**JNCC Report**
**No. 669a**

# Guidance for end users on DNA methods development and project assessment

**Jones, E.P., Adams, I.P., Walshaw, J., Benucci, M., MacArthur, R. Boonham, N. & Bryce S.**

**December 2020**



ISSN 0963-8091

**For further information please contact:**

Joint Nature Conservation Committee
Monkstone House
City Road
Peterborough PE1 1JY
https://jncc.gov.uk/

**This report should be cited as:**

Jones, E.P., Adams, I.P., Walshaw, J., Benucci, M., MacArthur, R., Boonham, N. & Bryce S. (2020) Guidance for end users on DNA methods development and project assessment. JNCC Report No. 669a. JNCC, Peterborough, ISSN 0963-8091.

**JNCC EQA Statement:**

This report is compliant with JNCC's Evidence Quality Assurance Policy

https://jncc.gov.uk/about-jncc/corporate-information/evidence-quality-assurance/

**Acknowledgments:**

This project was led and funded by the Defra DNA Centre of Excellence in 2019/20, which aims to transform Defra group science to support policy and delivery by sharing and demonstrating the wider potential of existing methods; enabling the rapid development and trialling of new DNA based method applications; providing a focal point for leadership in Defra, UK and across other government departments and building shared capability including skills and facilities.

## Summary


- There is increasing interest from environmental public bodies in applying DNA-based methods to improve the efficiency and comprehensiveness of environmental monitoring.

- The rapid evolution of DNA-based methods means there is a need to support non-specialist end users commissioning and evaluating studies using these methods.

- This document aims to provide guidance that can be used by public bodies commissioning the development of DNA-based monitoring methods, to help give confidence that projects are developed in a robust way with appropriate validation. The scope is DNA-based detection methods applied to monitor the environment: predominantly end point PCR and metabarcoding.

- The document includes explanations of key terms, followed by a (non-exhaustive) checklist of factors to consider in project design and development.

- The checklist contains the following, explained in more detail in this document:
    1. What is the purpose of the project (e.g. one-off research or a stage in developing a test intended for routine use)?
    2. What is the intended use of the test (e.g. standalone test vs. use to inform further surveys or support other evidence) and what are the impact of test results?
    3. How far is the test along the pathway to deployment?
    4. What level of sensitivity, specificity, accuracy and repeatability are required?
    5. Is there a current method that the DNA-based approach can be verified against?
    6. What scope is the test intended for (e.g. range of environmental conditions)?
    7. Is the sampling in the project appropriate given the distribution and rarity of the target organism(s)?
    8. How will sensitivity, specificity, accuracy, and repeatability be assessed in the project?
    9. Is it likely that contamination can be adequately controlled and monitored?
    10. Can the correct performance of the method or test be monitored?
    11. Can the accuracy of the method be assessed?
    12. Are there data confidentiality issues (e.g. personal data, location data for rare species)?
    13. For non-targeted testing, is there a possibility that the method will detect a notifiable organism?

- Additional practical constraints to consider are:
    1. The availability of the necessary DNA sequences.
    2. The availability of the necessary biological reference materials (e.g. to validate the test).
    3. Whether logistical constraints can be resolved (e.g. transporting preservatives and decontaminants into the field, seasonality and adverse weather affecting sample collection, *etc*.).

- Additional considerations specific to routine monitoring include:
    1. Sufficient laboratory capacity to process the samples that would be collected.
    2. Acceptable per sample cost.
    3. Sufficient benefits of method (e.g. cost, comprehensiveness, accuracy, *etc*.).
    4. Adequate turnaround time from sample submission to receiving results.

- Answers to these questions often requires technical expertise. If this is not available internally, it can be obtained as part of project development.

- Additional information can be found in the complementary JNCC Report No. 669b ‘End User Frequently Asked Questions on DNA-based Methods for Environmental Monitoring’.

- This guide should be revisited for potential update by 2025.

# Contents

# 1 Introduction

## 1.1 Background

The use of DNA-based methods to monitor the environment is increasing (e.g. Deiner *et al.* 2017), in part because these methods may offer cost or time savings over established approaches. They may also allow detection of organisms that are otherwise difficult to survey and determination of species composition in diverse communities. With the increased interest in DNA-based methods, there is a growing need for support to end-users commissioning and evaluating DNA-based studies, in order to improve the design and understanding of commissioned projects and to aid comparison of results and methods produced by different agencies and laboratories.

DNA-based analytical tests have been well established as standardised, accredited methods for over 20 years in many sectors and there are well established standards and frameworks that substantially overlap with the requirements for DNA monitoring of the environment (e.g. Bustin *et al.* 2009). The approach within this guidance note is to adopt the common language and frameworks from the plant health, animal health and GM sectors (all based in part on an ISO standard, ISO 17025 *General requirements for the competence of testing and calibration laboratories*), and adapt this to the area of Environment, making reference to other standards and frameworks and to ongoing work on validation and standardisation within the Environmental Monitoring area (e.g. DNAquaNet).

The existing frameworks and standards are well established for targeted detection tests, particularly using real time PCR platforms. However, very few standards exist for the application of High Throughput Sequencing (HTS) platforms (but see e.g. HTS for clinical tests, Aziz *et al*. 2014, or a developing ISO standard ISO/TC 34/SC 9/WG 25 for bacterial whole-genome sequencing for typing and genomic characterization), and none that we are aware of for metabarcoding. A number of working groups in other sectors (e.g. plant health, GM) will be reporting on this topic over the coming year or two; these may be applicable to environmental DNA. For example, the International Plant Protection Convention has a working group on “Next Generation Sequencing technologies as a diagnostic tool for phytosanitary purposes” due to report soon, and VALITEST (an ongoing EU project on validation of plant health diagnostic testing) is also producing guidelines on the use of HTS. It may be more cost and time effective to synthesise the guidance from these upcoming reports than to develop metabarcoding guidance for the Environment sector *de novo*.

The guidance given here is necessarily higher level and does not examine the technical detail of DNA detection methods and technologies. This is because the evaluation of the performance of methods is based on higher level concepts applicable to all methods (e.g. accuracy, repeatability, sensitivity) rather than technical details, and for the practical reason that it is not possible to comprehensively describe all relevant technical aspects for the range of sample types, target organisms, methodological steps and depth of audience technical expertise the document is intended to cover.

## 1.2 Scope

This document covers DNA-based detection methods applied to monitor the environment: predominantly end point PCR (such as real time PCR and conventional PCR visualised by gel electrophoresis) and metabarcoding using high throughput sequencing. Non-PCR based DNA detection methods (such as isothermal amplification methods like LAMP and RPA) exist but they are not the focus here. Also not in scope are DNA methods used to study the relationships between and within populations and species, such as population genetic and phylogenetic or phylogeographic studies.

This guidance note should be updated periodically, especially sections relating to HTS and metabarcoding. The use of DNA-based methods in the Environment sector is relatively new, and what is considered “good practice” is likely to change as the methods and applications mature. This document should be revisited for potential update by 2025.

## 1.3 Audience

As part of this project, we interviewed agencies and departments involved in monitoring the environment, who use DNA-based methods or may use them in the future. These ranged from agencies with one or two individuals with responsibility for exploring the use of DNA in environmental monitoring and relatively low molecular biology expertise, to agencies with multiple teams of experts in DNA methods.

This diversity of end user is reflected in how much the respondents would use a methods guidance document: for some, this would ideally be the central resource, while for others such as APHA, Fera and the Environment Agency Laboratories, existing standards would be referred to alongside or above this. This document is intended to align to these other standards, which to our current knowledge are ISO 17025 or local standards based around it.

The target audience for this document includes people with a broad range of technical knowledge but is focussed on those with some knowledge of DNA-based methods rather than experts or those with no molecular knowledge at all. As a starting point, the audience is assumed to be familiar with common DNA-based methods (e.g. real time PCR, metabarcoding from amplicon sequencing) and sample types (e.g. true ‘eDNA’, DNA from aggregated organisms) without necessarily an extensive knowledge of the technical details. Brief definitions of some of these technical terms are given in Section 2.2 below. The target audience also includes those solely commissioning work from external providers, and those partly involved in undertaking the project themselves.

## 1.4 Objectives

The objective of this document is to provide concise accessible guidance that can be used by public bodies commissioning the development of DNA-based monitoring methods to:

- give confidence that standardised protocols proposed for operational use have been developed in a robust way with appropriate validation;
- assess whether test methods are being deployed in a way that is fit for their particular purpose.

# 2 Explanation of key concepts and terms

This section defines how selected terms relevant to DNA-based methods are to be interpreted in the rest of this document, and also provides an introduction to key concepts.

## 2.1 General terms

### 2.1.1 Environmental DNA (eDNA)

True environmental DNA is DNA shed by an organism into its environment, rather than a sample composed of the organism itself. For example, fish DNA captured from a water

sample is eDNA, while bacteria captured from a water sample will primarily be composed of the bacteria (not eDNA shed by the bacteria). This definition is not always clear cut.

### 2.1.2 Polymerase Chain Reaction (PCR)

A PCR amplifies a target region of DNA exponentially, so that a small number of copies of target DNA at the start of the reaction can be amplified up to millions of copies of the target (called PCR products or amplicons) by the end of multiple cycles (typically 25+). The target DNA region to be amplified is determined by *primers*, which are synthetic stretches of DNA typically 18-25 bases long, that bind to conserved regions of DNA flanking the target sequence. All of the methods discussed in this document (barcoding, metabarcoding, species-specific real time PCR) are based on PCR.

### 2.1.3 DNA barcode region

A DNA barcode region is a specific region of DNA that has been selected as a target for sequencing, because for a given taxon (e.g. vertebrates) it is consistently and sufficiently dissimilar between species (or genera) to identify them correctly. Commonly used DNA barcode regions vary for different taxa, and more than one barcode may be necessary to identify some groups to species. The barcode region also has to be flanked on either side by relatively conserved regions so that primers can bind consistently across a taxonomic group.

### 2.1.4 Gel electrophoresis

Many applications need to determine that a PCR has successfully amplified the target DNA. This is most frequently done by running stained PCR products along a gel using electrophoresis, alongside a size ladder that can be used to determine the size (DNA length) of the PCR amplicon.

### 2.1.5 End point PCR

In this document, end point PCR is used to refer to methods in which amplification of the DNA target at the end of PCR is taken as the result (e.g. ddPCR, real time PCR, conventional PCR followed by gel electrophoresis). This is not a formal definition and the term can be used differently elsewhere. Further discussion of the sensitivity of PCR methods can be found in the complementary JNCC Report No. 669b covering Frequently Asked Questions (Jones *et al.* 2020).

