INTERNATIONAL STANDARD

ISO 22949-1

First edition
2021-10

Molecular biomarker analysis — Methods of analysis for the detection and identification of animal species in food and feed products (nucleotide sequencing-based methods) —

Part 1: General requirements

Analyse moléculaire de biomarqueurs — Méthodes d'analyse pour la détection et l'identification d'espèces animales dans les aliments et les aliments pour animaux (méthodes basées sur le séquençage des nucléotides) —

Partie 1: Exigences générales

Reference number
ISO 22949-1:2021(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 34, *Food products*, Subcommittee SC 16, *Horizontal methods for molecular biomarker analysis*.

A list of all parts in the ISO 22949 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

This document provides general guidance for deoxyribonucleic acid (DNA) sequencing for animal species detection or identification, or both, in food and feed products. DNA sequencing is the process of determining the order of the four nucleotide bases (adenine, guanine, cytosine and thymine) in a nucleic acid polymer. Nucleic acid polymers can range in length from a few nucleotides to hundreds of millions of bases.

Rapid DNA sequencing methods have been successfully validated and verified for detection and identification of animal species in food products[1]. Within the food industry, rapid, economical and high throughput access to whole genome sequences in foods has improved the control of both food quality and safety[2]. Two types of DNA sequencing methods are most widely used for food products: chain termination and high throughput sequencing[3][4][5][6][7][8].

Chain termination developed in 1977 by Frederick Sanger still bears his name. Sanger sequencing reactions can be prepared manually and electropherograms can be read directly by the user. Automated base calling capillary electrophoresis systems have mostly replaced manually read gels and rapid Sanger applications are being developed.

High throughput or next generation sequencing (NGS), including next generation short-read and third generation long-read methods, has reduced the cost of DNA sequencing, improved sequence readability and automated most of the steps from preparatory to bioinformatics. High throughput automated DNA sequencing applies base/wavelength specific fluorescence or ionic detection to determine the real-time enzymatic addition of nucleotides to a DNA template.

Sanger sequencing (dideoxy chain termination) generates high quality data for determining a single DNA sequence of an individual target. NGS, by contrast, can be used to assess millions of individual DNA fragments of mixed markers and targets at the same time.

Sanger sequencing and NGS can both be used to verify animal species composition in a food sample or compare DNA sequence results to previously defined databases to identify its animal origin, or both[9].

Molecular biomarker analysis — Methods of analysis for the detection and identification of animal species in food and feed products (nucleotide sequencing-based methods) —

Part 1: General requirements

1 Scope

This document specifies general requirements for DNA sequencing method performance in the detection and identification of animal species in food and feed products. Performance requirements are limited to Sanger and next generation sequencing (NGS), including second and third generation sequencing, for analysis of single species products and multispecies products.

This document is applicable to DNA sequences for mammals, birds, fish, molluscs, crustaceans, amphibians, reptiles and insects, and to the validation of the applicable methods.

Methods for DNA species quantification are not considered under the scope of this document.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 16577, *Molecular biomarker analysis — Terms and definitions*

ISO 20813, *Molecular biomarker analysis — Methods of analysis for the detection and identification of animal species in foods and food products (nucleic acid-based methods) — General requirements and definitions*

ISO 21571, *Foodstuffs — Methods of analysis for the detection of genetically modified organisms and derived products — Nucleic acid extraction*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 16577 apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

4 Performance characteristics of the methods

4.1 General

The identity of an organism at the species level is established by DNA sequencing through the comparison of DNA sequences obtained from samples with known reference sequences. Results can

also be obtained at multiple taxonomic levels (e.g. order, species, genus, family), depending on the type of database and DNA data analysis performed.

Methods used for animal species identification by DNA sequencing shall meet the performance guidance provided in this document. The method validation process described in this document, interlaboratory collaborative or single-laboratory studies or both shall be used to determine and elucidate sequencing method performance characteristics.

Single species products that have been produced from one piece of meat (e.g. fish fillet, beef tenderloin) are most appropriately analysed by Sanger sequencing, while NGS is the appropriate method for simultaneous multispecies identification.

4.2 Fitness for purpose of the method

The method shall be fit for purpose for the identification of organisms and their taxonomic relationship with other organisms. Identification at the sub-specific levels (e.g. breeds) can be included provided that the databases and bioinformatic analyses are available. Information regarding the applicability and limitations of the method shall be documented sufficiently to fulfil the criteria described in this document. The appropriate commercial DNA sequencing platform(s) suitable for use in all method procedures should be determined and identified. The target DNA can be located on the nuclear genome, but also on the mitochondrial genomes. A DNA barcoding or metabarcoding approach can be used based on one or more allelic regions to be sequenced[10][11].

4.3 Scientific basis

4.3.1 General

A general workflow of the laboratory analytical procedure is provided in Annex A.

4.3.2 Sanger sequencing

Sanger sequencing is based on a chain termination method for determining the nucleotide sequence of DNA. Elongation of the replicated strand of DNA by DNA polymerase is interrupted by the incorporation of non-elongatable nucleotides that terminate replication successively at each position of the sequenced fragment. Labelling the four non-elongatable nucleotides (dideoxy-A, G, C and T) permits determination of the molecular size of the replicated fragment by electrophoresis.

