763	Harbour Farm 1	120 ml	16.8
2764	Harbour Farm 1	160 ml	>20
2771	Harbour Farm 1	130 ml	>20
2774	Harbour Farm 1	120 ml	23.2
2775	Harbour Farm 1	120 ml	22.4
2777	Harbour Farm 1	120 ml	18.6
2779	Harbour Farm 1	130 ml	>20
2780	Harbour Farm 1	120 ml	>20
2788	Harbour Farm 1	170 ml	>20
2789	Harbour Farm 1	120 ml	>20
2762	Harbour Farm 2	150 ml	20.8
2765	Harbour Farm 2	120 ml	18.7
2773	Harbour Farm 2	120 ml	20
2778	Harbour Farm 2	120 ml	>20
2783	Harbour Farm 2	120 ml	>20
2769	Newtown Harbour	900 ml	13.3
2770	Newtown Harbour	480 ml	9.4
2772	Newtown Harbour	780 ml	13.1
2786	Newtown Harbour	480 ml	11.3
2790	Newtown Harbour	420 ml	1.4

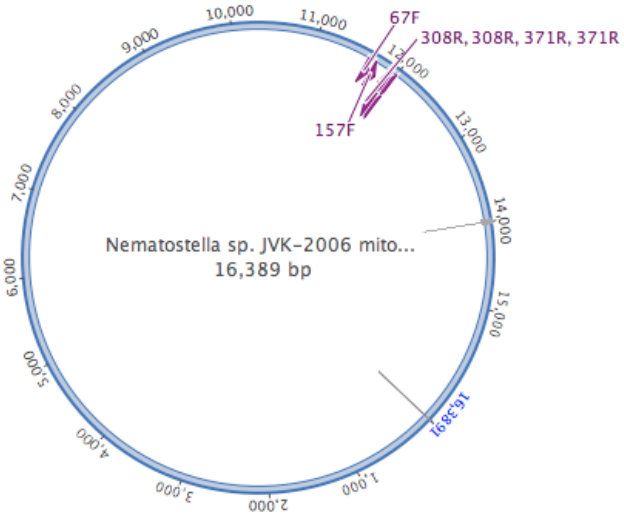
Results

Primer design and *in vitro* tests

Five different primers (Table 3) equating to three different primer pairs were designed (Figure 3). Primer pair 1 (using primers 157F and 470R) produces a product that is 314 bp long, Primer pair 2 (157F and 308R) produces a product that is 152 bp long, and Primer pair 3 (67F and 371R) produces a product that is 305 bp long. Each primer pair yielded PCR products for every tissue DNA extract (Figure 4). The PCR products were purified and sequenced. All of the sequences were identified as *N. vectensis*.

Table 3. Primers and their sequences

Primer	Direction	Sequence (5' - 3')
157F	Forward	CAC TGT CTC AAG AAG ACC CCC
67F	Forward	TGA ATG GCC GCG GTA AC
308R	Reverse	TAA CGG ATC CCA CCT TAT CT
371R	Reverse	CAA ACT CGC ATT GTC CCT AA
470R	Reverse	TTC GGG GTG TCC CCT CTC



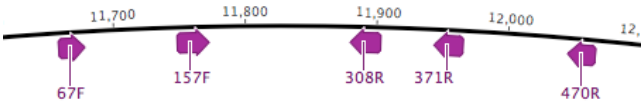


Figure 3. Schematic location of the primer sets on the *Nematostella* mitochondrial genome (DQ643835).

The most successful primer pair (157F - 308R), which tended to produce the cleanest PCR products (single bands of the correct product length, very few spurious bands in negatives), was focussed upon for eDNA trials.

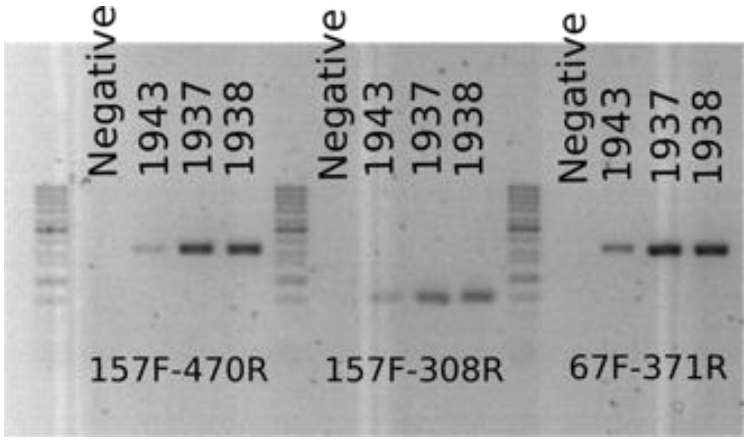


Figure 4. *In vitro* tests using the three primer pairs (157F - 470R, 157F - 308R, and 67F - 371R) using a template negative control (Negative) and three DNA tissues extracts (NM IDs: 1943, 1937, and 1938).

In vivo tests

DNA yields for the 31 samples were variable (Table 2), but high enough to proceed with *in vivo* tests. The volume of water filtered was consistently lower for the Harbour Farm samples, while this did not correlate with the total DNA concentration, it may had a disproportionate effect on DNA coming from rarer species, as increasing the volume of water filtered increases the detection rate.

PCR reactions were successful for 20 of the 31 samples. Nine of the 11 samples from Bembridge Harbour amplified consistently (Figure 5). NM2767 and NM2782 failed likely due to their low concentration (< 1 ng/µL). Samples from Harbour Farm 1 amplified in only a small number of replicates (Figure 6), and those that amplified did so at a low level (low intensity gel band – low concentration of PCR product). This inconsistent amplification may be due to lower species abundance in this lagoon, or sources of inhibition not present at the other sites. Samples from Harbour Farm 2 failed to amplify (NM2762, NM2765, NM2773, NM2778, and NM2783; Figure 7). This is due to the either absence of *Nematostella*, or possibly inhibition in the samples. Note that the positive well in NM2783 was a pipetting error that was repeated and found to be truly negative.

All 5 samples from Newtown Harbour amplified in the majority of replicates, a small number of failed replicates may be due to inhibition (NM2769, NM2770, NM2772, NM2786, and NM2790; Figure 7). For the positive replicates, electrophoresis bands from these samples were strong and of the expected size (Figure 5-7) and no repeat PCRs were necessary. For the negative samples, it is difficult to determine whether these results are due to inhibition or truly due to the species not being present in these samples, as the assay has not been fully

validated with a more extensive field survey and the limit of detection for *N. vectensis* using this assay is not known. Representative PCR replicates from the samples were sequenced. The quality of the sequences was fairly low, however the sequences were clear enough for identification in all but 1 sample. The sequence for sample NM2781 was very low quality and almost entirely ambiguous, it was therefore not possible to conclusively match the sequence to the NCBI database. Otherwise, the trimmed sequence length obtained for the remaining samples was sufficient for conclusive identification (~ 100 bp).

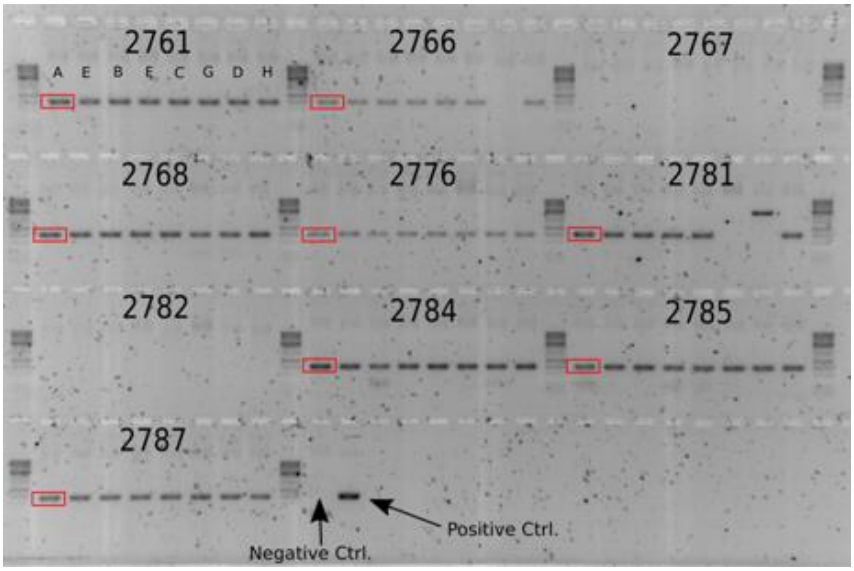


Figure 5. Replicate PCR gel images for the Bembridge Harbour samples. Red boxes indicate samples that were subsequently purified and sequenced.