### 2.1.6 Real time PCR and quantitative PCR (qPCR)

Real time PCR is a PCR method that detects the amplification of the DNA target in real time during the PCR cycles. This is done either by detecting fluorescent probes released during DNA amplification or by intercalating dyes that bind to any double-stranded DNA present in the reaction. The number of cycles at which exponential amplification starts (the cycling threshold value, ct value) can be determined and is strongly related to the starting number of copies of the DNA target. If a standard series of known copy number of targets is included in the reaction for comparison, then the starting quantity of DNA in the reaction can be determined. At this point, the method is called quantitative PCR (qPCR). However, for many applications, quantification is not determined, and standards are not included (i.e. the test is used to determine presence/absence of the target) and the method should be referred to as real time PCR. We do not use the abbreviation ‘rtPCT’ as it is ambiguous (it commonly refers to ‘reverse transcriptase’ PCR used to amplify RNA).

### 2.1.7 Droplet digital PCR (ddPCR)

ddPCR is a relatively new technique in which the target DNA is partitioned into several thousand individual droplets, which are then amplified by PCR. The amplification of the target DNA is then quantified for each droplet, and the overall number of positive and negative droplets quantified. The proportion of droplets in which amplification is detected can be used to quantify the number of starting copies of the target DNA.

### 2.1.8 High throughput sequencing

High throughput sequencing (HTS), also called next generation sequencing (NGS) or massively parallel sequencing, covers a range of methods and platforms that are capable of sequencing multiple DNA molecules in parallel, enabling hundreds of thousands or millions of DNA molecules to be sequenced at a time, from the same sample. The different methods and platforms produce differing data and have various advantages and disadvantages (see the complementary JNCC Report No. 669b covering Frequently Asked Questions (Jones *et al.* 2020).

### 2.1.9 Metabarcoding

Metabarcoding describes the process of generating barcode data from multiple different organisms from within the same sample using HTS methods. It uses broad-specificity primers that amplify a fragment of a gene of interest (typically a DNA barcode region) simultaneously from whole communities (e.g. bacterial rRNA encoding 16S). These primers can be designed to have varying degrees of specificity (e.g. to amplify only within a genus, or across a whole order).

## 2.2 Technical terms and concepts

### 2.2.1 Matrix

The matrix of a sample is the material the target DNA is contained within. Matrix types include soil, water, faeces, different tissue types, sediment, and whole organisms.

### 2.2.2 Method and test

A method can refer to a single step (e.g. DNA extraction methods, real time PCR methods) or to a combination of these steps in series. A test is the application of a specific method to detect a specific target (or targets) in a specific matrix (i.e. the detection of x in y using method z).

### 2.2.3 Positive control

A positive control is a ‘true positive’ – i.e. a sample known to contain the target (typically target DNA). These controls are included to identify false negatives (erroneous non-detection) and so contribute to monitoring that the method is performing as expected

Under certain circumstances (e.g. highly sensitive tests), the decision may be taken not to include a target DNA positive control because it may represent a contamination risk. An alternative is to include modified positive control DNA (or an endogenous control test; see 2.2.4, below), which shows that the test is functioning in the samples but with the benefit that any contamination from the positive control can be identified and excluded. Positive controls that contain the target DNA are not generally included during the sampling stage due to logistical and contamination constraints. Instead, degradation controls (see 2.2.4) or other

proxies to measure DNA degradation can be used to confirm the correct performance of this step in the test.

### 2.2.4 Degradation, inhibition and endogenous controls

Degradation of DNA after sampling or inhibition of DNA amplification during PCR may lead to false negative results. Control samples are used to check that these processes are not occurring. Degradation controls are used to monitor DNA degradation, typically during the sample collection, transportation and storage steps. They usually consist of a DNA marker added into samples at the sampling stage. The recovery of this DNA can confirm that degradation has not occurred.

Inhibition controls are used to monitor inhibition of the reaction during the PCR stage. They typically consist of DNA added into samples pre-PCR. Correct amplification of the inhibition control DNA confirms that samples are not excessively inhibited.

An endogenous control is typically an assay run in parallel to the main test to check performance of the method. For example, a species-specific test to detect an insect species in an assemblage may include an endogenous control assay that detects generic insect DNA. The endogenous control is run to show that the methodological steps up to the PCR stage have correctly produced high quality insect DNA.

### 2.2.5 Negative control/blank

A negative control, also referred to as a blank or a No Template Control (NTC) is a sample (or approximation of a sample) that does not contain the target DNA. If a negative control returns a positive result this is most commonly due to contamination.

Negative controls can be added at each methodological step to identify if and where contamination is occurring. For example, a water eDNA real time PCR test can include negatives at the sampling, filtering, extraction, and amplification/detection steps. Negative controls can be used at fewer steps (e.g. to reduce costs) but if contamination is present it may then be more difficult and costly to establish where in the process contamination has occurred. If a single negative control is used, it should be included at the earliest methodological step so that it can monitor contamination in the whole process. Negative controls are used to confirm the specificity of the method by demonstrating that the rate of false positive results is no higher than expected. They are used to reduce the false positive rate by identifying cases where contamination has occurred. **It is usually inappropriate to use a test or method without including negative controls, without convincing justification.**

For HTS, what constitutes an appropriate negative control may depend on the end-use of the results. If detecting rare species is important (or the data on the rare species will be interpreted as a true positive) then negative controls should contain little or no DNA (and certainly no DNA from the target species) to ensure that incidental contamination of the negative control with DNA from the rare species is detected. Note that using HTS data as a screen to identify potential positive results for rare species with a second better validated method to confirm the presence of the target circumvents the problem associated with appropriate negative controls for HTS in rare species detection (Fox *et al.* 2019; see also Section 2.2.15 on confirmatory testing).

Where the purpose of HTS is to build a community profile, it is not necessarily clear what constitutes an appropriate negative control. This is because negative control samples will typically contain some incidental contamination due to (i) the exponential amplification of a

PCR coupled with the difficulties of fully controlling contamination from DNA within a HTS workflow, and (ii) 'artefactual' contamination such as index hopping, over-clustering and chimeric products generated by primers binding to themselves during the PCR step.

Proposed 'negative controls' for community analysis include mock communities of species never to be encountered in the true sample or in the positive control; any contamination of the negative control with the species encountered in the true sample is then taken as an indication of genuine and problematic contamination. Another alternative proposed is true blanks, where the volume of amplicon spiked into the HTS run is not normalised (there is a normalisation step in most metabarcoding protocols so that approximately similar amounts of DNA go into the sequencing run – if this normalisation is applied to the negative control blank samples, an unrepresentatively large volume of blank is carried into the sequencing). The presence and read count of DNA within non-normalised blanks can be taken as an indicator of the potential threshold level of contamination.

### 2.2.6 Accuracy

**Accuracy** is the closeness of the observed value to the true value (e.g. closeness of the detection method results to the true number of presences or absences of the target organism in the environment sampled). In practice, the accuracy for environmental applications may be very difficult to determine as the true value (e.g. the number or presences or absences of the organism in the environment) may not be known.

### 2.2.7 Specificity

**Specificity** is the proportion of true negative samples that correctly produce a negative result. The inverse of specificity is the **false positive rate**, which is the proportion of samples that do not contain the target DNA but incorrectly give a positive result (called **false positive results**; Table 1). Tests with higher specificity have a corresponding lower false positive rate. **Within metabarcoding studies, the specificity will apply to each taxon separately**: a metabarcoding assay could therefore be highly specific for one taxon but have low specificity for another. Note that this definition of specificity covers the entirety of the test, beginning with sampling, and so includes sources of false positives such as contamination. Therefore, a test that is prone to higher levels of cross-contamination will have a lower specificity.

**Inclusive specificity** is the ability of the test to detect all targets, given the range of genetic variation expected within these targets. For example, a test for great crested newts should be able to detect individuals from all populations of great crested newts likely to be tested. Where a target species has a very large or under-sampled range, inclusive specificity may be difficult to demonstrate for the whole species, but it may be sufficient to show that all the populations likely to be tested are included. Inclusive specificity is generally demonstrated using samples of the target species taken to represent the available range of genetic variation, first *in silico* using reference sequences during primer design, then in the laboratory from DNA samples.

**Exclusive specificity** is the ability of the test not to return a positive result in the presence of non-target organisms and absence of the target. Exclusive specificity is typically demonstrated using a range of closely related organisms, and a range of organisms likely to be accidentally tested (or to be present in the test matrix). During the primer design stage of a project, exclusive specificity can be demonstrated *in silico*, but it must also be demonstrated on 'real' samples. An example of a test with poor exclusive specificity might be a great crested newt specific test that also returns a positive result in the presence of alpine newts.

Inclusive and exclusive specificity are typically considered to be attributes of PCR primers.

### 2.2.8 Sensitivity

The **sensitivity** of a method is the proportion of true positive samples that are correctly identified as positive. A true positive sample giving a negative result is known as a **false negative result** (the method incorrectly identified the target was absent from a sample when it was present). The sensitivity of the method determines the proportion of false negatives, called the **false negative rate**. Higher sensitivity therefore gives a corresponding lower false negative rate. **Within metabarcoding studies, the sensitivity will apply to each taxon separately**: a metabarcoding assay could therefore be highly sensitive for one taxon but have low sensitivity for another.

**Table 1:** Specificity and sensitivity in relation to positive and negative detections.

| | **Target is present** | **Target is absent** |
|---|---|---|
| **Test result positive** | True positive *(rate increased by high sensitivity and specificity)* | False positive *(rate increased by low specificity)* |
| **Test result negative** | False negative *(rate increased by low sensitivity)* | True negative *(rate increased by high sensitivity and specificity)* |

### 2.2.9 Limit of Detection and Limit of Quantification

In the laboratory, the Limit of Detection (LoD) is the smallest amount of DNA that has a high probability of giving a positive result if that amount of DNA is present in the sample analysed. This can also be referred to as 'analytical sensitivity'. The same concept can be applied to broader sampling and testing – e.g. the smallest amount of (e)DNA in the location sampled that has a high probability of giving a positive result.

Quantities of DNA below the LoD may still give valid positive results, but positive results become less likely as the quantity falls further below this limit. The LoD is also the highest quantity of DNA that can be present if a negative result has been observed (i.e. negative results are not usually obtained with DNA quantities above the LoD), and so it provides information necessary to interpret negative test results.

For methods that quantify the DNA in the sample (e.g. qPCR, ddPCR), the related Limit of Quantification (LoQ) should be determined. It is defined as the smallest amount of analyte (DNA) that can be measured and quantified with defined repeatability (precision) and accuracy under the experimental conditions by the method under validation. The LoQ can never be lower than the LoD.

### 2.2.10 Selectivity

**Selectivity** is the extent to which the variability of the matrix that the sample is found in (e.g. differences in soil type, water pH, *etc.*) affects the outcome of the test.