NOTE 1 Sanger sequencing can only be used when a sample contains tissue from a unique single species (e.g. one fish fillet, one steak, one shrimp), because multiple DNA target sequences in the same sample from more than one species can produce more than one end labelled chain terminated product of the same chain length. Multiple chain terminated fragments of the same chain length can present two or more terminated bases at each position and appear electrophoretically as convoluted or overlapping sequences. The electropherogram for the reaction is unlikely to be readable and the identification will be compromised.

Automated Sanger sequencing platforms have been available for many years and have appropriately been used for DNA barcoding to identify a single species, e.g. for fish species identification[12][13][14].

NOTE 2 For an example, see CEN/TS 17303:2019[15].

4.3.3 Next generation sequencing

NGS in combination with DNA metabarcoding enables the robust and reliable identification of species in complex food products (containing multiple species). NGS is a commercial term used to reference high throughput DNA sequencing methods (see Annex B)[16]. It is alternatively referred to as "massively parallel sequencing". Results usually take the form of a text file with millions of individual sequence reads. Commercially available platforms support second and third generation NGS sequencing methods.

Second generation sequencing is based on the analysis of short DNA fragments generated by a polymerase chain reaction (PCR) or a fragmentation process to produce fragments of up to 600 bp in

size. Third generation sequencing is capable of analysing much longer DNA fragments (e.g. bacterial genomes) and consequently is sometimes referred to as "long-read DNA sequencing".

4.4 Units of measurement

Sequence reads generated by Sanger sequencing or NGS methods are compared bioinformatically to previously determined and identified reference sequences in DNA databases. Percent similarity between the sample and reference sequences is calculated. Two sequences are homologous if they share more similarity than would be expected by chance ($p \leq 0{,}01$).

NOTE A percent similarity greater than 98 % is usually sufficient to identify sequences as the same species.

Identification data shall be provided as a qualitative result, in the form of list of appropriate taxa, their respective percent similarity and, as appropriate, the expected value.

4.5 Applicability

4.5.1 Meat or food product considerations

The method should contain a definition of the food or food product(s) that will be sequenced and, where appropriate, the raw materials from which it is derived and any necessary references to manufacturing processes. In addition, whether the sample(s) will be composed of a single species or mixed species or both.

4.5.2 Sampling plan

Representative samples taken through an applicable sampling plan should be provided to the sequencing laboratory. Appropriate methods for subsampling laboratory samples should continue to ensure the representativeness of the sequencing result. If applicable, statistical criteria for acceptance or rejection of the analytical result should include the sampling plan and subsampling. Requirements and limitations for the laboratory sample size and number of individual items forming the sample, and the handling of the sample and its storage, should be evaluated and described based upon:

— the degree of processing of the sample constituents;

— the different species and animal tissue types involved;

— the nature of the sample matrices;

— the preparation of the sample matrix for analysis.

4.5.3 Genomic considerations

When assessing whether a method is fit for purpose, the following aspects regarding the nature of the genomic region to be sequenced should be considered:

— the cellular and genomic disposition of the sequence, i.e. cellular: nuclear or mitochondrial, or genomic: allelic or gene specific;

— the length of the sequence that will be determined;

— the types and design of primer sequences used for sample and library preparation and sequencing.

4.5.4 Method testing

Applicability shall be tested by extracting DNA from matrix-matched test samples representative of the analytical scope of the method.

Highly processed food products can have fragmented DNA or low DNA concentration or both, making them difficult to sequence. To compensate, some strategies can be employed, e.g. process more samples, concentrate DNA extracts, work with smaller targeted genomic regions.

4.6 Primer selection and evaluation

Primer selection and evaluation is the first and most important step in the nucleotide sequencing analytical workflow when a PCR-based DNA sequencing technology is used. The design of new primers is a bioinformatics task performed using applicable software. There are many commercial and freely available primer selection tools, e.g. Primer-BLAST and Primer 3. The following primer design factors affect primer performance: secondary structures, partial sequence matches, hairpins, GC content, etc.

NOTE 1 Primer design and selection is not considered as part of the workflow to select genomic regions to be sequenced in third generation NGS technology and the DNA fragmentation strategy, e.g. second generation based NGS, shotgun sequencing.

Universal (consensus) primers can be defined for any taxonomic level (species, genus, family, order, class, domain, etc). The design of these primers is based on DNA sequence alignments and identification of consensus regions that are almost identical amongst the taxa to be identified. Degenerate primers can be designed to increase the universality of the primers. Determining the number of degenerations included in a single primer to avoid non-specific binding requires evaluation of unspecific annealing.

Primers should be evaluated in silico irrespective of their use in the laboratory workflow. They shall be assessed using appropriate bioinformatics tools for their match with the species (taxon) to be detected/identified.

NOTE 2 An example of an in silico workflow for primer set selection is shown in Annex C.

The final evaluation of the performance of each set of primers should be done by applying the laboratory workflow using reference DNA or tissue material (whenever it is available) representative of the identifiable group of species (taxa).

4.7 Database construction and evaluation

The DNA database used for species detection or identification or both will determine the number of different species identifiable by the laboratory workflow. Database construction and evaluation is a bioinformatics task and many different tools can be used, including in-house developed software. Publicly available databases can also be used, e.g. GenBank. The databases should have the following information available:

— the species list;

— the number of different entries for each species;

— the DNA region(s), gene(s) and length of DNA sequences;

— a measure of the validation quality of the sequence entry(ies).