No contaminating DNA sequences were found and no samples matched any other species at a higher % identity than to *N. vectensis* (Table 4). Where some results have comparatively low % ID, this is due to poor sequence quality. Figure 5 shows replicate PCR gel images for the 10 Bembridge Harbour samples. Red boxes indicate samples that were subsequently purified and sequenced. Results from the 11th Bembridge Harbour sample (NM1981), which was collected in the previous year, is shown in Figure 7.

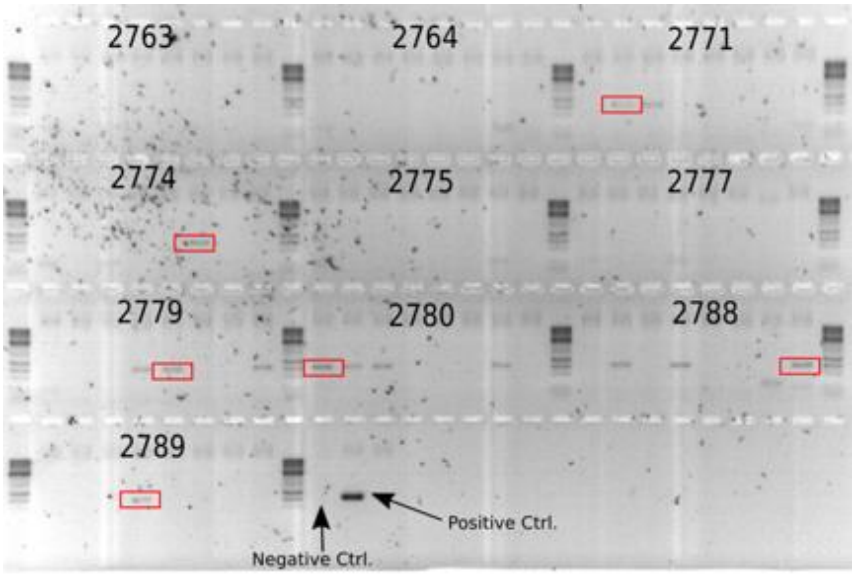


Figure 6. Replicate PCR gel images for the 10 Harbour Farm 1 samples. Red boxes indicate samples that were subsequently purified and sequenced.

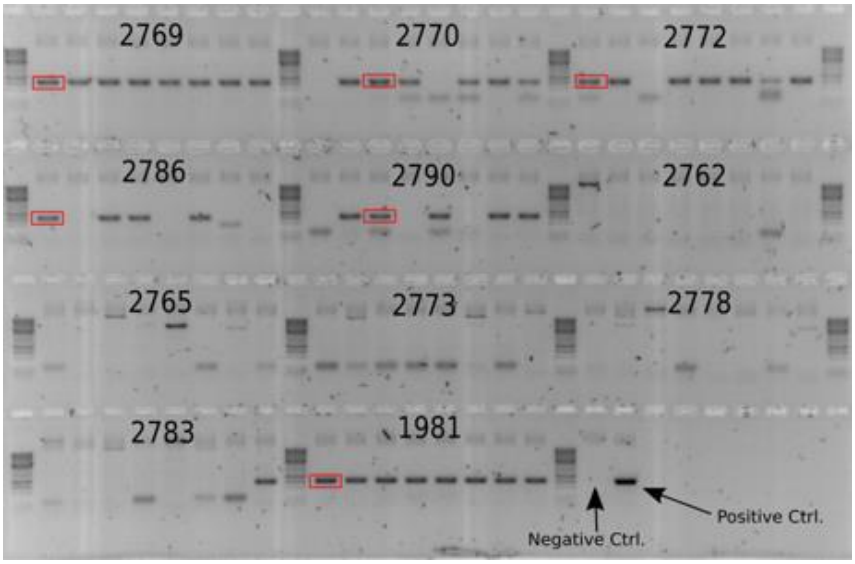


Figure 7. Replicate PCR gel images for the 5 Newtown Harbour samples (NM2769, NM2770, NM2772, NM2786, NM2790), 5 Harbour Farm 2 samples (NM2762, NM2765, NM2773, NM2778, and NM2783), and a single Bembridge Harbour sample (NM1981). Red boxes indicate samples that were subsequently purified and sequenced. Note that the positive well in NM2783 was a pipetting error that was repeated and found to be negative.

Table 4. The identification results are based on the percentage similarity (% ID) of the generated sequences to the reference database. In all other cases where "NA" is given, these samples did not produce PCR amplifications.

ID	Lagoon	Binomial	Vernacular species name	% ID
2766	Bembridge Harbour	*N. vectensis*	Starlet sea anemone	99
2768	Bembridge Harbour	*N. vectensis*	Starlet sea anemone	99
2785	Bembridge Harbour	*N. vectensis*	Starlet sea anemone	98
2784	Bembridge Harbour	*N. vectensis*	Starlet sea anemone	97
1981	Bembridge Harbour	*N. vectensis*	Starlet sea anemone	96
2761	Bembridge Harbour	*N. vectensis*	Starlet sea anemone	95
2787	Bembridge Harbour	*N. vectensis*	Starlet sea anemone	95
2776	Bembridge Harbour	*N. vectensis*	Starlet sea anemone	91
2782	Bembridge Harbour	NA	NA	NA
2767	Bembridge Harbour	NA	NA	NA
2781	Bembridge Harbour	NA	NA	NA
2774	Harbour Farm 1	*N. vectensis*	Starlet sea anemone	100
2789	Harbour Farm 1	*N. vectensis*	Starlet sea anemone	97
2771	Harbour Farm 1	*N. vectensis*	Starlet sea anemone	96
2780	Harbour Farm 1	*N. vectensis*	Starlet sea anemone	90
2788	Harbour Farm 1	*N. vectensis*	Starlet sea anemone	88
2779	Harbour Farm 1	*N. vectensis*	Starlet sea anemone	84
2763	Harbour Farm 1	NA	NA	NA
2777	Harbour Farm 1	NA	NA	NA
2775	Harbour Farm 1	NA	NA	NA
2764	Harbour Farm 1	NA	NA	NA
2765	Harbour Farm 2	NA	NA	NA
2773	Harbour Farm 2	NA	NA	NA
2762	Harbour Farm 2	NA	NA	NA
2778	Harbour Farm 2	NA	NA	NA
2783	Harbour Farm 2	NA	NA	NA
2790	Newtown Harbour	*N. vectensis*	Starlet sea anemone	98
2786	Newtown Harbour	*N. vectensis*	Starlet sea anemone	96
2769	Newtown Harbour	*N. vectensis*	Starlet sea anemone	93
2770	Newtown Harbour	*N. vectensis*	Starlet sea anemone	91
2772	Newtown Harbour	*N. vectensis*	Starlet sea anemone	87

Discussion

The starlet sea anemone, *Nematostella vectensis* has been previously recorded in Bembridge harbour and Harbour Farm using traditional monitoring methods, where it is consider superabundant and common/frequent, respectively. It has not been detected in Newtown Harbour. Table 5 summaries *Nematostella vectensis* detections following both traditional and molecular methods. Detections based on traditional and molecular methods are largely consistent for Bembridge Harbour and Harbour Farm Lagoons, which both had positive detections.

Table 5. *N. vectensis* recorded (e.g. SACFOR abundance scale) in the study sites in the past using traditional monitoring techniques. S = Superabundant, C = Common, F = Frequent. The eDNA scores indicate the number of positive detections out of the number of samples taken for those sites. Harbour Farm Lagoons 1 and 2 are lumped into a single complex because traditional monitoring data is not available for each lagoon. eDNA for Harbour Farm Lagoon 1 = 6/10 and Harbour Farm Lagoon 2 was 0/5.