### 2.2.11 Repeatability

**Repeatability** is the size of the variation observed in results from the same test run on the same sample by the same operator under identical analytical conditions within a single run.

It is frequently assessed using samples just above the LoD, as this gives a measure of repeatability under more challenging conditions for the test.

The number of repeats required to determine the repeatability of a test (during methods validation) will vary according the type of test and the operational purpose. To validate tests that produce a simple quantitative response (e.g. ct values from real time PCR assays, read number for a specific taxon), for each step of the method a recommended minimum of six replicates should be used (including samples across the range of DNA concentrations). This can be six replicates of the same sample, or a number of samples repeated to a total of six replicates. Where the response is not continuous (e.g. detected / not detected) or compositional (e.g. community metabarcoding data), the number of required repeats is likely to be higher.

When the repeatability of the method is already known (e.g. using a previously validated method), occasional duplicate samples should be used to check that the repeatability remains as expected.

For metabarcoding, repeatability can be quantified on different metrics. For example, it could be quantified per taxon, for the community of taxa as a whole, for metrics related to the metabarcoding runs (e.g. read numbers per sample, QC statistics), or for a metric derived from the community statistics (e.g. a quality assessment such as the Trophic Diatom Index). The ultimate use of the metabarcoding data should guide how the repeatability is assessed and interpreted.

### 2.2.12 Reproducibility

**Reproducibility** is the size of the variation observed in results produced by the same test applied to the same samples under different conditions. For many tests, it is necessary to demonstrate reproducibility across different laboratories, although it can also be demonstrated within a single lab (e.g. by different users, machines at different times).

Where test reproducibility is demonstrated in a *ring test* (in which multiple laboratories receive the same samples), different approaches to determining reproducibility can be taken according to the test. For tests that produce a simple quantitative response (e.g. ct values from real time PCR assays, read number for a specific taxon), a commonly used approach is to send five samples in duplicate to each participating lab. These are normally sent ‘blind’ (i.e. the participating lab does not know which sample is in each tube) and use samples from across the range of DNA concentrations, including blanks. Where the response is not continuous (e.g. detected / not detected) or compositional (e.g. community metabarcoding data), the number of required samples is likely to be higher.

### 2.2.13 Scope (of a test)

The **scope** of a validated test describes the range of conditions the test has been validated for. At the broadest level, it defines the target taxa to be detected, the matrices it can be detected in, and the method that will be used. It can encompass variations within the sampling period (e.g. the test has been validated for summer sampling for target A) or environment, or variations in the laboratory process (e.g. it may define which instruments can be used). Where a test result is obtained within the validated scope the uncertainty around the results is known, but where a test is used out of scope of the validation that assurance no longer exists.

### 2.2.14 Validation

Conceptually, **the process of validation progressively removes the unquantified uncertainty associated with a test result**. The result from a well-validated test will have a known uncertainty attached to the positive and negative results (e.g. 'a positive will be correct 99% of the time, and a negative 95% of time for samples at or above the LoD'), allowing a clear interpretation of the results.

A technical definition of validation is "*the process that demonstrates that a test consistently produces a result to within given performance parameters when the test is applied within its defined scope*". For the purposes of this guidance note, the minimum set of parameters that should be assessed are **sensitivity, specificity, repeatability,** and **Limit of Detection**, and may also need to include **selectivity** and **reproducibility**. Typically, these parameters can be assessed in the laboratory. In addition, the **accuracy** of the test (i.e. how the test result relates to the true distribution of the organism) needs to be assessed – this generally requires field sampling, although that may not always be possible if the target organism is not yet present in the environment.

The validation process has been described for targeted DNA-based tests in some areas (e.g. for GM – Marchezi *et al.* 2015; plant health – EPPO PM7/98; and aquatic eDNA – DNAqua-Net Working Group https://edna-validation.com/). However, the term 'validation' can be used in different ways in the scientific literature. It is important to understand what is meant by validation where it is applied, and whether it is a formal definition.

The validation framework described in Marchezi *et al.* (2015), EPPO PM7/98 and Council Directive 96/23/EC are similar and use similar descriptions of analytical parameters (e.g. accuracy, sensitivity). The AquaNet guidelines do not use all of these terms (although they use some of the concepts), potentially making it more user-friendly to non-technical audiences. In order to harmonise with other areas, we have followed the other sectors.

Note that a *test* can be validated ("the detection of x in y using method z") but not a method.

**Also note that the process of validation does not make a test fit for purpose; it provides estimates of performance parameters that can be used to tell whether a method is, or is not, fit for particular purposes.**

### 2.2.15 Confirmatory testing

If the result is of high importance or the test is not considered sufficiently reliable (e.g. low or unquantified specificity), a positive result may be confirmed using an additional method, often referred to as **confirmatory testing**. In principle, a confirmatory test should use a different biological attribute (e.g. morphological identification, an antibody-based test) from the original test) on the basis that the more dissimilar the confirmatory test is from the initial test, the more reliable it will be at excluding other reasons for a positive result. If such a distinct test is not possible, the confirmatory test should ideally target a different part of the genome, to exclude the possibilities that the chosen barcode region is not adequately discriminatory or that there is laboratory contamination (e.g. with the PCR amplicon of the specific region). DNA-based confirmatory testing should be undertaken with re-extracted DNA from the original sample if possible, or from freshly taken samples. Other troubleshooting should also check for evidence of field or laboratory contamination (e.g. checking for unexpected results within a single set of field samples, re-examining negative controls).

The outcome of confirmatory testing is likely to be interpreted on a case-by-case basis. In some scenarios a confirmatory test may exclude the initial finding, but in others (e.g. an initial positive result from a well-validated test in the absence of any contamination, followed by a negative result in the confirmatory test) the initial positive should not be ignored. **For standardised methods, the confirmatory testing and interpretation of results should be decided in advance** and written into Standard Operating Procedures, and the overall sensitivity and limit of detection of the combined tests should be assessed.

### 2.2.16 Positive and negative predictive values

These are more technical concepts but are important to understand when interpreting a test result. The positive predictive value is the proportion of positive results from real world samples that are true positives. A test with a positive predictive value of 0.1 means that 9 out of 10 positive results were incorrect (i.e. they were false positives). The negative predictive value is the proportion of negative results from real world samples that are true negatives.

The positive and negative predictive values are based on the incidence of the target in the environment; false positive and false negative rates are intrinsic to a specific method, but the real world incidence of false positives and false negatives in the test results depends on the prevalence of the target in the test samples. Even highly sensitive tests (e.g. sensitivity of up to 99%) will return more false positives than true positives when prevalence of the target is low (e.g. 1% or lower). This is a key consideration for surveys where the target species is expected to be either absent or at very low abundance.

**The basic rule of thumb for assessing the fitness-for-purpose of test methods is that the false positive rate should be much lower than the proportion of positives expected (or required) to be detected. Similarly, the false negative rate should be much lower than the expected proportion of true negative samples.** If these conditions are met, individual tests will have good predictive power.

For example, assume a diagnostic test for very rare species in the UK has been developed with a specificity of 99% (i.e. a false positive rate of 1%). The true prevalence of the species in the environment is 0.1%. 1000 tests are run, which return 11 positive results. Of these, ten are likely to be false positives (they are actually negative) and one is likely to be a true positive. In this example, positive results should be presumed false with additional testing/troubleshooting require to check whether they are genuine. The converse is also true: where prevalence is very high (>99%), a test with a seemingly high sensitivity of 99% (i.e. the false negative rate is 1%) will return more false negatives than true negatives. Figure 1 shows the relationship between the incidence of the target in the environmental samples and predictive values where sensitivity and specificity are 99%.

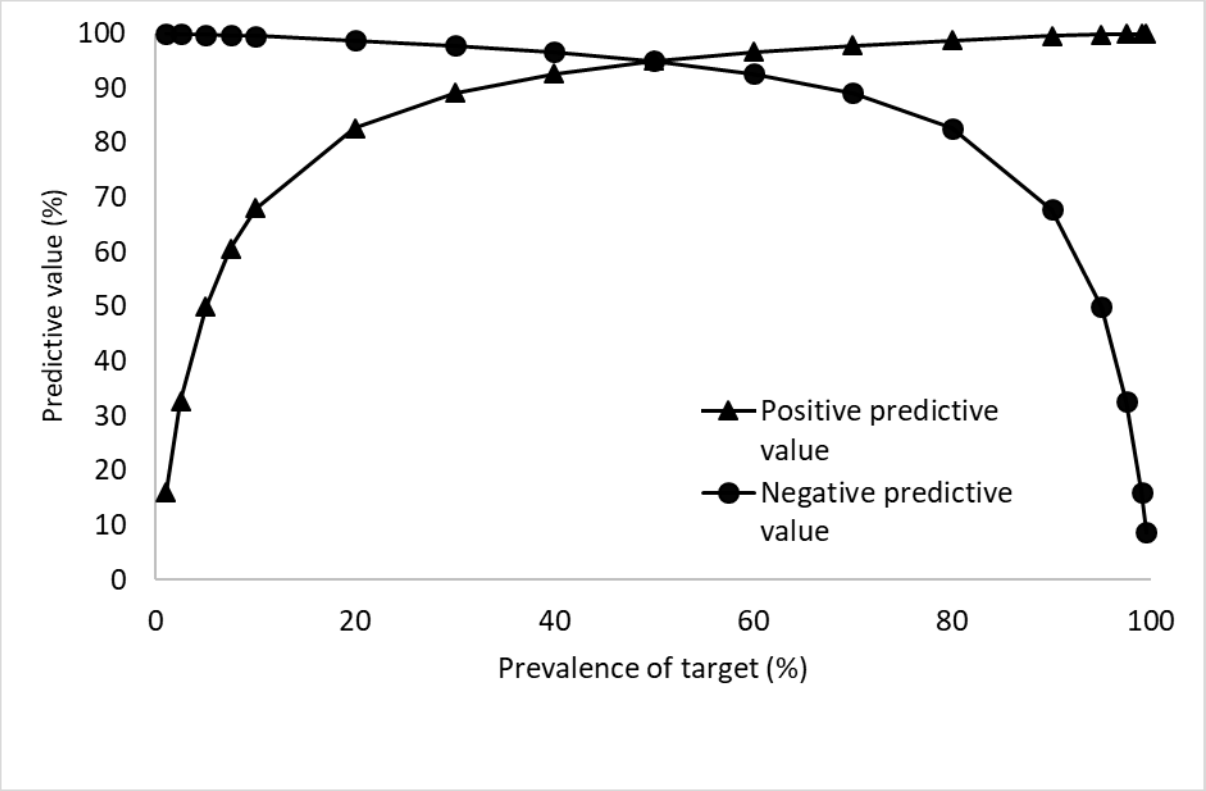


**Figure 1:** Positive and negative predictive values of a test with a sensitivity and specificity of 99%, at different levels of prevalence of the target.