The final performance evaluation of a bioinformatics search of a database shall be done by applying the laboratory workflow using reference material (whenever possible) that is representative of the species included in the database (vouchered). The use of higher taxonomic levels is appropriate when the identification at the species is inconclusive, i.e. when the sample is not differentiable by a more targeted search. In GenBank, the designation "*Bos* sp." can be used as the search term for all of the species of the genus *Bos*.

4.8 Selectivity and nucleotide sequence specificity

4.8.1 General

Excessive amplification of non-targeted species DNA can interfere with the detection of targeted species DNA, e.g. increased limit of detection (LOD) or critical value or both. Targeted (PCR-based) NGS and Sanger sequencing methods should provide experimental evidence for nucleotide specificity with non-targeted species (see ISO 20813). The degree of interference (selectivity) from non-targeted DNA should be determined. The nucleotide sequence specificity should be assessed in a two-step procedure: theoretical and experimental evaluation of the inclusivity and exclusivity. Experimental evaluation is done by PCR with the selected primers. For NGS, the performance of each set of primers should also be evaluated with adapters or barcodes or both. The minimum concentration of DNA required for specific sequence identity analysis can be determined as the probability of detection (POD) (see ISO/TS 16393) [17]. Because sequence data are used for the verification of animal speciation results, these data should be based on evaluated databases with due consideration of the timing of submission of individual entries and any subsequent changes in taxonomic classification or naming (i.e. provenance). In cases of unexpected results, further investigation should be carried out to confirm reference material identity. DNA amplification is not required when using a non-PCR based sequencing analysis, e.g. shotgun sequencing.

4.8.2 Requirements for inclusivity testing

Experimental results from testing a method with a target animal taxon should be provided. This testing should include relevant species for the scope of the method and the taxonomic level covered by the primers to be used. An appropriate number of different species should be used (see ISO 20813). For primers with less than 100 % sequence homology with target DNA species (based on in silico evaluation), the amplifiability of those species should be tested in the laboratory to define the limits of acceptable mismatches. The user shall define a theoretical list of species that can be identified and the list of species that were tested in the laboratory. It is known that depending on the DNA region that is targeted by the primers there can be very different conservation levels. Different genes have different rates of intraspecific DNA sequence variability. Appropriate numbers of individuals from each species should be tested accordingly.

Material for experimental inclusivity testing should contain sufficient DNA concentration for PCR and DNA sequencing, as described in ISO 20813. Replicates of each sample material shall be tested with controls appropriate to the different workflow steps, e.g. PCR. If present, sequence variants of the target animal species should be identified with comparable amplification efficiency.

4.8.3 Evaluation of non-targeted DNA interference

Experimental results from testing the sequencing method with non-target animal species shall be provided. This testing should include both taxonomically close and not-closely-related animal species. Animal species or taxonomic groups relevant to the scope of the method should be tested, e.g. species commonly used in food in general and particularly in matrices considered in the scope of the method. The method should clearly distinguish between target and non-target animal species. In silico analysis should be performed with the appropriate non-target taxa to theoretically evaluate the sequence homology of the universal primers. Sequence mismatches are good indicators of specificity. They indicate the degree to which a set of primers and probe will bind to unintended sequences to produce a false-positive result. A minimum number of mismatches should be defined as acceptable for in silico analysis. Exclusivity is determined more effectively when a higher number of mismatches is permitted in the analysis. Select an appropriate number of species that can cause interference with the target animal species present in the food test material (as described in ISO 20813). Examples of suitable organisms are listed in ISO 20813:2019, Annex A. Other species should be included if relevant, e.g. if there are sequence homologies of oligonucleotides to nucleic acid sequences. The suitability of the DNA used for amplification should be confirmed by using appropriate controls. Sufficient DNA should be used for experimental PCR exclusivity testing.

4.9 Sensitivity

4.9.1 General

Experimental results from testing the method at different DNA concentrations to evaluate the range of use of the method shall be documented and described in the validation report. Animal species should be detected at levels relevant for the interested party, e.g. the consumer.

4.9.2 Limit of detection

4.9.2.1 General

The LOD (see ISO 21569 and ISO 20813) and the POD (see ISO/TS 16393[17]) can vary significantly between matrices. Animal constituents at different processing steps in the same product can exhibit different levels of DNA degradation and possible DNA quality differences among ingredients and can contribute to sequence uncertainty. As an example, a product can be composed of different types of animal tissue containing different amounts of DNA. This imbalance can be further exacerbated if some ingredients have undergone pre-processing. Cooking or acid treatment of some ingredients will lower DNA quality compared to other ingredients that can be added in raw form or from a different process. Depending on the target gene used for identification, the number of copies of that gene can be very different. The determination of the LOD can be performed as described in ISO 20813:2019, 4.7.2.

4.9.2.2 Limit of detection for Sanger sequencing

The LOD determination should be performed as described for a PCR-based method in accordance with ISO 20813. The data resulting from Sanger sequencing is a DNA sequence obtained from a previously amplified PCR product. The LOD for this method is equivalent to the LOD obtained for the PCR reaction by using the universal primers defined for the specific taxon. It should be noted that LOD determination for Sanger sequencing is of particular importance when dealing with products containing processed ingredients, due to the potential for much lower DNA quality and quantity from those components.