Lagoon	Last recorded	SACFOR	eDNA
Bembridge	2013	Superabundant	8/11
Harbour Farm Lagoon	2013	Common / Frequent	6/15
Newtown Quay	N/A	-	5/5

Bembridge Harbour

The superabundant rating of *N. vectensis* in Bembridge Harbour in 2013 is associated with the eDNA detection level of 73% (8 out of 11 samples) in 2017. The amplification for *N. vectensis* DNA was largely consistent and strong, which is indicative of the presence of higher amounts of target eDNA. Failure to detect *N. vectensis* in NM2767, NM2781, and NM2782, may be a true negative in those particular samples or, at least in NM2767 and NM2782, may be due to the unexplainable lower amount of DNA extracted from these filter (0.7 ng/µl vs. average 16.5 ng/µl).

Harbour Farm Lagoon

Traditional monitoring methods have classified the *N. vectensis* populations at Harbour Farm Lagoons (a complex of lagoons) as common or frequent, the eDNA detection level for the site as a whole was 40% (6 out of 15 samples), or 60% for Harbour Farm Lagoon 1 (6 out of 10) and 0% for Harbour Farm Lagoon 2 (0 out of 5). Whether *N. vectensis* is still in the second lagoon is unknown given that the last traditional survey of the site as a whole was in 2013.

Interestingly the strength of the PCR amplifications is much fainter and more inconsistent among replicates compared to the Bembridge and Newtown samples (Figures 5-7), which

could indicate lower amounts of eDNA being present in the samples, this might be confounded by the typically lower amount of water processed in these samples (average ~130 ml for Harbour Farm vs. ~270 ml for Bembridge Harbour vs. ~610 ml for Newtown Harbour). Despite having the smallest average volume of water sampled among the sites, the total DNA concentration for these samples (including DNA from non-target species) was the highest (average >20 ng/µl for Harbour Farm vs. ~13.6 ng/µl for Bembridge Harbour vs. ~9.7ng/µl for Newtown Harbour), it should be noted however that the DNA concentration of each sample refers to the total DNA present in the sample and not the DNA concentration of the target organism.

Newtown Harbour

All traditional surveys have not detected *N*. *vectensis* in Newtown Harbour before (up to 2013). All 5 samples analysed here with eDNA were positive for *N*. *vectensis*. These particular samples were particularly amenable because they had the good level of total DNA concentration bolstered by the fact that a great deal of water had been sampled, which is likely to increase detection rates especially for rare species.

Gammarus insensibilis eDNA assay

Primer selection

There is a paucity of genetic data for which to design bespoke *Gammarus insensibilis* primers. As of the 16th of August 2018, NCBI has a total of 4 COI sequences. COI has well known issues with eDNA analysis from filters, specifically COI primers tend to amplify bacterial species that are not represented on the NCBI database and are therefore unforeseeable. We targeted the 16S mtDNA region of *Gammarus insensibilis* because the region is highly variable allowing for species-specific primers, but also avoids the issue of bacterial amplification common with COI mtDNA primers. A literature search identified two studies in which 16S had been amplified and sequenced for other species of *Gammarus* (Table 6) and trialled the primers used in those studies. LSU-oni-F + LSU-oni-R produces a product 360 - 415 bp long and while LR-J-Gf + LR-N-Gf produces a product 395 bp long.

Table 6. Primers, their sequences, and their source

Primer	Sequence (5' - 3')	Source
LSU-oni-F	CGC CTG TTT AAC AAA GAC AT	Michel-Salzat and Bouchon (2000). Life Sciences 323: 827–837
LSU-oni-R	TCG GTC TGA ACT CAG ATC ACG T	Michel-Salzat and Bouchon (2000). Life Sciences 323: 827–837
LR-J-Gf	AAG GTT GAA CAA ACC CTC TAC T	Müller (2000). Molecular Phylogenetics and Evolution 15(2): 260–268
LR-N-Gf	AAG TAA AAC CTG CCC GGT GCT T	Müller (2000). Molecular Phylogenetics and Evolution 15(2): 260–268

In vitro tests

Primer pairs were tested against 6 tissue *G. insensibilis* samples collected from Eight Acre Pond, Saltern Lagoon, Lymington-Keyhaven on the 27th September 2016. The concentration of the DNAs ranged from 20 ng/µL to 94 ng/µL.

PCR conditions were as stated in the respective published articles. Despite several tests aiming to optimise several aspects of the PCR (including: annealing temperature, elongation time, primer concentration, cycle number, etc.), it was not possible to obtain a high-quality single-band PCR product. Bands tended to be very weak, and the primers often produced multiple bands for each sample (Figure 8). Consequently this assay could not be used to determine the presence or absence of *G. insensibilis* in the Newtown Harbour and Bembridge lagoons.

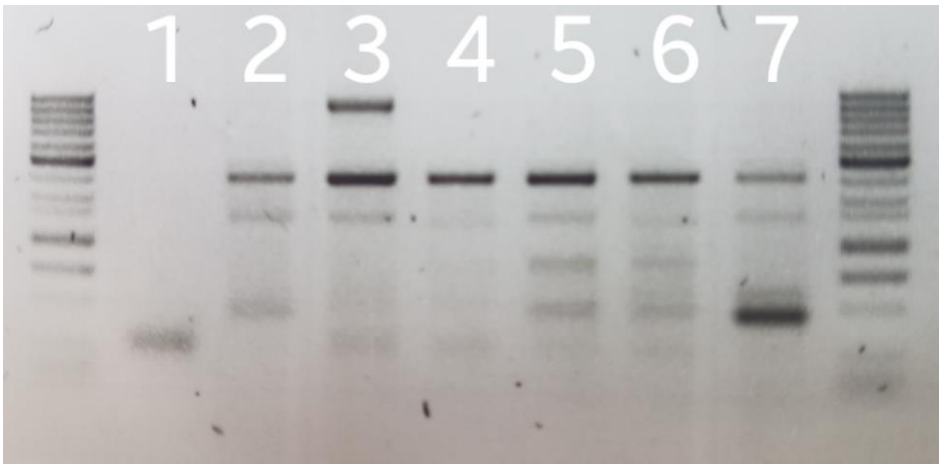


Figure 8. Gel electrophoresis of LR-J-Gf + LR-N-Gf. The expected band size is 395 bp, which is consistently dark band present in 2-7. Lane 1 is the template negative control, while 2-7 are the 6 *G. insensibilis* tissue DNA extracts. Multiple bands are present.

Future developments

Nematostella vectensis eDNA assay

Here we have developed a set of primers that specifically amplify *Nematostella vectensis* and started the initial stages of their validation. These primers seem specific to *N. vectensis* but require a greater deal of validation, including:

1. More extensive field testing at sites of known presence and absence of *N. vectensis.* These tests need to be corroborated with traditional field tests so that an idea of abundances and ecology can be obtained.
2. A series of artificial or laboratory dilution assays to determine proper limits of detection.
3. An improvement on the efficiency of the sequencing reaction needs to be made as well because the quality of these sequences, while high enough to identify them as *N. vectensis*, were low. Multiple sequences could be assembled to improve the quality.

As it stands, a positive sample that has been sequenced can be interpreted as a positive detection, but more work is required to confirm whether negative assays equate to truly negative samples especially if a small volume of water has been processed.

Gammarus insensibilis eDNA assay

The failure to design specific primers is down to the fact that there is a paucity of genetic reference data available for this species. Obtaining the first reference sequences relies on the liberal use of universal primers and a great deal of sequencing effort. More time is required to get this started from the ground up. Alternative future work on this project could involve building an alignment of other sequenced *Gammarus* species from which to design genus-specific, rather than species-specific primers.

Metabarcoding of invertebrate samples from vane traps and from riparian pitfall traps

Scope

This report forms the metabarcoding component of a Natural England project seeking to understand the potential and limits of applying metabarcoding protocols to monitoring arthropods. Two sets of samples were collected in a woodland but used different trap types (Burghley House: 10 Vane traps with 1 Owen trap; King's Wood: 10 Flight Intercept traps with 1 Owen trap), whilst the third set of samples were riparian pitfall traps (River Frome). For all three datasets the specimens were identified by Natural England staff prior to DNA extraction to enable a capture efficiency comparison between metabarcoding and morphology-based methods for analysing bulk invertebrate samples.