### 2.2.17 Inter-laboratory studies: methods to ensure equivalence of results

Exchanging samples between labs can be used to check the equivalence or robustness of methods, or the equivalence of results. The most common terms and approaches to describe these are outlined below.

**Test Performance Studies/Collaborative Trials**
Studies designed to assess the performance of a method or test. Various terms (collaborative trials, test performance studies, ring tests, method performance tests) are used in different disciplines, with variations in what is assessed with the study/ trial.

Within a typical analytical *test performance study*, all samples and reagents are distributed from a central laboratory, and each receiving lab performs the same test under the same conditions using the same samples and reagents. Test performance studies are done as part of method validation and produce estimates of the performance of the test under defined conditions. However, this approach often does not reflect how the method will be used in 'real world' conditions. A better approach for the purposes of the Environment sector would be less controlled test performance studies where each laboratory uses the same method, but with local conditions and reagents.

**Proficiency Test**
*Proficiency testing* assesses the ability of a laboratory to competently perform a particular test. In a proficiency test, samples are circulated to all laboratories from a central competent lab. These can then be tested by the receiving labs using any method (although the methods are often similar – e.g. all might use real time PCR) and the equivalence of the results can be compared. Participating laboratories can use this to ensure they are achieving expected results, and take remedial action as required.

**Proficiency testing alone does not ensure that laboratories meet minimum requirements** as there is generally no requirement to report the result, nor is external action taken in response to a poor outcome. The only check a proficiency test provides in isolation is that end users can request the results and choose not to use a laboratory that has performed poorly in the test. Proficiency testing as part of an accreditation scheme provides a better route to control quality. If the laboratory is certified to a particular quality assurance accreditation (e.g. ISO17025), the accreditation body (UKAS in the UK) will check the results of the proficiency test during the auditing process and may ultimately withdraw accreditation if proficiency testing results are consistently poor.

# 3 Pre-commissioning guidance

This guidance aims to provide a (non-exhaustive) checklist of factors to consider (i) when deciding whether to proceed with a project (a form of project risk assessment), (ii) during the project design, and (iii) if developing an invitation to tender and evaluating responses.

Although some factors are difficult to assess, these should be considered as far as possible, and it may be advisable to seek additional technical advice to support this. Tender documents can also request that providers consider some of the factors, for example by requesting a justification of key aspects the methods and a description of what validation will be undertaken on the method (e.g. if and how sensitivity will be determined).

Taken in isolation, very few of the factors listed would demonstrate a project should not proceed. However, some factors may make a project at high risk of failure (e.g. non-availability of test materials). Note that the questions will not be universally relevant given the guidance is to cover a range of project types, from exploratory pilot projects through to finalising methods for deployment. These questions are partly based on the ‘Risk analysis before performing validation’ checklist contained in the document EPPO PM7/98, with additional points specific to environmental samples.

A flow chart summarising the steps is given in Figure 2 – this can be used to inform the project development process.

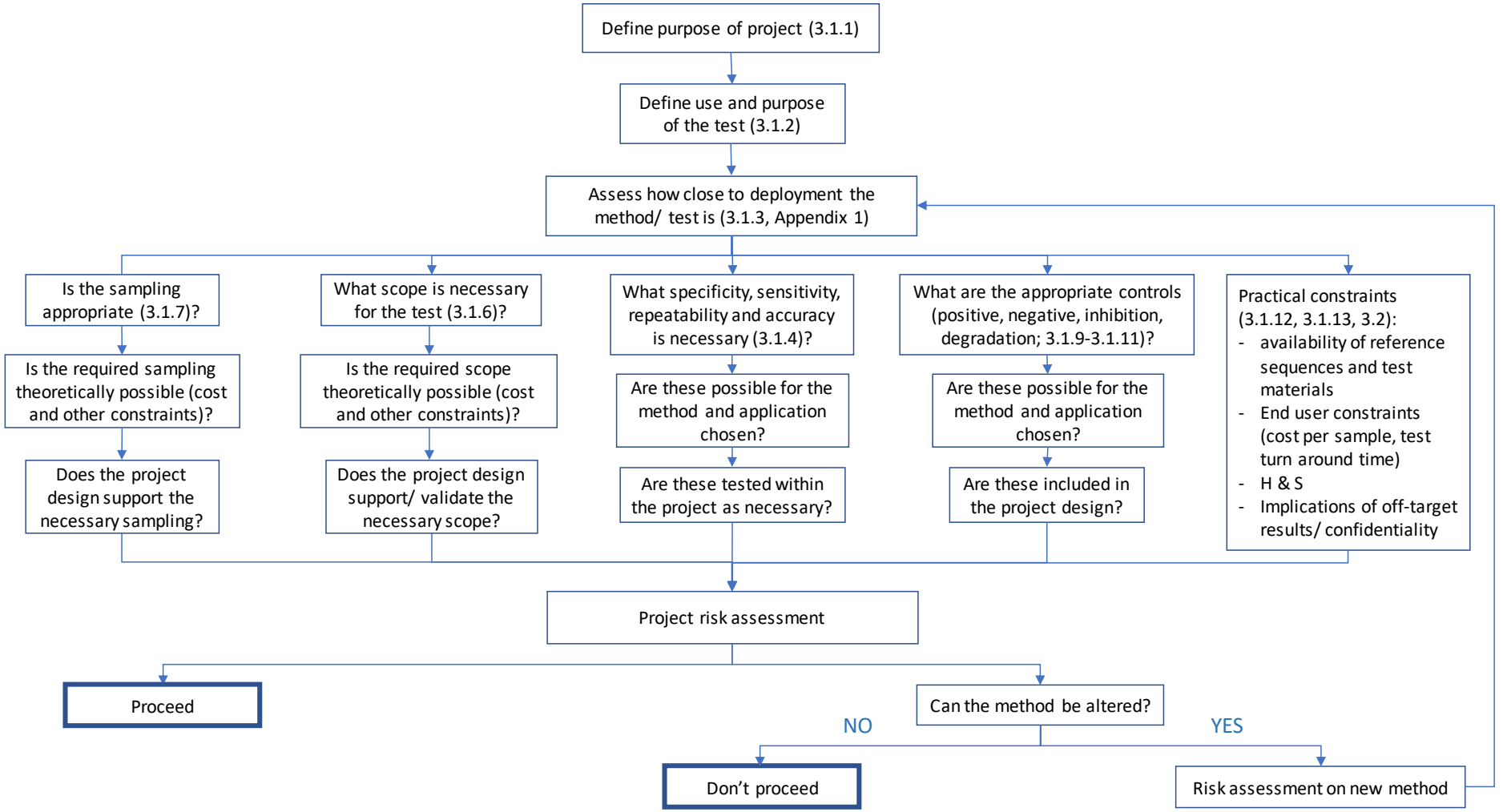


**Figure 2:** Flow diagram incorporating key stages and questions in decision-making process for DNA-based methods.

## 3.1 End user requirements

### 3.1.1 What is the purpose of the project?

Broadly, projects intended to run once (e.g. a project to test for differences between two or more habitats) can be considered research projects, while projects to develop a test for routine use that produces comparable results when performed at different times can be considered development projects.

Raw data produced by different research projects are only directly comparable if there is sufficient information about method performance (e.g. **scope**, **false positive** **and** **false negative** **rates**) and method deployment (e.g. the number and location of sites sampled, the number of samples per location, quantity of material used during the test). If this information is available, results from each project can be interpreted and compared by taking into account the method performances and method deployments. For example, if different methods were used to identify the presence of species at the same site, and for both methods the confidence in the positive results was high but for one or both methods the confidence in the negative results was low or unquantified (i.e. the non-detection of a species does not mean it is absent), it would be appropriate to compile an aggregated list of species presences across the methods, but it would not be appropriate to combine or compare the absence data. This kind of comparison will be facilitated by common standards of reporting.

If a project aims to extend the range of areas surveyed for a particular organism or community and produce data comparable with previous results, this should only be done as part of a 'method development' project with calibration against previous methods - otherwise it is more difficult to determine whether findings are genuine or a consequence of methodological changes.

Many of the sections of the guidance are more directly applicable to development projects (e.g. sections 3.1.2 and 3.1.3).

### 3.1.2 What is the intended use and consequence of the method/ test?

This question is potentially more relevant for projects developing a method or test than to research-focussed projects.

End users should define the purpose of the test and the potential impact of the test results. These factors inform how high the confidence needs to be in the test results and the extent to which unquantified uncertainty is acceptable. Example test purposes include:

- Confirmatory in support of other evidence
- Screening to identify samples/locations for further testing
- Standalone test that will be the sole basis for actions or reporting

Where the test result is used in isolation (i.e. there is no other evidence to support or contradict the result), greater confidence in the result is required

Examples of results that have high potential impact could be the first notifiable presence of a high-risk invasive species or a result that demonstrates a change in status of a protected site, while a lower impact result might be the routine survey of a widely distributed species. In the first two cases, each test result must have a high confidence of being accurate: a low false negative rate (i.e. erroneous non-detection of invasive or protected species) to mitigate the high risk of impacts from the invasive or from failing to adequately protect a site, and a

low false positive rate (i.e. erroneous detection of an invasive or a protected species) to reduce the cost of incorrectly taking action. Conversely, in the case of routine surveys the results from *individual tests* can have lower confidence provided there is sufficient accuracy across samples to detect changes and differences (e.g. spatial or temporal trends).

The combination of the purpose of the test and the impact of the result inform how precise estimates of false positive and false negative rates need to be, to support acceptable positive and negative predictive values (2.2.16). This then defines the amount of work needed to validate the method adequately and guide how positive and negative controls should be applied.

**A minimum level of validation is usually required** (e.g. repeatability, sensitivity, specificity demonstrated under laboratory conditions, some steps towards determining the accuracy). It can be acceptable to have more unquantified uncertainty where the test result will be of lower impact, and/ or where it is one piece of evidence in an overall assessment (e.g. other sources of confirmatory evidence will be used).

### 3.1.3 How far is the test along the pathway to deployment?

This question applies to projects intending to take a method or test and move it towards routine or repeated deployment. It is less applicable to one-off research projects.

A common framework for assessing the maturity of different technologies was developed by NASA in the 1970s and has been adopted widely in many disciplines. This Technology Readiness Level (TRL) tool allows non-technical users to understand how close a technology or method is to deployment, allowing more informed decisions on funding and commissioning projects. A version more appropriate to analytical methods was adapted by the EU to understand research carried out under the Horizon 2020 Research and Innovation programme (see Figure 3 for a modified version of the EU TRL framework). Although the framework is commonly presented as a linear process for clarity, in practice there are many feedback loops between steps. Using the TRL framework to assess progress towards deployment can be valuable because it is applicable across most disciplines, it gives non-technical users a good understanding of the method maturity, and it has built-in expectations about the degree of unquantified uncertainty that is acceptable at different stages.