4.9.2.3 Limit of detection for NGS

For PCR-based NGS, LOD is based on the LOD of the PCR system used (see ISO 20813). It is calculated as described for Sanger sequencing.

When no PCR is performed, e.g. shotgun sequencing, the LOD is based on the integrity (sequenceability) of the input DNA and the NGS system used.

In all NGS methods, an additional LOD should be defined based on the lowest number of reads required to find a species match from the database used.

4.10 Robustness

4.10.1 General

The capacity of a sequencing method to remain unaffected by small variations in method parameters (e.g. variation in concentration of kit components, variation in apparatus) shall be verified. The evaluation of robustness can be empirically performed.

4.10.2 Robustness determination by inter-laboratory study

Robustness can be determined by performing an interlaboratory study[18]. A robust method is one where the results from different laboratories do not vary significantly. Robust methods should be selected for use wherever possible.

4.10.3 Robustness determination by a multifactorial orthogonal test design

The test method should be carried out using an orthogonal multifactorial approach where several alterations are assessed (see ISO 20813) including, but not limited to, variations in mastermix component concentration, reaction volume, primer concentrations, annealing temperature and thermocycler platform.

5 Single-laboratory validation

An analysis method should have been sufficiently tested within a laboratory to disclose the required specification prior to an interlaboratory study[19]. See ISO 13495[20] and CEN/TS 17329-1[21] for guidance. Reference materials or certified reference materials (CRMs) should be used when available for the validation of DNA sequencing methods for nucleic acids.

6 Interlaboratory study (collaborative study)

Information about the collaborative study (organizer, protocol, number of participating laboratories, etc.) and the performance data obtained by the study shall be reported with appropriate references to the relevant documents[19]. Collaborative studies for the validation of PCR methods for detection, identification and quantification of specific DNA sequences can be performed in accordance with other relevant documents, e.g. Codex Alimentarius CAC/GL 74-2010[22] and CEN/TS 17329-2[23].

NOTE 1 A small-scale collaborative study (pre-validation study involving, for example, two to four laboratories) can be performed to test the general transferability of the method before the expense of organizing a large-scale study is incurred.

For precise validation, data shall be collected from multiple laboratories having facilities and competence in molecular biology. ISO 13495:2013 requires eight and four participating laboratories for international and national levels of validation, respectively. CEN/TS 17329-2 specifies a minimum of 12 laboratories. In accordance with AOAC International[24], the required number is eight laboratories. Statistical analysis for accuracy should be calculated based on ISO 5725-1:1994[25].

NOTE 2 Traditional nonparametric 5 % false positive and 5 % false negative rates reflect PODs of 5 % and 95 %.

7 General laboratory and procedural requirements

7.1 General

The procedure shall be documented to include the following steps:

— ensuring a representative sample;

— preparation of the test samples and subsamples;

NOTE If the test sample is not the whole laboratory sample, the laboratory sample is homogenized and test samples are obtained in accordance with relevant International Standards, including grinding, homogenization and preparation of test portions

— extraction of DNA (if necessary, determination of DNA concentration and purity by appropriate methods)[26];

— DNA sequencing;

— interpretation and reporting of results.

The manufacturers' safety recommendations shall be followed.

7.2 Facilities, materials and equipment

The work area in the laboratory should be designed to prevent accidental DNA contamination originating from dust, human material and spreading aerosols. Consideration shall be given to:

— the systematic containment of the method steps involved in the production of the results;

— a forward flow principle for sample handling.

Separation (temporal and/or physical) of work is required to prevent contamination. Designated contained or dedicated work areas or both with their own apparatus are recommended, as follows:

a) a work area for grinding and homogenization;

b) a work area for extraction of the nucleic acid from the test material;

c) a work area dedicated to the setup of PCR amplification reactions;

d) a work area dedicated to DNA sequencing.

If contamination from human DNA is detected by the method, additional prevention measures (e.g. the use of masks, gloves and disposable coats) should be taken to prevent false-positive results during analysis.

Physical separation or using different rooms or both is the most effective and preferable way of preventing carryover contamination, but other methods can be used provided the effectiveness is comparable. Air flow for the laboratory should be directed to prevent intrusion of dust and amplicons from work areas with higher contamination risk to work areas with lower contamination risk, i.e. forward flow.

Unless otherwise stated, only analytical grade reagents suitable for molecular biology, free from DNA and DNases should be used. Reagents and solutions should be stored at room temperature, unless otherwise specified. PCR reagents should be stored in small aliquots to minimize the risk of contamination. Water for analysis shall be double-distilled, deionized (≥ 18 MΩ) or of comparable quality. Solutions should be prepared by dissolving the appropriate reagents in water and autoclaving, unless specified differently. Sterile filtration devices (possibly 0,22 µm pore size) can be used when autoclaving is not possible.

To avoid contamination, sterile techniques should be adopted in the PCR set-up area, e.g. powder-free gloves, sterilized plasticware, autoclaved reagents, disposable plasticware and aerosol-protected, DNA/RNA free and DNase/RNase free filtered pipette tips should be used.

Materials and all containers and disposables containing reagents shall be protected from any contaminating agent, e.g. dust.

The manufacturers' recommendations for the use of reagents should be followed. Appropriate controls can be used to assess the integrity of reagents and the absence of DNase.