We tested two different primer sets on the Burghley House and King's Wood samples as previous experience had shown that the standard *Leray* primers produce datasets with a significant proportion of non-target taxa. This can be extremely wasteful as large amounts of data may be discarded after sequencing. The arthropod-specific *Zeale* primers were therefore used in parallel to determine whether these would make a more cost-effective alternative. Both primer sets target the metazoan 'barcode' region so there should be no difference in the availability of reference sequences for identification.

The King's Wood samples were size sorted prior to DNA extraction to test whether this could reduce the biasing effect of large specimens on the results. These samples were analysed with the *Leray* primer set only. Hereafter the three datasets will be referred to as *BurghleyHouse*, *RiverFrome* and *KingsWood*, with the marker under discussion appended, e.g. *RiverFrome-Leray*.

Methods

Laboratory

BurghleyHouse and *RiverFrome*

Each trap sample was split into smaller subsamples, each of which was homogenised with a pestle and mortar using liquid nitrogen. DNA from each homogenate was extracted using a commercial DNA extraction kit with a protocol modified to increase DNA yields. DNA extracted from each subsample was pooled in equal volumes. Purified DNAs were amplified with 12 replicate PCRs for two different sections the mitochondrial cytochrome c oxidase

subunit 1 gene:

1. *Leray*: A 313 bp segment located within and at the 3' end of the standard 658 bp COI barcode region. The primers were designed to broadly amplify in Metazoans (Leray et al. 2013. Frontiers in Zoology 10:34).
2. *Zeale*: A 157 bp segment located within and at the 5' end of the standard 658 bp COI barcode region. The primers were designed to broadly target Arthropods (Zeale et al. 2011. Molecular Ecology Resources 11:236–244).

All PCRs were performed in the presence of both a negative control and a positive control sample (mock community with a known composition). Amplification success was determined by gel electrophoresis. PCR replicates were pooled and purified, and sequencing adapters were added. Success was determined by gel electrophoresis. Amplicons were purified and checked by gel electrophoresis, these were then quantified using a Qubit high sensitivity kit (Table 1) according to the manufacturer's protocol. All purified index PCRs were pooled into a final library with equal concentrations. The final library was sequenced using an Illumina MiSeq V2 kit at 15 pM with a 10% PhiX spike in.

KingsWood

Specimens from each trap sample were size sorted to limit the effect of larger individual specimens overwhelming the DNA of smaller individuals. Specimens longer than approximately 1 cm were removed and from these, the head and prothorax was removed and returned to the bulk sample for inclusion in the DNA extraction, the body was retained and stored. This coarse size sorting was seen as a good compromise between thoroughness and eventual applicability of these techniques. Each sample was then homogenised in a Precellys homogeniser. DNA extraction, PCRs, pooling, purification and sequencing all followed the same procedure as previously except that only the *Leray* primer set was used.

Bioinformatics

All *Leray* datasets were processed together using a custom bioinformatics pipeline for quality filtering, denoising, OTU clustering at 98%, and taxonomic assignment. The same procedure was also applied independently to the *Zeale* datasets. After OTU clustering, taxa were given a preliminary identification with MEGAN, based on blastn searches against the NCBI *nt* database (Huson et al., 2016:https://doi.org/10.1371/journal.pcbi.1004957). Species and genus level identifications were made where possible for all Arthopoda OTUs via a two-step process. Firstly, the OTUs were queried against the Barcode of Life Database (BOLD) and species-level assignments made where there were matches at 98% or better to a single species. Genus-level assignments were made where there were matches at 97% or better to a single genus. Secondly, the original blastn search results were filtered to retain only high-quality hits (≥98% identity and ≥90% amplicon length covered). Species- or genus-level assignments were made where all hits agreed. For both blastn- and BOLD-based assignments, where there was species-level agreement at ≥99% but only genus-level agreement at ≥98%, the species-level assignment was accepted. For the remaining OTUs

the MEGAN identification was used but limited to family level only. The same procedure was applied to both the *Leray* and *Zeale* analyses.

The *RiverFrome* Rushton subsamples (A & B; NMID2334 and NMID2335) were combined together for all further analyses. After taxonomic assignment the OTU table was filtered to ensure that very low abundance OTUs, likely to represent cross-contamination between samples, were removed. OTUs had to accrue at least 0.015% (*Leray*) or 0.085% (*Zeale*) of the reads mapped to a sample to be accepted as present in that sample. These levels were set to ensure that the minimum number of reads for any OTU in any sample was not less than 10.

Comparison with morphology

The morphological results provided by Natural England were cleaned up to include a single count per sample per species for those specimens included in the DNA extraction. The resultant data table was then merged on species name with that obtained from each of the metabarcoding analyses. In the latter, any OTUs with the same species-level identification were collapsed.

Results

A total of 26 samples in three datasets were analysed, with the first two datasets (Burghley House and River Frome) processed together in the laboratory and the third dataset (King's Wood) processed later.

Laboratory

Burghley House and River Frome samples were amplified with two different primer sets targeting the cytochrome *c* oxidase subunit I (COI) 'barcode' region – *Leray* for Metazoa and *Zeale* for Arthropoda. King's Wood samples were amplified with the *Leray* primers only. Sequencing data from all three *Leray* datasets was analysed together for comparability and any non-arthropod taxa were removed from the final results.

DNA yields were as expected (Table 1). PCR reactions were consistently successful for all 26 samples. Electrophoresis bands were strong and of the expected size and no repeat PCRs were necessary. Electrophoresis showed that all samples were successfully indexed, with strong bands all of the expected size. All amplicons were successfully purified and were of high yield (Table 1).

Table 1. Concentration of DNA extracts and purified index PCRs.

NM ID	Sample ID	DNA (ng/µl)	Index (ng/µl) - Leray	Index (ng/µl) - Zeale
2323	Tree 1 Burghley	45	10.3	2.5
2324	Tree 2 Burghley	31.85	6.68	2.2
2325	Tree 3 Burghley	23.7	7.6	1.9
2326	Tree 4 Burghley	70.2	12	1.6
2327	Tree 5 Burghley	71.3	10.5	1.68
2328	Tree 6 Burghley	32.25	9.44	1.86
2329	Tree 7 Burghley	32.6	11.4	2.64
2330	Tree 8 Burghley	12.17	11.5	2.1
2331	Tree 9 Burghley	46.9	9.8	2.08
2332	Tree 10 Burghley	12.21	6	1.26
2333	Owen trap Burghley	62.55	12.3	2.1
2334	R. Frome - Rushton Riparian A	2.235	1.27	1.6
2335	R. Frome - Rushton Riparian B	5.22	7.76	1.12
2336	R. Frome - Moreton Riparian	32.6	9.48	1.73
2337A	R. Frome - Woodsford Riparian	29.2	8.84	1.87
2807	Beech 31	7.34	9.76	NA
2808	Oak 18	> 20	9.42	NA
2809	Oak 2	> 20	9.1	NA
2810	Oak 29	> 20	10.4	NA
2811	King's Wood Combined	> 20	9.12	NA
2812	Oak 4	> 20	10.2	NA
2813	Oak 17a	> 20	11.2	NA
2814	Oak 28	> 20	9.24	NA
2815	Oak 15	> 20	9.08	NA
2816	Owen trap	> 20	9.98	NA
2817	Oak 20	> 20	9.76	NA

Bioinformatics

The MiSeq paired-end sequencing for the *Leray* amplicon yielded 1,434,877 reads for *BurghleyHouse*, of which 96.0% passed our internal quality filter. For *RiverFrome* the yield was 417,748 with 96.6% retained for analysis. For *KingsWood* the yield was 1,319,195 with 86.4% passing our quality filter. These statistics are indicative of a high-quality data. The positive and negative controls performed as expected. The final *Leray* dataset includes 316 OTUs (75%) and 99.22% of mapped reads (86 OTUs representing 0.72% of reads, either non-Arthropoda or not identified to family, were discarded; 19 Arthropoda OTUs and 0.06% of reads were removed by filtering).