DNAqua-Net have also developed a framework that can be used to understand the level of validation applied to end point detection PCR assays for use on eDNA samples (https://www.biorxiv.org/content/10.1101/2020.04.27.063990v3). The stages in this framework (numbered 1-5) do not fully map onto the TRL scales because they deal with validation rather than deployment (such that the method could be validated to stage 5 but no assessment has been made that it is fit for purpose for deployment). However, the stages approximately map across to TRLs 5 to 7/8.

To briefly describe the TRL steps:

- Steps 1-2 (invention). The general principle of the concept has been formulated.
- Steps 3-4 (concept demonstration). Experimental proof of the method is generated in the laboratory and is being applied to a particular use (e.g. detecting organism A in water). End users have been consulted on requirements. The different possible platforms (e.g. real time PCR, ddPCR) and methods (e.g. water filtration or precipitation) are being evaluated. At the end of step 4, the method and platform have been decided (e.g. organism A will be detected by real time PCR from ethanol precipitated samples). At this point, a decision is

made on whether to carry the application of the method forward into validation and field testing.

- Steps 5 to 8. The test is evaluated in the laboratory, and in the field by a range of users, including the end users (e.g. field ecologists). Step 8 is operational validation, involving piloting in active use by the end users. At the end of step 8, the test has been shown to work during small scale deployment and can (theoretically) be scaled up to full delivery.
- The decision to proceed from 8 to 9 is often a business/ policy decision.

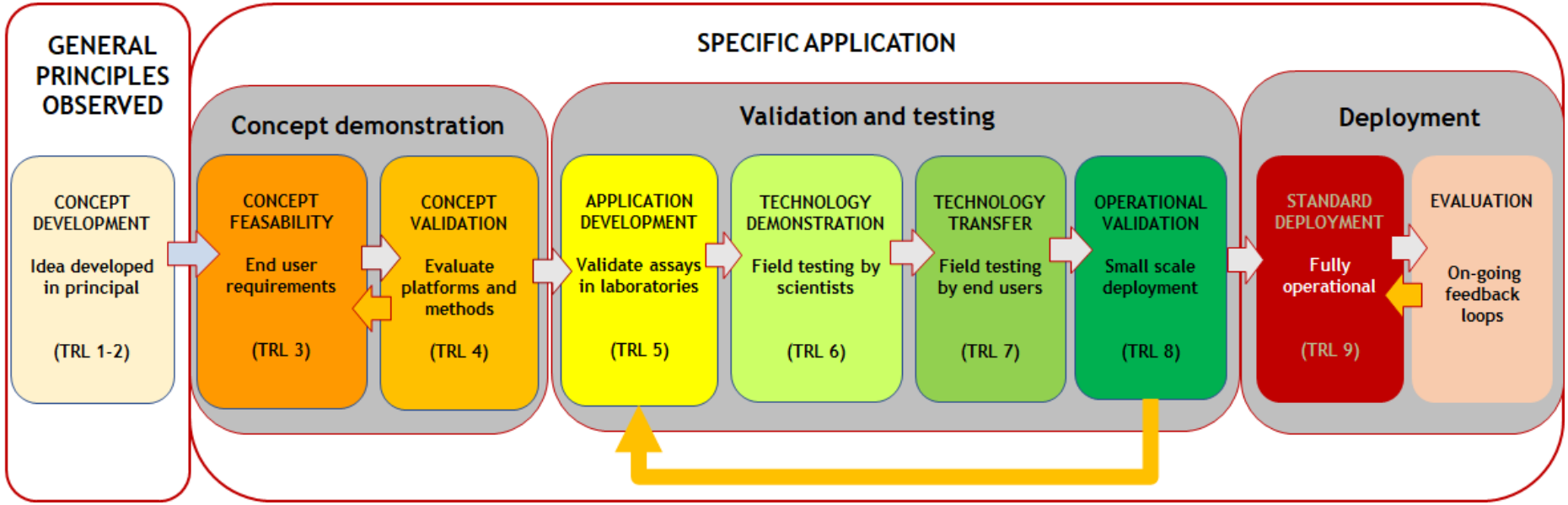


**Figure 3.** Visualisation of the Technology Readiness Levels, adapted to molecular detection tests for monitoring the environment. Levels are clustered by broader concepts (grey boxes). Commonly encountered feedback loops are shown as additional yellow arrows.

See https://www.nasa.gov/directorates/heo/scan/engineering/technology/txt_accordion1.html for NASA version and https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf for EU version.

During the early development stages (TRL3-5) the potential and limitations of the methods are being explored, therefore unquantified uncertainty is more acceptable, and validation is of lesser importance. For later stages (TRL6-9), unquantified uncertainty becomes less acceptable and the project(s) will increasingly be focused on validation (in the broad sense), logistics and policy. For many DNA-based applications, much of the proof of concept and similar in-field applications will have been demonstrated and so the starting point will often be TRL3-5

**Projects can span more than one stage, but it is highly unlikely that a single project will successfully deliver a test that is finalised and fit for deployment.**

### 3.1.4 What sensitivity, specificity, accuracy and repeatability are required?

Defining the required level of sensitivity, specificity, accuracy and repeatability in advance provides a benchmark against which to define the success of the test, and can be used as a target for laboratory optimisation. In practical terms, this definition could include e.g. at what population threshold the test should reliably detect a target organism (or what amount of DNA must be detectable within a sample), what the standard deviation of the test results can be on the same sample, and whether the level of misidentification of certain non-target taxa is acceptable.

**This assessment should help to guide the selection of methods to be used, and is therefore expected to be an activity at the lower stages of method development (TRL4-5)**

An informal assessment of whether the test can achieve the required performance (assessed by the positive and negative predictive values) can be made where the chosen method and matrix are well understood (i.e. the sensitivity and specificity can be estimated) and the prevalence of the target in the environment is understood. This can be used to refine the choice of method (e.g. selecting real time PCR or ddPCR if sensitivity is of utmost importance) or to indicate that the project has a higher risk of failure. Appendix 1 provides a preliminary modelling framework developed in this project to formally assess the fitness for purpose of a test method.

### 3.1.5 Is there a current survey test to verify a DNA-based method against?

In some cases, existing tests can be used to determine the presence of the target organism in the environment and verify the performance of the novel DNA-based test. This potentially avoids the need for validation based on the assessment of each methodological step, with the caveat that the existing test should itself have been validated to an acceptable level.

Validating against an existing test involves demonstrating that the new test returns sufficiently similar results to the existing test to be acceptable. This should be done by paired sampling, with each sample run for both tests, across the full scope. The repeatability of the new test should also be demonstrated. The availability of an existing test to survey for the target organism can be built into the project design.

### 3.1.6 What scope of test is necessary?

The scope of the test should define the minimum and desirable range of environmental and temporal conditions the test should be effective for. For example, the scope of a test might be to describe invertebrate assemblages in lowland arable soils in summer, or it might be to describe invertebrate assemblages across all UK soil types at any time of year. The wider the range of conditions, the greater the amount of validation required (the test performance

should be assessed for all conditions, but this can be prohibitively expensive). Project development should consider the required scope of the test, and whether tender proposals will be able to assess performance across this scope. For example, validation may consist of testing a reduced set of conditions exploring the limits of the method.

At earlier development stages the scope of the test can be left loosely defined, or a subset of the total scope can be assessed. At the later development stages, the full range of the scope should be accounted for.

### 3.1.7 Is the sampling appropriate?

Sampling should be representative of the distribution of the target organism(s) in the environment (or its eDNA). **Key factors to consider are heterogeneous distributions and rarity**. In principle, heterogeneity in the distribution of DNA (which may have contributions from low prevalence of organisms and variation in the density of organisms) is mitigated by increasing the number of primary samples taken; a low *average* quantity of DNA across primary samples is mitigated by increasing the total amount of test portion from which DNA is extracted and/or applying a more sensitive DNA test method (Appendix 1).Where the sensitivity of the method is well characterised, it should be possible to infer the extent of sampling required to detect an organism at a given abundance threshold (Appendix 1).

At the early stages of method development (TRL3-5), a project may well focus on determining an appropriate sampling strategy. By the later development stages (TRL 7-8) the uncertainties around sampling regimes must be appropriately quantified and the sampling strategy well established. How much uncertainty is acceptable, and what the threshold is for detection, are defined by the end user based on the intended use and consequence of the method/test (Section 3.1.2). Also see the complementary JNCC Report No. 669b on frequently asked questions (FAQ8) (Jones *et al*. 2020).

### 3.1.8 How will repeatability, sensitivity, specificity and accuracy be addressed?

The use and impact of the results (3.1.2), and how far along the pathway to deployment the method or test is (3.1.3) informs how much emphasis should be placed on determining repeatability, sensitivity, specificity and accuracy. Invitation to tender documents should make it clear if including these factors is desirable/required, and responses should be clear on how they will be addressed.

Projects should include sufficient numbers of repeats to assess how much results can be expected to vary, as well as positive and negative controls. For projects with higher per sample cost (e.g. metabarcoding) it can be attractive to maximise the number of sites sampled at the expense of controls and repeats. However, this is likely to be a false economy if the validity of results or the test cannot be assessed at the end of the project.

### 3.1.9 Is it likely that contamination can be adequately controlled and monitored?

The potential for contamination should be considered in all experimental design, but its effect will depend on the methods used and outcomes intended. Metabarcoding is particularly sensitive to contamination because the target primers are designed to amplify from a wide range of different organisms and because of the need for multiple post amplification processing steps. For this reason, it is important to design a range of negative controls to monitor for contamination. End point PCR visualised by gel electrophoresis also suffers from the risks of amplicon contamination, and therefore needs carefully designed controls.

The incidence and expected levels of the target can also influence the risks of contamination. For low incidence targets a positive result may be more likely to be a false positive due to contamination (see 2.2.16: Positive and Negative Predictive Values) than a true positive. Where the target DNA is likely to be present in large amounts, this can increase the possibility of cross-contamination, both during field sampling and within the laboratory.

The risks posed by contamination and how this can be controlled or monitored need to be part of the risk assessment for the project.

### 3.1.10 Can the correct performance of the method or test be monitored?

The correct performance of a methodological step is monitored by including positive controls (Section 2.2.3), endogenous controls or tests, degradation controls and inhibition controls (Section 2.2.4). It can also be monitored by inspecting the results during each metabarcoding step. The two principal reasons to monitor the performance of a method or test are (i) to help troubleshoot when results are not as expected, and (ii) to discriminate false negatives (where the method has failed) from true negatives (the target is absent). For example, the correct functioning of a real time PCR can be monitored by including a positive control. If the positive control has not amplified, there is likely to be something amiss with the PCR. If there was no positive control, the samples that failed to amplify at the PCR stage would be interpreted as a negative result. Including a positive control therefore acts to reduce the false negative rate.