No unintended enzyme activities (e.g. exonuclease) that can interfere with PCR and DNA sequencing shall be present in the preparations.

7.3 Sample preparation

7.3.1 General

The requirements for sample preparation depend on the type of sample that is analysed. Samples can be divided into three different categories (A, B and C) as shown in 7.3.2 to 7.3.4.

7.3.2 Category A: single species sample consisting of a single piece

EXAMPLE One sample = one fish fillet or one beef steak.

A portion of the laboratory sample (single piece) can be taken for analysis by Sanger sequencing or NGS. To minimize the risk of detecting adhering contaminants, test sample material shall not be taken from the surface of the laboratory sample.

7.3.3 Category B: single species product composed by several pieces or units of the same type of tissue

7.3.3.1 General

EXAMPLE One sample = package with 10 fish fillets or 20 meat pieces.

Depending on the sample preparation, either NGS or Sanger sequencing can be used as described in 7.3.3.2 and 7.3.3.3.

A portion of the laboratory sample (piece or unit) can be taken for analysis or all pieces or units can be collected and mixed to produce a composite sample. The composite sample should be homogenized as described to guarantee the representativeness of each portion collected.

7.3.3.2 Preparation of a representative test sample for NGS

There are two options:

— Take at least one equivalent sized portion of each fish fillet/meat piece (same type of tissue and portion size). All portions collected are mixed to produce a composite sample of the food product received. The composite sample should be homogenized to uniformity to ensure representativeness of each portion collected.

— Homogenize the entire laboratory sample to uniformity.

7.3.3.3 Preparation of a test sample for Sanger sequencing

Proceed as described in 7.3.2 for each individual piece composing the laboratory sample.

7.3.4 Category C: multiple mixed species processed product including those of non-meat origin or different types of tissues as ingredients

EXAMPLE Lasagne, pizza, seafood cocktail, minced meat or as unwanted contamination (food containing trace species amounts or cross-contamination or both).

Samples are totally ground and homogenized to uniformity.

Because the complexity of samples in this category precludes effective use of Sanger sequencing, category C products should be analysed exclusively by NGS[27].

The laboratory can subsample a sample containing readily separable ingredients when appropriate. In this case, the subsampling procedure and associated limitations should be clearly documented with the subsampling step indicated in the final report. The decision to subsample should preferably be made when the sample is submitted for analysis.

Grinding and homogenization should be sufficient to allow a representative sample to be taken. The procedure shall be validated for the species and tissue type.

NOTE Where applicable, products can be separated into their distinct component parts which can be a single species. Each part can be considered category A or B samples.

7.4 DNA extraction

The general requirements for DNA extraction and quantification are described in ISO 21571.

DNA concentration should be determined using an appropriate method, as excessive amounts can later inhibit PCR amplification or sequencing steps or both[26].

7.5 DNA sequencing workflow

7.5.1 General

Workflow shall follow the Sanger or NGS sequencing method used and the manufacturer's instructions for specific instrumentation. Sequences are usable if they comply with the quality control parameters specified by the instrument manufacturer and for individual project requirements.

7.5.2 Sanger sequencing method

7.5.2.1 General

This method produces a single electropherogram. It is most appropriate for use with samples that consist of a single species, e.g. one fish fillet, one steak, one shrimp. Therefore, Sanger sequencing is appropriate for foods included in categories A and B. After DNA extraction, the Sanger sequencing workflow includes:

— PCR;

— purification of PCR product;

— asymmetric amplification with dideoxy termination (PCR);

— purification of sequencing fragments;

— DNA sequencing.

7.5.2.2 Assessing raw sequence quality from PCR amplified fragments

A universal primer pair is used for PCR amplification, targeting specific genes known to be useful for species identifications. Commonly referred to as "DNA barcoding", there are national and internationally agreed standard sets of universal primers and targets recommended for use. Cytochrome C oxidase subunit 1 (cox1), cytochrome B (cytB), eukaryotic mitochondrial small ribosomal subunit RNA(12S rDNA) genes and eukaryotic large ribosomal subunit RNA (18S rDNA) nuclear genes are routinely used in animal species identification. Universal primers for barcoding from the 16S ribosomal RNA gene have also been developed[28]. However, additional primers can be designed using primer design bioinformatics tools[29]. The primers are designed to anneal to highly conserved DNA regions that are identical or highly similar in a large group of taxa (species, genus, family, order, domain, etc.). The degree of universality is established according to the DNA regions (genes) where the primers anneal and is assessed in a first step by bioinformatics tools. Degenerate nucleotides are used at key primer nucleotide positions to maximize the number of species detected. An in silico universality analysis of the primers used should be done to compile the list of taxa where the primers can anneal.

After PCR amplification, the amplicon can be visualized by agarose gel electrophoresis or other means. This step is used to confirm the expected size of the amplicon and the absence of additional unwanted non-specific amplicons that can be produced during the PCR reaction.

7.5.2.3 PCR purification

The purpose of this step is to separate the amplicon from PCR reagents, leftover primers and nucleotides in particular that can interfere with downstream sequencing steps. PCR purification is normally done using commercially available kits. However, another suitable purification procedure can be used.