For the *Zeale* amplicon there were 226,264 reads for *BurghleyHouse*, with 95.1% passing the quality filter. For *RiverFrome* there were 61,208 reads and 94.8% were retained. The negative control performed as expected, however the positive control recovered only one of the expected species. The final dataset includes 67 OTUs (87%) and 88.54% of mapped reads (9 OTUs representing 11.39% of reads were discarded because they were not

identified to family; 1 OTU and 0.06% of reads were removed by filtering).

Taxonomic composition

In all three *Leray* datasets the majority of OTUs and reads were Coleoptera (overall: 95.27% OTUs and 99.94% reads; *BurghleyHouse*: 97.12% OTUs and 99.99% reads; *RiverFrome*: 92.17% OTUs and 99.75% reads; *KingsWood*: 98.37% and 99.94% reads).

In *BurghleyHouse-Leray* the single largest OTU was identified as *Melanotus castanipes* (45.64% reads), the next largest OTU was identified as *Harmona axyridis* (10.96% reads). All samples were dominated by a single OTU (ranging from 48.72% to 91.54%). In all but one sample the dominant OTU was identified to species. For 6 of 11 samples the OTU was identified as *Melanotus castanipes*, in the remaining samples the dominant OTUs were variously Cantharidae, Elateridae, Curculionidae, and Coccinellidae. See Appendix 1 for a full taxon list. In *RiverFrome-Leray* the largest OTU was identified as *Notaris acridulus* (22.33%), followed by *Platambus maculatus* (18.65%). All three samples were strongly dominated by a single OTU but the species was different in each case. Rushton was dominated by *Notaris acridulus* (51.84%), Moreton by *Nebria brevicollis* (20.11%), and Woodsford by *Platambus maculatus* (44.26%). See Appendix 2 for a full taxon list.

In *KingsWood* the largest OTU was identified as *Melanotus castanipes* (25.70%), followed by *Athous subfuscus* (11.55%). All samples were strongly dominated by a single OTU (ranging from 34.87% to 94.12%) and in all cases the OTU was fully identified to species. In 8 of 11 samples the dominant OTU was an Elateridae, in the remaining samples they were identified as Cerambycidae, Curculionidae, and Ptinidae. See Appendix 3 for a full taxon list.

In the two *Zeale* datasets all but one of the OTUs were Coleoptera (overall: 98.5% OTUs and 99.50% reads; *BurghleyHouse*: 96.43% OTUs and 99.25% reads; *RiverFrome*: 100% OTUs and 100% reads). In *BurghleyHouse-Zeale* the single largest OTU was identified as *Phyllobius pyri* (39.79%), followed by *Phyllobius* sp. (32.19%) and *Tachyporus hypnorum* (6.60%). Most of the samples were dominated by a single OTU (ranging from 54.12% to 94.35%). In 8 of the samples the dominant taxon was either *Phyllobius pyri* or *Phyllobius* sp., and one sample was dominated by *Tachyporus hypnorum*. Of the remaining two samples, one had over 80% of the reads split between *Phyllobius pyri* or *Phyllobius* sp., whereas in the other sample over 99% of the reads were split between *Phyllobius pyri*, *Cantharis livida* and *Brassicogethes aeneus*. See Appendix 4 for a full taxon list.

In *RiverFrome-Zeale* the largest OTU was identified as *Pterostichus* sp. (36.27%), followed by *Carabus granulatus* (14.57%) and *Elaphrus cupreus* (12.36%). Both the Moreton and Woodsford samples were strongly dominated by *Pterostichus* sp. (over 80% of reads in each), whereas four different taxa accrued at least 10% of the reads in the Rushton sample – *Carabus granulatus*, *Pterostichus* sp., *Elaphrus cupreus*, and *Paranchus albipes*. See Appendix 5 for a full taxon list.

Table 2. Taxonomic richness among the samples with *Leray*. Some taxa could not be identified to species or genus and so a higher taxonomic resolution is given in Appendices 1-3.

NM ID	Sample ID	Order	Family	Genus	Taxa (IDed to species)
2323	Tree 1 Burghley	1	8	14	22 (14)
2324	Tree 2 Burghley	3	8	9	15 (9)
2325	Tree 3 Burghley	1	9	15	21 (15)
2326	Tree 4 Burghley	1	5	5	15 (5)
2327	Tree 5 Burghley	1	11	17	28 (17)
2328	Tree 6 Burghley	2	10	12	25 (12)
2329	Tree 7 Burghley	1	5	7	17 (7)
2330	Tree 8 Burghley	1	9	10	18 (10)
2331	Tree 9 Burghley	1	8	9	15 (9)
2332	Tree 10 Burghley	1	9	10	18 (10)
2333	Owen trap Burghley	2	11	15	29 (15)
2334-5	R. Frome - Rushton Riparian	4	16	27	53 (27)
2336	R. Frome - Moreton Riparian	4	12	34	73 (34)
2337A	R. Frome - Woodsford Riparian	2	10	23	57 (23)
2807	Beech 31	1	5	7	27 (7)
2808	Oak 18	1	10	13	28 (13)
2809	Oak 2	1	9	9	18 (9)
2810	Oak 29	1	12	16	36 (16)
2811	Kings' Wood Combined	2	11	24	58 (24)
2812	Oak 4	1	9	11	15 (11)
2813	Oak17a	1	6	7	15 (7)
2814	Oak 28	1	8	12	27 (12)
2815	Oak 15	1	2	2	6 (2)
2816	Owen trap	1	6	10	31 (10)
2817	Oak 20	2	7	13	22 (13)

Table 3. Taxonomic richness among the samples with Zeale. Some taxa could not be identified to species or genus and so a higher taxonomic resolution is given in Appendices 4-5.

NM ID	Sample ID	Order	Family	Genus	Taxa (IDed to species)
2323	Tree 1 Burghley	1	3	3	5 (3)
2324	Tree 2 Burghley	1	2	3	4 (3)
2325	Tree 3 Burghley	1	3	4	5 (4)
2326	Tree 4 Burghley	1	2	3	3 (3)
2327	Tree 5 Burghley	1	5	6	8 (6)
2328	Tree 6 Burghley	1	5	5	7 (5)
2329	Tree 7 Burghley	1	3	4	5 (4)
2330	Tree 8 Burghley	1	3	4	5 (4)
2331	Tree 9 Burghley	1	3	4	5 (4)
2332	Tree 10 Burghley	2	5	6	7 (6)
2333	Owen trap Burghley	1	4	3	7 (3)
2334-5	R. Frome - Rushton Riparian	1	1	5	23 (5)
2336	R. Frome - Moreton Riparian	1	1	5	24 (5)
2337A	R. Frome - Woodsford Riparian	1	3	7	29 (7)

The combined *Leray* analysis detected 316 taxa in 8 orders: Coleoptera (42 families; 302 taxa), Diptera (4 families; 5 taxa), Hymenoptera (3 families; 3 taxa), Hemiptera (1 family; 2 taxa), Psocodea (1 family; 1 taxon), Sarcoptiformes (1 family; 1 taxon), Isopoda (1 family; 1 taxon), Amphipoda (1 family; 1 taxon). Overall, 194 taxa were identified to 167 species. The average taxon richness per sample was 28, ranging from 6 (Oak 5 Kings) to 73 (R. Frome - Moreton Riparian).

The *Zeale* analysis detected 68 taxa in 2 orders: Coleoptera (12 families; 66 taxa) and Lepidoptera (1 family; 1 taxon). Overall, 36 taxa were identified to 32 species. The average taxon richness per sample was 10, ranging from 3 (Tree 4 Burghley) to 29 (R. Frome - Woodsford Riparian).

There were striking differences in composition between samples, both within each site and between *Leray* and *Zeale*. Figures 1a and 1b summarise the recovered composition at family level for *Leray* and Zeale respectively for the Burghley House samples. Equivalent information is shown for the River Frome samples in Figures 2a and 2b, and the *Leray* results for King's Wood are shown in Figure 3.