**All tests and projects where the detection of PCR products is deemed a positive and no amplification is a negative should include positive controls (and degradation / inhibition / endogenous controls as appropriate)** - e.g. real time PCR, conventional PCR, ddPCR. These controls should cover the extraction and PCR/detection steps and should be run regardless of the stage of development of the method or test (i.e. at all TRLs).

**For metabarcoding, it is highly desirable to include a positive control** (typically a synthetic positive) and a sequencer performance control such as PhiX. Again, these should be included at all stages of development of the method or test (all TRLs). It is good practice to include a positive control at the PCR stage when undertaking Sanger sequencing. When assessing a tender proposal, consider whether the described method includes sufficient controls to ensure the method works correctly.

### 3.1.11 Can the accuracy of the method be assessed?

When the test is in the laboratory stages of development (i.e. it is not being used in the field), accuracy is effectively the sensitivity and specificity. It should be possible to assess the accuracy of the overall test – i.e. whether the test correctly determines the presence/absence of the target organism in the environment. For targeted assays, the ideal scenario is to have known true positive and true negative samples (e.g. sites where the organism is known to be present at a range of abundances, and sites where it is known to be absent).

For metabarcoding, measuring accuracy is more complicated because a sample may contain multiple targets at different proportions, and frequently many of the targets have never been previously quantified. One alternative is to assess test accuracy using a side-by-side comparison with results from an existing detection method – the reliability and usefulness of this comparison will depend on the repeatability, sensitivity and accuracy of the existing method and the number of samples in the comparison.

In some cases, the accuracy of the test cannot be directly assessed in the field (e.g. invasive species not yet present in the environment), but an approximation of accuracy can be established using mock positives. These are sometimes equivalent to true positives (e.g. spiking in a target organism to a field sample of aggregated invertebrates) but in other cases it may be extremely difficult to find a similar mock positive (e.g. potentially eDNA results for an invasive species). Mesocosm results can be unrepresentative of environmental results because the concentration of DNA can be widely different between the two.

Where accuracy cannot be assessed, it is still possible to optimise and validate other aspects of the test (e.g. sensitivity, specificity, repeatability for a given quantity of DNA), but the practical performance will remain uncertain and interpretation of results needs to take this into account. Modelling can also be used to estimate the degree of confidence that can be placed in a negative result if there are reasonable *a priori* estimates of the presence of the target organism.

Demonstrating method accuracy in the earlier stages of development (e.g. TRL 4-6) may not be necessary, although gaining a rough idea would be desirable. Accuracy should be more fully determined by the later TRL stages.

### 3.1.12 Are there issues related to data confidentiality?

Consideration needs to be given to the types of data collected during the project and any unintended consequences of the use or release of these data. The first issue revolves around the use and collection of personal data. Data are considered personal if they relate to a person who can be identified from the data or from other data held by the data holder or addresses of people involved in any part of the process. This could include data from samples taken on private land identifiable by GPS co-ordinates. GDPR and other legal requirements must be met, and data properly anonymised prior to release.

A second issue is the sensitivity of location data for rare, protected or commonly persecuted species. From the early stage of the project, consideration should be given to how to keep these data confidential where necessary and prevent accidental release. Data confidentiality should be taken into consideration for any stage of method development but will only apply once the methods are being tested on real samples.

### 3.1.13 Is there a possibility of detecting notifiable organisms?

Environmental samples can contain a range of organisms (possible pests, pathogens, diseases, rare species) that are: "*subject to statutory control*" (Plant Health); "*notifiable*" (Animal Health) or "*protected*" (Wildlife and Countryside Act 1981). **Where non-targeted testing is used, it is important to consider how to respond to the potential detection of such organisms, as their presence can have serious consequences for the owner of the land on which they are detected.**

For example, a potato growing landowner allows a soil sample to be collected for soil metabarcoding, but the results indicate the potential presence of *Synchytrium endobioticum*, a fungus causing potato wart disease. The presence of potato wart on agricultural land leads to a potentially decades-long ban on the growing of potatoes in an area. There may be a statutory obligation to report even the suspicion of the presence of specific pathogens or potential novel pathogens. It is important to discuss or be ready to discuss with the relevant authorities any findings, and to know what the degree of uncertainty in those findings is and how to communicate them.

The detection of notifiable organisms should be taken into consideration for all stages of method development but will only apply once the methods are being tested on real samples.

## 3.2 Practical constraints of the project

### 3.2.1 Are the necessary DNA sequences available?

To design a PCR-based test, it is strictly necessary to have a reference sequence for that species for the selected barcode or other gene region. It is highly desirable to have a range of sequences from across the geographic range of the species, to test inclusive specificity (2.2.7). To test exclusive specificity, it is highly desirable to use samples from all taxa closely related to the target organism that are likely to be present in the environment. This is used to confirm that there is no cross-amplification of non-target species and is essential for high impact projects nearing final application.

The availability of DNA sequences is equally relevant for many metabarcoding studies. For these studies, there needs to be a suitable database of relevant DNA sequences or plans to create one. The database should contain multiple reference sequences from the target species, along with similar sequences of related species that may be encountered in samples. It is also important to confirm that the region of the chosen barcode can provide the level of taxonomic discrimination required for the intended use. Very few (if any) barcodes will distinguish all taxa to species.

### 3.2.2 Are all the necessary biological reference materials available?

Just as DNA sequences are required to design PCR based assays, appropriate biological reference materials are critical for validating an assay to test the practical sensitivity and inclusive and exclusive specificity.

To assess sensitivity, DNA of the target organism can be extracted and run as a dilution series. To assess the inclusive specificity, reference materials should be from individuals across the range of the species (or taxon of interest) to ensure that the assay works in all relevant populations. For a UK-only test involving the commonly used barcode regions (e.g. COI) it is unlikely that the within-species variability will be very high, and the inclusive specificity can be tested on a few individuals (or even one). To assess exclusive specificity, it is important to have biological samples of all related species that may be encountered, and to test a range of other species likely to be present in real samples.

This will apply during the earlier stages of method development (e.g. TRL5-6).

### 3.2.3 Logistical constraints

A range of logistical constraints may affect the project. These should be first considered at the early stages of method development (TRL3-5), and whilst they do not have to be fully resolved at this stage, they should not be insurmountable. By the later TRL stages (7-9), the logistical problems should be resolved. Commonly encountered difficulties include:

- Transporting preservatives and decontamination products to the field.
- Laboratory sampling methods not translating easily to the field, and causing delays (e.g. filtration of water, transfer of samples into preservative).
- Seasonality and adverse weather affecting sample collection.
- Inability to identify sites that are true positives and true negatives to test the method.

## 3.3 Risks more specific to routine monitoring

These considerations are most relevant for development projects that are at the higher TRL stages, but are worth considering at earlier stages of development:

- There should be appropriate laboratory capacity (or plans to create capacity) to deliver the number of tests required if the method moves into routine deployment. There should be sufficient field staff to collect the samples in the required time window.
- The acceptable per sample cost for the application should be determined by the end user (e.g. the new test must be the same cost or lower than an existing test), and the method proposed should be able to meet that cost requirement.
- The proposed test should provide sufficient benefits (e.g. cost, ease of use, accuracy, utility of results) over the existing test or over taking no action for it to have a reasonable likelihood of adoption.
- The turnaround time from sample submission to receiving results should be estimated for the new method and should fulfil the end user requirements.

# 4 Generic guidance to assess a project based on DNA methods

End users also need to evaluate project reports and other outputs (e.g. scientific publications). This applies both to outputs that are the direct result of commissioning by the end user, and to outputs produced by others that may be relevant. This section aims to aid this evaluation. Note that the guidance is based on the assessment of the high-level performance of the method and not the technical details.

## 4.1 Assessment of results

### 4.1.1 Did the project achieve its goals?

To answer this question, the project needs to have a defined aim. For one-off research projects (e.g. comparing two or more habitats), the aim will often be self-evident.

For projects to develop a test for routine use that produces comparable results when performed at different times (development projects) the aim of a specific project can be harder to define. In such cases, the level of validation and performance of the test can be assessed as appropriate for the TRL stage, the intended use of the methods, and the impact of the result. The TRL pathway can also be used to determine if the project has progressed the method along the TRL stages or whether it is necessary to go back to a previous stage and refine/change an aspect of the method (see 4.1.2).

### 4.1.2 Has the test progressed on the pathway to deployment?

This question applies to projects that intend to move a method or test towards or into routine or repeated deployment. It is less applicable to one-off research projects. It is informative to assess how far along the pathway to deployment a method is, and to understand how much further the project has progressed the method/test or whether it is necessary to return to a previous stage in the pathway.

A modified version of a widely used framework to assess how close a method is to deployment (the Technology Readiness Framework; TRL) is given in Figure 3, discussed in Section 3.1.3. For example, the project may have been designed to move the test from use by scientists to use by end users (moving from TRL 6 to 7). However, the project determined

that the complexity of the steps within the method are unsuited and too prone to error. The next project would therefore have to return to TRL 4 or 5 to develop and validate an alternative method.

### 4.1.3 Is the uncertainty acceptable for the intended use of the method/ test?

The end user should have defined the purpose of the test and the potential impact of the results; both of these factors inform the degree of confidence needed in the test results and the extent to which unquantified uncertainty is acceptable. Examples of the purpose of a test are: a confirmatory test in support of other evidence, a screening test to identify samples/locations for further testing, a standalone test. Where the test result is used in isolation, there must be less uncertainty about the result.

Results with a high potential impact might be those that identify the first notifiable presence of a high risk invasive species or results that demonstrate the need for a change in status of a protected site, while individual test results of lesser impact might be the routine survey of a widely distributed species. In the first two cases, each test result must provide a high confidence of being accurate, while in the latter case the individual results can have lower confidence provided the results are generally accurate (e.g. trends over time across many sites). **The combination of these two factors (purpose of the test, impact of the result) inform decisions about how high the confidence needs to be in the test results, which in turn informs the acceptable level of the positive and negative predictive values, the depth of validation required and the level of acceptable uncertainty.**

A minimum level of validation is usually required (e.g. some indication of the sensitivity, specificity and repeatability). It is difficult to envision a project where these are not important considerations. However, where the method is still in development (i.e. it is at the lower TRLs), lower levels of validation and associated unquantified uncertainty are generally acceptable provided they will be addressed in later studies.