7.5.2.4 Asymmetric amplification

This step is done using a kit from the DNA sequencer manufacturer. Depending on the Sanger sequencing instrument, each brand has its own consumables and instructions. The manufacturer's instructions should be followed. Target sample DNA is replicated from a single forward or reverse primer in the presence of common dNTPS and fluorescence-labelled ddNTP substrates. Extension of individual DNA strands from the primer is stopped at each site of random integration of a ddNTP, creating a population of fragments created from the same template but with different lengths and differentially labelled according to the last nucleotide addition.

7.5.2.5 DNA sequence resolution

The mixture of fragment lengths produced during the asymmetric amplification are separated by charge and molecular mass using a high resolution (single nucleotide) electrophoresis system that can detect and differentiate the end labels. Capillary or polyacrylamide gel electrophoresis is preferable, producing an electropherogram of the different DNA fragment sizes obtained each with their respective terminator. There are different DNA sequencing instrument and equipment providers. Instruments should be calibrated and used following the manufacturer's instructions.

Although very limited in read length, Sanger sequencing can be performed using four separate termination reactions: one for each of the respective bases, ACTG in separate wells or tubes. The same label which is often radioactive may be used for each base. Polyacrylamide gel electrophoresis is used to separate the fragments from each of the four reactions. They are loaded in consecutive wells on the gel so that the chain lengths from each reaction can be compared. When a radioactive label is used the gel is exposed to X-ray film for sequence determination.

7.5.2.6 Assessing raw sequence quality

This is usually done by visual examination of the electropherogram obtained. Each corresponding sequenced nucleotide should present a unique curve. Background noise should be clearly distinguishable from the real curves obtained for each nucleotide. Where background interference prevents a clear sequence read, the sequence should be discarded. Factors that affect the quality of Sanger sequencing are commonly related to the quantity of DNA used, primer annealing, nucleotide sequence composition and the presence of interfering DNA.

7.5.3 Next generation sequencing method

7.5.3.1 General

This method generates a file containing multiple complete DNA sequences, usually thousands or millions of sequences. Specific base elongation is incrementally detected and recorded for each fragment, each DNA polymer generating a unique DNA sequence.

NGS can be used for any sample regardless of whether it is composed of one or multiple species. It is most appropriate for use with samples that contain multiple species because it can identify each species present in a sample.

After DNA extraction, NGS workflow includes:

— DNA library preparation;

— NGS sequencing;

— assessing raw read data quality.

7.5.3.2 DNA library preparation

7.5.3.2.1 General

DNA libraries are prepared from extracted DNA. This process can be performed directly or include an optional PCR step targeting specific gene region(s).

There are several different sequencing platforms available and it is important that requirements for DNA library construction specified by the instrument manufacturer are met, i.e. DNA library concentration, average amplicon size, range of amplicon size, etc.

7.5.3.2.2 DNA library preparation without PCR or fragmentation

DNA library construction performed directly and without preparatory PCR or fragmentation consists of the following steps:

— ligation of sample index (when sample pooling is to be done);

— quantification, normalization and quality control assessment of the resulting DNA library;

— pooling of libraries for multiplexed sequencing runs.

All these steps can be performed with reagents available for each NGS instrument and following manufacturer instructions. Appropriate controls and adjustments for these steps should be done as recommended for each instrument brand and according to the technology used, e.g. long-read technology optionally requires DNA fragmentation and ligation of sample indices.

NGS methods produce sequences of millions of individual DNA molecules simultaneously. For some applications, pooling of DNA libraries is advantageous because it allows sequencing of DNA from more than one sample in the same sequencing run. Pooling of DNA libraries is accomplished through the use of a sample index that adds a unique identifier to the DNA of each sample and is typically registered to allow association of the resulting sequence data with the correct metadata. The number and type of sample index depends on the NGS instrument used and the type of DNA library produced.

NOTE Sample barcoding (sample indexing) of fragments of less than 600 bp provides an example for library preparation without PCR and without fragmentation. For some specific workflows, a DNA fragmentation can be part of the procedure.

7.5.3.2.3 DNA library preparation with PCR

DNA library construction with preparatory PCR consists of the following steps:

— target-specific PCR (single or multiplex);

— ligation of sample index (when sample pooling is to be done);

— quantification and purification;

— pooling of libraries for multiplexed sequencing runs.

Universal primers are used for a PCR-based approach as described for Sanger sequencing. Multiplexing strategies can be incorporated during the DNA library preparation and can be done at the following levels:

— sample multiplexing (pooling more than one sample in the same sequencing run);

— gene or amplicon multiplexing or both (e.g. pooling more than one gene-specific amplicon in the same sequencing run);

— sample multiplexing and gene or amplicon multiplexing or both (pooling more than one sample and more than one specific gene or amplicon or both in the same sequencing run).

Sample indices are added to the marker amplicon during or after PCR.

When using a gene-specific PCR-based approach, there are three ways of sample indexing the amplicons for sequencing.

a) It can be done as a specific step after performing PCR where the amplicons can be mixed with specific reagents that add a unique sample index to each amplicon.

b) It can be added during the PCR reaction by using fusion primers for amplification. Fusion primers are long primers, usually 50 to 70 nucleotides long and need to be designed to fit with the NGS instrument brand that will be used.