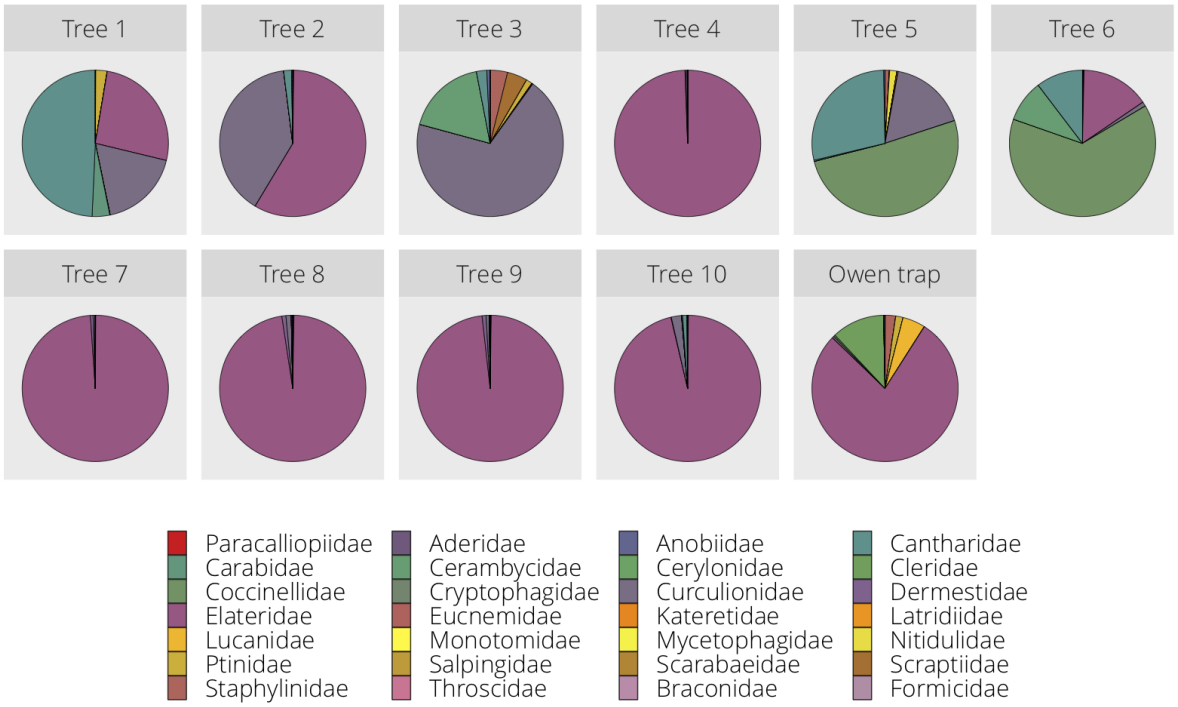


Figure 1a. Proportion of the sequencing output per family with the *Leray* primers for the Burghley House samples. Most samples are dominated by Elateridae; Tree 1 is dominated by Cantharidae; Tree 3 is dominated by Curculionidae; Trees 5 and 6 are dominated by Coccinellidae. The colours used to represent each family are consistent throughout the report.

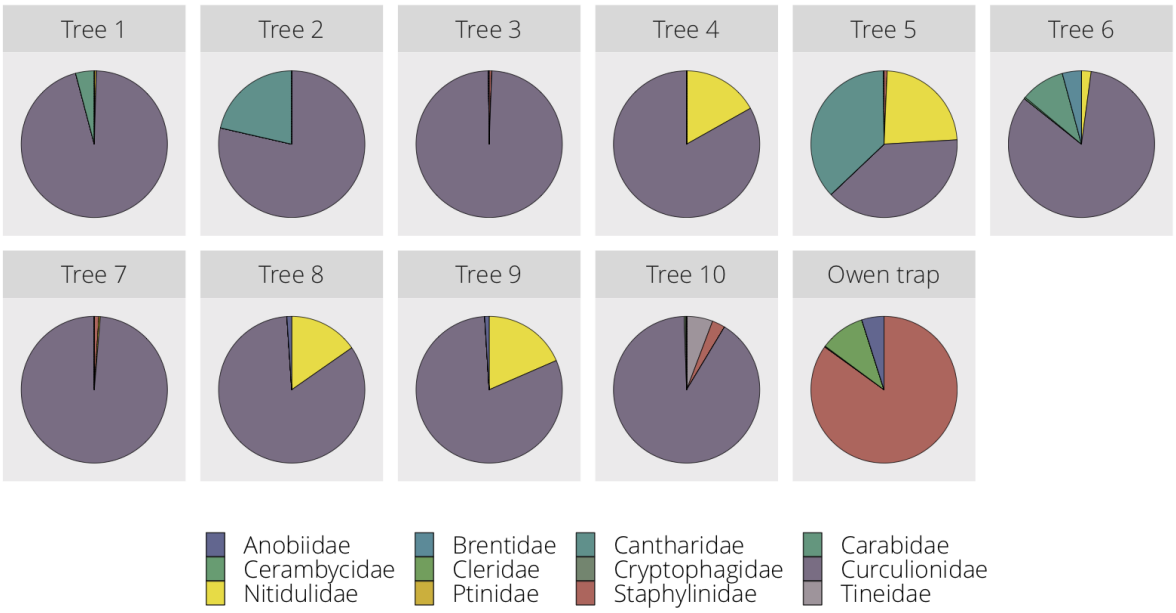


Figure 1b. Proportion of the sequencing output per family with the *Zeale* primers for the Burghley House samples. All FIT samples dominated by Curculionidae whereas the Owen trap is dominated by Staphylinidae. Other significant groups include Cantharidae and Nitidulidae in some FIT samples.

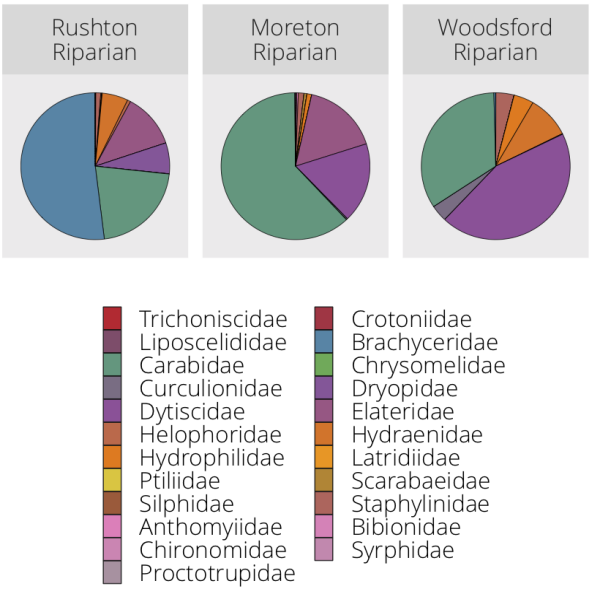


Figure 2a. Proportion of the sequencing output per family with the *Leray* primers for the River Frome samples. Rushton is dominated by Brachyceridae followed by Carabidae; Moreton is dominated by Carabidae followed by Dytiscidae and Elateridae; Woodsford is dominated by Dytiscidae followed by Carabidae

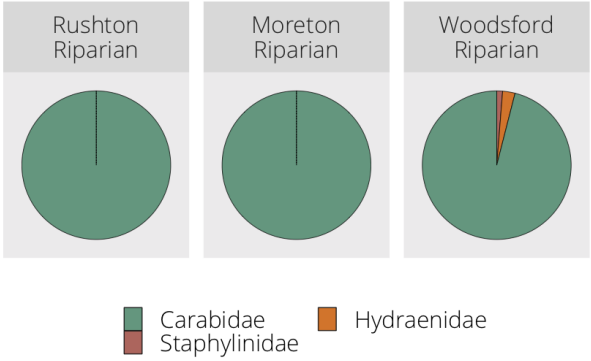


Figure 2b. Proportion of the sequencing output per family with the *Zeale* primers for the River Frome samples. All samples are dominated by Carabidae.

.