### 4.1.4 Repeatability, reproducibility, sensitivity and specificity

As discussed in 4.1.1-4.1.3, the level of validation of a test and the acceptable performance depends on how far the test is along the pathway to deployment, the intended use of the methods, and the impact of the result. With this in mind, end users evaluating or interpreting project reports should consider the following to an appropriate degree:

- **Repeatability**: Check that repeatability of the method was assessed, and that the variation measured (typically expressed as the standard deviation) was acceptable.
- **Reproducibility**: If the method is to be transferred between labs, there should be some demonstration of inter-laboratory reproducibility, and the variation measured should be acceptable.
- **Sensitivity**: Check that the sensitivity (can be expressed as the false negative rate) has been assessed if necessary and is acceptable. For later stages of method development, the limit of detection (“analytical sensitivity”) should also have been determined and be acceptable.
- **Inclusive and exclusive specificity:** check that the assay has been demonstrated *not* to react in the presence of relevant non-target species (the exclusive specificity), and that a sufficient range of individuals from different populations of the target species have been included to demonstrate the inclusive specificity. If these weren’t included, consider whether these were likely to be important to the application. Depending on the stage of method development, the specificity should have been tested *in silico* or *in vivo*. If specificity was only demonstrated *in silico*, further work will need to demonstrate specificity *in vivo*.

### 4.1.5 Accuracy

Consider how the accuracy of the method was assessed and whether that was fit for the project aims. **Key considerations are whether the study used known positive and negative sampling locations or used spiked samples**. Known positives and negatives should ideally cover the range of target DNA concentrations (e.g. sites with low abundance of the target organism, sites with high abundance, many negative sites). If spiked samples were used, consider whether spiking with DNA was at a sufficiently early stage in the method to assess test effectiveness and whether the amount of DNA spiked in seems a good proxy for the expected levels of DNA found in a real sample. If it was not possible to use either known positive samples, or appropriate spiked samples, assess whether the other possible causes of the results have been appropriately accounted for (e.g. laboratory contamination, non-specific primers, inhibition, degradation, *etc.*).

### 4.1.6 Method reported to be inappropriate for use

In some cases, a project may conclude that a method or test is inappropriate for the intended application due to e.g. insufficient sensitivity, specificity, repeatability, or accuracy. However, it is important to try to understand if this reflects the choice of methods or poorly optimised methods, or if it is genuinely a case that this particular application cannot work. This is challenging to answer with limited technical knowledge, but factors to look for are:

- Inappropriate sample matrix (e.g. using eDNA from water or buffer when aggregated organisms are available).
- Insufficient sampling (e.g. filtering low volumes of water for eDNA studies).
- Inappropriate DNA extraction method (e.g. choosing a simple extraction method for a complex matrix).
- Poorly designed primers (e.g. real time PCR primers should achieve a LoD of around 10 copies of DNA).
- Inappropriate choice of barcoding region (e.g. sequences cannot distinguish between relevant taxa).
- Inappropriate choice of end point detection method (e.g. failing to detect DNA from a sample with low concentration of DNA, but using end point PCR visualised on a gel rather than real time or ddPCR).
- Issues with the reference sequences (e.g. the sequences are being compared to an online database with inappropriate levels of quality control to check for incorrect reference sequences).

An additional ‘flag’ would be evidence that the method works for similar organisms or systems but not in this particular case, without an obvious reason.

### 4.1.7 Documentation and reporting

Project reports (and academic papers) should contain enough detail to reproduce the study. As a general point, reports should state the methodological steps (including any deviations) and covering the following:

- Details of primary sampling (e.g. number and size of primary samples, sample locations, duration any traps were left in the field). Variations in the primary sampling method between sites or deviations from the defined method should be recorded.
- For eDNA, details of DNA capture method, including any variation in the quantity of matrix sampled per site (e.g. volume of water, weight of soil). The period between sampling and DNA capture should be reported.

- Details of the DNA extraction method, the total quantity of sample put into the DNA extraction, and any deviations from this. For novel applications, reasons for selecting the method should be given. For projects approaching deployment, evidence demonstrating the chosen extraction method should be available, either within the current project or other projects.
- Details of the PCR stage, including primers, reagents, details of PCR master mixes, PCR cycling conditions, and instruments used. For all studies it is desirable to include primer and probe sequences as a table, or that a correctly cited publicly accessible reference is given for the primer and probe sequences.
- For Sanger sequencing, details should be given of the number of sequences generated per sample (e.g. forward and reverse primers were both sequenced), how the sequences were aligned and consensus sequences generated (e.g. software used), and what quality control was used to determine that sequences were of sufficient quality. The final sequences should have been made publicly available.
- For real time PCR, the interpretation of results should be reported for each sample (e.g. 'positive', 'inconclusive', 'negative') alongside the criteria used to inform the interpretation (including ct thresholds, results from associated control samples). The raw ct values should be available in the report or on request, as should illustrative graph files. For conventional PCR results visualised by gel electrophoresis, the results per sample should be reported (including corresponding controls) and the gel images should be available on request or included in an appendix.
- Where the assay has been developed and validated within a project or publication, full details of the sequences used to design the assay primers should be provided (ideally as an alignment but could be references to publicly accessible sequences). As discussed above, these should be sufficiently detailed to demonstrate the inclusive and exclusive specificity. For *in vivo* validation, details of the biological reference materials and their geographic provenance should be provided.
- Where field sampling is used to demonstrate the accuracy of a test, details should be given on how the field sites were identified as positive and negative for the target organism(s).
- The report should state what positive and negative controls were used (including inhibition, degradation, or endogenous controls) and the results of these should be reported and their significance interpreted.
- If the project includes other quality control (QC) steps, these should be described and the outcomes (passed/failed QC) reported.
- For both DNA barcoding and metabarcoding, the DNA database used to assign the sequence to species (or equivalent) should be given, along with the criteria used to positively identify to species (e.g. barcode gap, pairwise sequence similarity, sequence overlap).
- For metabarcoding, full description of the methods should include all primer details, PCR cycling conditions, sequencing platform, and reagents. All quality control criteria (including positive and negative controls) should be given, and it should be reported that all samples have passed QC. Typically, read numbers should be reported pre- and post-QC for each sample, and data should be visible for the controls. The output should be fit for purpose (e.g. taxonomic assignment to required taxonomic level). Raw reads and QC data should be available within the report or on request.
- Bioinformatics pipelines should be available, and the report/paper should give a breakdown of all software used, including settings and versions. Where purpose written code is used, this should be archived and available in an appropriate location or format. A description of the pipeline and what different stages within it are doing is desirable.

### 4.1.8 Method comparison within or between projects

Studies that compare two methods (e.g. sensitivity of different platforms, performance of different primer pairs) should present evidence to show this has been done appropriately – e.g. by holding all other variables static and testing at the lower end of the Limit of Detection. The latter is important because if the performance of different methods are only compared at the higher end of detectability, the comparison cannot demonstrate equivalence at the lower limit of detectability, where the differences between the methods are likely to be.

# 5 Acknowledgements

We are very grateful to end users from environmental public bodies for discussions and comments, in particular the following: Colin Bean, Katie Clark, Joan Cottrell, Willie Duncan, Debbie Leatherland, Iveta Matejusova, Mike Nelson, Paul Woodcock and Kerry Walsh.

This project was led and funded by the Defra DNA Centre of Excellence in 2019/20, which aims to transform Defra group science to support policy and delivery by sharing and demonstrating the wider potential of existing methods; enabling the rapid development and trialling of new DNA based method applications; providing a focal point for leadership in Defra, UK and across other government departments and building shared capability including skills and facilities.

# 6 References

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L. & Vandesompele, J. (2009). The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. *Clinical Chemistry*, **55 (4)**, 611–622.

Deiner, K., Bik, H.M., Machler, E., Seymour, M., Lacoursiere-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D.M., de Vere, N., Pfrender, M.E. & Bernatchez, L. (2017) Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Molecular Ecology*, **26**, 5872-5895.

EPPO. (2016) PM 7/129 (1) DNA barcoding as an identification tool for a number of regulated pests. *EPPO Bulletin*, **46**, 501-537.

EPPO. (2018). PM 7/76 (4) Use of EPPO diagnostic protocols. *EPPO Bulletin*, **47**, 7-9.

Fox, A., Fowkes, A.R., Skelton, A., Harju, V., Buxton-Kirk, A., Kelly, M., Forde, S.M.D., Pufal, H., Conyers, C., Ward, R. & Weekes, R. (2019). Using high-throughput sequencing in support of a plant health outbreak reveals novel viruses in *Ullucus tuberosus* (Basellaceae). *Plant Pathology*, **68(3)**, 576-587.

Jones, E.P., Adams, I.P., Walshaw, J., Benucci, M., MacArthur, R., Boonham, N. & Bryce S. (2020) End-user Frequently Asked Questions on DNA-based methods for environmental monitoring. JNCC Report No. 669b. JNCC, Peterborough, ISSN 0963-8091.

OIE International Office of Epizootics. (2019) *Manual of Diagnostic Tests and Vaccines for Terrestrial Animals 2019*. Office International Des Epizooties. https://www.oie.int/standard-setting/terrestrial-manual/access-online/.

Marchezi, U., Mazzara, M., Broll, H., De Giacomo, M., Grohmann, L., Herau, V., Holst-Jensen, A., Hougs, L., Hübert, P., Laurensse, E. & Morisset, D. (2015). Definition of minimum performance requirements for analytical methods of GMO testing. *European Union Reference Laboratory for GM Food and Feed and European Network of GMO Laboratories*.

**Other sources of information:**

DNAquaNet eDNA validation table for end point PCR https://edna-validation.com/.

NASA TRL descriptions:
https://www.nasa.gov/directorates/heo/scan/engineering/technology/txt_accordion1.html.

EU Horizon 2020 TRL description:
https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf.

# 7 Appendix 1: Preliminary framework for judging the likely fitness for purpose of an eDNA detection method to detect species presence or absence in a specific location

Fera Science Report, April 2020

Author: MacArthur, R.M.

The basic assumption is that if a species is present at a location (e.g. if a newt is present in a pond, or a bacterium is present in soil), then DNA will be also be present. It may be present because it is shed by an organism or is otherwise dispersed throughout the environment: e.g. newts; or it may be present contained within whole organisms: e.g. insects. Here, we give a model that can be used to assess whether a method for testing DNA is likely to be fit for purpose when applied to true eDNA detection (e.g. newt DNA in ponds) or in scenarios close to true eDNA (e.g. bacterial species in soil).

The limit of detection (LoD) for DNA in a laboratory extract is fit for the purpose of determining whether a species is present or absent at a location if it provides a sufficiently low false positive rate and low false negative rate at some target limit of detection. For the generation of *presence only* information a method with a sufficiently low false positive rate can provide fit for purpose observations.

The model given here shows how the number of samples taken, the amount of sample tested, and the performance of the laboratory test is related to the LoD for eDNA at the sampling site. It is consistent with and an extension of models used by Wilcox et al. (2018) for fish eDNA.