   Long or fusion primers can produce artefacts and interferences during the PCR reaction. Different fusion primers should be tested and evaluated for their impact on the PCR evaluation quality parameters as described in ISO 20813. This testing and evaluation can involve a set of reference DNAs amplified with the different fusion primers to assess the amplicon intensity obtained. These experiments should be done with long primers before any routine use.

c) It can be added with a second PCR reaction containing the sample index. This process is accomplished with kits that are specific for each type of NGS instrument used. This is a post-PCR process to be performed according to the manufacturer's instructions, including all the quality controls accompanying those instructions, e.g. oligonucleotide ligation, nested or second PCR step.

NOTE Eukaryote genomes are generally of sufficient size to limit the number of individual samples that can be analysed simultaneously even with NGS technology. PCR is used to enrich representation of a specific genetic regions, thereby increasing the capacity for multiple species detection and the number of samples that can be analysed in one run. PCR can also bias which species are amplified preferentially within a mixed sample (see 4.6).

7.5.3.3 NGS sequencing

Procedures vary depending upon the platform used. The instrument manufacturer's instructions should be followed.

7.5.3.4 Raw read processing and sequence quality assessment

Sequence read files are generated using instrument-specific software or instrument-specific pipelines or both. Several physical measures such as signal-to-noise ratio shall be considered. Their measurements should be monitored during the sequencing experiment. Sequence read files should be configured in the appropriate file format, which contains the compilation of individual sequence reads, each with its own identifier, and an associated base quality score for each nucleotide. FASTQ format is the most common file format obtained at the end of an NGS run and contains the reads and all of its quality metrics. The quality should be evaluated using appropriate software, e.g. the FASTQC tool is commonly used to evaluate the quality of sequencing results for FASTQ[30]. Raw data pre-processing can include removal/trimming of low-quality sequences/bases, demultiplexing when mixtures of samples are used in a same run, removal of adapters/primer, trimming of reads to a fixed length (when a PCR-based approach is used). Additional computational steps can be added according to the equipment/technology used.

Quality assessment includes but is not limited to determining quality scores (both nucleotide and full sequence Q-(Phred) scores), length distribution, and GC content (not needed for a PCR-based NGS). A defined Q-score threshold should be established, e.g. Q-score of 20. Based on the processing of raw data and the quality criteria defined, sequences not complying should be discarded. A FASTA file is commonly originated containing all the good-quality sequences. The number of reads of the file used for database comparison should be documented. With a multiplex approach based on a barcoding PCR-based NGS with mixed samples or different DNA regions or both, the number of reads for each combination sample/gene should be documented. It is preferable to establish a minimum number of reads for each combination, e.g. 1 000 reads per sample per region. However, this threshold should be established according to the criteria defined by each laboratory. The use of a non-PCR NGS approach will not permit obtaining this type of information at this stage. However, after comparison of the reads obtained with the database these parameters can be calculated.

7.6 DNA sequence data analysis and interpretation

In this step, the DNA sequences obtained are compared with the defined database used for identification. When multiple sequences are obtained as output, the DNA sequences are normally sorted first into clusters and the cluster master sequence is used for comparison with the DNA database. The identification result is defined as the entry with the highest match (sequence homology) in the database. The percent homology found between the DNA sequence(s) obtained and the highest database match should be reported together with the size (in number of nucleotides) of the amplicon used for sequence homology. For Sanger sequencing, it is recommended to obtain the sequences of both DNA strands (forward and reverse).

For NGS, the number of reads can be reported according to each species identified as a proportion of the total number of different species found. The comparison of the sequence(s) obtained with reference databases should be done with appropriate bioinformatics tools. Applicable tools can be developed in-house or can be obtained through public media or commercially, e.g. the basic local alignment search tool (BLAST)) can be used as long as the criteria defined for data analysis are met.

7.7 Expression of results

The results shall be reported as a list of taxa identified. The results shall include the most detailed taxonomic name, e.g. family, genus, species, sub-species or breed name, that can positively be identified using an appropriate DNA database and bioinformatic analyses.

8 Validation of the NGS bioinformatics pipeline

8.1 General

Laboratories should perform their own validation of the bioinformatics pipeline. The performance requirements for the assay shall be established during the validation procedure and the same specifications shall be used to monitor the performance of the assay each time a sample is processed.

Specific quality control and quality assurance parameters shall be evaluated during validation and used to determine satisfactory performance. Each laboratory shall define the criteria and a means to monitor all quality metrics to ensure optimal analytical performance.

Laboratories shall include specific measures to ensure that each data file generated in the bioinformatics pipeline maintains its integrity, and provides alerts that prevent the use of data files that have been altered in an unauthorized or unintended manner.

Supplemental validation is required whenever a significant change is made to any component of the bioinformatics pipeline.

8.2 Quality metrics

Laboratories shall establish acceptable raw base call quality score thresholds for the assay during validation. Pre-processing methods to remove low-quality base calls should be established to reduce the false positive species ID rate.

Parameters for mapping quality shall be established in a validation plan and should demonstrate that the test only analyses reads that map to the regions targeted by the assay, e.g. only DNA regions homologous with the database should be compared for matching and species ID. If applicable, steps should be taken to filter reads that map to non-homologous regions.

When using a non-PCR NGS approach, coverage shall be defined to achieve adequate sensitivity and specificity in the regions of interest. Each laboratory shall establish the minimum criteria for the depth of coverage characteristic of a specific DNA region under standard assay conditions depending on the aim of the sequencing run.