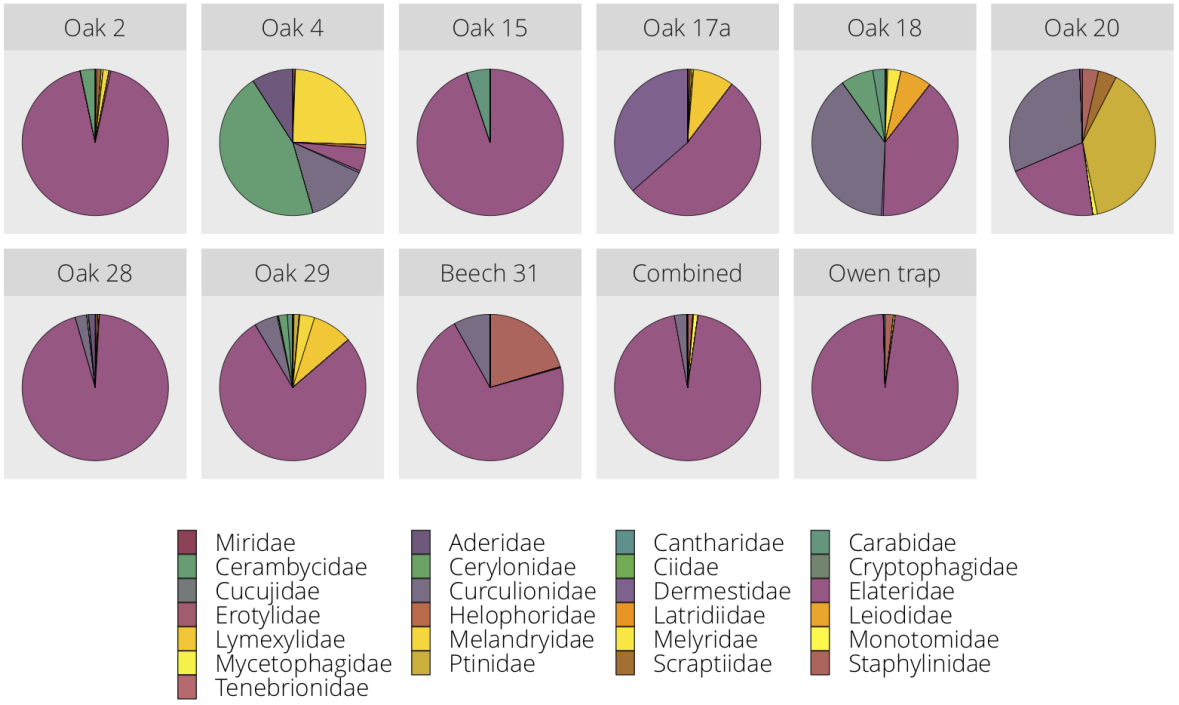


Figure 3. Proportion of the sequencing output per family with the *Leray* primers for the King's Wood samples. Most samples are dominated by Elateridae; Oak 4 is dominated by Cerambycidae followed by Melandryidae; Oak 20 is dominated by Ptinidae followed by Curculionidae and Elateridae; Oak 17a has a large proportion of Dermestidae; Oak 18 has a large proportion of Curculionidae; Beech 31 has a large proportion of Staphylinidae.

Comparison with morphological data

Morphological identifications are available for all three sets of samples. As with the metabarcoding data, not all specimens were fully identified so the following comparison is limited to the specimens and OTUs with species level labels

BurghleyHouse

In this dataset 97 species were identified from expert examination of the specimens whereas 63 and 20 species were identified from the *Leray* and *Zeale* metabarcoding data respectively. Between the three datasets there were a total of 122 species with 42 shared between specimens and metabarcoding. This represents a recovery rate by metabarcoding of 43.3% relative to the expected species list. Of these, 29 were recovered by *Leray* only, 2 were recovered by *Zeale* only, and 11 were recovered by both markers.

Several Coleoptera species were unique to the metabarcoding datasets: *Cantharis cryptica, Leistus spinibarbis, Phyllobius roboretanus, Brassicogethes aeneus,* and *Sepedophilus testaceus* were recovered by both *Leray* and *Zeale*; *Dromius meridionalis, Adalia decempunctata, Cryptophagus schmidti, Anthrenus verbasci, Agriotes pallidulus, Brachypterus urticae, Mycetophagus quadripustulatus, Epuraea aestiva, Hedobia imperialis, Ptinus fur, Aleochara stichai, Atheta deformis, Carpelimus corticinus, Mocyta orbata, Philhygra palustris,* and *Trixagus carinifrons* were recovered by *Leray* only; and *Cryptophagus pallidus* was recovered by *Zeale* only.

RiverFrome

In this dataset there were 95 species in the morphology-based taxon list, with 60 and 12 species in the *Leray* and *Zeale* analyses respectively. In total there were 109 species identified between the three datasets, of which 47 were shared between specimens and metabarcoding. This represents a recovery rate by metabarcoding of 49.5% relative to the expected species list. Of these, 37 were recovered by *Leray* only and 10 were recovered by both markers. *Zeale* did not recover any additional species from the specimen list.

Several Coleoptera species were unique to the metabarcoding datasets: *Pterostichus anthracinus* was recovered by both *Leray* and *Zeale*; *Melanotus castanipes, Enicmus transversus, Ptenidium intermedium, Aleochara lanuginosa, Anotylus sculpturatus,* and *Carpelimus bilineatus* were recovered by *Leray* only; *Pterostichus rhaeticus* was recovered by *Zeale* only.

KingsWood

In this dataset 94 species were identified in the morphological assessment while 60 species were identified from the metabarcoding data. Between the two datasets there were a total of 105 species, with 49 shared. This represents a recovery rate by metabarcoding of 52.1 % relative to the expected species list.

Several Coleoptera species were unique to the metabarcoding dataset: *Dromius agilis, Nebria salina, Cryptophagus scanicus, Helophorus brevipalpis, Elateroides dermestoides, Conopalpus testaceus, Anaspis maculata, Gabrius breviventer, Gyrohypnus angustatus, Habrocerus capillaricornis*.

Discussion

Comparison of markers

In the present study two alternative markers were trialled because previous experiments with the *Leray* primers have suffered large data losses due to significant non-target amplification. This does not appear to have been an issue herein, possibly because the target DNA in a homogenised specimen sample is so much more abundant than any non-target DNA. For this sample type the choice of marker should therefore be driven by the accuracy of the recovered species compositions.

In the two datasets for which comparative data are available, *Leray* outperformed *Zeale* with regards to both the proportion of expected species recovered and overall diversity. The amount of data was highly unequal between the two markers, with the number of *Zeale* sequences approximately 15% of that available for *Leray.* However, subsampling of the *Leray* data to match *Zeale* produces almost identical species richness estimates as the full dataset (for both *BughleyHouse* and *RiverFrome*), suggesting that the low observed diversity in the *Zeale* datasets is not an artefact of low sampling effort.

The *Leray* primers contain more degenerate nucleotides (6 on the forward primer; 7 on the reverse) than the *Zeale* primers (1 on the forward; 2 on the reverse), having been designed for Metazoa and Arthropoda respectively. Consequently, the likelihood of mismatch between the primers and the DNA of any particular species in the sample is much higher for the *Zeale* primers unless the binding site is very conserved between taxa. This does not appear to be the case, with Figures 4 and 5 showing that the number of mismatches between the primers and a sample of reference sequences is always higher for *Zeale*, but that some taxonomic groups are affected worse than others. For example, the likelihood of mismatch appears to be much greater for Coleoptera than Lepidoptera (Figure 4) and for Elateridae than Carabidae (Figure 5). Such differences are likely to explain the strongly contrasting compositions observed between the primer sets in Figures 1a, 1b, 2a and 2b. This is because, in mixed samples with variable mismatch between primer and target DNA, the species with the fewest mismatches are likely to be preferentially amplified.