The shape of eDNA sampling and testing is as follows:

1. **Primary sampling**: a number of samples **N** is taken from the location within which the presence of a species is to be determined
2. **Composite sample**: Primary samples are combined and homogenised
3. **Laboratory sample**: some or all of the composite sample is sent for analysis
4. **Test portion**: part or all of the laboratory sample of size (mass or volume **S**) is taken for analysis.
5. **Analysis:** a proportion $\mathbf{p}_1$ of the DNA in the test portion is of sufficient quality and is extracted (possibly with dilution, concentration and clean-up steps) into a volume $\mathbf{V_1}$ of DNA solution. A volume $\mathbf{V_2}$ of the solution is tested by a method with a known, estimated or planned limit of detection of $\mathbf{L_D}$ copies of DNA for which the probability of detection is $\mathbf{P_D}$.

Hence, sampling and testing of a location can be described by seven numbers:

N: the number of primary samples
S: the size (mass or volume) of the test portion
$p_1$: the proportion of DNA in the test portion of that is that is extracted into DNA solution
$V_1$: the volume of DNA solution into which DNA is extracted
$V_2$: the volume of DNA solution tested
$L_D$: the limit of detection of the DNA test
$P_D$: the probability of detection at $L_D$

The scenario in which eDNA is being used can be described as:

The presence of a species at the target level **y** e.g (1 individual per pond, 1 insect per kg soil) that should be detected with a high probability is known, estimated or assumed to result in the presence of eDNA in the sampling target at an average level of **μ** copies per quantity (mass or volume) of environmental matrix (e.g. soil or water). The expected concentration varies across different parts of the sampling target with a relative standard deviation **R.** Hence, the scenarios to which we apply eDNA sampling and testing can be described by three numbers:

y: the target presence of species that we want to detect with a high probability (individuals per unit of environmental matrix)

μ: the expected average quantity of eDNA across the whole sampling target associated with presence y in the sampling target

R: the size local variation in the quantity of target DNA across the sampling target expressed as a relative standard deviation

Then by applying a negative binomial model to eDNA presence and an independent binomial model for the probability of detection of DNA by the test method we can estimate the probability of detecting species presence as:

$$P_{+ve} = 1 - \left( 1 + \frac{R^2.S.\mu.p_1.V_2.\left(1-(1-P_D)^{\frac{1}{L_D}}\right)}{N.V_1} \right)^{-\frac{N}{R^2}}$$

*Equation 1*

We can estimate the quantity of DNA we require in the environment (per unit of environmental matrix in the same units as the quantity of analytical test portion) to provide a probability of detection $P_{+ve}$of detection using:

$$\mu = \frac{N.V_1.\left((1-P_{+ve})^{-\frac{R^2}{N}}-1\right)}{R^2.S.p_1.V_2\left(1-(1-P_D)^{\frac{1}{L_D}}\right)}$$

*Equation 2*

This particular approach is generally applicable to a wide range of scenarios where DNA is dispersed in the environment (such as newt DNA in pondwater). Where we are using DNA that is extracted directly from larger organisms which may be present in the primary samples then we need to express μ as average number organisms per unit mass across the sampling target and apply different scenario-specific models for DNA extraction and detection.

## 7.1 Illustrative examples

### 7.1.1 Fitness for purpose of a eDNA test for the presence of newts

The aim of the test is to establish the presence or absence of great crested newts in ponds. The test is applied to ponds of up to 1 HA area and 1m depth.

### 7.1.2 Method

- **20** primary samples are taken from different locations around the edge of the pond and combined and mixed into a composite sample which is sent to the laboratory.
- A **90 ml** test portion sample is taken from the laboratory sample.
- DNA is extracted from the test portion (assumed **80% efficiency**) into 100 µl of extract.
- **3 µl** aliquots are each placed into **12 PCR** tests for the presence of great crested newt DNA. Newt DNA is reported as detected if any of the 12 tests give a positive response.

Hence,

N = 20 (the number of primary samples)
S = 90 ml (the size of the test portion)
$p_1$ = 80% (efficiency of the DNA extraction)
$V_1$: 100 µl (volume of DNA solution)
$V_2$: 36 µl (volume of DNA solution tested)

Getting to the limit of detection of the DNA test applied to the 36 µl of DNA extract requires some thought in this case because *the test* consists of 12 independent DNA tests, each of which has a limit of detection estimated to be five copies of DNA.

Assuming that for each of the twelve tests detection is an event that happens to each DNA copy independently of the presence of other copies within the same test or in other tests, then from the binomial distribution the limit of detection of five copies applies to the 36 ul of DNA solution, i.e. the probability of detecting five copies in one test is the same as one copy in each of five tests; two in each of two tests and one in one test[1], *etc*.

Hence,

$L_D$=5 copies
$P_D$=95%

The remaining factor that we have not considered is the size of the between-location variation R. Where no estimates of parameters that affect detections are available, we can explore scenarios. Initially we set:

- R=0.0001 (first scenario R=0, but we can't use exactly zero, this is close enough)

Lastly, we want a 95% probability of detecting the newt DNA in the pond:

- $P_{+ve}$=0.95

Applying Equation 2 gives:

$$\mu = \frac{20 \times 100 \times \left( (1-0.95)^{-\frac{0.0001^2}{20}} - 1 \right)}{0.0001^2 \times 90 \times 0.8 \times 36 \left( 1 - (1-0.95)^{\frac{1}{5}} \right)} = 0.256$$

[1] With these assumptions the limit of detection of the 12 tests each with a LoD of five copies in 3ul is simply five copies in total of 36 ul. However, if there is an appreciable analytical false positive rate, this may be amplified by using multiple tests: better sensitivity may come at the cost of worse specificity.

This is a measure of the concentration in the sampling target with units equal to the units of the limit of detection of the test method (in this case, copies of DNA) per units in which the size of the test portion was expressed (in this case, ml). Hence, the limit of detection (for 95% probability of detection) for DNA in the pond from which samples were taken is an average of 0.256 copies/ml if the DNA is known to be distributed homogenously (R=0) in the pond.

In practice we may not expect the DNA to be distributed homogenously at the same concentration in every part of the pond; the concentration will vary between sampling locations. Figure 4 shows the effect of increasing variation on the average concentration of newt DNA in the whole pond required to provide a 95% probability of detection. We can see that the LoD is no worse than double the homogenous LoD as long as the variation in concentration has a relative standard deviation no greater than approximately 300% (actually 290%). But the LoD increases much more rapidly thereafter. Hence, it may be sensible to describe the limit of detection as less than 513 copies of newt DNA per litre of pond where between-location relative standard deviation in DNA concentration is no greater than 290%. Increasing the number of primary samples mitigates the effect of variation on the LoD but may increase costs.

The LoD tells us what a negative test result means. In this instance a negative test result tells us that, subject to our modelling assumptions, we can be confident that there is less than 513 copies of newt DNA per litre of water in the pond from which samples are taken. For the largest pond to which we might apply the test (1ha×1m deep = 10 million litres) this is means we are confident that there are fewer than approximately 5 billion cells shed by newts in the pond given a negative test result. Hence, IF the biology and ecology of newts tells us that any meaningful newt presence would result in the presence of at least 5 billion newt cells in the pond then the test method is fit for the purpose of determining newt presence and absence.

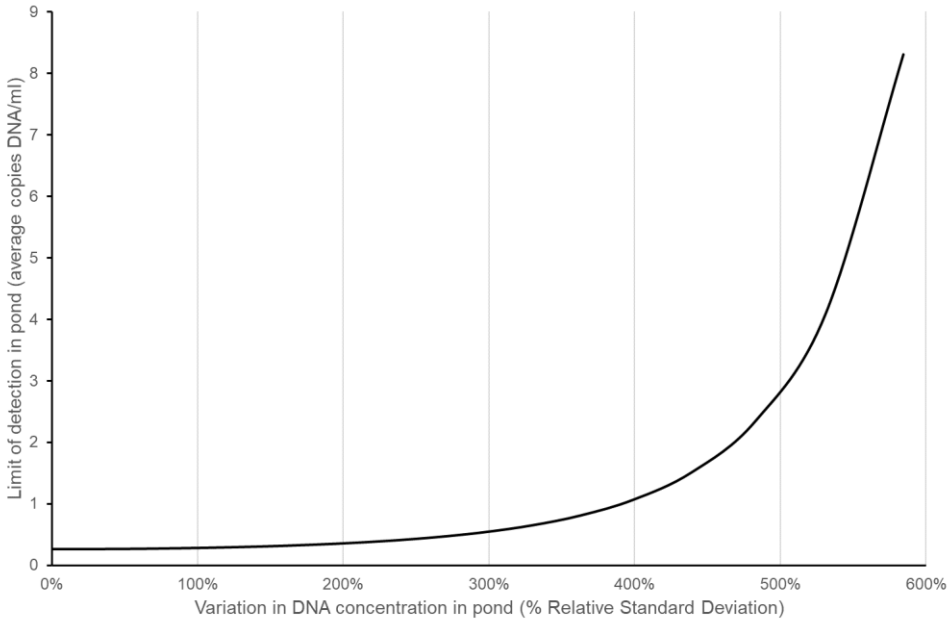


**Figure 4:** Effect of variation in the concentration of DNA in a pond on the limit of detection for newt DNA expressed as average concentration in whole pond.

In this scenario we can see from Equation 2 that, all other things remaining equal, the LoD is inversely proportional to the volume of DNA solution tested, and also inversely proportional to the volume of pond water from which we are able to extract DNA. Crucially, the limit of detection (expressed as total number of newt cells in a pond) is proportional to the volume of

pond to which a single test is applied. Finally, we can consider the effect of improving the performance of the underlying test method to reliably detect the presence of a single copy of DNA as a best possible case for the technology.

For example, doubling the volume of water and testing all of the DNA extract reduces the LoD for DNA in the pond by a factor of 5.6; improving the LoD of the underlying DNA test from five copies to one copy reduces the LoD for DNA in the pond by a further factor of 2.1. Hence, these improvements may reduce the LoD by a factor of 12.1, at some cost, from five billion copies of DNA in a large pond to 430 million copies.

A key feature of this scenario is that it applies a targeted method of analysis that only amplifies a target species DNA. Hence, in terms of DNA presence, only the absolute numbers of newt DNA in the analytical test portion affects the probability of detection Where non-targeted methods are used, the competition between different DNA may be an issue.

**Reference**

Wilcox, T.M., Carim, K.J., Young, M.K., McKelvey, K.S., Franklin, T.W. & Schwartz, M.K. (2018) The importance of sound methodology in environmental DNA sampling, *North American Journal of Fisheries Management,* **38**, 592–596.