9 Test report

Test reports shall contain at least the following:

— all the information needed to identify the laboratory sample (including the size of the laboratory sample);

— any particular information relating to the laboratory sample;

— all information related to the test sample (size of the sample particle, if ground);

— a reference to this document, i.e. ISO 22949-1;

— a statement about the date and type of sampling procedure(s) used;

— the date of receipt;

— the storage conditions, if applicable;

— the analysis start and end dates, if applicable;

— the person responsible for the analysis;

— the results according to the requirements of the specific method and the units used to report the results and the calibrators and the calculation method used;

— any particular observations made during testing;

— any deviations, additions to, or exclusions from the test specification;

— the sequencing method used (Sanger or NGS);

— for NGS, the corresponding platform and strategy, i.e. short fragments, long fragments, PCR-based or other;

— the length of the sequence/reads used for identification;

— the gene(s) used for identification in the database;

— the size of the amplicon used when a PCR-based approach is used (both for Sanger and NGS);

— the DNA sequence database(s) used (including version(s));

— the percent sequence homology and sequence length used for each species identified;

— the date the test was performed.

NOTE For NGS, laboratories can choose to indicate the percentage of reads per species over the total number of reads obtained.

Annex A
(informative)

General workflow of the laboratory analytical procedure

Refer to Figure A.1.

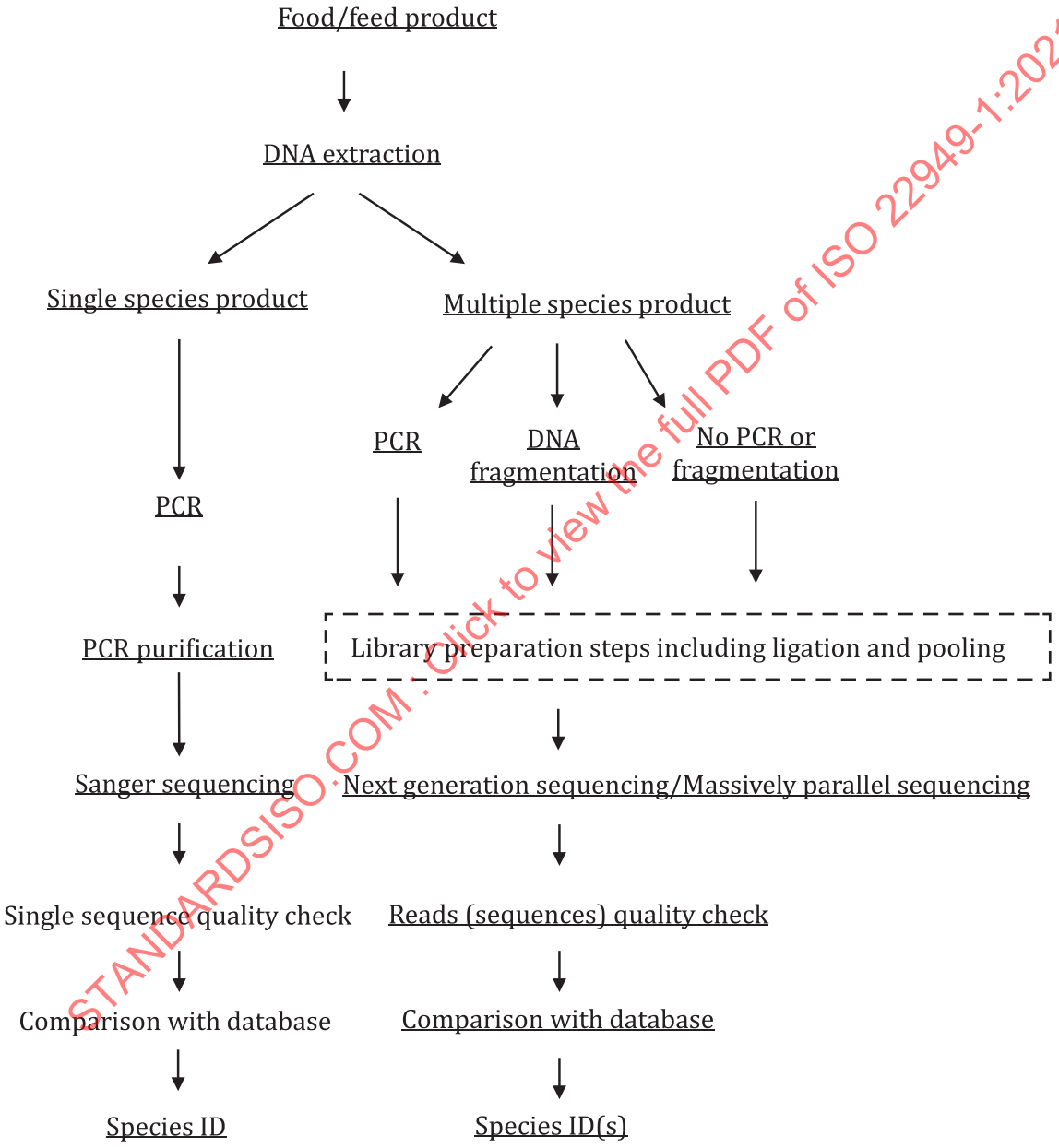

Figure A.1 — General workflow of the laboratory analytical procedure