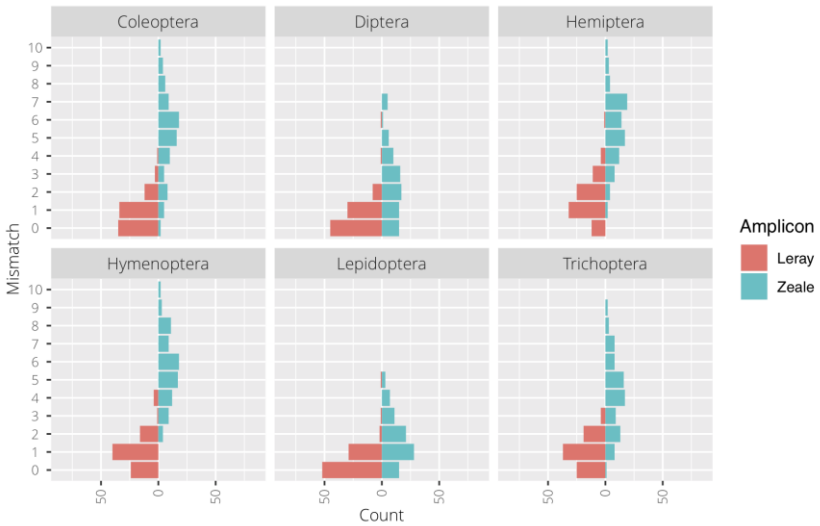


Figure 4. Frequency distribution of the total number of mismatches between a sample of reference sequences and the *Leray* and *Zeale* primers for six insect orders.

Comparison with morphology

When compared with the morphological assessment of the same samples it is apparent that the metabarcoding approach has underperformed across all three datasets. Taking both *Leray* and *Zeale* together, 43.4% of the expected species were recovered for Burghley House, 49.5% for River Frome, and 52.1% for King's Wood. Where both *Leray* and *Zeale* were available, the majority of recovered species were found in the *Leray* datasets.

The samples contained specimens with a large range of body sizes and it is likely that the method of DNA extraction biased the results towards higher biomass species in each sample. For the *KingsWood* samples an attempt was made to reduce the effect of large specimens by reducing the amount of tissue that was included in the homogenisation. The slightly increased rate of expected species recovery may indicate that this strategy was partially successful, however further refinement is required. For example, a more graduated approach to size sorting with multiple size classes may be more appropriate, and the reduction in tissue may need to be more significant for the very largest specimens, e.g. a single leg.

Another aspect of the experimental design that could contribute to bias in the metabarcoding data is the PCR conditions that are used to amplify *Leray*. Future work should test whether a reduction in the number of PCR cycles can decrease the bias towards a small number of species. PCR amplification leads to exponential increases in the number of sequences for the amplified fragments so the species that are amplified first (due to primer binding efficiency and/or abundance of DNA) may rapidly come to dominate the sample over successive cycles, to the exclusion of lower biomass species or those with more divergent sequences at the primer-binding sites.

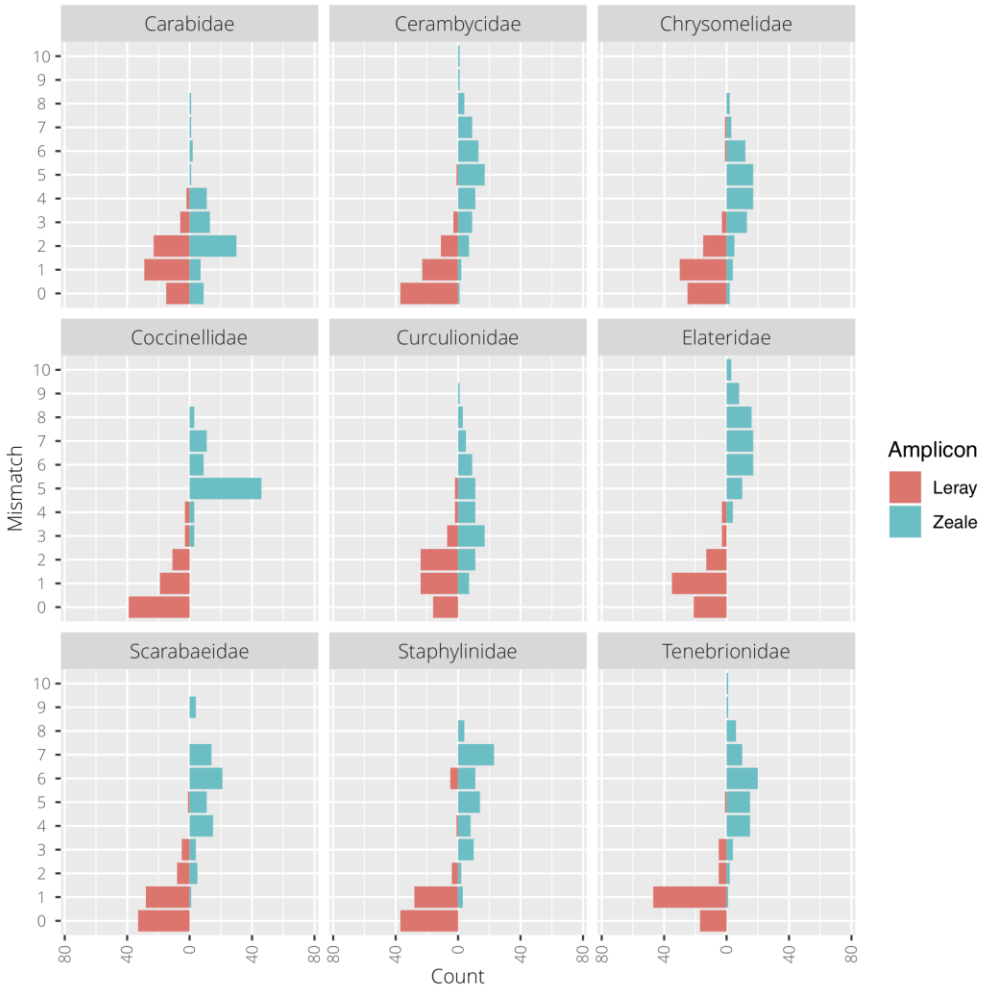


Figure 5. Frequency distribution of the total number of mismatches between a sample of reference sequences and the *Leray* and *Zeale* primers for nine families of Coleoptera.

Conclusion

From the present study we conclude that the *Leray* primer set is the more promising for bulk invertebrate samples because the potential for taxonomic bias introduced by differential primer binding efficiency is reduced relative to the *Zeale* set. It is unlikely that changes to PCR conditions would significantly improve the recovery of expected species with the *Zeale* primers due to the variation in how well they match to the target DNA of different taxa. In contrast, the number of mismatches observed in *Leray* is consistently low, in particular in the reverse primer where there were no mismatches in the majority of cases.

Although the comparison between the morphological and metabarcoding results indicated that the latter has underperformed, there is some indication herein that size sorting could be a valuable tool to combat the dominance of species with a large input biomass. Further experiments with different size sorting strategies should be prioritised, along with careful examination of the effects of *Leray* PCR conditions on the distribution of sequences between recovered taxa.

Appendices are available as separate downloads

Appendix 1
Proportion of the sequencing output allocated to the different taxa in the *BurghleyHouse-Leray* dataset. 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. Darker shades of blue correspond to higher proportion of sequence output per site.

Appendix 2.
Proportion of the sequencing output allocated to the different taxa in the *RiverFrome-Leray* dataset

Appendix 3.
Proportion of the sequencing output allocated to the different taxa in the *KingsWood* dataset.

Appendix 4.
Proportion of the sequencing output allocated to the different taxa in the *BurghleyHouse-Zeale* dataset.

Appendix 5.
Proportion of the sequencing output allocated to the different taxa in the *RiverFrome-Zeale* dataset.

Appendix 6.
Frequency of occurrence of all detected families obtained from the *BurghleyHouse* dataset (*Leray* and *Zeale*). Numbers correspond to the number of taxa belonging to those families in those samples.

Appendix 7.
Frequency of occurrence of all detected families obtained from the *RiverFrome* dataset (*Leray* and *Zeale*). Numbers correspond to the number of taxa belonging to those families in those samples.

Appendix 8
Frequency of occurrence of all detected families obtained from the *KingsWood* dataset (*Leray*). Numbers correspond to the number of taxa belonging to those families in those samples.